UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

EXTRACCIÓN Y ANÁLISIS DE VARIACIÓN DENOMINATIVA DE TÉRMINOS EN DOMINIOS DE ESPECIALIDAD

TESIS
QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN CIENCIAS (COMPUTACIÓN)

PRESENTA:
JORGE ARMANDO ÁVILA ESTRADA

DIRECTOR DE LA TESIS:
DR. GERARDO EUGENIO SIERRA MARTÍNEZ....
INSTITUTO DE INGENIERÍA

CODIRECTORA DE LA TESIS:
DRA. SOFÍA NATALIA GALICIA HARO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

CIUDAD UNIVERSITARIA, CD. MX., OCTUBRE 2016

CONTENIDO

1. INTRODUCCIÓN 8

PLANTEAMIENTO DEL PROBLEMA 8

OBJETIVO 10

ORGANIZACIÓN DE LA TESIS 10

2. ANTECEDENTES 13

TERMINOLOGÍA 14

Terminología y las áreas de especialidad 15

INGENIERÍA DE SOFTWARE Y LA BIODIVERSIDAD 17

Los términos en la Ingeniería de Software 17

Los términos en la Biodiversidad 20

Lingüística Computacional 22

Variaciones en los términos 25

CORPUS LINGÜÍSTICOS 27

Clasificación de corpus lingüísticos 29

Necesidad de un corpus en un área de especialidad 32

EXTRACCIÓN TERMINOLÓGICA 32

Extractores automáticos de términos 34

SEMÁNTICA DISTRIBUCIONAL 38

Modelos de semántica distribucional 40

3. LOS CORPUS DE ESPECIALIDAD: INGENIERÍA DE SOFTWARE Y BIODIVERSIDAD. 43

Descripción del corpus de Ingeniería de Software 43

Descripción del corpus de Biodiversidad 45

Etiquetado de partes de la oración (POS) 46

4. LA EXTRACCIÓN DE TÉRMINOS EN EL ÁREA DE INGENIERÍA DE SOFTWARE 49

TERMEXT 49

TERMOSTAT 52

TÉRMINOS OBTENIDOS 55

5. DETECCIÓN DE VARIANTES DENOMINATIVAS 62

METODOLOGÍA PARA LA OBTENCIÓN DE TÉRMINOS SIMILARES 62

Matriz de coocurrencia *64*
Obtención de distancias *72*
Agrupación de términos similares *74*
TÉRMINOS SEMÁNTICAMENTE SIMILARES 77
6. RESULTADOS Y EVALUACIÓN 87
7. CONCLUSIONES 127
TRABAJO FUTURO 128
BIBLIOGRAFÍA 130
APÉNDICES 135
ETIQUETAS UTILIZADAS (TAGSET) 135
LISTADO DE PALABRAS FUNCIONALES (STOPLIST). 140
TERMINOLOGÍA DE LA INGENIERÍA DE SOFTWARE 143
VARIANTES DENOMINATIVAS DE LA INGENIERÍA DE SOFTWARE (TERMOSTAT-PMI) 151
VARIANTES DENOMINATIVAS DE LA INGENIERÍA DE SOFTWARE (TERMEXT-PMI) ... 154
TERMINOLOGÍA DE LA BIODIVERSIDAD 158
VARIANTES DENOMINATIVAS DE LA BIODIVERSIDAD (TERMOSTAT-PMI) 165
VARIANTES DENOMINATIVAS DE LA BIODIVERSIDAD (TERMEXT-PMI) 168
CONTEXTOS DE EVALUACIÓN PARA LOS TÉRMINOS DE INGENIERÍA DE SOFTWARE. 172

Índice de figuras

Figura 1. Nivel de uso de los términos: computadora, ordenador y computador **9**
Figura 2. Extracto del Glosario de Términos de la Ingeniería de Software de la IEEE **19**
Figura 3. Extractores terminológicos y su filtro **36**
Figura 4. Extractores terminológicos, dominio y características **36**
Figura 5. Representación distribucional **39**
Figura 6. Obtención de resultados Termostat **55**
Figura 7. Funcionamiento programa de obtención de Matriz de Coocurrencia **69**
Figura 8. Funcionamiento programa Distancias Coseno **72**
Figura 9. ASV Toolbox **75**
Figura 10. Dendograma términos Termostat PMI **77**
Figura 11. Ejemplo de términos proporcionados por CONABIO **93**
Figura 12. Dendograma - Evaluación #1 **103**
Figura 13. Dendograma - Evaluación #2 **111**
Figura 14. Dendograma ejemplo distancias coseno **116**
Figura 15. Gráfica mapa 2D de niveles similitud **116**

Índice de tablas

Tabla 1. Conformación del corpus Ingeniería de Software **44**
Tabla 2. Conformación del corpus Biodiversidad..... **45**
Tabla 3. Requerimientos de uso Termext **50**
Tabla 4. Número de términos obtenidos. **56**
Tabla 5. Evaluación entre extractores terminológicos..... **58**
Tabla 6. Comparativa de términos en Ingeniería de Software **58**
Tabla 7. Comparativa de términos en Biodiversidad..... **59**
Tabla 8. Límite de extracción de términos. **62**
Tabla 9. Formato de lista de términos y lista de palabras..... **70**
Tabla 10. Número de pares de variantes denominativas obtenidas utilizando Termostat en Ingeniería de Software **78**
Tabla 11. Número de pares de variantes denominativas utilizando Termext en Ingeniería de Software **78**
Tabla 12. Similitud de términos utilizando Termext – PMI **79**
Tabla 13. Similitud de términos utilizando Termext – Frecuencia..... **80**
Tabla 14. Similitud de términos utilizando Termostat – PMI **80**
Tabla 15. Similitud de términos utilizando Termostat – Frecuencia..... **81**
Tabla 16. Número de variantes denominativas utilizando Termostat en Biodiversidad..... **82**
Tabla 17. Número de variantes denominativas utilizando Termext en Biodiversidad..... **82**
Tabla 18. Similitud de términos utilizando Termext - PMI..... **83**
Tabla 19. Similitud de términos utilizando Termext – Frecuencia..... **84**
Tabla 20. Similitud de términos utilizando Termostat – PMI **84**
Tabla 21. Similitud de términos utilizando Termostat – Frecuencia..... **85**
Tabla 22. Evaluación de variantes denominativas - Ingeniería de Software.... **88**
Tabla 23. Relaciones entre términos comunes y científicos proporcionados por CONABIO del corpus de Biodiversidad..... **97**
Tabla 24. Total de Variantes Denominativas - Evaluación #1..... **98**
Tabla 25. Variantes Denominativas con PMI - Evaluación #1..... **99**
Tabla 26. Variantes Denominativas con Frecuencia - Evaluación #1 ..... **101**
Tabla 27. Total de Variantes Denominativas - Evaluación #2..... **104**

Tabla 28. Variantes Denominativas obtenidas con PMI - Evaluación #2 ..... **105**
Tabla 29. Variantes Denominativas obtenidas con Frecuencia - Evaluación #2 ..... **108**
Tabla 30. Base de datos proporcionada por CONABIO **112**
Tabla 31. Porcentaje de términos encontrados en corpus con respecto a los proporcionados por la CONABIO **113**
Tabla 32. Variantes Denominativas con PMI - Evaluación #3 ..... **115**
Tabla 33. Ejemplo de diversos niveles de similitud **115**
Tabla 34. Ejemplo clasificaciones taxonómicas con API de CONABIO **123**
Tabla 35. Totales y rangos de similitud de variantes denominativas con términos de CONABIO **124**
Tabla 36. Total variaciones denominativas utilizando términos del experto en Biodiversidad **126**

Agradecimientos

Esta tesis se llevó a cabo con el apoyo de la beca de maestría otorgada por el CONACyT durante el periodo 2013-2015. Además, fue completada gracias al proyecto CONACyT “Detección y medición automática de similitud textual” con número interno 178248.

Agradezco a mis tutores por toda la paciencia y el gran apoyo que me han brindado a través del desarrollo de esta tesis.

Al Dr. Alejandro Molina Villegas, investigador de la Comisi. n de Nacional para el Conocimiento y Uso de la Biodiversidad, por la asesoría brindada durante la realización de este trabajo en el tema de minería de textos de biodiversidad.

1. Introducción

Hoy en día, a diario se elaboran una gran cantidad de documentos como revistas o periódicos, además se publican diferentes artículos, comentarios y notas en las redes sociales o páginas web, es decir, se genera una gran cantidad de información de diferentes ámbitos, por lo que se ha vuelto muy importante y necesario el manejo de dicha información de manera eficiente. Sin embargo, al ser una enorme cantidad de documentos, se vuelve una tarea sumamente complicada para un ser humano. Por lo que es necesaria la utilización de diversas herramientas automáticas capaces de procesar grandes cantidades de texto y obtener ciertas características o información específica de ellos.

Para manipular, procesar y obtener información de un gran conjunto de documentos hacemos uso de la Lingüística Computacional y de diversas herramientas que nos ayudan al manejo eficiente de grandes cantidades de textos, como extractores automáticos de términos, extractores de definiciones en internet, sistemas automáticos para estudios estilométricos, entre otros.

Específicamente en este trabajo de investigación hicimos uso de una gran cantidad de textos (corpus), los cuales fueron recopilados cuidadosamente dentro de dos áreas de especialidad y para lograr nuestro objetivo principal, desarrollamos una herramienta mediante la cual identificar las posibles variantes denominativas en un área de especialidad. Esto apoyándonos de otras herramientas como extractores automáticos de términos y etiquetadores morfosintácticos, logrando resultados bastante favorables.

Planteamiento del problema

Dentro de las áreas de especialidad existe ambigüedad entre términos, aunque no debería, ya que un término tiene un único concepto asociado a él dentro del dominio de especialidad que se trate, pero ocurre con frecuencia debido a que existen palabras polisémicas, es decir, palabras que pueden tener distintos significados dependiendo del uso que se le esté dando en un determinado

momento, en este caso nos referimos a los términos involucrados dentro de un dominio de especialidad.

Un ejemplo de este problema es el uso de diversos términos dentro de una misma área de especialidad, los cuales hacen referencia a un mismo concepto de la realidad, tal es el caso de los términos “Computadora”, “PC”, “Ordenador” y “Computador” en el área de la Ingeniería de Software, todos ellos se refieren a una máquina electrónica con ciertas características específicas. Sin embargo, su uso depende de varios factores, principalmente el área geográfica.

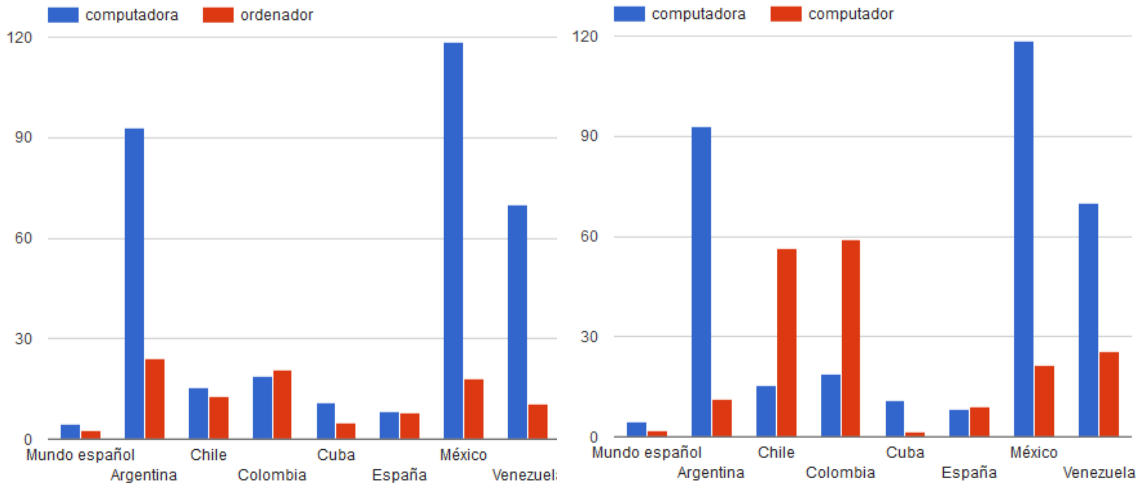


Figura 1. Nivel de uso de los términos: computadora, ordenador y computador

En las gráficas anteriores utilizamos la herramienta Diatopix[1], la cual nos permite observar la distribución geográfica del uso de palabras, términos y expresiones, y mostramos las diferencias entre el uso de los términos “computadora”, “ordenador” y “computador”, donde claramente se nota el predominante uso del término “computadora” en países como Argentina, México y Venezuela, mientras que en España, Chile y Colombia se utiliza más el término “ordenador” o “computador”.

Otro caso que ocurre en algunas áreas de especialidad como la Biología, es el uso de nombres científicos y nombres comunes asociados a un mismo

[1] http://olst.ling.umontreal.ca/diatopix/?lg=es

elemento de la naturaleza, este caso ocurre con algunas frutas o verduras, tal es el caso de la “aceituna”, la cual tiene asociado el nombre científico “Olea Europea” pero comúnmente se le conoce como “olivo”, “aceituno” u “olivera”.
La identificación de este tipo de variantes denominativas es un problema actual, que aunque existen diversas investigaciones relacionadas con este tipo de problemas, no han sido del todo resueltos y en este trabajo de investigación proponemos una solución a este tipo de problema.

Objetivo

El objetivo principal de este trabajo de investigación es obtener, analizar y representar las variantes denominativas dentro de la terminología utilizada en áreas de especialidad de manera automática mediante el uso de diferentes extractores terminológicos y dos corpus especializados de diferentes áreas en idioma español.

Objetivos Específicos

- Conocer la terminología de áreas de especialidad, así como su similitud y variabilidad entre ellas.
- Creación de corpus especializados de diferentes áreas de especialidad (Ingeniería de Software y Biodiversidad).
- Mostrar que el modelo de distribución semántica y el análisis de variantes denominativas se puede aplicar sobre otras lenguas y dominios de especialidad.

Organización de la tesis

En el capítulo 2 presentamos el estado actual de la investigación realizada acerca de la terminología y su relación con la lingüística, así como con la lingüística computacional, mostrando sus diferentes aspectos y características que las conforman, de igual manera mostramos los conceptos y las propiedades principales de los corpus lingüísticos, que son parte esencial para muchos de los trabajos realizados en el área de Procesamiento de Lenguaje Natural, además presentamos una descripción sobre la extracción

terminológica y las diferentes maneras en que es posible realizarla y las tecnologías que se utilizan o se han utilizado hasta el momento.
A lo largo del capítulo mencionamos trabajos realizados por diferentes autores en estas áreas con la finalidad de mostrar un panorama amplio de los antecedentes que sirven de referencia al trabajo de investigación que presentamos en esta tesis. Este capítulo también se enfoca en la semántica distribucional, donde mostramos su origen, sus características, sus variantes, los diferentes modelos que existen actualmente y que se han empleado en diferentes trabajos de investigación, así como algunas de las aplicaciones en las que se ha empleado y/o se emplea actualmente.

En los capítulos 3 y 4 presentamos la extracción terminológica automática que se realizó, comenzando con el proceso de recopilación de los textos que conforman el corpus de trabajo. Abordamos temas como la recuperación de los textos desde la web y la fase de limpiado de dichos textos para su procesamiento. Además, también contemplamos la fase de etiquetado de partes de la oración, ya que hacemos uso de dos etiquetadores automáticos distintos, TreeTagger y Freeling. Finalmente describimos las herramientas utilizadas para la obtención de los términos (Termext y Termostat) y presentamos los términos monoléxicos y poliléxicos obtenidos.

El capítulo 5 contiene el bagaje teórico necesario para sustentar la obtención de las variantes denominativas dentro de las áreas de Ingeniería de Software y la Biodiversidad, mostrando la metodología completa y las diferentes actividades que llevamos a cabo, hasta lograr obtener los términos con diferentes niveles de similitud entre ellos.

En el capítulo 6, Resultados y evaluación, nos concentramos en mostrar la evaluación realizada a la metodología propuesta y utilizada en esta tesis a través del uso de variantes denominativas identificadas previamente por un experto en el área de especialidad de la Biodiversidad, donde aplicamos la metodología descrita en los capítulos anteriores e hicimos uso del corpus de esta misma área.

Finalmente en el capítulo 7, una vez realizadas las diferentes pruebas y de haber sido evaluada la metodología propuesta, presentamos las conclusiones finales con base en los resultados obtenidos y presentados en esta tesis, además mencionamos el posible trabajo a futuro que puede ser realizado dentro del Grupo de Ingeniería Lingüística de la UNAM.

2. Antecedentes

El uso del lenguaje es parte esencial de nuestra vida diaria debido a nuestra necesidad de comunicarnos, y debido a que existen diferentes idiomas, cada uno con características propias por lo que son muy diferentes entre ellos. Por lo anterior, se ha vuelto cada día más importante el estudio sobre diversas características del lenguaje, tanto oral como escrito.

Desde hace ya varios años se han realizado diferentes trabajos de investigación dentro de la UNAM y en diferentes partes del mundo a los diferentes niveles del lenguaje, desde la fonética y la fonología, es decir, desde los sonidos o fonemas, hasta la pragmática que se refiere al estudio que considera el contexto lingüístico. Para ello se favorecen del uso de la tecnología, es decir, el uso de las computadoras, y también se apoyan de otras áreas relacionadas como la ingeniería, la informática y la computación, en lo que se denomina Ingeniería Lingüística, la cual es de gran ayuda en el estudio del lenguaje en sus diferentes niveles, y con la que es posible estudiar grandes cantidades de información.

El empleo de los corpus lingüísticos para diversas investigaciones es uno de los principales elementos dentro de la Ingeniería Lingüística, ya que contienen grandes cantidades de texto con diferentes características, de los cuales es posible obtener rasgos específicos del lenguaje utilizando diversos métodos matemáticos y lingüísticos, como es la clasificación de documentos, la recuperación de información y en este caso la extracción terminológica. El uso de términos es muy importante en las áreas de especialidad, ya que permite una comunicación clara, sin permitir rasgos de ambigüedad. Sin embargo, existe una gran variedad de términos que se refieren a un mismo concepto, en cuanto a su uso y significado, incluso aunque se encuentren dentro de un dominio especializado. A esta variedad de términos se les ha llamado variantes denominativas.

Es necesario, antes de comenzar a explicar la metodología propuesta para la obtención de variantes denominativas dentro de áreas de especialidad,

mencionar algunos conceptos esenciales para su comprensión, así como brindar un panorama general sobre los diversos trabajos relacionados, que utilizamos como referencia para algunos aspectos de la metodología.

Terminología

Primeramente, para poder definir que es la terminología, es necesario establecer claramente la definición y la diferencia entre dos elementos muy importantes, ¿Qué es una palabra? y ¿Qué es un término? Una palabra es una unidad descrita por un conjunto de características lingüísticas sistemáticas y dotada de la propiedad de referirse a un elemento de la realidad, mientras que un término es de igual manera una unidad de características lingüísticas similares, pero el cual es utilizado dentro de un dominio de especialidad, por tanto, se le considera término a toda palabra que forme parte de un ámbito especializado y que, en su interior, tenga un concepto asociado preciso.

Aunque los términos pueden clasificarse de diferentes maneras, existen términos conformados por una sola palabra, llamados monoléxicos, o por más de una palabra, llamados poliléxicos. Pero sin importar el tipo de término del que se esté tratando, todo término es totalmente dependiente de su contexto, es decir, dependiente del área de especialidad donde se encuentre, ya que como se mencionó anteriormente, un término tiene asociado un concepto preciso, el cual no puede ser trasladado a otra área de especialidad. (Cabré M. T., 1993).

Para definir lo que es la terminología es necesario considerar la polisemia de este término, lo cual nos remite por lo menos a tres nociones, a la disciplina, a la práctica y al producto generado. Primeramente, la terminología se concibe como la disciplina que se encarga de los términos especializados. Como práctica es el conjunto de directrices o principios que rigen la recopilación de términos, y finalmente concebida como producto generado, se refiere al conjunto de términos de un área de especialidad.

La terminología como disciplina, posee bases teóricas delimitadas, y un objeto de estudio definido, y como cualquier otra, cuenta con una vertiente teórica y

una práctica. Aunque no es una disciplina estrictamente original, ya que toma elementos de la lingüística, la lexicografía, la morfología y la semántica, estos elementos los reconfigura construyendo su propio campo de estudio, con objetivos y métodos diferentes. Como práctica, o como disciplina aplicada, se diferencia de otras como la lexicografía aplicada por su metodología para la recopilación de datos y su presentación en forma de glosarios de términos (Cabré M. T., 1999).

Finalmente podemos decir que la terminología estudia subconjuntos bien definidos de la lengua (al interior de textos de especialidad), en los que se hallan palabras que se utilizan para nombrar objetos (lógicos y físicos), procesos, acciones, y diversos elementos más, dentro de algún área en particular; en este trabajo, en la Ingeniería de Software y en la Biodiversidad, que son las áreas de especialidad que consideramos a lo largo del desarrollo de esta tesis.

Terminología y las áreas de especialidad

La terminología, vista como un producto, contiene un conjunto de términos de un área de especialidad, por lo que sólo está restringido a un lenguaje de especialidad y debe ser monosémico y carente de sinonimia. Un ejemplo es la palabra *síntesis*, la cual presenta un aspecto polisémico, puesto que su significado depende del contexto, es decir, del área de especialidad en que se encuentre; dentro del campo de la filosofía, su significado se remite a operaciones mentales, en el campo psicológico, representa una conjunción de elementos psíquicos y finalmente en el campo de la química se refiere a la formación de sustancias (Reyes, 2002).

Como se puede observar en el ejemplo anterior, la terminología únicamente puede existir dentro de las áreas de especialidad.

La situación del lenguaje en el mundo ha cambiado drásticamente en las últimas décadas. Los idiomas están sufriendo un doble movimiento de modo aparentemente contradictorio. Por un lado, existe una tendencia a la uniformización lingüística, principalmente en áreas de especialidad, con el

objetivo de asegurar una comunicación eficiente. Por otro lado, también existe una tendencia a mantener y defender la diversidad lingüística respecto a su área geográfica, ambas tendencias constituyen un cambio obvio que está creciendo en nuestros días.

Sin embargo, los cambios lingüísticos, o al menos la mayor parte de los que se han producido hasta ahora, son la consecuencia de cambios sociales y, como consecuencia, son resultado de factores externos que no están directamente vinculados con el lenguaje. Para mostrar un enfoque general de los rasgos más importantes que caracterizan la situación actual de los idiomas, hay que tener en cuenta varios aspectos diferentes:

- **Aspectos socioculturales:** Los idiomas son sistemas culturales que implican una percepción específica del mundo y, al mismo tiempo, establecen una jerarquía de diferentes grupos sociales de acuerdo a su uso del lenguaje.
- **Aspectos psicosociales:** Las personas constituyen una comunidad, y tienen la sensación de integración gracias al idioma que comparten.
- **Aspectos políticos:** La identidad del lenguaje refleja la presencia de una comunidad particular que se hace más fuerte mediante el uso de su idioma.
- **Aspectos económicos:** El idioma es tanto la representación de la situación económica de una sociedad y también uno de los principales factores en la balanza económica del mundo.
- **Aspectos gramaticales y/o lingüísticos:** El idioma es un sistema de recursos. En teoría, este sistema de recursos tiene posibilidades infinitas (todos los idiomas son capaces de expresar todos los conceptos), pero, en la práctica, no todos los idiomas están igualmente desarrollados en el sentido de tener los mismos recursos para lograr la comunicación en todos los ámbitos y en todas las situaciones (Cabré M. T., 1997).

En el caso de esta investigación propiamente, nos enfocamos en los aspectos socioculturales (dialectos geográficos, sociales y cronológicos) y los aspectos

gramaticales y/o lingüísticos (variaciones funcionales) de los términos de dos áreas de especialidad diferentes, tratando de obtener dichas diferencias, con base en el contexto en el que se encuentran dentro de grandes cantidades de textos especializados.

Finalmente, como mencionamos anteriormente, la terminología estudia a los términos y los términos conforman a la terminología, pero para ello debe existir un dominio de especialidad, por lo que su relación es muy importante para esta investigación y debemos conocer más acerca de las áreas de especialidad con las que trata esta investigación.

Ingeniería de Software y la Biodiversidad

La ingeniería de Software y la Biodiversidad son dos áreas de especialidad totalmente distintas, mientras que una se refiere a elementos tecnológicos, administrativos y de uso de la información, la otra hace referencia a elementos de la naturaleza como plantas y animales, entre otros. Esta gran diferencia entre ambas áreas de especialidad, es parte medular para la obtención de variantes denominativas, puesto que la terminología utilizada por cada una de ellas es contrastante y es posible evaluar los resultados de manera clara, con base en nuestra experiencia para el caso de Ingeniería de Software y por parte de expertos de CONABIO para el caso de Biodiversidad.

Antes de comenzar a explicar los antecedentes para la obtención de variantes denominativas y los trabajos realizados dentro del área de Extracción Terminológica y Semántica Distribucional, es necesario tener más información acerca de las áreas de especialidad con las que se realiza el presente trabajo.

Los términos en la Ingeniería de Software

La noción de Ingeniería de Software fue propuesta inicialmente en 1968 durante una conferencia para discutir lo que en ese entonces se llamó la “crisis del software”, la cual fue resultado de la llegada de nuevas computadoras, basadas en circuitos integrados. El software resultante cada vez era de órdenes de magnitud más grande y complejo, aumentando rápidamente su costo y fue necesario el desarrollo y uso de nuevas técnicas y métodos para

controlar su inherente complejidad, lo que dio origen a lo que se conoce como Ingeniería de Software.

Sin embargo existen diferentes conceptos acerca de lo que es la Ingeniería de Software, pero antes de eso, es necesario establecer claramente el concepto de Software, ya que muchas veces este término se asocia únicamente a los programas de computadora, pero su definición va más allá, el software no se conforma únicamente de programas de computadora, sino también los documentos asociados a él y la configuración que se requiere para que funcionen correctamente, por lo que Somerville indica que el software son programas de computadora, así como su documentación asociada, el cual es posible desarrollar para un cliente particular o un mercado general (Sommerville, 2005).

Una vez que se conoce qué es el software, el término Ingeniería de Software se define como una disciplina de la ingeniería que comprende todos los aspectos de la producción de software, desde las etapas iniciales de la especificación del sistema hasta el mantenimiento de éste después de que se utiliza (Sommerville, 2005). Sin embargo, existen otras definiciones, una de ellas dice que la Ingeniería de Software es la disciplina que estudia la naturaleza del software, enfoques y metodologías para el desarrollo de software a gran escala, así como las teorías y leyes que fundamentan el comportamiento de software y de las prácticas de ingeniería de software (Wang, 2007). Finalmente, de acuerdo a la IEEE[2] se define como la aplicación de un planteamiento sistemático, disciplinado y cuantificable al desarrollo, operación y mantenimiento de software (IEEE, 2002).

Con base en las definiciones anteriores podemos decir que la Ingeniería de Software tiene como objetivo el desarrollo de software de calidad y para ello se

[2]El Instituto de Ingeniería Eléctrica y Electrónica es una asociación mundial de ingenieros dedicada a la estandarización y el desarrollo en áreas técnicas. Con cerca de 425 000 miembros y voluntarios en 160 países, es la mayor asociación internacional sin ánimo de lucro formada por profesionales de las nuevas tecnologías.

basa en diversos enfoques y/o metodologías, por lo que su terminología utilizada es bastante amplia. Actualmente existe un glosario de términos publicado en 1990 y actualizado en el 2002, este glosario cuenta con cerca de 1300 términos definidos dentro del campo de la Ingeniería de Software que incluye varios temas como la administración de la configuración, arquitectura de software, el proceso de desarrollo de software y herramientas de software, entre otros. De este glosario de términos fueron excluidos aquellos que son parte especifica de un grupo u organización, una compañía propietaria o una marca registrada y aquellos términos multipalabra cuyo significado puede deducirse directamente del significado de las palabras que los componen (IEEE, 2002).

Este glosario se encuentra en idioma inglés y actualmente no existe un glosario de términos que se encuentre en idioma español, por ello es importante el recopilar textos en español de diversas áreas geográficas, no con la finalidad de obtener un glosario automático, sino conocer los términos que se utilizan y las variantes denominativas que existen entre ellos.

addressing exception. An exception that occurs when a program calculates an address outside the bounds of the storage available to it. *See also:* **data exception; operation exception; overflow exception; protection exception; underflow exception.**

afferent. Pertaining to a flow of data or control from a subordinate module to a superordinate module in a software system. *Contrast with:* **efferent.**

algebraic language. A programming language that permits the construction of statements resembling algebraic expressions, such as $Y = X + 5$. For example, FORTRAN. *See also:* **algorithmic language; list processing language; logic programming language.**

algorithm. (1) A finite set of well-defined rules for the solution of a problem in a finite number of steps; for example, a complete specification of a sequence of arithmetic operations for evaluating sine x to a given precision.

(2) Any sequence of operations for performing a specific task.

algorithmic language. A programming language designed for expressing algorithms; for example, ALGOL. *See also:* **algebraic language; list processing language; logic programming language.**

Figura 2. Extracto del Glosario de Términos de la Ingeniería de Software de la IEEE

Dentro de esta área de especialidad existen trabajos relacionados con el uso de términos y con el área de Procesamiento de Lenguaje Natural (PLN). Por PLN se entiende la habilidad de una máquina para procesar la información comunicada, no sólo las letras o los sonidos del lenguaje (Gelbukh, 2007). Por

ejemplo la asistencia al proceso de Interpretación de textos en la Ingeniería de Software, el cual se basa en la generación automática, a partir del texto, de un esquema conceptual utilizado en ingeniería de software llamado diagrama Entidad-Relación (ER) y los resultados obtenidos muestran cómo el diagrama ER puede ser una valiosa herramienta de apoyo al proceso de interpretación, gracias a las inferencias que, de manera automática, se realizan a través de él. (Jaramillo, Zapata, & Isaza, 2007). Otro trabajo realizado dentro del área es la detección de reportes de defectos duplicados, dichos informes se generan a partir de diversas pruebas y actividades de desarrollo en la ingeniería de software y algunas veces se describe el mismo problema más de una vez, llevando a informes duplicados. Estos informes están escritos en su mayoría en lenguaje natural y la evaluación muestra que aproximadamente 2/3 de los duplicados, posiblemente, se pueden encontrar utilizando técnicas del procesamiento del lenguaje natural (Runeson, Alexandersson, & Nyholm , 2007).

En este trabajo de investigación obtenemos variantes denominativas, como por ejemplo el término “computadora” utilizado en México, el cual también es conocido y utilizado como “ordenador” en otros lugares como España; otras variantes denominativas para éste mismo término son: “computador” o “PC”, dependiendo la región geográfica. Estas variantes identificadas están relacionadas a un mismo concepto físico y cualquiera de ellos es correcto. En este trabajo de investigación obtuvimos diversas variantes denominativas entre los términos extraídos automáticamente con diferentes herramientas de extracción terminológica automática.

Los términos en la Biodiversidad

Existen diferentes conceptos acerca de lo que es la Biodiversidad en México; para la SEMARNAT (Secretaria de Medio Ambiente y Recursos Naturales), la biodiversidad es un concepto que abarca a toda la variedad de la vida, incluyendo a los ecosistemas y a los complejos ecológicos de los que forma parte. Por lo que tiene tres escalas *grosso modo*: ecosistemas, especies y genes (INECC, 2013). Mientras que para la Comisión Nacional para el Conocimiento y Uso de la Biodiversidad, la biodiversidad o diversidad biológica

se refiere a la variedad de la vida. Este reciente concepto incluye varios niveles de la organización biológica y abarca a la diversidad de especies de plantas, animales, hongos y microorganismos que viven en un espacio determinado, a su variabilidad genética, a los ecosistemas de los cuales forman parte estas especies y a los paisajes o regiones en donde se ubican los ecosistemas[3], incluyendo los procesos ecológicos y evolutivos que se dan a nivel de genes, especies, ecosistemas y paisajes (CNCUB, CONABIO, 2012).

Como ya se mencionó, en el área de Biodiversidad existen una gran cantidad de elementos que la integran, por lo que existe una gran cantidad de términos, ya que se nombra de manera muy específica a cada uno de los elementos que la conforman y requieren tener un orden, por lo que se recurre a la Taxonomía y a la Nomenclatura como parte de la organización de los términos en esta área. La Taxonomía es una disciplina científica que propone una clasificación científica de los organismos vivos en unidades biológicas particulares llamados taxones o taxa. La Nomenclatura es una disciplina de la taxonomía que incluye la teoría y reglas para nombrar a dichos taxones o taxa y que resulta en la aplicación de las reglas de nomenclatura de la taxonomía de un grupo de organismos (Dubois, 2012).

Los niveles taxonómicos según la CONABIO son:

1. Reino
2. División
3. Clase
4. Subclase
5. Orden
6. Suborden
7. Familia
8. Subfamilia
9. Tribu
10. Subtribu
11. Género
12. Subgénero
13. Sección
14. Subsección
15. Serie
16. Subserie
17. Especie
18. Subespecie
19. Variedad
20. Subvariedad
21. Forma
22. Subforma

[3] El ecosistema es el conjunto de especies de un área determinada que interactúan entre ellas y con su ambiente abiótico; mediante procesos como la depredación, el parasitismo, la competencia y la simbiosis, y con su ambiente al desintegrarse y volver a ser parte del ciclo de energía y de nutrientes. Las especies del ecosistema, incluyendo bacterias, hongos, plantas y animales dependen unas de otras.

Estos niveles taxonómicos van de lo general a lo más específico y no se aplican para todos los organismos vivos, sólo en algunos grupos grandes y complejos llegan a utilizarse el total de las categorías. Estos niveles taxonómicos dan orden y accesibilidad a la clasificación de los organismos y proporciona un importante sistema de entrada y recuperación de información (CAT, CONABIO, 2013).

Por ejemplo, el Leopardo (nombre común) pertenece al reino Animalia, a la división Chordata, a la clase Mammalia, del orden Carnivora, de la familia Felidae, del género Panthera y finalmente a la especie Panthera Pardus. El conocer dichos niveles taxonómicos nos permite conocer la relación que existe entre los elementos de la biodiversidad. En este trabajo son vistos como términos, que los expertos han realizado con el paso del tiempo y mediante la metodología propuesta en este trabajo de tesis, nos permite validar y/o comparar si los resultados obtenidos como variantes denominativas se acercan a los establecidos por los expertos en esta área de especialidad.

Debido a su ubicación geográfica y a su diverso relieve, México tiene una gran diversidad de ecosistemas, que van desde lo más alto de las montañas hasta los mares profundos, pasando por desiertos y arrecifes de coral, bosques nublados y lagunas costeras. Esto influyó a que la CONABIO pudiera recopilar textos de diversas regiones del país, en los que se contiene información acerca de la amplia biodiversidad del país.

Lingüística Computacional

Ahora que se ha profundizado un poco más sobre la terminología y las áreas de especialidad con las que se trabajará, es necesario conocer acerca de lo que es la lingüística computacional, esto para entender como mediante está disciplina es posible lograr el objetivo de este trabajo de investigación: obtener variantes denominativas entre los términos de las áreas de especialidad de Ingeniería de Software y la Biodiversidad.

La Lingüística Computacional constituye un campo científico de carácter interdisciplinario entre la lingüística y la informática, cuyo fin fundamental es la elaboración de modelos computacionales que reproduzcan distintos aspectos del lenguaje humano y que faciliten el tratamiento informatizado de las lenguas. A este amplísimo campo de estudio también se le denomina lingüística informática, procesamiento del lenguaje e incluso ingeniería lingüística.

A pesar de constituir una disciplina relativamente reciente, en este campo se reconocen dos líneas fundamentales de actuación e investigación: la lingüística computacional teórica y la lingüística computacional aplicada, a las que se le añade una tercera: la informática aplicada a la lingüística.

La primera de estas líneas, la lingüística computacional teórica, se encarga de la consecución de tres objetivos complementarios: la elaboración de modelos lingüísticos en términos formales e implementables, la aplicación de dichos modelos a los diferentes niveles de descripción lingüística, y la comprobación computacional de la congruencia del modelo y de sus predicciones. En segundo lugar, la lingüística computacional aplicada, supone una orientación más tecnológica de la lingüística computacional, que se centra en el diseño de sistemas informáticos capaces de administrar, comprender, producir y traducir enunciados orales y escritos en lenguaje natural, por lo que desarrolla aplicaciones informáticas que pueden agruparse en cuatro grandes categorías:

1) Sistemas de consulta a bases de datos a través del lenguaje natural.
2) Aplicaciones de las tecnologías del habla, como los sistemas de conversión de texto a voz.
3) Herramientas para el procesamiento de textos para la elaboración, gestión y revisión de documentos, los programas de generación automática de resúmenes, los sistemas de extracción de información y los de catalogación documental automatizada.
4) Las herramientas orientadas al procesamiento de más de una lengua o de una lengua extranjera, como son las aplicaciones didácticas para la

enseñanza de las lenguas, las herramientas de traducción (semi)automática, las memorias de traducción y las bases de datos terminológicas.

Finalmente, el campo de trabajo que se caracteriza por la aplicación de computadoras a la investigación lingüística, es decir, el estudio científico del lenguaje, se denomina informática aplicada a la lingüística o lingüística informática, la cual puede aplicarse en un sentido amplio a todas las disciplinas de la lingüística que usan herramientas informáticas para sus estudios (Perez & Moreno, 2009).

Actualmente no se sabe muy bien cómo las personas producen y comprenden el lenguaje, aún nuestra comprensión de los mecanismos del procesamiento del lenguaje humano sigue creciendo. Los modelos de estos mecanismos a través de herramientas computacionales, también nos ayudan a descubrir y describir propiedades ocultas del lenguaje humano que son relevantes para cualquier tipo de procesamiento del lenguaje incluyendo a las aplicaciones de software. El objetivo de la Lingüística computacional es la comprensión profunda del lenguaje humano, además el desarrollo de aplicaciones lingüísticas inteligentes, sin embargo, incluso las tecnologías del lenguaje de hoy que hacen uso de técnicas de procesamiento de poca profundidad se pueden convertir en productos de software de gran utilidad (Uszkoreit, 2000).

Las prácticas científicas y técnicas utilizadas en la lingüística computacional provienen de diversas áreas de especialidad, como la lingüística, la informática, la psicología y las matemáticas. En esta tesis, para la obtención de variantes denominativas, hacemos uso de la lingüística computacional, específicamente de la lingüística (terminología), la informática (elaboración de programas de automatización de tareas) y las matemáticas (Semántica Distribucional).

Variaciones en los términos

El español es una lengua muy rica en su léxico y su morfología. Esto implica que el español tiene una gran productividad y flexibilidad en sus mecanismos de formación de palabras y a su vez en la formación de términos (Vilares, Cabrero, & Alonso, 2001). Como ya se mencionó anteriormente, los términos están asociados a un concepto preciso dentro de un área de especialidad, sin embargo, al momento de referirse a ellos de manera escrita, existen algunas variaciones entre un mismo término, considerando los siguientes tipos de variaciones:

- Ortográfica: como el uso de guiones y barras (Aminoácidos y amino-ácidos), el uso de mayúsculas y minúsculas (NF-KB y NF-kb), ciertas variaciones ortográficas (“Tumor” y “El tumor”), diferentes transcripciones, etc.
- Morfológicas: las variaciones más simples son en relación con los fenómenos de inflexión (por ejemplo: singular, plural), además algunas transformaciones derivativas pueden conducir a variantes en algunos casos (muchachas, muchachada).
- Léxica: auténticos sinónimos léxicos, que pueden ser utilizados de forma intercambiable (carcinoma y cáncer, hemorragia y pérdida de sangre).
- Estructural: por ejemplo, el uso de sustantivos posesivos utilizando preposiciones (clones de humano y clones humanos), variantes prepositivas (células en la sangre y células de la sangre), coordinaciones (glándulas suprarrenales y gónadas).
- Acrónimos y abreviaturas: son fenómenos frecuentes de variación de un término técnico (ADN para el ácido desoxirribonucleico).

Estas variaciones escritas de los términos no cambian su significado y continúan haciendo referencia a un mismo concepto. En diversos trabajos de investigación, la obtención de estas variantes se trata como una tarea independiente a la

extracción y/o identificación de términos (Nenadié, Ananiadou, & McNaught, 2004).

Durante los últimos años, se han realizado diferentes trabajos, tratando de obtener diversas variantes de términos, como es la obtención y expansión de enlaces de hiperónimos[4] como parte de la llamada terminología computacional para el diseño de herramientas para la adquisición, la estructuración y la explotación de datos terminológicos. Estas herramientas en primer lugar, presentan un sistema para la adquisición basada en el corpus de las relaciones terminológicas a través de patrones discursivos, y en segundo lugar, se muestra cómo enlaces de hiperónimos entre los términos de una sola palabra pueden extenderse a enlaces semánticos entre términos de varias palabras a través de la extracción basada en corpus de variantes semánticas (Morin & Jacquemin, 2004).

Otros trabajos se han enfocado a la obtención de sinónimos entre términos, lo cual no es una tarea fácil, enfocándose a términos multipalabra a partir de corpus especializados en el área de energía de diferentes idiomas (español, francés e inglés), donde proponen un método semicomposicional no supervisado que hace uso de la semántica distribucional (Hazem & Daille, 2014). En otro trabajo donde de igual manera se busca conocer si dos significantes tienen el mismo o muy similar contenido semántico, demuestra la forma en que sinónimos de términos médicos se pueden extraer de forma automática a partir de un amplio corpus de textos clínicos utilizando también la semántica distribucional (Henriksson, Moen, Skeppstedt, Eklund, Daudaravicius, & Hassel, 2012).

Aunque el presente trabajo de investigación no está enfocado a la obtención de sinonimia entre términos, los trabajos previos realizados en esta área son una

[4] Palabra cuyo significado está incluido en el de otras, por ejemplo: pájaro es hiperónimo de jilguero y de gorrión.

base firme para la elaboración de la metodología propuesta para la obtención de variantes denominativas de términos.

Corpus lingüísticos

Definir de forma clara qué es un corpus lingüístico no es una tarea sencilla, etimológicamente proviene del latín corpus que se refiere a un cuerpo, por lo que se puede definir como un conjunto de textos de lenguaje natural e irrestricto, almacenados en un formato electrónico homogéneo, seleccionados y ordenados de acuerdo con criterios explícitos para ser utilizados como modelo de un estado o nivel de lengua determinado, en estudios o aplicaciones relacionados en mayor o menor medida con el análisis lingüístico (Perez & Moreno, 2009).

De igual manera encontramos otra definición muy parecida que indica que un corpus lingüístico consiste en la recopilación de un conjunto de textos de materiales escritos y/o hablados, agrupados bajo un conjunto de criterios mínimos, para realizar ciertos análisis lingüísticos (Sierra, 2008). Finalmente cada vez es mayor la necesidad de utilizar recursos informáticos en diversas áreas de investigación, y para ello es necesario disponer de materia prima para realizar dichos análisis lingüísticos, en este caso textos, ya sean orales o escritos, los cuales debidamente recopilados, forman un corpus lingüístico (Torruella & Llisterri, 1999).

Como podemos observar, son dos las definiciones anteriores que son muy similares, en esencia podemos decir que un corpus lingüístico es un conjunto de textos, orales o escritos, recopilados bajo ciertas características específicas. La función principal de un corpus, tanto escrito como oral, es establecer la relación entre la teoría y los datos. Un corpus tiene que mostrar a pequeña escala cómo funciona una lengua natural pero para ello es necesario que esté diseñado correctamente sobre unas bases estadísticas apropiadas que aseguren que el resultado sea efectivamente un modelo de la realidad. Teniendo siempre presente que un corpus nunca puede ser la realidad sino solamente un modelo

de ésta, el cual debe mostrar sus aspectos más destacados y más característicos, ya que cuanto más grande sea el corpus y el número de características de los textos que lo integran, más posibilidades habrá de asegurar la presencia de todos los aspectos de la lengua y, por lo tanto, de acercarse a la realidad. Finalmente, un corpus siempre tiene que ser selectivo ya que no es posible recopilar todo lo escrito y/o hablado de una lengua, y operativamente, es preferible un corpus bien seleccionado y representativo a un corpus exhaustivo.

Actualmente existe un gran número de corpus, muy variados con lo que respecta a su extensión, a su diseño y a sus finalidades. El hecho es que los corpus han demostrado ser una herramienta excelente para muchos tipos de investigaciones, principalmente en el campo de la investigación lingüística porque, como ya se mencionó, proporcionan bases mucho más reales para el estudio de las lenguas que los métodos intuitivos tradicionales.

Algunos corpus existentes que han sido y pueden ser usados para diferentes investigaciones son: Brown Corpus, Corpus del Español Mexicano Contemporáneo (CEMC), Corpus Diacrónico del Español (CORDE), Corpus de Referencia del Español Actual (CREA), Corpus del Español de Mark Davies, además de otros corpus creados dentro del Grupo de Ingeniería Lingüística como el Corpus Lingüístico en Ingeniería (CLI), Corpus de las Sexualidades en México (CSMX), Corpus Histórico del Español en México (CHEM) o el Corpus de Contextos Definitorios (CORCODE).

Una de las muchas aplicaciones que existen para los corpus es la enseñanza de diferentes idiomas, como es la enseñanza de inglés para fines específicos, por ejemplo el caso de inglés para Turismo. Para conseguir una amplia muestra de variedades geográficas del inglés es necesario diversificar el origen de las muestras del corpus dividiéndolas en varias zonas geográficas, además de obtener textos de diversas fuentes como libros, periódicos, revistas, folletos,

guías de viaje, correspondencia formal, y para el caso de lenguaje oral, radio, televisión y entrevistas formales. De esta forma es posible conformar una muestra más cercana a la realidad del uso del idioma inglés en el turismo, tal como lo muestra el trabajo de Rafael Rocamora dentro de la Universidad de turismo de Murcia (Abellán, 1999).

Otro de los campos de investigación y de estudio que más se beneficia de la información que los corpus textuales y los corpus de lengua oral aportan es la terminología. Estos corpus son de gran ayuda para configurar el lemario de los diccionarios (tanto para incluir nuevas palabras como para excluir aquellas en desuso), así como para separar las distintas acepciones de cada lema, para detectar las palabras coocurrentes, las combinaciones sintácticas, entre otras tareas más. En el caso específico de este trabajo de investigación, servirá para obtener la terminología utilizada en diferentes áreas de especialidad, y posteriormente para identificar de manera automática variantes denominativas entre dichos términos.

Clasificación de corpus lingüísticos

Los corpus se pueden clasificar de diferentes maneras, principalmente existen dos grandes clasificaciones, los “corpus orales” y los “corpus escritos”. Los corpus escritos se refieren claramente a muestras de la lengua escrita (libros, revistas, periódicos), mientras que los corpus orales pueden referirse solo a transcripciones de la lengua hablada o incluso a conjuntos de grabaciones con sus transcripciones ortográficas, las cuales pueden haber sido recopiladas a través de diferentes medios, lugares y en diversas circunstancias (entrevistas espontáneas, radio, televisión).

Otra manera en que los corpus se pueden clasificar, es en función de los parámetros que se quieran utilizar:

- Según el porcentaje y la distribución de los diferentes tipos de textos que lo componen, es decir, pueden ser grandes, donde no se plantea el límite

del volumen de textos que ha de contener o que, si se lo plantea, lo cuantifica en un número de palabras muy elevado sin tener en cuenta cuestiones de equilibrio, de representatividad. Puede ser equilibrado con diferentes variedades de textos distribuidos cuantitativamente en proporciones parecidas para cada variedad, o finalmente pueden ser multilingües o paralelos con textos traducidos a una o varias lenguas.

- Según la especificidad de los textos que lo componen, pueden ser generales, los cuales pretenden reflejar la lengua común en su ámbito más amplio, o pueden ser específicos, los cuales se oponen al corpus general ya que recogen textos que puedan aportar datos para la descripción de un tipo particular de lengua.
- Según la cantidad de texto que se recoge en cada documento, se dividen en corpus textuales, de referencia y léxicos, donde el primero recoge íntegramente todos los textos de los documentos que lo constituyen, el segundo está formado por fragmentos de los textos de los documentos que lo constituyen y el tercero por fragmentos de textos muy pequeños y de longitud constante de cada documento.
- Según la codificación y las anotaciones añadidas a los textos, donde únicamente puede ser un corpus anotado o no.
- Según la documentación que le acompañe, donde puede o no contener documentación.

Un corpus bien formado, al menos debe corresponder a alguna de las clasificaciones descritas anteriormente. Estos tipos de corpus, además, pueden contener características específicas que dependen del tipo de investigación o el nivel de análisis de la lengua que se quiera realizar (Torruella & Llisterri, 1999).

Los corpus lingüísticos, como ya se mencionó, son un recurso medular dentro de la lingüística computacional, el cual debe cumplir una serie de características esenciales como representatividad, variedad y equilibrio.

Representatividad

Para alcanzar los objetivos en los diferentes proyectos que se realizan en el área de la lingüística computacional, dado que no es posible obtener o recolectar la totalidad de los textos que hacen referencia a una lengua, una región geográfica o un periodo de tiempo determinado, se debe obtener únicamente una muestra que sea representativa, es decir, es necesario tener cierta organización o estructura sobre lo que se desea analizar o estudiar.

Variedad

La variedad de un corpus está estrechamente ligada a los objetivos a alcanzar, es decir, dependiendo de que se desea obtener o que se desea analizar se puede conformar el corpus con textos de diferentes tópicos, regiones, o distinto tiempo, logrando así una variedad en el corpus.

Equilibrio

El equilibrio significa que el material contenido para cada uno de los rubros sea relativamente proporcional, evitando ser tendencioso a una parte únicamente (Vivaldi, Cabrera-Diego, Sierra, & Pozzi, 2012), la representatividad requiere de variedad de textos, pero se debe evitar que esos textos únicamente pertenezcan a un tópico y/o rubro o exista una cantidad mayor de textos sobre ellos, ya que de ser así los resultados estarán sesgados y pueden ser poco precisos.

Un corpus que ejemplifica claramente estas características es el Corpus Nacional Británico, el cual es una colección de 100 millones de palabras de muestras del lenguaje hablado y escrito a partir de una amplia gama de fuentes, cuyo objetivo es representar una amplia sección del inglés británico, este corpus está clasificado como:

- Monolingüe: Se trata únicamente del idioma inglés británico, no incluye otros idiomas. Sin embargo palabras de otra lengua aparecen en el corpus.

- Sincrónico: Cubre el Inglés Británico de finales del siglo XX, no el desarrollo histórico que lo produjo.
- General: Incluye muchos estilos y variedades diferentes, y no se limita a ningún campo, tema o género en particular, además de contener ejemplos tanto de lenguaje hablado como escrito.
- Léxico: Para las fuentes escritas, las muestras de 45.000 palabras están tomadas de diferentes partes de los textos de un solo autor; los textos más cortos, hasta un máximo de 45.000 palabras, o textos de varios autores, revistas y periódicos, se incluyen en su totalidad.

Además, este corpus está formado por dos etapas principales: la planificación (etapa de diseño) y la ejecución (etapa de creación) (Burnard, 2009).

Necesidad de un corpus en un área de especialidad

Como ya se mencionó anteriormente, los corpus lingüísticos son un elemento medular para la lingüística computacional, específicamente para el análisis de diversas características dentro de la lengua. Actualmente existen corpus generales y especializados de diversas áreas, pero no en la Ingeniería de Software y la Biodiversidad, mucho menos investigaciones dentro de estas áreas de especialidad, es por ello que es necesaria la recopilación y conformación de un corpus de cada área de especialidad antes mencionada.

Al contar con ambos corpus de especialidad es posible realizar la extracción terminológica y posteriormente la obtención de las variantes denominativas entre los términos obtenidos de cada una de las áreas, siguiendo la metodología propuesta en este trabajo de investigación.

Extracción terminológica

Como ya mencionamos anteriormente, un término es toda palabra que forme parte de un ámbito especializado y que, en su interior, tenga un concepto preciso asociado, considerando también que un término no está limitado a estar

constituido por una sola palabra. La identificación u obtención de los términos es una tarea que hasta hace unos años únicamente podía ser realizada por expertos de un área de especialidad, lo cual generalmente es muy costoso y toma mucho tiempo, pero con la ayuda de la lingüística computacional y el uso de corpus lingüísticos, se han realizado diversas investigaciones para lograr identificar, de manera automática, los términos o la terminología relacionada a un área de especialidad y en diferentes idiomas como inglés, francés y español.

La extracción terminológica es una actividad dentro de la extracción de información, que consiste en la obtención automática de posibles unidades terminológicas (candidatos a término) a partir de un corpus especializado, respondiendo a las necesidades de investigación en los ámbitos de lexicografía y terminología, además contribuye al desarrollo de numerosas aplicaciones en Procesamiento del Lenguaje Natural, tales como la construcción de glosarios, vocabularios y diccionarios de especialidad, la indexación de textos, la traducción automática, etc. (Sierra, 2008).

Cada día las necesidades de información son mayores y se han desarrollado diversas aplicaciones, como los rastreadores web[5] o los sistemas de recomendación[6] que basan su funcionamiento en la web semántica[7], pero para ello es necesario conocer el vocabulario de términos utilizado en un dominio

[5] Programa que inspecciona las páginas web de forma metódica y automatizada, cuyo funcionamiento normal es: dado un grupo de direcciones electrónicas iniciales, el programa descarga las direcciones, analiza las páginas y busca enlaces a páginas nuevas.

[6] Programa de filtro de información, que presenta distintos tipos de temas de información como películas, música, libros, noticias, imágenes, páginas web que son del interés de un usuario en particular.

[7] http://www.w3c.es/Divulgacion/GuiasBreves/WebSemantica La Web Semántica es una Web extendida, dotada de mayor significado en la que cualquier usuario en Internet podrá encontrar respuestas a sus preguntas de forma más rápida y sencilla gracias a una información mejor definida.

especializado, por lo que el desarrollo en el área de extracción de terminológica es esencial para el avance y el crecimiento en la lingüística computacional.

Extractores automáticos de términos

Desde hace ya varios años se ha trabajado en la extracción terminológica, aunque continúa sin ser una tarea sencilla. Lingüistas computacionales, lingüistas, traductores, intérpretes, periodistas científicos e ingenieros computacionales se han interesado por aislar automáticamente la terminología de los textos, esto ha llevado a diferentes grupos de profesionales a diseñar herramientas de software con el fin de extraer directamente la terminología de los textos con diferentes finalidades como: la construcción de glosarios, vocabularios y diccionarios terminológicos; la indexación de texto; traducción automática; construcción de bases de datos de conocimiento; construcción de sistemas de hipertexto; construcción de sistemas expertos y análisis del corpus.

Basados en Lingüística

Este tipo de extractores basan al menos alguna de sus etapas, en la explotación de conocimiento lingüístico para la adquisición de los candidatos a ser términos, como lo es el etiquetado de las partes de la oración (POS); entre ellos se encuentra Lexter[8], un extractor de términos para el procesamiento de documentos en francés que fue programado en lenguaje C[9], basado en conocimiento lingüístico que requiere contar con un conjunto de términos del área de especialidad tratada.

Basados en Estadística

Estos basan su funcionamiento exclusivamente en conocimiento estadístico, como la frecuencia de las palabras o la información mutua, que calcula matemáticamente la similitud entre dos conjuntos de palabras, entre ellos se

[8] (Bourigault, Gonzalez-Mullier, & Gros, 1996)

[9] C es un lenguaje de programación originalmente desarrollado por Dennis M. Ritchie entre 1969 y 1972 en los Laboratorios Bell. Es un lenguaje de tipos de datos estáticos, débilmente tipificado, de medio nivel pero con muchas características de bajo nivel.

encuentra ANA[10], un extractor que al no requerir conocimiento lingüístico, es independiente de la lengua tratada.

Híbridos

Este tipo de extractores no se encuentran basados exclusivamente en lingüística o estadística y se ha demostrado que la combinación de ambas técnicas genera mejores resultados, ya que logran localizar términos que por un sólo método no se hubieran detectado y, sobre todo, descartan candidatos a términos que en realidad no lo son (Sierra, Villaseñor, & Barrón, 2006), entre ellos se encuentran ACABIT[11], TermoStat[12], TerMine[13] y TERMEXT.

En un estudio reciente se ha hecho una descripción de algunos de los extractores de términos que se han desarrollado para varios idiomas como el inglés, francés y alemán, desde la aparición de TERMINO[14] (el primer detector terminológico conocido) en 1990, hasta otros extractores automáticos que utilizan diversas técnicas y/o metodologías para la obtención de términos en diferentes áreas de especialidad con la finalidad de apoyar en actividades lexicográficas o de traducción, (Cabré, Estopà, & Vivaldi, 2001).

[10] (Enguehard & Pantera, 1995)

[11] (Daille, 1996)

[12] http://termostat.ling.umontreal.ca/doc_termostat/doc_termostat.html

[13] http://www.nactem.ac.uk/software/termine/

[14] (David & Plante , 1990)

System		Term Filtering					
	Name /Author	Freq.[10]	Linguistic	statistical + linguistic	linguistic + statistical	reference terms	user defined
1	ANA					X	
2	CLARIT			X	X		
3	Daille				X		
4	FASTR					X	
5	Heid		X				
6	LEXTER		X				
7	Naulleau						X
8	NEURAL				X		
9	NODALIDA-95		X				
10	Termight	X	X				
11	TERMINO		X				
12	TERMS	X	X				

Figura 3. Extractores terminológicos y su filtro

En la figura 3 (Cabré, Estopà, & Vivaldi, 2001) se muestran algunos extractores automáticos de términos, indicando cuales son los filtros que utiliza para su obtención (frecuencia, lingüístico, estadístico y lingüístico, etc.). Sin embargo, poco se ha hecho sobre extracción automática de términos en documentos en español, como se muestra en la figura 4 (Cabré, Estopà, & Vivaldi, 2001).

System		Test corpora			Terms %	
	Name /Author	Domain	Language	Size.[Kw.]	precision	recall
1	ANA	Aviation engineering	French	120		?
		Acoustics	English	25	?	?
2	CLARIT[11]	News	English	240 Mb	-	81.6
3	Daille	Telecommunications	French	800	?	?
4	FASTR	Medicine (abstracts)	French	1.560	86.7	74.9
5	Heid	Engineering	German	35	?	?
6	LEXTER	Engineering	French	3.250	95	?
7	Naulleau	Technical	French	?	?	?
8	NEURAL	Medicine	English	55	?	70
9	NODALIDA-95	Cosmology Technical text	English	20	95-98	98.5-100
10	Termight	Computer science	English	?	?	?
11	TERMINO	Medicine	French	?	72	70-74
12	TERMS	Statistics	English	2.3	77	
		Semantics		6.3	86	
		Chromatography		14.9	96	

Figura 4. Extractores terminológicos, dominio y características

Existen algunas investigaciones, las cuales proponen diversas metodologías para su extracción, haciendo uso de técnicas lingüísticas, estadísticas e hibridas y están inmersas en diferentes áreas de especialidad.

Uno de los trabajos relacionados es el realizado por Jorge Lázaro quien propone una metodología de trabajo para la extracción de términos en el área de la sexualidad utilizando un corpus de textos extraído de Internet, comenzando con la eliminación de las palabras funcionales y posteriormente haciendo uso de la herramienta WordSmith[15] y análisis manuales para obtener el "Vocabulario básico de las sexualidades en México" (Sierra, Medina, & Lázaro, 2009). Siguiendo la misma línea de investigación, se encuentra el trabajo de Antonio Reyes, que de igual forma hace uso de WordSmith, con la diferencia que realiza la comparación de listas de palabras utilizando textos de dos diferentes áreas de especialidad, la Física y la Literatura, en este caso Octavio Paz, con lo que obtiene buenos candidatos a términos (Reyes, 2002).

Otro trabajo de investigación, fue el realizado por Luis Adrián Cabrera, cuyo propósito fue obtener el vocabulario científico básico en el español de México, donde como parte del trabajo crea el Corpus Básico Científico del Español de México (COCIEM)[16] que utiliza junto con el extractor terminológico YATE[17], obteniendo resultados favorables al evaluar sus términos con Wikipedia (Barrón, Sierra, & Villaseñor, 2006) (Vivaldi, Cabrera-Diego, Sierra, & Pozzi, 2012)

Finalmente unos de los trabajos que nos sirvió de base para el desarrollo de este trabajo de investigación, es decir, la obtención de variantes denominativas, fue el realizado por Luis Alberto Barrón, quien propuso una adaptación a la etapa C-value del algoritmo C-value/NC-value, que se ha implementado exitosamente para la extracción automática de términos sobre documentos en inglés, para la extracción de términos monoléxicos y poliléxicos en español, utilizando patrones sintácticos como sustantivo, sustantivo + adjetivo, sustantivo + sustantivo, entre otros (Barrón, Sierra, Drouin, & Ananiadou, 2009).

---

[15] http://www.lexically.net/wordsmith/

[16] www.corpus.unam.mx/cociem/

[17] http://eines.iula.upf.edu/cgi-bin/Yate-on-the-Web/yotwMain.pl?lang=ES

Como parte del desarrollo de este trabajo se creó la herramienta TERMEXT[18], el cual es un extractor terminológico automático que está compuesto por un conjunto de programas individuales que realizan cada una de las etapas necesarias en el proceso de búsqueda de términos en un conjunto de documentos. El proceso va desde el etiquetado morfosintáctico hasta la generación final de la lista de términos (Barrón, Sierra, & Villaseñor, C-value aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español., 2006).

Semántica Distribucional

Una importante necesidad en la terminología, como ya se mencionó, es lograr identificar las diversas variantes denominativas, que existen dentro de las áreas de especialidad, un elemento importante para identificarlas es la Semántica Distribucional. La Semántica Distribucional es un área de investigación que desarrolla y estudia las teorías y métodos para cuantificar y clasificar similitudes semánticas entre elementos lingüísticos en función de sus propiedades de distribución en grandes cantidades de datos. Su idea básica se puede resumir en la llamada hipótesis de distribución: “Los elementos lingüísticos con contextos similares tienen significados similares.” (Harris, 1954), y basa su uso en el álgebra lineal como herramienta computacional y marco de representación.

Harris describe cómo cada idioma puede describirse en términos de una estructura distributiva, es decir, en términos de la aparición de elementos lingüísticos en relación con otros elementos lingüísticos, y cómo esta descripción es completa sin la intrusión de otras características tales como su historia o su significado. La distribución de un elemento se entiende como la suma de todos sus contextos, es decir, la estructura distribucional de un elemento lingüístico no

[18] http://www.corpus.unam.mx/termext/index.php

es otra cosa que ese elemento lingüístico dentro de un arreglo que especifique los elementos que lo rodean y sus posiciones determinadas (Harris, 1954).

Un ejemplo claro de aplicación de esta hipótesis distribucional de Harris es el mostrado en el trabajo de Stefan Evert (Stefan, 2010) donde dado un elemento lingüístico, en este caso la palabra “bardiwac” cuyo significado se desconoce, y una serie de enunciados, es posible conocer el significado de dicha palabra.

- He handed her glass of bardiwac.
- Beef dishes are made to complement the bardiwacs.
- Nigel staggered to his feet, face flushed from too much bardiwac.
- Malbec, one of the lesser-known bardiwac grapes, responds well to Australia’s sunshine.
- I dined off bread and cheese and this excellent bardiwac.
- The drinks were delicious: blood-red bardiwac as well as light, sweet Rhenish.

⇒ Finalmente se concluye que la palabra “bardiwac” se refiere a una bebida alcohólica de color rojo, elaborada con uvas.

Siguiendo con la hipótesis de Harris, una manera de representar la estructura distribucional de un elemento es mediante una matriz de coocurrencia en la que se especifican los elementos que lo rodean.

	get	see	use	hear	eat	kill
knife	51	20	84	0	3	0
cat	52	58	4	4	6	26
dog	115	83	10	42	33	17
boat	59	39	23	4	0	0
cup	98	14	6	2	1	0
pig	12	17	3	2	9	27
banana	11	2	2	0	18	0

Figura 5. Representación distribucional

En la figura 5 (Stefan, 2010) mostramos un ejemplo de la representación de la estructura distribucional de la palabra “dog”, en la que se encuentran las palabras con las que se relaciona (get, see, use, hear, eat, kill), así como la cantidad de ocasiones en que lo hace, con lo que es posible obtener vectores de contexto (Widdows & Cohen, 2010). Posteriormente es posible darle una interpretación geométrica mediante el uso de diferentes modelos, los cuales se explican a continuación.

Modelos de semántica distribucional

La semántica distribucional favorece el uso de álgebra lineal como herramienta computacional y marco de representación. Diferentes tipos de similitudes se pueden extraer dependiendo de qué tipo de información sobre la distribución de los elementos lingüísticos se utiliza para construir los vectores.

La idea básica de una correlación entre la similitud de distribución y semántica puede hacerse operativa de muchas maneras diferentes, ya que hay una rica variedad de modelos computacionales de aplicación semántica distribucional, incluyendo Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997) (Landauer, Foltz, & Laham, 1988), Hyperspace Analogue to Language (HAL) (Lund & Burgess, 1996) y el modelo de espacio de palabras (Word Space) (Schütze, 1993) (Sahlgren, 2006).

Estos modelos de distribución semántica se basan en la suposición de que el significado de una palabra puede (al menos hasta cierto punto) derivarse de su uso, es decir, de su distribución en el texto. Por lo tanto, estos modelos construyen dinámicamente representaciones semánticas a través de un análisis estadístico de los contextos con los que coocurre cada palabra. Dichos modelos son una técnica prometedora para resolver el cuello de botella de la adquisición del léxico mediante el aprendizaje no supervisado y su representación distribuida ofrece una arquitectura robusta y flexible para la organización y el procesamiento de la información semántica (Stefan, 2010).

Sahlgren en su trabajo describe el espacio de palabras, un modelo computacional que utiliza patrones de distribución de las palabras a lo largo de grandes cantidades de datos, para representar la similitud semántica entre las palabras en función de la proximidad espacial. Este modelo ha sido utilizado por más de una década, y ha demostrado buenos resultados en numerosos experimentos y aplicaciones, e incluso pasar de entornos de investigación a la implementación práctica en sistemas comerciales (Sahlgren, 2006).

Aunque este modelo ha sido ampliamente utilizado e intensamente investigado, nuestra comprensión teórica de este modelo sigue siendo poco clara. Otro ejemplo del uso de la semántica distribucional es la creación automática de taxonomías, demostrando que dado un conjunto de sustantivos con distintos hiperónimos, es posible su agrupación bajo el hiperónimo correcto por procedimientos estadísticos que prácticamente no requieren conocimiento explícito de la lengua analizada (Nazar & Renau, 2012).

Los diferentes modelos de distribución semántica difieren primordialmente con respecto a los siguientes parámetros:

- Tipo de contexto (regiones texto vs elementos lingüísticos).
- Contexto de la ventana (tamaño, extensión, dirección).
- Ponderación de frecuencia (entropía, PMI).
- Reducción de la dimensión (random indexing, singular value decomposition).
- Medidas de similitud (similitud coseno, distancia minkowski).

Así como los diferentes modelos de semántica distribucional difieren entre sí, también han sido aplicados con éxito para diversas tareas (Kao & Poteet, 2007), como:

- Búsqueda de similitud semántica entre las palabras y expresiones multipalabra.
- Agrupación de palabras basada en la similitud semántica.

- Creación automática de diccionarios bilingües y tesauros.
- Resolución de la ambigüedad léxica.
- Expansión de las consultas de búsqueda utilizando sinónimos y asociaciones.
- La definición del tema de un documento.
- Agrupamiento de documentos para la recuperación de la información.
- Minería de datos y reconocimiento de entidades con nombre.
- La creación de mapas semánticos de diferentes dominios en cuestión.
- Paráfrasis.
- Análisis de los sentimientos.

En el caso de este trabajo de tesis, utilizamos el modelo de espacio de palabras, enfocado a identificar variantes denominativas en dos áreas de especialidad diferentes: Ingeniería de Software y Biodiversidad.

3. Los corpus de especialidad: Ingeniería de Software y Biodiversidad.

Como ya mencionamos anteriormente, un corpus debe mantener las características de representatividad, variedad y equilibrio, y siguiendo esas características logramos reunir un corpus especializado, es decir, cuyo dominio o tema se encuentre restringido y claramente definido, en este caso la Ingeniería de Software. El corpus de textos especializados del área de la Ingeniería de Software está conformado por textos provenientes de diferentes regiones geográficas donde se habla el idioma español, en este caso España, México y Latinoamérica, ya que aunque hablan el mismo idioma, muestran algunas diferencias en su uso.

Por otro lado, la CONABIO, en conjunto con el Grupo de Ingeniería Lingüística, se encuentra trabajando en diferentes proyectos de investigación dentro del área de la Lingüística Computacional, por lo que nos proporcionó un corpus especializado del área de Biodiversidad, el cual ha sido recopilado desde hace ya varios años y ha logrado conjuntar textos de diferentes partes de la República Mexicana donde existen diferentes variantes terminológicas.

Descripción del corpus de Ingeniería de Software

Este corpus está conformado por varios artículos de diferentes ediciones trimestrales publicadas de la revista mexicana "Software Gurú"[19] a partir del año 2008, una revista especializada en Tecnologías de Información; Además consta de ejemplares completos de la Revista Latinoamericana de Ingeniería de Software (RELAIS)[20], una publicación técnica bimestral que forma parte de la Red de Ingeniería de Software de Latinoamérica (RedISLA)[21] que se encarga, entre otras cosas, de difundir artículos técnicos, empíricos y teóricos, dar a

[19] http://sg.com.mx/revista

[20] http://sistemas.unla.edu.ar/sistemas/redisla/ReLAIS/index.htm

[21] http://sistemas.unla.edu.ar/sistemas/redisla/index.htm

conocer los desarrollos científico-tecnológicos y contribuir a la promoción de la cooperación institucional entre universidades y empresas, todo para el fortalecimiento de la Ingeniería de Software Latinoamericana como disciplina y profesión.

El corpus cuenta también con ejemplares de la Revista Española de Innovación, Calidad e Ingeniería del Software (REICIS)[22], la cual pretende canalizar las contribuciones de conocimiento práctico y original que se produce en España en el campo de la Ingeniería y la Calidad del Software, así como la innovación en su desarrollo y mantenimiento.

Finalmente decidimos agregar el libro "Ingeniería de Software. Un Enfoque Práctico", (Pressman & Troya, 1988) el cual es una guía especializada para el área de Ingeniería de Software, la más vendida tanto para estudiantes como para profesionales de esta industria.

Fuente	Elementos	Tokens
Revista Española de Innovación, Calidad e Ingeniería del Software (REICIS).	22 ejemplares.	280,198
Revista Latinoamericana de Ingeniería de Software (RELAIS).	11 ejemplares.	413,089
Revista "Software Gurú".	293 artículos.	325,675
Ingeniería de Software. Un Enfoque Práctico.	1 libro.	366,408
TOTAL	327 archivos.	1,385,370

Tabla 1. Conformación del corpus Ingeniería de Software

Todos los archivos recolectados fueron obtenidos en formato PDF[23], por lo que tuve la necesidad de transformarlos a un formato de texto plano para procesarlo

[22] http://www.ati.es/reicis

[23] Portable Document Format (Formato de documento portable) es el formato de archivos desarrollado por Adobe Systems

adecuadamente, para ello hice uso de un OCR[24], el cual mostró algunas dificultades durante este proceso, ya que debido al formato de las revistas, sobretodo Software Gurú, tuvimos que realizar varios ajustes al texto obtenido de manera manual, ya que mostraron varios saltos de línea o espacios en blanco.

El corpus en su totalidad está conformado de 327 archivos en formato de texto plano, logrando un total de 1,385,370 tokens[25] y 36,433 types[26].

Descripción del corpus de Biodiversidad

El corpus está conformado por textos de diferentes partes de la república, Aguascalientes, Campeche, Chiapas, Chihuahua, Estado de México, Michoacán, Morelos, Quintana Roo y Yucatán.

Estado	*Elementos*	*Tokens*
Aguascalientes	85 archivos.	132,389
Campeche	121 archivos.	174,671
Chiapas	133 archivos.	267,994
Chihuahua	102 archivos.	139,626
Estado de México	39 archivos.	96,350
Michoacán	31 archivos.	44,731
Morelos	64 archivos.	66,856
Quintana Roo	12 archivos.	25,962
Yucatán	172 archivos.	211,269
TOTAL	759 archivos.	1,145,959

Tabla 2. Conformación del corpus Biodiversidad

[24] Optical Character Recognition, (Reconocimiento Óptico de Caracteres**)**. Término que se utiliza en la informática para nombrar a un procedimiento que permite digitalizar un texto a través de un escáner.

[25] El token o palabra es cada una de las formas que aparecen en el texto, sin importar cuántas veces ocurra cada una. El número total de tokens define el tamaño del corpus.

[26] El type se refiere a cada una de las formas o palabras diferentes que aparecen en un texto.

Está conformado por un total de 759 archivos en formato de texto plano, haciendo un total de 1,145,959 tokens y 52,267 types.

Estos textos se encuentran organizados de acuerdo al estado del que proceden y aunque en su totalidad corresponden a la biodiversidad de la República Mexicana, tienen diferentes divisiones incluyendo plantas, animales, insectos, hongos, entre otros (INECC, 2013).

Etiquetado de partes de la oración (POS)

Una parte importante para la obtención de las variantes denominativas es resolver las ambigüedades léxicas presentes en los textos que conforman ambos corpus obtenidos. Esto lo llevamos a cabo con el etiquetado de partes de la oración (POS)[27], es decir, la obtención de una etiqueta para cada una de las palabras del corpus, la cual indica la categoría gramatical a la que pertenece según la función que desempeña dentro de cada una de las oraciones. Dichas etiquetas indican, entre otras cosas, si es sustantivo, verbo, pronombre, preposición, adverbio, conjunción, modal o artículo (Martin & Jurafsky, 2000). Además de la categoría gramatical, se obtiene el lema de la palabra, es decir, su forma canónica, eliminando partes no esenciales de la misma como son los sufijos o los prefijos.

El primer paso que realizamos para el etiquetado de las partes de la oración fue revisar que la codificación de los archivos fuera utf-8, una de las 3 diferentes formas de codificación para la representación de caracteres que en nuestro idioma son necesarios, como el uso de acentos, diéresis, entre otros (Allen, 2014). Este etiquetado lo realizamos haciendo uso de la herramienta de software libre Freeling[28], un etiquetador morfosintáctico para el procesamiento multilingüe automático, que proporciona una amplia gama de servicios de análisis lingüísticos para diversos idiomas. Las etiquetas que utiliza Freeling se

[27] Part of Speech: Proceso de etiquetado con el fin de resolver ambigüedades léxicas.

[28] http://nlp.lsi.upc.edu/freeling

basan en las etiquetas propuestas por el grupo EAGLES[29] para la anotación morfosintáctica de corpus.

Freeling requiere que los archivos que se desean etiquetar se encuentren en texto plano y con la codificación *UTF-8* para que funcione correctamente o de lo contrario el etiquetado presentará errores. En el caso de nuestro corpus hubo la necesidad de cambiar dicha codificación ya que algunos de los archivos contenían la codificación *ISO 8859-1*, este cambio lo hicimos haciendo uso del sistema operativo Linux y una vez que los archivos tenían la codificación correcta se procedió a realizar el etiquetado.

El etiquetado se realiza para cada palabra de cada archivo del corpus con el formato:

```
Palabra<b>lema<b>etiqueta (POS)
```
[30]

Para la oración “El gato come pescado.”, obtenemos un etiquetado donde podemos observar que en cada línea del archivo de salida se encuentra la palabra tal cual se encuentra en el texto, posteriormente el lema de la misma y finalmente su categoría gramatical.

```
El el DA0MS0
gato gato NCMS000
come comer VMIP3S0
pescado pescado NCMS000
. . Fp
```

En la línea número uno vemos la palabra “`El`” seguida de su lema “`el`” con la etiqueta “`DA0MS0`” que de acuerdo a las etiquetas propuestas por el grupo EAGLES nos indica que se trata de un determinante(D) de tipo articulo (A), no posesivo (0), masculino (M), singular (S), sin distinción de los referentes (0); En

[29]Expert Advisory Group on Language Engineering Standards
http://www.ilc.cnr.it/EAGLES96/intro.html

[30] https://talp-upc.gitbooks.io/freeling-user-manual/content/tagsets/tagset-es.html

la segunda línea se muestra la palabra “gato” con el lema “gato” y la etiqueta “NCMS000” que indica que es un sustantivo (N) común (C), masculino (M), singular (S).

De igual manera llevamos a cabo esta tarea para cada uno de los archivos que conforman ambos corpus, de esta forma es posible identificar palabras que contienen el mismo lema como los plurales o algunas conjugaciones verbales.

4. La extracción de términos en el área de Ingeniería de Software

Como ya vimos en el capítulo 2, existen diferentes metodologías para la extracción de términos y diferentes extractores terminológicos, en este trabajo optamos por utilizar dos de ellos: Termext y Termostat, ya que implementan metodologías diferentes y se pueden comparar los resultados obtenidos.

Termext

Termext es un extractor terminológico de tipo hibrido, basado en el algoritmo c-value/nc-value cuya metodología se lleva a cabo mediante una técnica estadística y una lingüística (Barrón, Sierra, Drouin, & Ananiadou, 2009). Termext está compuesto por diferentes programas elaborados con el lenguaje de programación Python[31], donde cada uno de ellos realiza actividades específicas y requiere de algunos elementos de software específicos, los cuales deben ser instalados y configurados previamente para su correcto funcionamiento (Barrón Cedeño, 2008).

Requerimiento	Descripción	Versión requerida	Versión instalada
GNU/Linux[32]	Sistema Operativo abierto basado en UNIX.[33]		OSX 10.10.1
Python	Lenguaje de programación Orientado a Objetos.	2.5+	2.7.9

[31] Python es un claro y poderoso lenguaje de programación orientado a objetos.

[32] Linux es un sistema operativo: un conjunto de programas que le permiten interactuar con su ordenador y ejecutar otros programas.

[33] Unix es un Sistema Operativo multiusuario y multitarea ampliamente utilizado en servidores o estaciones de trabajo, además de ser utilizado como un estándar.

Requerimiento	Descripción	Versión requerida	Versión instalada
NLTK[34]	Módulos para la investigación y desarrollo en procesamiento natural.	0.9.5+	3.0.1
Mumby[35]	Paquete para cómputo científico (matemático).	1.1.1+	1.9.2
Treetagger	Etiquetador morfosintáctico basado en arboles de decisión.	3.2+	3.2

Tabla 3. Requerimientos de uso Termext

Cabe mencionar que la versión de Python requerida para su correcto funcionamiento no se encuentra disponible actualmente, por lo que no fue posible de obtener. Al instalar la versión 2.7.9 tuvimos que realizar los cambios necesarios para su funcionamiento debido a que algunas de sus funciones han sido modificadas y/o sustituidas.

Cada uno de los elementos requeridos para su funcionamiento los obtuvimos de manera gratuita a través de su dirección electrónica oficial y los instalamos conforme a su documentación y/o manuales correspondientes. Posteriormente procedimos a realizar la extracción terminológica automática haciendo uso de los corpus de Ingeniería de Software y Biodiversidad y mediante la ejecución de los programas que componen en su totalidad a Termext.

Primero se realiza el etiquetado morfosintáctico a través de la herramienta Treetagger, el cual a diferencia de Freeling, se basa en la probabilidad, utilizando árboles de decisión (de ahí el nombre del etiquetador), éste genera como salida cada uno de los archivos originales con la extensión ".tag" para identificarlo del archivo original. Un ejemplo de la salida del etiquetado se puede observar a continuación:

[34] Natural Language Toolkit es una plataforma líder para la creación de programas con Python para trabajar con los datos del lenguaje humano.

[35] Numpy es el paquete fundamental utilizado para la computación científica en Python.

```
CALIDAD_calidad/NC EN_en/PREP PSPVDC_<unknown>/NC
La_el/ART calidad_calidad/NC en_en/PREP
PSP_<unknown>/NC incluye_incluir/VLfin :_:/COLON
•_<unknown>/VLfin estimar_estimar/VLinf el_el/ART
número_número/NC total_total/ADJ de_de/PDE
defectos_defecto/NC inyectados_inyectar/VLadj y_y/CC
removidos_remover/VLadj •_<unknown>/VLfin
estimar_estimar/VLinf el_el/ART número_número/NC
de_de/PDE defectos_defecto/NC inyectados_inyectar/VLadj
y_y/CC removidos_remover/VLadj por_por/PREP
fase_fase/NC •_<unknown>/VLfin estimar_estimar/VLinf
el_el/ART tiempo_tiempo/NC requerido_requerir/VLadj
por_por/PREP fase_fase/NC Se_se/SE
presenta_presentar/VLfin en_en/PREP esta_este/DM
sección_sección/NC las_el/ART
modificaciones_modificación/NC al_al/PAL
PSP_<unknown>/NC para_para/CSUBI
planificar_planificar/VLinf la_el/ART
calidad_calidad/NC en_en/PREP PSPVDC_<unknown>/NP
._./FS
```

Una vez que Termext realizó el etiquetado morfosintáctico, hace uso del algoritmo C-value/NC-value[36] modificado, que consiste en la generación de una lista de posibles términos a través de patrones sintácticos que más comúnmente aparecen en los términos del área de especialidad y la eliminación de palabras funcionales[37] con el fin de eliminar candidatos que contengan palabras que no

[36] C-value/NC-value es un método hibrido para la extracción de términos que fue originalmente diseñado para obtener candidatos a término del área de biomedicina en inglés. (Barrón, Sierra, Drouin, & Ananiadou, 2009)

[37] Palabras funcionales son palabras muy comunes que parecen varias veces en un texto y aportan poco valor para la obtención de los términos, (ver apéndice “Listado de palabras funcionales”).

se espera que formen parte de los términos del área.

Posteriormente Termext realiza el cálculo del NC-value de cada candidato a término para conocer si debe ser considerado como un verdadero término, esto por medio de su longitud y frecuencia de aparición dentro del corpus. Finalmente Termext considera que el contexto en el que aparecen los términos es significativo, por lo que también considera la vecindad de cada uno de los candidatos, lo que mejora la calidad de los resultados obtenidos.

Como salida obtuvimos un archivo con un formato específico, indicando el término obtenido, el NC-value y el C-value asociados al mismo.

```
término ### NC-value ### C-value
```

A continuación mostramos algunos términos que obtuvimos con Termext:

```
proceso ### 169.284763015 ### 202.768545221
software ### 153.745726495 ### 186.965847136
proyecto ### 137.052564103 ### 167.658745965
```

Termostat

Termostat es una herramienta automática de extracción terminológica que hace uso de corpus especializados y no especializados para la identificación de los términos[38], actualmente la versión disponible en línea es la versión 3.0 que soporta los idiomas inglés, francés, español, italiano y portugués.

En Termostat, su versión en línea, únicamente es necesario seleccionar el archivo que se desea procesar como entrada y se obtiene como resultado principal una lista de términos, los cuales pueden ser monoléxicos o poliléxicos, es decir, compuestos de una o varias palabras.

Termostat contiene corpus de referencia para cada uno de los idiomas soportados, en este caso el corpus de referencia del español es de

[38] http://termostat.ling.umontreal.ca/doc_termostat/doc_termostat.html

aproximadamente 30 millones de palabras, lo que corresponde a cerca de 527,000 palabras diferentes y se trata de un corpus no técnico o especializado proveniente de la Asamblea Parlamentaria Europea[39].

Cada término recibe un valor basado en su frecuencia dentro del corpus analizado, y su frecuencia dentro del corpus de referencia, por ello es que la diversidad de los temas tratados dentro de este corpus es crucial y necesaria. Termostat considera el establecimiento de un corpus más equilibrado a partir de muestras de documentos de diferentes áreas y documentos más generales.

Básicamente Termostat funciona en tres diferentes etapas:

1. *Texto etiquetado*

   Al igual que Termext, Termostat inicialmente realiza un etiquetado de partes de la oración utilizando Treetagger, con el objetivo de eliminar la ambigüedad de palabras que podrían recibir más de una categoría gramatical.

2. *La extracción de palabras que corresponden a un conjunto de reglas predefinidas*

   Una vez que se realizó el etiquetado de partes de la oración, Termostat aplica un filtro mediante expresiones regulares para extraer palabras o grupos de palabras que corresponden a ciertos patrones que se encuentran comúnmente en los términos del idioma español (Drouin, 2003).

   Sustantivo
   Sustantivo + Adjetivo
   Sustantivo + Preposición + Sustantivo
   Sustantivo + Preposición + Sustantivo + Adjetivo

[39] La Asamblea Parlamentaria del Consejo de Europa está compuesta por parlamentarios de los cuarenta y siete Estados miembros del Consejo de Europa, http://assembly.coe.int/nw/Home-EN.asp.

Sustantivo + Verbo (pasado participio)
Sustantivo + Adjetivo + Preposición + Sustantivo

Adjetivo
Adverbio
Verbo

(*)Donde “+” indica concatenación

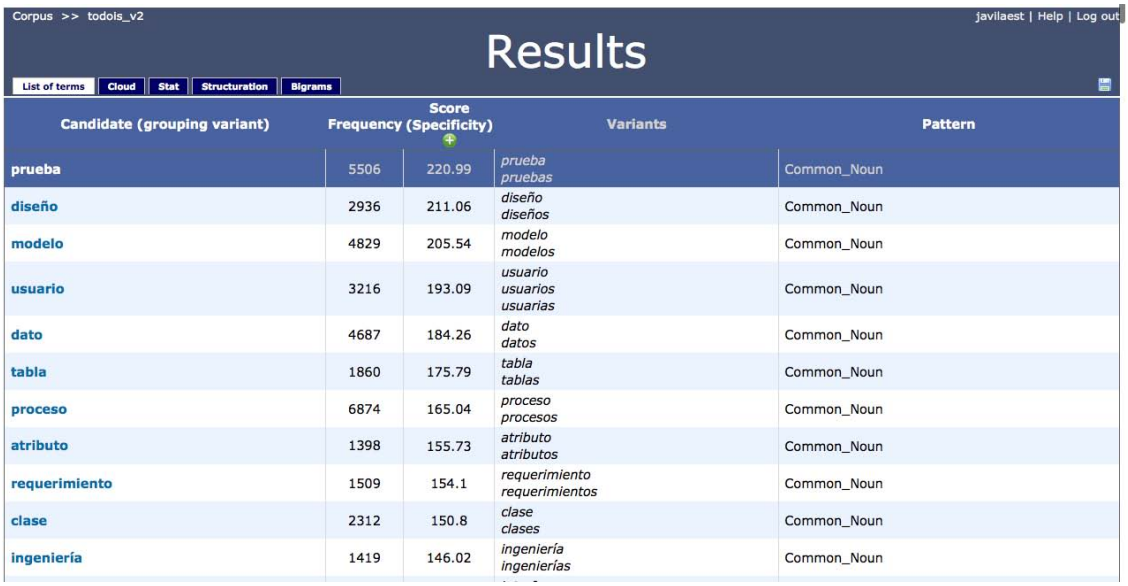
3. *Ponderación y selección de términos*

   Cada candidato a término recibe una puntuación basada en el nivel de especificidad[40] y aquellos candidatos a términos que recibieron los niveles más altos son considerados los más relevantes en el texto; un umbral de aceptabilidad permite excluir palabras o frases que no se consideran parte de la terminología en el corpus.

   Este método permite comparar el comportamiento de las unidades léxicas mediante la fusión del corpus de referencia y el corpus de análisis para comprobar si el léxico de este último se comporta como el léxico del primero.

Finalmente Termostat muestra la salida de los términos obtenidos en línea, con las características de frecuencia, especificidad, variantes (básicamente plurales), y el patrón sintáctico al que corresponde, pudiendo almacenarlos en formato de texto plano (ver Figura 6).

[40] El cálculo de la especificidad propuesto por Lafon (1980) para identificar el vocabulario específico en un sub-corpus con respecto a todo un corpus.



Corpus >> todols_v2 javilaest | Help | Log out

Results

List of terms | Cloud | Stat | Structuration | Bigrams

Candidate (grouping variant)	Score Frequency	(Specificity)	Variants	Pattern
prueba	5506	220.99	prueba pruebas	Common_Noun
diseño	2936	211.06	diseño diseños	Common_Noun
modelo	4829	205.54	modelo modelos	Common_Noun
usuario	3216	193.09	usuario usuarios usuarias	Common_Noun
dato	4687	184.26	dato datos	Common_Noun
tabla	1860	175.79	tabla tablas	Common_Noun
proceso	6874	165.04	proceso procesos	Common_Noun
atributo	1398	155.73	atributo atributos	Common_Noun
requerimiento	1509	154.1	requerimiento requerimientos	Common_Noun
clase	2312	150.8	clase clases	Common_Noun
ingeniería	1419	146.02	ingeniería ingenierías	Common_Noun

Figura 6. Obtención de resultados Termostat

Posteriormente obtuvimos el archivo en texto plano para su manejo al obtener las variantes denominativas. Cabe mencionar que Termostat a diferencia de Termext, nos muestra las variantes del término obtenido, en este caso los plurales.

```
Candidate(grouping variant)  Frequency  Specificity
Variants  Pattern

prueba     5506  220.99     prueba___pruebasCommon_Noun
diseño     2936  211.06     diseño___diseñosCommon_Noun
modelo     4829  205.54     modelo___modelosCommon_Noun
```

Términos obtenidos

Los términos que obtuvimos mediante ambos extractores son diferentes, aunque no del todo, ya que podemos observar que algunos de ellos se encuentran dentro de los resultados de ambos extractores; únicamente cambia el valor de la medida utilizada por cada uno de ellos, siendo para Termostat el nivel de especificidad y para Termext el NC-value.

La cantidad de términos obtenidos también es diferente. Al utilizar el corpus de Ingeniería de Software, Termext obtuvo un listado de 19,800 términos, mientras que Termostat obtuvo únicamente 18,094 términos. Esto debido a que Termostat obtiene las variantes de cada uno de los términos, básicamente los plurales, por lo que el número de resultados se reduce.

Al utilizar el corpus de biodiversidad, Termext obtuvo un listado de 49,121 términos, mientras que Termostat obtuvo únicamente 18,137 términos.

Ingeniería de Software			
Extractor Terminológico	**Términos Monoléxicos**	**Términos Poliléxicos**	**Total de Términos**
Termext	5,605	14,195	19,800
Termostat	5,958	12,136	18,094

Biodiversidad			
Extractor Terminológico	**Términos Monoléxicos**	**Términos Poliléxicos**	**Total de Términos**
Termext	13,753	35,367	49,121
Termostat	7,189	10,947	18,137

Tabla 4. Número de términos obtenidos.

Como se puede observar en la tabla, el número total de términos cambia para cada uno de los corpus y varía dependiendo del extractor que se utilice; también para el número de términos poliléxicos, que es mayor en ambos casos.

Además del corpus de biodiversidad, se obtuvieron 759 archivos con términos obtenidos mediante la "Biblioteca para la minería de textos de biodiversidad en español" (Barrios, Molina, Sierra-Alcocer, & Zenteno-Jimenez, 2015). Entre los datos extraídos se encuentra el nombre científico de las especies detectadas, así como el offset y número de línea en el que aparecen, como mostramos en el extracto a continuación:

```
[[5, "Crustacea", [12, 21]], [13, "Eumalacostraca",
[143, 157]], [13, "Stomatopoda", [165, 176]], [13,
"Decapoda", [204, 212]], [16, "Lysiosquilla
campechiensis", [106, 132]], [16, "Pseudorhombila
ometlanti", [183, 207]], [16, "Batodaeus adanad",
[241, 257]], [20, "Bullis", [215, 221]], [20,
"Bullis", [229, 235]], [23, "Squilla empusa", [109,
123]], [23, "Callinectes sapidus", [134, 153]], [23,
"Sycionia brevirostris", [172, 193]], [26, "Raninoides
louisianensis", [61, 85]], [26, "Raninoides lamarcki",
[87, 98]], [26, "Persephona mediterranea", [100,
123]], [26, "Iliacantha subglobosa", [125, 146]], [26,
"Stenorhynchus seticornis", [148, 172]], [26,
"Anasimus latus", [175, 189]], [27, "Porcellana
sayana", [74, 91]], [27, "Moreiradromia antillensis",
[93, 118]], [27, "Hypoconcha sabulosa", [120, 139]],
[27, "Calappa sulcata", [141, 156]], [27, "Calappa
flammea", [158, 168]], [27, "Hepatus epheliticus",
[170, 189]], [27, "Libinia dubia", [192, 205]], [45,
"Costera", [84, 91]]]
```

Para el objetivo de obtener las variantes terminológicas únicamente son necesarios los términos, los cuales los obtuvimos mediante el uso de una expresión regular (*(["'])(?:(?=(\\?))\2.)*?\1*), obteniendo un total de 6,292 términos.

Como última actividad dentro la extracción automática de términos y para validar la confiabilidad de los extractores utilizados, realizamos una comparación con los términos obtenidos con Termext y Termostat, con la finalidad de conocer cuántos de ellos se encuentran dentro del listado de términos proporcionado por la CONABIO, obteniendo los siguientes resultados:

Extractor Terminológico	Términos Monoléxicos	Términos Poliléxicos	Total de Términos	Términos encontrados	Porcentaje
Termext	13,753	35,367	49,121	2726 de 6,292	43.32%
Termostat	7,189	10,947	18,137	931 de 6,292	14.87%

Tabla 5. Evaluación entre extractores terminológicos

En la tabla 5 se puede observar claramente que Termext encontró un mayor número de términos correspondientes al listado proporcionado por la CONABIO, lo cual no indica que un extractor sea mejor que otro, sino que para este tipo de textos, Termext funciona de manera más favorable.

Ingeniería de Software		
Termext	**Termostat**	
software	Prueba	prueba___pruebas
información	Diseño	diseño___diseños
proyecto	Modelo	modelo___modelos
sección	usuario	usuario___usuarios___usuarias
calidad	dato	dato___datos
datos	tabla	tabla___tablas
organización	proceso	proceso___procesos
empresas	atributo	atributo___atributos
procesos	requerimiento	requerimiento___requerimientos
proceso	clase	clase___clases
sistema	ingeniería	ingeniería___ingenierías
desarrollo	interfaz	interfaz___interfaces
trabajo	proyecto	proyecto___proyectos
parte	negocio	negocio___negocios
pruebas	herramienta	herramienta___herramientas
cliente	sección	sección___secciones
empresa	arquitectura	arquitectura___arquitecturas
tiempo	figura	figura___figure
desarrollo de software	equipo	equipo___equipos
producto	cliente	cliente___clientes
personas	diagrama	diagrama___diagramas
problemas	patrón	patrón___patrones
figura	interacción	interacción___interacciones
proyectos	técnica	técnica___técnicas
años	componente	componente___componentes

Tabla 6. Comparativa de términos en Ingeniería de Software

Biodiversidad		
Termext	**Termostat**	
especies	especie	especie___especies
estado	conservación	conservación
conservación	vegetación	vegetación
méxico	suelo	suelo___suelos
parte	área	área___áreas
biodiversidad	biodiversidad	biodiversidad
años	municipio	municipio___municipios
estado de méxico	manejo	manejo___manejos
región	ecosistema	ecosistema___ecosistemas
chiapas	bosque	bosque___bosques___bosquecillo
estado de campeche	selva	selva___selvas
península de yucatán	planta	planta___plantas___plantilla
país	sierra	sierra___sierras
suelo	ave	ave___aves
recursos naturales	fauna	fauna
especie	cuadro	cuadro___cuadros
entidad	figura	figura___figuras
figura	género	género___géneros
número de especies	diversidad	diversidad
yucatán	hongo	hongo___hongos
distribución	Superficie	superficie___superficies
mayoría	aprovechamien to	aprovechamiento___aprovechamiento s
cuadro	agua	agua___aguas
estado de aguascalientes	hábitat	hábitat
agua	río	río___ríos___riachuelo

Tabla 7. Comparativa de términos en Biodiversidad

En las tablas 6 y 7 mostramos los primeros 25 términos que obtuvimos con cada uno de los extractores utilizando ambos corpus recopilados, donde podemos observar, por ejemplo, que el término "proceso" aparece en ambos resultados y que además en Termostat se obtuvieron sus variantes en plural, hecho que no ocurre en Termext ya que muestra por separado los términos "proceso" y

“procesos”. Sabemos que ambos términos (“proceso” y “procesos”) se refieren a un mismo término, únicamente que uno de ellos se encuentra en plural.

Esto mismo ocurre para el corpus de Biodiversidad, como es el caso del término “especies” que aparece en ambos resultados, únicamente que el resultado utilizando Termostat contiene sus variantes en plural (especie, especies).

Una manera de evitar este problema fue elaborando un programa para obtener las variantes de un término en plural, para ello se hizo uso del corpus etiquetado con Freeling, el cual contiene el lema de cada palabra, con lo que sin importar el tiempo o conjugación en que se encuentre el término, en caso de ser verbo, el lema de ellos siempre será el mismo. Lo mismo ocurre para los sustantivos, ya que si un término se encuentra en plural, el lema siempre se encuentra de manera singular, de esta manera obtenemos las variantes en plural de cada término.

```
un uno DI0MS0                    la el DA0FS0
proceso proceso NCMS000          capa capa NCFS000
que que PR0CN000                 de de SPS00
permite permitir VMIP3S0         negocio negocio NCMS000
determinar determinar VMN0000    y y CC
el el DA0MS0                     la el DA0FS0
nivel nivel NCMS000              tecnología tecnología NCFS000
de de SPS00                      sus su DP3CP0
usabilidad usabilidad NCFS000    procesos proceso NCMP000
```

En el texto anterior mostramos dos fragmentos etiquetados, donde del lado izquierdo se puede observar la palabra “proceso“ cuyo lema es “proceso“, y del lado derecho la palabra “procesos“ cuyo lema es “proceso“, por lo que mediante el programa realizado se obtiene que tanto la palabra “proceso“, como “procesos“ hacen referencia a un mismo término.

Esto ocurre de igual manera para los términos poliléxicos, únicamente se obtienen los plurales de cada una de las palabras que conforman dicho término, exceptuando las palabras funcionales, tal es el caso del término "proceso de desarrollo", donde obtenemos los plurales de las palabras "proceso" y "desarrollo" y posteriormente se combinan para obtener los términos en plural "proceso de desarrollo", "procesos de desarrollo", "proceso de desarrollos" y "procesos de desarrollos".

```
la el DA0FS0                    evaluar evaluar VMN0000
mayoría mayoría NCFS000         la el DA0FS0
de de SPS00                     usabilidad usabilidad NCFS000
los el DA0MP0                   durante durante SPS00
procesos proceso NCMP000        el el DA0MS0
de de SPS00                     proceso proceso NCMS000
desarrollo desarrollo NCMS000   de de SPS00
. . Fp                          desarrollo desarrollo NCMS000
```

En el texto anterior mostramos otros dos fragmentos etiquetados, donde del lado derecho se puede observar el término "proceso de desarrollo" y del lado izquierdo el término "procesos de desarrollo" cuyos lemas son exactamente los mismos, por lo que hace referencia a un solo término poliléxico. Una desventaja de utilizar este método, es que al realizar la combinación de las palabras y sus plurales, algunos de los términos obtenidos no se encontraron dentro del corpus.

5. Detección de variantes denominativas

Para la obtención de las variantes denominativas, primeramente, contemplamos la posibilidad de no hacer uso de la totalidad de los términos obtenidos por cada extractor, sino únicamente utilizar aquellos cuya medida de extracción superaba o igualaba cierto límite establecido, como se muestra en la siguiente tabla:

Extractor	Límite de extracción	Ingeniería de Software	Biodiversidad
Termext	NC-value >= 40.00	197 términos	376 términos
Termostat	Especificidad >= 50.00	154 términos	153 términos

Tabla 8. Límite de extracción de términos.

Al momento que utilizamos únicamente este número de términos, había la posibilidad de que no existieran variantes denominativas entre ellos, ya que aunque sean los términos más relevantes dentro del corpus, podrían existir variantes con otros términos menos relevantes y/o frecuentes, por lo que decidimos hacer uso de la totalidad de los mismos, evitando así la exclusión de la mayoría de los mismos al basarnos sólo en su medida de extracción ya sea el NC-value o su nivel de especificidad y no su contexto de aparición.

Metodología para la obtención de términos similares

La extracción terminológica, como ya lo mencionamos en el capítulo 2, es un problema en el que se continúa trabajando y la Semántica Distribucional es un modelo que se utiliza en diferentes áreas de la lingüística computacional, en este caso específicamente, lo utilizamos para la obtención de variantes denominativas en dominios de especialidad.

Hasta este punto únicamente hemos obtenido los términos de dos dominios de especialidad, la Ingeniería de Software y la Biodiversidad. Ahora explicaremos cómo, con la ayuda de la Semántica Distribucional realizamos una comparación entre los diferentes términos obtenidos para conocer el nivel de similitud o el nivel de variación que existe entre ellos con base en el contexto de cada uno de

ellos, es decir, las palabras más cercanas con las que se encuentra cada término dentro de los corpus.

Una vez que obtuvimos los términos a utilizar dentro del área de Ingeniería de Software y la Biodiversidad, fue posible obtener las palabras más cercanas a cada uno de ellos, es decir, su contexto y la frecuencia con que aparecieron en el texto, pero sobre todo obtener la frecuencia con que los términos se relacionan con cada una de las palabras dentro de cada corpus. Para ello seguimos la metodología definida por la Semántica Distribucional.

La metodología utilizada para la obtención de las variantes denominativas se basa en la hipótesis de distribución de Harris: “Palabras con significado similar tienden a ocurrir en contextos similares” (Harris, 1954). La puesta en práctica más común para este método es la Semántica Distribucional (Schutze , 1992) donde primeramente se debe obtener una Matriz de Coocurrencia, es decir, una matriz que indica para cada término cada una de las diferentes palabras del corpus con las que se relacionan. Una vez realizada dicha matriz, cada término es visto como un vector, donde cada una de las dimensiones del mismo corresponde a cada una de esas palabras del corpus. Para cada una de esas palabras se calcula una medida de cercanía (frecuencia, PMI) que indique qué tan estrecha es la relación de la palabra con el término. En seguida, se calcula la distancia (distancia coseno, Jaccard, Dice) que existe entre pares de vectores obtenidos (Manning, Raghavan, & Schütze, 2009). Finalmente con base en las distancias obtenidas para cada par de términos, se hace uso de un algoritmo de agrupamiento para conocer cuáles de ellos tienen una distancia menor, es decir, cuáles de ellos son candidatos a ser variantes denominativas ya que su distancia es menor.

La metodología que mencionamos anteriormente es general y se utiliza o se puede utilizar para diferentes tipos de investigaciones en el área de la lingüística computacional (Nazar & Renau, 2012) (Dang, Xue, & Croft, 2009) (Galicia-Haro

& Gelbukh, 2016). Sin embargo, es posible realizar algunas adecuaciones dentro de cada una de las etapas que la conforman para fines específicos, y en este trabajo de investigación no fue la excepción. Entre las adecuaciones que se realizaron se encuentra el uso de la medida de asociación Pointwise Mutual Information (PMI) y no únicamente la frecuencia de coocurrencia al obtener la matriz. PMI es una medida de asociación entre dos características (comúnmente palabras) que cuantifica la diferencia entre la probabilidad de su coincidencia dada su distribución conjunta y sus distribuciones individuales, suponiendo independencia, en este caso es entre el término y su contexto (dependiendo del tamaño de la ventana de contexto utilizada). Además de utilizar un corpus etiquetado previamente para de esta manera considerar los lemas de las palabras dentro del contexto de los términos, este tipo de adecuaciones se detallan en cada una de las etapas a continuación.

Matriz de coocurrencia

Para la obtención de la matriz de coocurrencia, primeramente se debe establecer una ventana de contexto, es decir, definir el tamaño del contexto para cada término obtenido, esto se refiere al número de palabras. En este trabajo establecimos una ventana de 5 palabras ya que se han realizado diferentes trabajos donde al hacer uso de este tamaño de ventana se han obtenido resultados favorables (Nazar & Renau, 2012); (Nenadié, Ananiadou, & McNaught, 2004); (Barrón Cedeño, 2008), sin embargo es posible utilizar un tamaño de ventana diferente; Esta ventana considera el número de palabras que se consideran como contexto en ambas direcciones del término, es decir, 5 palabras a la izquierda y 5 a la derecha, considerándose dicho conjunto de palabras como el contexto del término.

En esta primera etapa para identificar las variantes denominativas, específicamente en la matriz de coocurrencia, cada una de las columnas (*m*) corresponde a una palabra dentro del corpus, y las filas (*n*) corresponden a cada

uno de los términos obtenidos anteriormente. Cada una de las entradas en la tabla indica la frecuencia con que coocurren las palabras y los términos.

Términos/ Palabras	**Palabra 1**	**Palabra 2**	**…**	**Palabra n**
Término 1	# coocurrencias			
Término 2				
…				
Término m				

En este caso se obtienen cuatro matrices de coocurrencia, ya que por cada corpus se obtienen dos de ellas, cada una correspondiente a cada uno de los extractores utilizados. Todas las matrices contienen el mismo número de columnas pero diferente número de filas, puesto que la cantidad de términos obtenidos por cada extractor terminológico utilizado es diferente.

- Obtenemos una matriz de tamaño *m*n*

  donde *m* = # de términos

  n = # de palabras

$$\mathbf{M} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

- Vector de distribución x_i = i-ésima fila de **M** ($x_2 = x_{software}$)

- Componentes $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$ = frecuencias de coocurrencia del i-ésimo término

Para la obtención de esta matriz realizamos un programa utilizando el lenguaje de programación Java[41], el cual fue elaborado en diferentes etapas con la finalidad de evitar errores al momento de obtener los resultados. En una primera

[41] Java es un lenguaje de programación y plataforma de computación lanzada por Sun Microsystems en 1995.

etapa fueron contempladas, además de los términos ya obtenidos, todas las palabras del corpus, obteniendo como resultado vectores muy grandes y difíciles de manejar computacionalmente, ya que contenían demasiadas dimensiones, 52,267 para el caso del corpus de Biodiversidad y 36,433 para el de Ingeniería de Software.

Una manera de reducir el tamaño de los vectores fue eliminando las palabras funcionales del corpus, las cuales no aportan valor dentro del contexto de los términos (Nazar & Renau, 2012). En la mayoría de los vectores, por no decir que en todos, se encuentran una o varias de ellas y su frecuencia de coocurrencia es muy alta, y al ser consideradas obtuvimos que varios de los términos eran poco distantes entre ellos debido a que sus contextos eran muy similares al coocurrir con las mismas palabras funcionales.

> ...bien—. La estructura incluye un **proceso**, un conjunto de métodos y...
> ...de gran número de las **especies** presentes en el Estado; 2) presencia...

En el texto anterior mostramos dos fragmentos de los corpus, donde se observan los términos "proceso" y "especies", así como las palabras que los rodean, es decir, su contexto. Las palabras subrayadas son consideradas funcionales ("de", "la", "las", "un", "en", "el", "y") y no aportan valor al contexto ya que son pronombres, artículos y conjunciones. Estas mismas palabras coocurren con muchos otros términos.

Para obtener mejores resultados, un menor tiempo de procesamiento y menos uso de recursos computacionales, decidimos utilizar una lista de palabras funcionales para el español proveniente del proyecto Snowball[42] y de la empresa

[42] http://snowball.tartarus.org/algorithms/spanish/stop.txt

Ranks NL[43], la cual contiene 465 palabras[44], las cuales no fueron consideradas al momento de obtener el contexto de los términos dentro de la matriz de coocurrencia.

> "el hardware definiendo tiempos (medios)[5] (antes)[4] de (fallo)[3]. Desde el (punto)[2] de (vista)[1] del software estos (niveles)[1] (representan)[2] la (peligrosidad)[3] o (nivel)[4] de (riesgo)[5] que produciría un fallo de software. "

En el extracto anterior se muestra el contexto para el término "software"(subrayado), las 5 palabras a la izquierda y derecha consideradas como su contexto (entre paréntesis). Las palabras que se encuentran sin paréntesis entre las palabras de contexto son palabras funcionales que no se consideran parte de su contexto..

Cabe mencionar que estas palabras funcionales no se eliminaron del corpus, ya que algunos de los términos obtenidos contienen dichas palabras como "Ingeniería de software" o "Diseño de pruebas" las cuales contienen la palabra "de" pero en este caso forma parte del propio término.

Primeramente hicimos uso únicamente de los términos monoléxicos obtenidos tanto con Termext como con Termostat, posteriormente se localizaron los términos poliléxicos, y una vez identificado cada uno de ellos se obtuvieron las palabras con las que coocurren o que se encuentran dentro de la ventana de contexto, es decir, las 5 palabras que se encuentran a la izquierda del término y las 5 palabras a la derecha del mismo, sin considerar las palabras funcionales.

---

[43] http://www.ranks.nl/stopwords/spanish

[44] Ver apéndice "Lista de palabras funcionales"

Otro aspecto importante que consideramos es el hecho de que puede existir más de una vez una palabra de contexto con el mismo lema dentro del contexto de un término, para ello hicimos uso de ambos corpus etiquetados con Freeling, esto para identificar que nuestro programa considerara la frecuencia de aparición de dicha palabra dentro del contexto del término una sola vez.

```
lograr lograr VMN0000
una uno DI0FS0
prueba prueba NCFS000
efectiva efectivo AQ0FS0
, , Fc
se se P00CN000
debe deber VMIP3S0
usar usar VMN0000
un uno DI0MS0
plan plan NCMS000
de de SPS00
pruebas prueba NCFP000
creado crear VMP00SM
previamente previamente RG
. . Fp
```

El texto anterior es un extracto del corpus de Ingeniería de Software etiquetado, donde cada línea corresponde a una palabra, en el cual si consideramos el término "plan", podemos observar que dentro de las palabras de contexto a la izquierda se encuentra la palabra "prueba" y a la derecha se encuentra la palabra "pruebas", pero aunque sean palabras diferentes, su lema es el mismo en ambos casos, por lo que establecemos que son la misma palabra y consideramos que coocurre dos veces con la palabra "prueba".

El programa que realizamos funciona de la siguiente manera:

1. Requiere un conjunto de entradas:

a. Corpus etiquetado (Directorio que contiene los archivos etiquetados mediante Freeling).
b. Lista de términos (Archivo con listado de términos obtenidos con Termext o Termostat).
c. Palabras funcionales (Archivo con lista de palabras funcionales).
d. Tamaño de la ventana de contexto (Número de palabras a considerar a la derecha e izquierda de cada término).

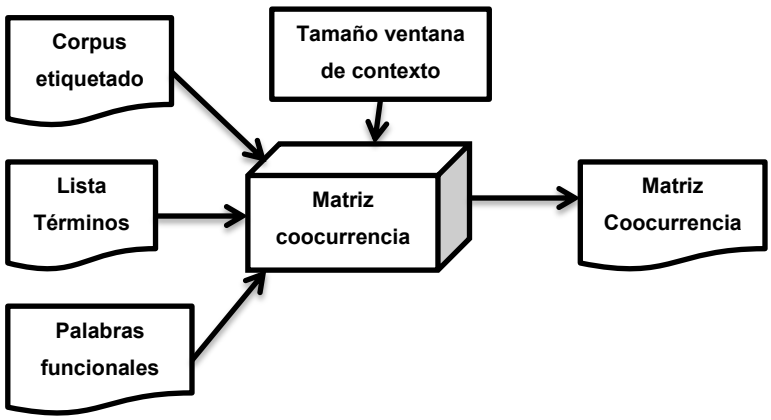


Figura 7. Funcionamiento programa de obtención de Matriz de Coocurrencia

2. Lee cada uno de los archivos etiquetados del corpus para obtener la totalidad de las palabras que aparecen en el corpus, asignando un índice a cada una de ellas y generando un archivo con el listado de dichas palabras y sus índices, de igual manera para cada uno de los términos, se asigna un índice, se obtiene la frecuencia de aparición en el corpus y se genera un archivo.

   Dicho archivo contiene únicamente el índice asignado, la palabra encontrada y la frecuencia, es decir, el número de ocasiones en que apareció dentro del corpus, tal como se muestra en la Tabla 9.

Listado de Términos <índice>#<término>#<frecuencia>	Listado de Palabras <índice>#<palabra>#<frecuencia>
351#Acaciaconstricta#5 1181#Acaciaconstrictavarvernicosa#1 1154#Acaciacornigera#2 666#Acaciafarnesiana#18 1677#Acaciagaumeri#4 214#Acaciapennatula#8 5536#Acaciaretinodes#0	453#MINAS#14 454#EXTRAEN#33 455#PLATA#17 456#COBRE#17 457#PLOMO#19 458#ZINC#10

Tabla 9. Formato de lista de términos y lista de palabras

3. Extrae los contextos de cada término y se obtiene un archivo, el cual contiene todos los términos obtenidos y sus contextos. El número de palabras a la izquierda y a la derecha depende del tamaño de ventana de contexto que se desea y se obtiene con el siguiente formato:

```
<Término> <índice>_<palabra> ##  <índice>_<palabra>
```

Ejemplo:

```
Acacia cornigera      16709_PENTANDRA 7679_FICUS
      1184_ESPECIES   3048_ÁRBOLES    3166_COMUNES    ##
      1222_ACACIA     5667_PENNATULA  16093_ANNONA
      18749_RETICULATA5660_BURSERA
Acacia cornigera      3110_G     31121_SEPIUM
      1184_ESPECIES   412_PRESENTES   32339_HIZCANAL  ##
      16771_GUAZUMA   16772_ULMIFOLIA 1649_CACTUS
      1625_OPUNTIA    32341_PUBERULA
```

En el ejemplo anterior se muestran dos de los contextos obtenidos para el término "Acacia cornígera", donde se pueden observar las 5 palabras de contexto a la izquierda y a la derecha del mismo, separadas por el símbolo ##.

4. Ordena alfabéticamente los contextos obtenidos de los términos y una vez estando ordenados, se obtiene la frecuencia de aparición entre cada término y cada palabra de sus contextos, además de realizar el cálculo de PMI para cada uno de los términos y las palabras en el corpus.

$$\mathrm{PMI}(x, y) = \log \frac{p(x, y)}{p(x)p(y)}$$

Donde ***p(x,y)*** se refiere a la coocurrencia del término(**x**) con una palabra(**y**) dentro de la ventana de contexto establecida, obteniendo un archivo de salida con el siguiente formato:

```
<Término> [<Indice_palabra>, <Palabra>,
                    <Frecuencia_de_coocurrencia>, <PMI>]
```

Ejemplo:

```
Acacia cornigera
[1222,ACACIA,1.0,12.082509448276229]
[16093,ANNONA,1.0,14.220012972026165]
[5660,BURSERA,1.0,12.878976054191096]
[1649,CACTUS,1.0,14.95697856619237]
[3166,COMUNES,1.0,11.028871484674156]
[1184,ESPECIES,2.0,6.204690849849235]
[7679,FICUS,1.0,12.997620550689716]
[3110,G,1.0,10.35211650803351]
[16771,GUAZUMA,1.0,14.039440726384344]
[32339,HIZCANAL,1.0,18.12690356763468]
[1625,OPUNTIA,1.0,12.700638812932585]
[5667,PENNATULA,1.0,14.541941066913527]
[16709,PENTANDRA,1.0,14.95697856619237]
[412,PRESENTES,1.0,10.023615759222661]
[32341,PUBERULA,1.0,18.12690356763468]
[18749,RETICULATA,1.0,15.126903567634685]
[31121,SEPIUM,1.0,15.541941066913527]
```

```
[16772,ULMIFOLIA,1.0,13.878976054191098]
[3048,ÁRBOLES,1.0,9.316331932893537]
```

En el ejemplo anterior se muestra un extracto del archivo obtenido, donde se observa el término "Acacia cornígera" y las 19 palabras con las que coocurre este término dentro del corpus de Biodiversidad, además de la frecuencia con las que estas coocurren y el valor de PMI calculado para cada una de ellas.

Obtención de distancias

Al igual que en la obtención de la matriz de coocurrencias, para la obtención de las distancias entre los vectores, elaboramos un programa con el lenguaje de programación Java, dicho programa recibe como entrada el archivo con la matriz de coocurrencias generada por el programa anterior.

El programa que elaboramos funciona de la siguiente manera:

1. Como ya se mencionó, el programa que realizamos requiere como entrada el archivo generado por el programa anterior, es decir, la matriz de coocurrencias, además del listado de palabras, archivo generado de igual manera por el programa anterior, además del listado de términos.

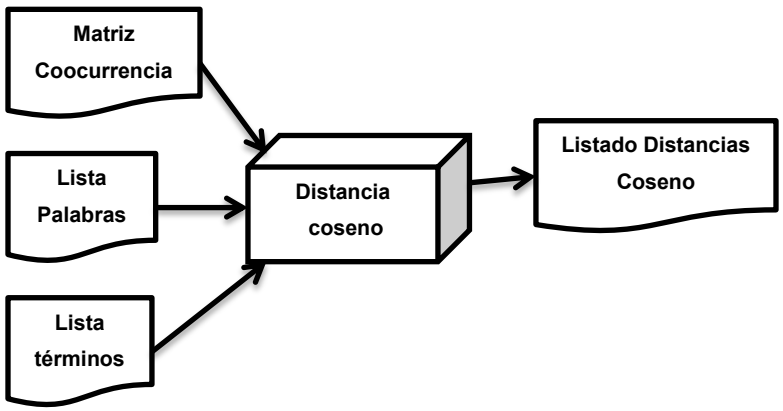


Figura 8. Funcionamiento programa Distancias Coseno

2. Obtiene el vector de distribución de cada uno de los términos, utilizando la frecuencia y PMI, con el siguiente formato:

```
<Término> <# de columna 1> <Valor medida 1>...
<# de columna n> <Valor medida n>
```

Ejemplo:

```
Acacia_cornigera       413       10.023615759222661       1185
6.204690849849235      1223      12.082509448276229       1626
12.700638812932585      1650      14.95697856619237       3049
9.316331932893537       3111      10.35211650803351       3167
11.028871484674156      5661      12.878976054191096      5668
14.541941066913527      7680      12.997620550689716      16094
14.220012972026165      16710      14.95697856619237      16772
14.039440726384344      16773     13.878976054191098      18750
15.126903567634685      31122     15.541941066913527      32340
18.12690356763468 32342 18.12690356763468
```

Como se puede observar en el ejemplo anterior, el vector contiene un formato simplificado[45], lo que hace que los vectores sean compactos y sencillos de manejar, únicamente indicando el número de columna a la que corresponde cada valor proporcionado dentro de la matriz de coocurrencia. En este caso, se obtuvieron dos tipos de vectores, uno indicando el valor de frecuencia y otro indicando el valor de PMI.

3. Una vez obtenidos los vectores, se procede a obtener las distancias entre cada uno de los pares resultantes, para ello hicimos uso de la Distancia Coseno[46], esto debido a que en la mayoría de los trabajos de investigación donde se ha utilizado, los resultados son favorables (Matsuo & Ishizuka, 2004).

---

[45] Los valores del vector se interpretan por pares. El primero de cada par indica una posición en el vector de características (o índice de elementos). El segundo muestra representa el valor del elemento vector de características de esa posición. Los valores de los elementos no especificados son interpretados como cero. http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/HAC.html#ahc-vectors

[46] Ángulo entre los vectores como medida de similitud. Si los documentos son iguales, el ángulo vale 0 y el coseno 1. En cambío si son ortogonales el coseno vale 0. (Kao & Poteet, 2007)

$$D_{coseno}(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum x_j^2}\sqrt{\sum y_k^2}}$$

4. Finalmente la salida de este programa es un archivo con un listado de los términos y las distancias coseno que existen entre ellos, mostrando primeramente aquellos cuya distancia coseno es mayor, lo que indica que su nivel de similitud es mayor.

```
[<Término 1>, <Término 2>] = <Distancia Coseno>
```

Ejemplo:

```
[Anaxyrus_compactilis,Bufo_compactilis]=1.0
[C_jejuni,C_upsaliensis]=0.951245623028531
[Pterocereus,Pterocereus_gaumeri]=0.9421887887545919
[Nelsonia,Nelsonia_neotomodon]=0.9404714594287151
[Manilkara,Manilkara_zapota]=0.9390105363492958
```

En el ejemplo anterior mostramos los primeros 5 resultados obtenidos, es decir, los 5 primeros pares de términos cuya distancia coseno es mayor, donde podemos concluir que los contextos de los términos "`Anaxyrus_compactilis`" y "`Bufo_compactilis`" son exactamente iguales para esta medida.

Agrupación de términos similares

Una vez que obtuvimos las distancias coseno entre todos los términos, realizamos un agrupamiento entre los términos, para conocer aquellos cuyos contextos son similares, agrupando las posibles variantes denominativas.

Para ello hicimos uso del proyecto ASV Toolbox[47], el cual es una colección modular de herramientas para la exploración de datos del lenguaje escrito: Uno de sus módulos es el agrupamiento jerárquico, para el cual únicamente es

[47] http://wortschatz.uni-leipzig.de/~cbiemann/software/toolbox/

necesario indicar el nombre del archivo que contiene los vectores de los términos obtenidos en formato simplificado, la medida de Distancia del Vector que se desea utilizar (Coseno, Dice o Jaccard). En nuestro caso utilizamos la distancia Coseno y el Método de Agrupamiento del algoritmo jerárquico que se desea: Single Linkage, Complete Linkage, AverageLinkage (Manning, Raghavan, & Schütze, 2009).

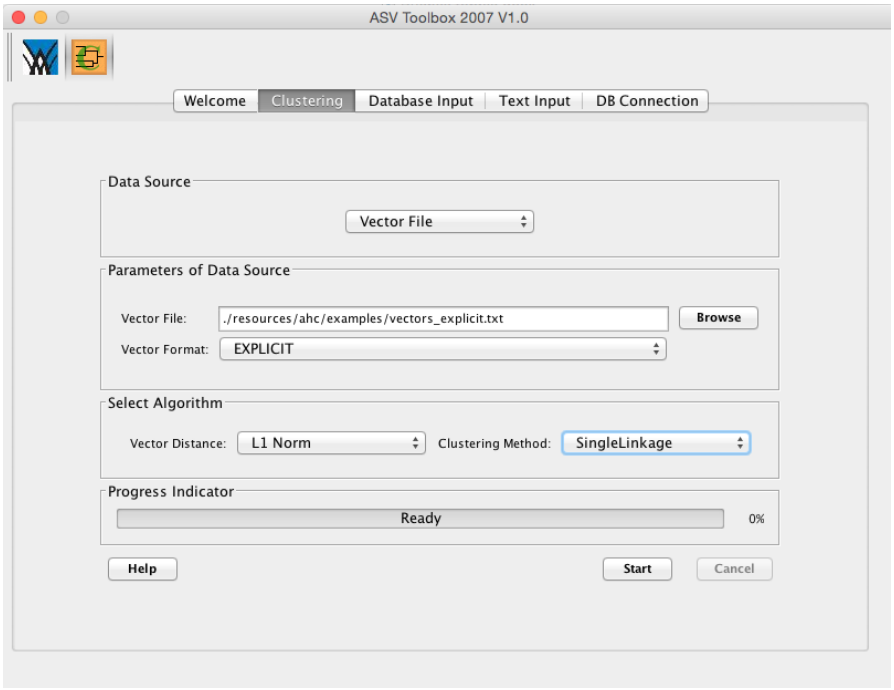


Figura 9. ASV Toolbox

```
java -Djava.ext.dirs=./lib de.uni_leipzig.asv.toolbox.hac.TextUI
--vectorfile
"/Users/Armando/Desktop/RES_TERMOSTAT/results_MatrizCoocurrencia_
n5_todoIS_FRECUENCIA_termostat.txt" --explicit -o
"/Users/Armando/Desktop/RES_FREC" -x -v Cosine -c SingleLinkage
```

En la figura 9 se puede observar la visualización de la herramienta ASV Toolbox, utilizada de manera gráfica y su uso a través de línea de comando, así como las opciones del Método de Agrupamiento y la Distancia del Vector que se desean para la realización del agrupamiento jerárquico.

La herramienta generó un archivo en XML en el cual se indican los agrupamientos obtenidos, además de un dendograma, el cual es un diagrama bidimensional utilizado para representar la clasificación jerárquica realizada y seguir de forma gráfica el procedimiento de unión de los términos, mostrando qué términos se unen, en qué nivel concreto lo hacen, así como el valor de la distancia coseno entre ellos cuando se agrupan (Guevara, 2011) (Henriksson, Moen, Skeppstedt, Eklund, Daudaravicius, & Hassel, 2012). Cabe mencionar que esta herramienta cuenta con una interfaz gráfica, además de poder hacer uso de ella mediante línea de comando, lo cual fue más conveniente para nosotros ya que fue posible utilizarla junto con los programas que hemos realizado anteriormente.

En la Figura 10 mostramos un extracto del dendograma que obtuvimos al aplicar el algoritmo de agrupamiento para los términos obtenidos con Termostat utilizando la medida de asociación PMI para el corpus de Biodiversidad; donde se pueden observar algunos términos y su nivel de similitud, la cual puede estar entre 0 y 1. Obtuvimos un total de 4 dendogramas, cada uno correspondiente a los términos obtenidos con los extractores utilizados (Termext y Termostat) en conjunto con las diferentes medidas de asociación utilizadas (PMI y Frecuencia de coocurrencia).

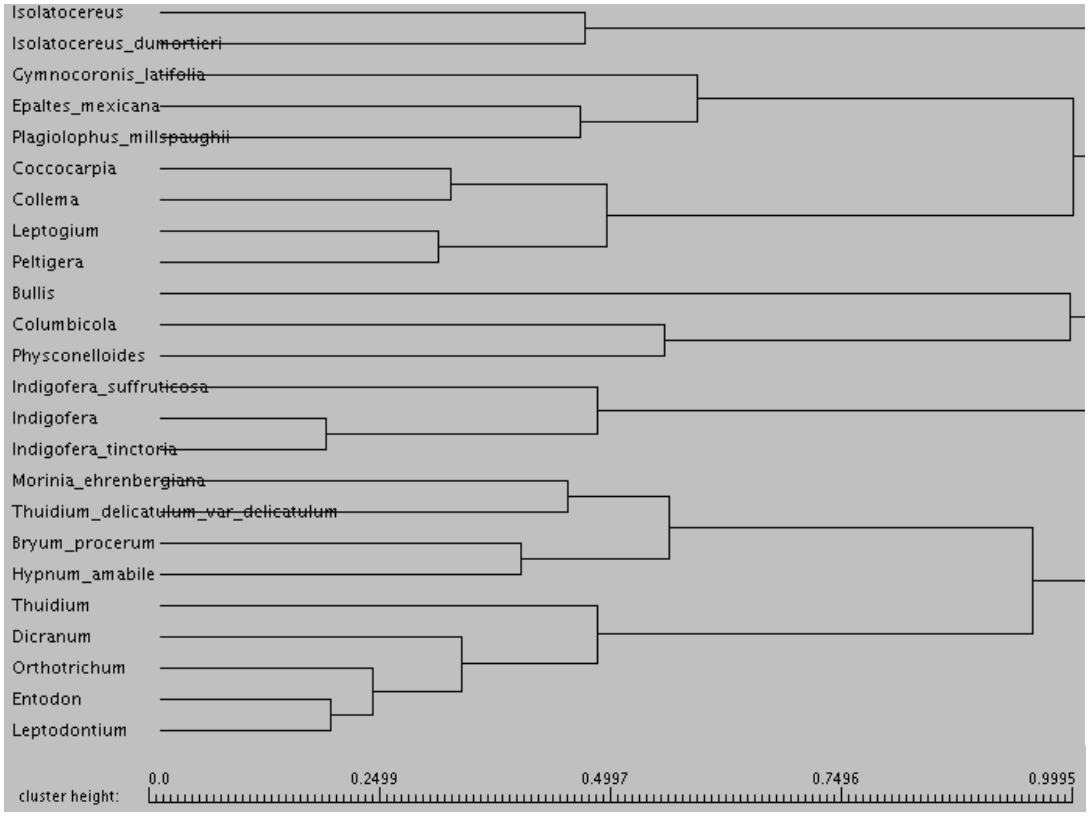


Figura 10. Dendograma términos Termostat PMI

Términos semánticamente similares

Una vez que realizamos la metodología descrita anteriormente para la obtención de variantes denominativas utilizando ambos corpus, obtuvimos un listado con todos los pares posibles de términos y su nivel de similitud existente entre ellos.

Realizamos una comparación entre los diferentes resultados obtenidos, donde únicamente consideramos los pares de términos cuyo nivel de similitud fue mayor a 0.5, esto con la finalidad de reducir el número de candidatos a ser variantes denominativas. Primero hicimos uso del corpus de Ingeniería de Software.

Nivel de similitud	Termostat[48]	
	Con Frecuencia	Con PMI
1.0 a 0.9	340	222
0.9 a 0.8	1,342	295
0.8 a 0.7	7,872	475
0.7 a 0.6	29,452	732
0.6 a 0.5	82,461	1,286
TOTAL	**121,468**	**3,011**

Tabla 10. Número de pares de variantes denominativas obtenidas utilizando Termostat en Ingeniería de Software

Nivel de similitud	Termext[49]	
	Con Frecuencia	Con PMI
1.0 a 0.9	38	21
0.9 a 0.8	308	26
0.8 a 0.7	2575	111
0.7 a 0.6	12,539	386
0.6 a 0.5	40,644	1,005
TOTAL	**56,104**	**1,549**

Tabla 11. Número de pares de variantes denominativas utilizando Termext en Ingeniería de Software

En las tablas 11 y 12, se muestra el número de pares de variantes denominativas obtenidas utilizando las medidas de asociación de frecuencia y de PMI, clasificadas de acuerdo al nivel de similitud obtenido. En ellas claramente se puede observar que al utilizar la medida de asociación PMI, el total de resultados se reduce en un gran porcentaje ya que esta medida no considera únicamente la frecuencia de coocurrencia, sino también considera la frecuencia de ocurrencia de manera independiente en el corpus para cada término; además obtuvimos una mayor cantidad de resultados, es decir, posibles variantes denominativas al utilizar los términos que obtuvimos con el extractor automático Termostat.¿

[48] Apéndice “Listado completo de niveles de similitud Ingeniería de software-Termostat”.

[49] Apéndice “Listado completo de niveles de similitud Ingeniería de software-Termext”.

TERMINO 1	TERMINO 2	SIMILITUD
INGENIERÍA INFORMÁTICA	INGENIERÍA INFORMÁTICA	1
CONJUNTO DE REGLAS SINTÁCTICO-SEMÁNTICAS	REGLAS SINTÁCTICO-SEMÁNTICAS	0.966173986
ARQUITECTURA DE INFORMACIÓN AUTOR	INFORMACIÓN AUTOR	0.965477056
OR SETTINGS	OR TEAM OR FIRMS	0.959786489
FÁBRICA DE TELARES	PRODUCCIÓN DE MATERIA PRIMA	0.939283641
LÍDERES DE GRUPO	LÍDERES SERVIDORES	0.927649156
TASA DE ACEPTACIÓN	TASA DE RECHAZO	0.910237048
FLUJO DE VALOR	MAPA DE FLUJO	0.909542727
EXPORTADORES DE SERVICIOS	PAÍSES EXPORTADORES	0.907787424
CONTROLES DE FORMULARIOS	POSICIÓN DE CONTROLES	0.904502606
FLEXIBILIDAD DE USO	MITIGACIÓN DE RIESGOS AMBIENTALES	0.894741424
GESTO DISCRETO	TIPOS DE GESTOS	0.885160007
ADMINISTRADOR DE BASE	CERTIFICADO DE ADMINISTRADOR	0.88083403
GESTIÓN DE RECURSOS	GESTIÓN DE RECURSOS HUMANOS	0.874275016
LIDERAZGO RESPONSABLE	TI ESPECIALIZADO	0.867647072
BACKLOG DE PROYECTOS	LISTA PRIORIZADA DE PROYECTOS	0.863096235

Tabla 12. Similitud de términos utilizando Termext – PMI

TERMINO 1	TERMINO 2	SIMILITUD
CONCEPTOS DE ADMINISTRACIÓN	CONCEPTOS DE ADMINISTRACIÓN	1
DE SOFTWARE	SOFTWARE	0.976099964
ABRIL	NOVIEMBRE	0.952318044
ARQUITECTURA DE INFORMACIÓN AUTOR	INFORMACIÓN AUTOR	0.935414347
DESARROLLO	SOFTWARE	0.926737425
CONJUNTO DE REGLAS SINTÁCTICO-SEMÁNTICAS	REGLAS SINTÁCTICO-SEMÁNTICAS	0.925
FÁBRICA DE TELARES	PRODUCCIÓN DE MATERIA PRIMA	0.919866211
OR SETTINGS	OR TEAM OR FIRMS	0.919866211
AGOSTO-OCTUBRE	MAYO-JULIO	0.918090567

NORMA	NORMA ISO	0.916937707
DE SOFTWARE	DESARROLLO	0.915671964
AGOSTO-OCTUBRE	SEPTIEMBRE	0.913171556
CASOS DE PRUEBA	PRUEBA	0.911544049
OPERADORES	OPERADORES DE MUTACIÓN	0.910432736
MEJORA	PROCESOS	0.909922646

Tabla 13. Similitud de términos utilizando Termext – Frecuencia

En las tablas 13 y 14 mostramos los pares de términos obtenidos con mayor nivel de similitud utilizando Termext como extractor terminológico. Existen pares de términos iguales como es el caso de "INGENIERÍA INFORMÁTICA", esto debido a que durante la extracción de los mismos, se realiza un etiquetado de partes de la oración y se asignó más de una categoría gramatical a dicho término, haciendo que éste se repita. Sin embargo, una vez realizado el proceso de obtención de variantes denominativas, su nivel de similitud es de 1, ya que es el mismo término.

TERMINO 1	TERMINO 2	SIMILITUD
F-SECURE	F-SECURE	1
CATEGORÍA DE DESOCUPACIÓN	DESOCUPACIÓN	0.999765245
DEFECT REMOVAL	DEFECT	0.999729679
DISEÑO POSMODERNO	POSMODERNO	0.999147833
DEFECTOS CERRADOS	TOTAL DE ANTIGÜEDAD	0.999133576
FORMA|SOMBRERO	FORMA|TRONCO	0.999077242
INGENIERÍA|INGENIERÍAS	INGENIERÍA	0.998371272
MÉTODO ÁGIL|MÉTODOS ÁGILES	MÉTODOS ÁGILES	0.997240164
REALIZAR|REALIZARA|REALIZARTE|REALIZARLOS|REALIZAREMOS	REALIZAR	0.996930495
OBJETIVO|OBJETIVOS|OBJETIVA	OBJETIVO|OBJETIVOS	0.996798287
CONCEPTOS|HERMANOS	CONCEPTOS	0.996472709
EXPERTO|EXPERTA|EXPERTAS	EXPERTO|EXPERTA	0.994251752
EMPRESAS TRACTORAS	TRACTORAS	0.991838354
DEFINIR|DEFINIRÁ|DEFINIRLA|DEFINIRLOS|DEFINIREMOS	DEFINIR	0.991338934
ARCHIVAR|ARCHIVO	ARCHIVO	0.991109347

Tabla 14. Similitud de términos utilizando Termostat – PMI

TERMINO 1	TERMINO 2	SIMILITUD
VARIATION	VARIATION	1
INGENIERÍA|INGENIERÍAS	INGENIERÍA	0.99996351
OBJETIVO|OBJETIVOS|OBJETIVAS	OBJETIVO|OBJETIVOS	0.999856713
REALIZAR|REALIZARA|REALIZARTE|REALIZARLOS|REALIZAREMOS	REALIZAR	0.999824193
DISEÑO|DISEÑOS	DISEÑO	0.999823412
FORMA|SOMBRERO	FORMA|TRONCO	0.999734961
CONCEPTOS|HERMANOS	CONCEPTOS	0.999569894
DESARROLLO|DESARROLLOS	DESARROLLO	0.999521656
DEFINIR|DEFINIRÁ|DEFINIRLA|DEFINIRLOS|DEFINIREMOS	DEFINIR	0.998884077
VALORES NULOS	VALORES NULO|VALORES NULOS	0.998876463
PLANIFICACIÓN|PLANIFICACIONES	PLANIFICACIÓN	0.998749185
DESARROLLAR|DESARROLLARÁ|DESARROLLÉ|DESARROLLARLOS|DESARROLLAREMOS	DESARROLLAR	0.998731496
CONTROLAR|CONTROLARÁ|CONTROLARON|CONTROLARLOS	CONTROLAR	0.998616728
MUESTRA|MUESTRAS	MUESTRA	0.998576554
VALOR BAJO|VALORES BAJOS	VALOR BAJO	0.998563011

Tabla 15. Similitud de términos utilizando Termostat – Frecuencia

En las tablas 15 y 16 mostramos los primeros resultados obtenidos utilizando Termostat como extractor terminológico, donde de igual manera existen pares de términos iguales debido al etiquetado de partes de la oración, pero al ser la misma palabra, sus contextos son iguales y por lo tanto su nivel de similitud es el máximo posible.

Otro aspecto importante a notar es que la mayoría de las relaciones cuyo nivel de similitud es 1 o cercano a 1, corresponden a términos donde uno de ellos forma parte del otro, por ejemplo los términos "CATEGORÍA DE DESOCUPACIÓN" y "DESOCUPACIÓN" cuyo nivel de similitud es del 0.999765245, muy cercana a uno, donde el término "DESOCUPACIÓN" forma parte del término "CATEGORÍA DE

DESOCUPACIÓN", y por el hecho de encontrarse juntos en el texto, sus contextos son muy similares.

Este mismo procedimiento lo realizamos utilizando el corpus de Biodiversidad, junto con los términos obtenidos con ambos extractores donde obtuvimos los siguientes resultados:

Nivel de similitud	Termostat[50]	
	Frecuencia	PMI
1.0 a 0.9	112	128
0.9 a 0.8	442	250
0.8 a 0.7	1,757	486
0.7 a 0.6	6,212	760
0.6 a 0.5	19,631	1,275
TOTAL	**28,154**	**2,899**

Tabla 16. Número de variantes denominativas utilizando Termostat en Biodiversidad

Nivel de similitud	Termext[51]	
	Frecuencia	PMI
1.0 a 0.9	126	104
0.9 a 0.8	1,151	417
0.8 a 0.7	3,662	1,593
0.7 a 0.6	11,325	3,734
0.6 a 0.5	29,893	7,521
TOTAL	**46,157**	**13,369**

Tabla 17. Número de variantes denominativas utilizando Termext en Biodiversidad

En las tablas 17 y 18, mostramos el total de variaciones denominativas obtenidas, donde el número total de resultados fue mayor al utilizar Termext como extractor terminológico automático, esto debido a que el número de

[50] Apéndice "Listado completo de niveles de similitud Biodiversidad-Termostat".

[51] Apéndice "Listado completo de niveles de similitud Biodiversidad-Termext"

términos obtenidos también es mayor. Además se observa que al utilizar PMI como medida de asociación, el número de resultados disminuye.

TERMINO 1	TERMINO 2	SIMILITUD
PROGRAMA DE PROTECCIÓN	PROGRAMA DE PROTECCIÓN	1
ARCO IRIS	TRUCHA ARCO IRIS	0.986522501
ANAGNOSTIDIS	KOMÁREK	0.985283176
CONSERVACIÓN EX	CONSERVACIÓN EX SITU	0.97833553
AGUA PEQUEÑOS	CUERPOS DE AGUA PEQUEÑOS	0.9710916
AESCULIFOLIA	CEIBA AESCULIFOLIA	0.970599565
EDUCADORES	EDUCADORES AMBIENTALES	0.967909271
NATURALES PROTEGIDAS	ÁREAS NATURALES PROTEGIDAS	0.96581132
WIMMERIA	WIMMERIA PERSICIFOLIA	0.963351192
INSTITUTO NACIONAL DE ESTADÍSTICA	NACIONAL DE ESTADÍSTICA	0.962682097
LAGUNCULARIA	LAGUNCULARIA RACEMOSA	0.96191689
REGIONES PRIORITARIAS TERRESTRES	REGIONES PRIORITARIAS TERRESTRES DE CONSERVACIÓN	0.960097327
PRECIPITACIÓN TOTAL	PRECIPITACIÓN TOTAL ANUAL	0.958628725
CHIPE MANGLERO	MANGLERO	0.957088867
ANIM	TERRÍCOLA WUTZ ANIM	0.953570997

Tabla 18. Similitud de términos utilizando Termext - PMI

TERMINO 1	TERMINO 2	SIMILITUD
PROGRAMA DE PROTECCIÓN	PROGRAMA DE PROTECCIÓN	1
PDF	PDF PAG	0.978402298
PATAS BLANCAS	RATONES DE PATAS BLANCAS	0.973879114
PAG	PDF	0.97011676
MG	PS	0.969949714
BRITTON ET ROSE	ET ROSE	0.958333333
TIPO CRÓNICO	TIPO CRÓNICO DEGENERATIVO	0.958333333
ANAGNOSTIDIS	KOMÁREK	0.953462589
ANIM	TERRÍCOLA WUTZ ANIM	0.95
NAL TEL	TEL	0.95
PURPURA	PURPURA PATULA	0.95

LOPHOCAMBA CIBRIANI	LOPHOCAMBA CIBRIANI BEUTELSPACHER	0.949070753
PAG	PDF PAG	0.94496784
AGREGADO	VALOR AGREGADO	0.943481037

Tabla 19. Similitud de términos utilizando Termext – Frecuencia

En las tablas 19 y 20 mostramos los pares de términos obtenidos con mayor nivel de similitud utilizando Termostat como extractor terminológico.

TERMINO 1	TERMINO 2	SIMILITUD
MESETA CENTRAL|MESETAS CENTRALES	MESETA CENTRAL	1
PIM PIM	PIM	0.99902876
ASENTAMIENTO HUMANO|ASENTAMIENTOS HUMANOS	ASENTAMIENTOS HUMANOS	0.992742529
LEÑA|LEÑAS	LEÑA	0.990875142
TANGENTES TRIFÁSICOS	TRIFÁSICOS	0.990682757
ACTIVIDADES DE COLECTA CIENTÍFICA	ACTIVIDADES DE COLECTA	0.990610679
RECURSO NATURAL|RECURSOS NATURALES	RECURSOS NATURALES	0.989534075
RATÓN DE PATAS BLANCAS|RATONES DE PATAS BLANCAS	RATÓN DE PATAS|RATONES DE PATAS	0.989184695
PARTE CENTRAL|PARTES CENTRALES	PARTE CENTRAL	0.987883749
BROWNORUM	LITHOBATES BROWNORUM	0.984676903
PRECIPITACIÓN MEDIA ANUAL	PRECIPITACIÓN MEDIA	0.981104292
BALUMILAL TSELTAL	YAXCHI	0.979938132
POBLACIÓN URBANA	POBLACIÓN URBANA	0.979285099
ZONAS PROTECTORAS FORESTALES	ZONAS PROTECTORAS	0.977832937
REGAR|RIEGAN|RIEGO	RIEGO|RIEGOS	0.977805894

Tabla 20. Similitud de términos utilizando Termostat – PMI

TERMINO 1	TERMINO 2	SIMILITUD
MESETA CENTRAL|MESETAS CENTRALES	MESETA CENTRAL	1
RECURSO NATURAL|RECURSOS	RECURSOS NATURALES	0.998872412

NATURALES		
LEÑA|LEÑAS	LEÑA	0.997060076
ASENTAMIENTO HUMANO|ASENTAMIENTOS HUMANOS	ASENTAMIENTOS HUMANOS	0.996591956
TALAR|TALA	TALA|TALAS	0.99462079
REGAR|RIEGAN|RIEGO	RIEGO|RIEGOS	0.993924579
ESPECIE INVASORA|ESPECIES INVASORAS	ESPECIES INVASORAS	0.991657769
CAMBIO CLIMÁTICO|CAMBIOS CLIMÁTICOS	CAMBIO CLIMÁTICO	0.989849772
COLECTAR|COLECTA	COLECTA	0.989597501
INCENDIO FORESTAL|INCENDIOS FORESTALES	INCENDIOS FORESTALES	0.987812307
MONITOREAR|MONITOREO	MONITOREO	0.986600185
PALO|PALOS|PALILLOS	PALO	0.986206599
PARTE CENTRAL|PARTES CENTRALES	PARTE CENTRAL	0.984285158
BALSAS	BALSA|BALSAS	0.982074389
TANGENTES TRIFÁSICOS	TRIFÁSICOS	0.978019294

Tabla 21. Similitud de términos utilizando Termostat – Frecuencia

En las tablas 21 y 22 mostramos los primeros resultados obtenidos utilizando Termostat como extractor terminológico, donde de igual manera existen pares de términos iguales debido al etiquetado de partes de la oración y por lo tanto su nivel de similitud es el máximo posible.

Finalmente se observa que el número los resultados utilizando el extractor terminológico Termostat son mucho menores en comparación a Termext para ambos corpus, además de que se reducen aún más al utilizar PMI como medida de asociación entre los términos y las palabras de sus contextos.

Específicamente para el corpus de Ingeniería de Software y con base en mi experiencia y el conocimiento adquirido durante mi formación en dicha área de especialidad, las variantes denominativas obtenidas, son poco claras y poco

coherentes. En cambio para el corpus de Biodiversidad los resultados obtenidos son mucho mejores, ya que dichas variantes muestran términos específicos de dicha área, aunque desconocemos el significado de los mismos.

Finalmente es necesario evaluar y comprobar si los términos obtenidos de manera automática y las variantes denominativas obtenidas de acuerdo a su grado de similitud son correctos, y realmente equivalen a variantes denominativas en sus respectivas áreas de especialidad. Esta evaluación la realizamos utilizando el corpus de Biodiversidad únicamente, ya que los resultados que obtuvimos utilizando el corpus de Ingeniería de Software fueron poco favorables.

6. Resultados y Evaluación

Como ya mencionamos en el capítulo anterior, necesitamos realizar una validación de los resultados que obtuvimos y comprobar si son correctos y realmente equivalen a variantes denominativas con cierto grado de similitud, además de verificar si la metodología descrita anteriormente funciona correctamente. Para ello elaboramos y realizamos un proceso de validación haciendo uso de uno de los corpus especializados que ya habíamos obtenido y utilizado anteriormente.

Primero solicitamos una serie de términos a expertas en el área de Ingeniería de Software, las cuales nos brindaron 6 variantes denominativas claramente establecidas dentro de dicha área:

1. Gestión - Administración
2. Computadora - Ordenador
3. Requerimientos - Requisitos
4. Archivos - Ficheros
5. Pruebas - Ensayos
6. Plantillas – Formularios

Al observar los resultados obtenidos en la extracción automática de términos, tanto con Termext como con Termostat y al realizar la obtención de las variantes denominativas, pudimos observar que algunas de ellas no aparecieron dentro de los resultados, o incluso no fueron obtenidos como términos por ninguno de los extractores automáticos, tal como se muestra en la siguiente tabla.

Término	Similares	Posición Termostat	Posición Termext	Especificidad	NC-Value	Frec
Gestión	Gestión	-	109	-	60.03	1451
	Gestión de proyecto	178	84	45.65	76.81	130
	Proceso de gestión	265	453	38.69	20.67	94
Administración	Administración	-	202	-	36.86	660
	Administración de	443	559	29.82	16.37	52

Término	Similares	Posición Termostat	Posición Termext	Especificidad	NC-Value	Frec
	proyecto					
	Proceso de administración	475	7777	28.60	2.58	48
Computadora	Computadora	-	476	-	20	243
	Computadora personal	2835	3050	11.16	4	8
Ordenador	-	-	-	-	-	-
Requerimientos	Requerimientos	-	77	-	82.23	1431
	Análisis de requerimiento	848	1283	21.51	7.75	27
	Requerimiento funcional	430	1992	30.14	11.5	52
Requisitos	Requisitos	-	158	-	46.95	867
	Análisis de requisitos	1163	456	17.89	20.67	19
	Requisitos de software	10441	723	4.40	12.92	2
Archivos	Archivos	-	708	-	13.28	213
	Archivo de dato	1217	6192	17.43	13.28	20
	Archivo lógico	3140	-	10.33	-	7
Ficheros	-	-	-	-	-	-
Pruebas	Pruebas	-	15	-	187.76	3026
	Casos de prueba	37	273	93.64	31.01	494
	Proceso de pruebas	171	71	47.20	87.88	126
	Pruebas unitarias	213	160	42.21	47	101
Ensayos	Ensayo	-	-	-	-	-
	Centro de ensayo	10428	5746	4.40	2.58	2
Plantillas	Plantillas	-	7626	-	4	138
Formularios	-	-	-	-	-	-

Tabla 22. Evaluación de variantes denominativas - Ingeniería de Software.

En la tabla 22 se muestra el término indicado por el especialista, algunos términos similares encontrados, así como la posición en la que se encontró dentro del listado de términos con Termostat y con Termext, sus medidas de extracción y la frecuencia con que aparecen en el corpus.

Podemos observar que algunos términos sugeridos por el especialista no se encuentran dentro de los resultados obtenidos mediante Termostat. Sin embargo, algunos términos poliléxicos similares sí se encuentran en el listado de

ambos extractores pero con una medida de extracción y una frecuencia muy baja, además no obtuvimos estas variantes denominativas entre los resultados con un nivel de similitud mayor a 0.5.

Otro aspecto importante en el caso de los términos “Formularios”, “Ensayos”, “Ficheros” y “Ordenador”, es que no fueron obtenidos por los extractores como términos del área. Sin embargo, el especialista sí los consideró como tal, por lo que decidimos obtener los contextos en los que aparecen cada uno de los términos sugeridos y de esta manera poder analizar si realmente son variantes denominativas dentro del corpus de Ingeniería de Software recopilado[52].

Observamos en los contextos de los términos indicados por los expertos, que dichos términos realmente corresponden a variantes denominativas, aunque algunos de ellos más claramente que otros tras la revisión humana. Para el área de especialidad de Ingeniería de Software, en algunos casos los extractores terminológicos no obtuvieron los términos indicados por los expertos, y en algunos otros la metodología propuesta no funcionó correctamente.

Para la evaluación en el área de la Biodiversidad, como ya mencionamos en el capítulo anterior, la CONABIO nos proporcionó una serie de archivos con términos identificados de manera automática mediante la “Biblioteca para la minería de textos de biodiversidad en español” (Barrios, Molina, Sierra-Alcocer, & Zenteno-Jimenez, 2015), la cual ha sido utilizada por la CONABIO en diversos proyectos.

Según la documentación de la biblioteca[53], el módulo de “Extracción de información de usos tradicionales de especies” detecta fragmentos textuales que contienen fragmentos textuales de este tipo. El módulo identifica los fragmentos

[52] Ver apéndice “Contextos de evaluación para los términos de Ingeniería de Software.”

[53] https://bitbucket.org/conabio_cmd/text-mining

que mencionan el nombre de alguna especie entre otros patrones sintácticos como son: localidad, verbos funcionales, partes y estructuras, nombres comunes, formas de aplicación, entre otros.

```
[[5, "Crustacea", [12, 21]], [13, "Eumalacostraca",
[143, 157]], [13, "Stomatopoda", [165, 176]], [13,
"Decapoda", [204, 212]], [16, "Lysiosquilla
campechiensis", [106, 132]], [16, "Pseudorhombila
ometlanti", [183, 207]], [16, "Batodaeus adanad",
[241, 257]], [20, "Bullis", [215, 221]], [20,
"Bullis", [229, 235]], [23, "Squilla empusa", [109,
123]], [23, "Callinectes sapidus", [134, 153]], [23,
"Sycionia brevirostris", [172, 193]], [26, "Raninoides
louisianensis", [61, 85]], [26, "Raninoides lamarcki",
[87, 98]], [26, "Persephona mediterranea", [100,
123]], [26, "Iliacantha subglobosa", [125, 146]], [26,
"Stenorhynchus seticornis", [148, 172]], [26,
"Anasimus latus", [175, 189]], [27, "Porcellana
sayana", [74, 91]], [27, "Moreiradromia antillensis",
[93, 118]], [27, "Hypoconcha sabulosa", [120, 139]],
[27, "Calappa sulcata", [141, 156]], [27, "Calappa
flammea", [158, 168]], [27, "Hepatus epheliticus",
[170, 189]], [27, "Libinia dubia", [192, 205]], [45,
"Costera", [84, 91]]]
```

Dichas características se extrajeron de manera automática, y encontramos diferencias entre cada uno de los archivos recibidos, como se puede observar en el siguiente ejemplo, donde ambos archivos cuentan con el nombre del archivo y el fragmento del texto al que se hace referencia, pero el archivo #1 cuenta con características (partes_o_estructuras, verbo_aprovechar, forma_de_aplicacion, verbo_utilizar, alimentación) diferentes al archivo #2 (verbo_aprovechar, combustible, agrosistemas, verbo_usar).

Archivo 1: tika-yucatan-parte-3-cap7-diversidad-de-plantas-forrajeras-pdf.seg.txt

Fragmento: Los pobladores de las comunidades estudiadas utilizan una gran variedad de plantas procedentes de selvas, vegetación secundaria, milpa y huertos familiares (Acosta y otros, 1998), de las cuales aprovechan: hojas, raíces, toda la planta, frutas, semillas, cáscara para alimentar gallinas, patos, puercos, vacas, caballos, cabras, ovejas, a veces pavos, venados y puercos de monte.

partes_o_estructuras
- hojas
- raíces
- toda la planta
- semillas
- cáscara

verbo_aprovechar
- aprovechan
- aprovechan

commonName
- milpa
- Arellano
- Sosa

forma_de_aplicacion
- introducidos

verbo_utilizar
- utilizada
- utilizan
- utilizan

alimentacion
- frutas

Archivo 2: tika-campeche-capitulo4-reino-vegetal-manglar-pdf.seg.txt

Fragmento: Por otra parte existen normas que deben ser tomadas en cuenta para la protección del manglar: la noM-001-seMarnat-1996 que establece los límites permisibles de contaminantes en las descargas de aguas residuales y bienes nacionales, la noM-012-recnat-1996, donde se instituyen los criterios y especificaciones para efectuar el aprovechamiento de leña para uso doméstico, noM-022-seMarnat-2003 que menciona los lineamientos para la preservación, conservación y restauración de los humedales

costeros en zonas de manglar, la noM- 059-seMarnat-2001 que establece la conservación o uso sustentable de mangle, las noM-060-seMarnat-1994, noM-061-seMarnat-1994 y noM-062-seMarnat-1994 que detallan criterios para mitigar los efectos adversos ocasionados en los suelos y cuerpos de agua, ?ora y fauna por aprovechamiento forestal y en la biodiversidad por cambio de uso de suelo de terrenos forestales agropecuarios, respectivamente.

```
verbo_aprovechar                    mangle
     aprovechamiento                noM
     aprovechamiento                noM
combustible                         noM
     leña                      agrosistemas
commonName                          agropecuarios
     noM                       verbo_usar
     noM                            uso
     noM                            uso
     noM                            uso
```

Varios de los términos identificados dentro de los archivos contienen las características de nombre común y nombre científico, lo cual fue de gran utilidad para validar los resultados que obtuvimos anteriormente, así como validar la metodología propuesta en este trabajo de investigación. Para lo anterior únicamente fue necesario el uso de los términos identificados por el experto, que obtuvimos mediante el uso de una expresión regular[54], resultando en un total de 6,292 términos.

[54] Una expresión regular es una secuencia de caracteres que forma un patrón de búsqueda, principalmente utilizada para la búsqueda de patrones de cadenas de caracteres, la expresión regular utilizada fue (*(["'])(?:(?=(\\?))\2.)*?\1*)

AVES
TOLUCA
OCHROMA PYRAMIDALE
SAPIUM LATERIFLORUM
TORREZ
TARASCO
CORREA
COSTERA
TETIS
ESCALERA
MOTA
SUMIDERO
CRACIDAE
PENELOPINA NIGRA
PHASIANIDAE
COLINUS VIRGINIANUS
PSITTACIDAE
ARATINGA HOLOCHLORA
TROCHILIDAE
TILMATURA DUPONTI
TYRANNIDAE
XENOTRICCUS
CALLIZONUS
PARULIDAE
OPORORNIS TOLMIEI
THRAUPIDAE
TANGARA LEUCOPTERA
CARDINALIDAE
PASSERINA CAERULEA
SARCORAMPHUS PAPA
MADERA

Figura 11. Ejemplo de términos proporcionados por CONABIO

Como podemos observar en la Figura 11, obtuvimos términos poliléxicos y monoléxicos, específicamente 4,258 poliléxicos y 2,034 monoléxicos, que al igual que los términos obtenidos mediante los extractores automáticos en el capítulo anterior, los utilizamos para obtener las variantes denominativas, haciendo uso de la misma metodología.

Una vez que obtuvimos los términos procedimos a realizar la metodología propuesta, haciendo uso del corpus de Biodiversidad etiquetado con Freeling para evitar que el contexto de un término incluyera varias palabras con el mismo lema.

Realizamos varias pruebas con diferentes tipos de términos obtenidos, con la finalidad de comprobar que la metodología propuesta en este trabajo de investigación funciona correctamente y que realmente se obtienen las variantes denominativas correspondientes dentro del área de especialidad.

En todas las pruebas que realizamos, seguimos la metodología descrita, obtuvimos los contextos de cada uno de los términos proporcionados por la CONABIO, conforme a la ventana establecida, es decir, 5 palabras a la derecha y 5 palabras a la izquierda de cada uno de los términos, además obtuvimos un listado de palabras y términos con sus frecuencias de aparición en el corpus.

Cabe mencionar que una vez realizado este procedimiento de obtención de contextos, pudimos observar que 624 de los 6,292 términos totales identificados por los expertos, no fueron encontrados en el corpus una vez que fue realizado el etiquetado de las partes de la oración con Freeling.

Una vez que obtuvimos los contextos, procedimos a obtener las matrices de coocurrencia utilizando la frecuencia de coocurrencia y PMI como medidas de asociación entre los términos utilizados y las diferentes palabras del texto, excluyendo a las palabras funcionales. Finalmente calculamos las distancias coseno entre cada par de términos, y con ello sus niveles de similitud, considerando únicamente aquellas variantes cuyo nivel de similitud resultó mayor a 0.5. A continuación presentamos las evaluaciones que realizamos:

Evaluación #1

Entre las características de la "Biblioteca para la minería de textos de biodiversidad en español" (Barrios, Molina, Sierra-Alcocer, & Zenteno-Jimenez, 2015), se encuentra la detección de nombres comunes y nombres científicos.

Relaciones entre nombres proporcionadas por CONABIO		
No.	**Nombre(s) común(es)**	**Nombre(s) Científico(s)**
1	colorín árnica laurel cuero	Erythrina flabelliformis Prosopis laevigata Serjania schiedeana
2	chit tasiste cocoyol	Thrinax radiata Coccothrinax readii Acrocomia aculeata
3	guayacán	G. sanctum
4	guayaba	Psidium guajava
5	peyote	Stenocereus thurberi S. montanus Opuntia spp

Relaciones entre nombres proporcionadas por CONABIO		
No.	**Nombre(s) común(es)**	**Nombre(s) Científico(s)**
		Opuntia ficus-indica
6	tinto chi ich venado jabalí	Cameraria latifolia Coccoloba cozumelensis Haematoxylum campechianum Hyperbaena winzerlingii Neomillspaughia emarginata
7	Encino Pino palo blanco	Quercus spp Ipomoea arborescens Acacia pennatula
8	grado Campeche guayacán	g. mangle
9	frío pino	P. michoacana P. montezumae P. pseudostrobus P. teocote
10	guayacán cedro	Cedrela odorata Bursera simaruba
11	palo dulce guamúchil tehuistle cubata cuahulote copal copal chino brasil guaje huizache cirian ciruelo ocote ayacahuite abeto aile	Eysenhardtia polystachya Pithecellobium dulce Acacia bilimekii Lysiloma acapulcensis L. Divaricata Acacia cochliacantha Prosopis laevigata Guazuma ulmifolia Bursera glabrifolia B. bipinnata Gliricidia sepium Leucaena esculenta Acacia farnesiana Crescentia alata Spondias mombin Pinus montezumae

Relaciones entre nombres proporcionadas por CONABIO		
No.	**Nombre(s) común(es)**	**Nombre(s) Científico(s)**
	madroño	P. teocote P. pseudostrobus r. cornuta Abies religiosa Quercus spp Arbutus xalapensis
12	cedro pino oyamel	Abies religiosa
13	anona palo de corcho corcho	A. glabra
14	yuca palmilla ocotillo candelilla gobernadora	Simmondsia chinensis
15	carrizo café	Chusquea longifolia
16	árbol de Navidad	Pinus oocarpa Abies religiosa Pinus pseudostrobus P. douglasiana Quercus laurina Q. rugosa
17	pino encino	Pinus montezumae P. leiophylla P. teocote P. pringlei P. oocarpa P. michoacana r. cornuta Quercus rugosa Q. obtusata

Relaciones entre nombres proporcionadas por CONABIO		
No.	**Nombre(s) común(es)**	**Nombre(s) Científico(s)**
		Q. laurina Q. castanea Q. crassifolia
18	carrizo café	Chusquea longifolia
19	pino encino táscate pino piñonero	Quercus arizonica Q. emoryi Q. grisea Juniperus deppeana J. communis
20	Guatemala ahuehuete sabino	J. comitana J. gamboana J. deppeana Taxodium mucronatum
21	Aile huejote cedro blanco cucharillo oyamel árbol de navidad	Pinus spp Quercus spp Alnus acuminata Arbutus spp Salix paradoxa Clethra mexicana
22	cangrejo ermitaño mangle	R. mangle
23	achiote Campeche	Bixa orellana
24	café chintok iich	B. simaruba Krugiodendron ferreum Malmea depressa Diospyros cuneata Bauhinia divaricata Carica papaya H. albicans

Tabla 23. Relaciones entre términos comunes y científicos proporcionados por CONABIO del corpus de Biodiversidad

Esta primera prueba la realizamos utilizando únicamente los nombres científicos identificados por los expertos y se esperaba que los 98 nombres científicos obtenidos se encontraran dentro de la lista de 6,292 términos proporcionada por la CONABIO, pero únicamente se encontraron 78 nombres científicos. Aplicando la metodología propuesta, obtuvimos los siguientes resultados:

Nivel de similitud	Resultados	
	Frecuencia	PMI
1.0 a 0.9	0	0
0.9 a 0.8	0	0
0.8 a 0.7	8	0
0.7 a 0.6	29	3
0.6 a 0.5	36	6
TOTAL	73	9

Tabla 24. Total de Variantes Denominativas - Evaluación #1

Como podemos observar son muy pocas las variantes denominativas obtenidas y sus niveles de similitud son bajos, sin embargo, las variantes obtenidas corresponden en su mayoría a las identificadas y proporcionadas previamente por la CONABIO.

A continuación mostramos las variantes denominativas obtenidas cuyo nivel de similitud es mayor a 0.5 y aunque su nivel de similitud es bajo corresponden a variantes denominativas, por ejemplo el término "P_montezumae" y el término "P_pseudostrobus" son variantes de un nombre común, en este caso "pino", lo mismo ocurre con los términos "Q_emoryi" y "Quercus_arizonica".

PMI		
TERMINO 1	TERMINO 2	NIVEL SIMILITUD
P_montezumae	P_pseudostrobus	0.627472855
P_oocarpa	P_pringlei	0.620628183

Q_crassifolia	Q_laurina	0.604596021
P_leiophylla	P_teocote	0.581395909
Q_castanea	Q_crassifolia	0.539734712
Q_emoryi	Quercus_arizonica	0.529482192
P_michoacana	P_pseudostrobus	0.528176811
P_michoacana	P_teocote	0.523075158
P_oocarpa	P_teocote	0.509329117

Tabla 25. Variantes Denominativas con PMI - Evaluación #1

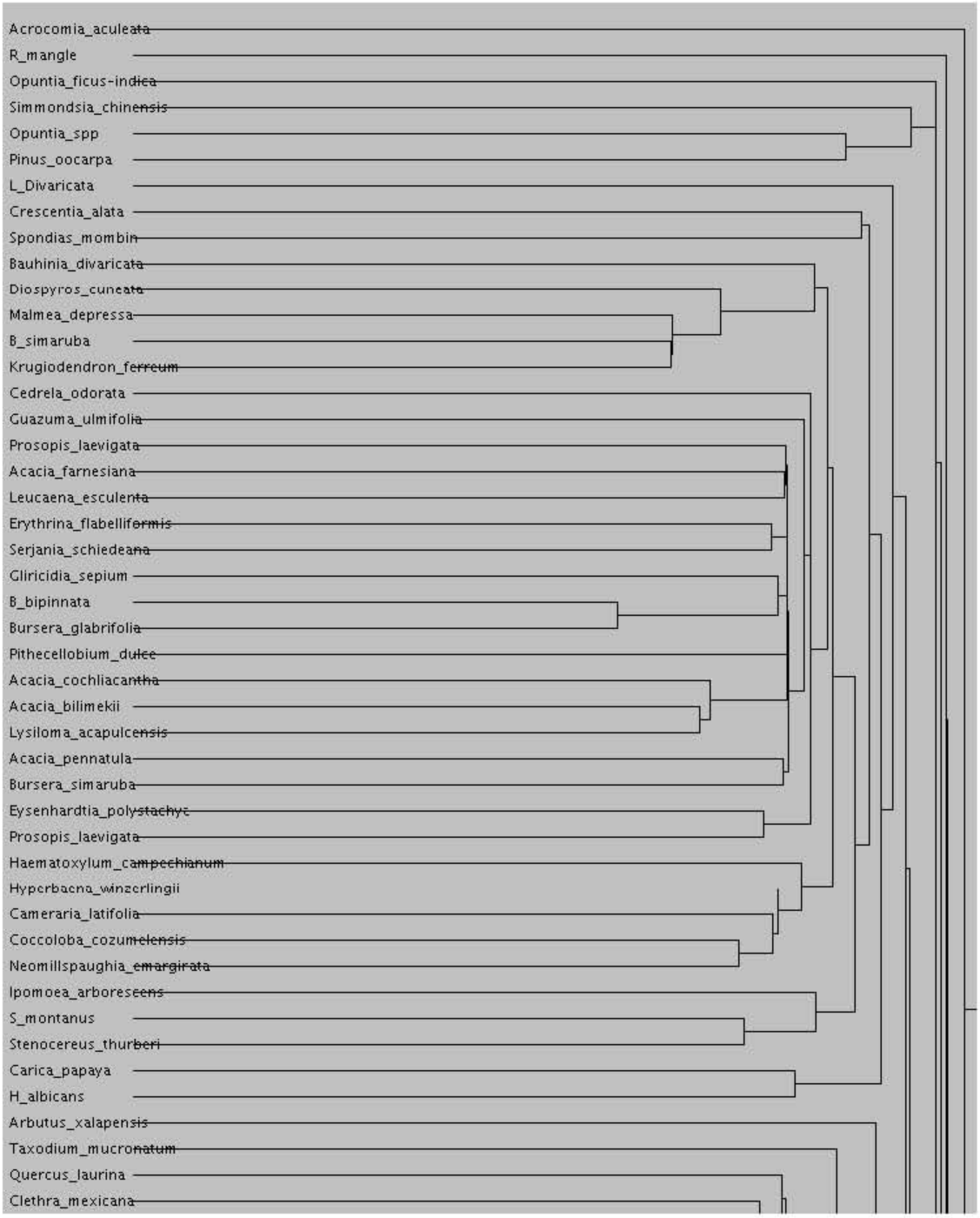
Frecuencia		
TERMINO 1	TERMINO 2	NIVEL SIMILITUD
Q_castanea	Q_obtusata	0.756351406
Q_laurina	Q_obtusata	0.751789979
P_montezumae	P_pseudostrobus	0.723572866
Q_crassifolia	Q_laurina	0.722222222
Q_castanea	Q_laurina	0.719864255
P_leiophylla	P_teocote	0.71022252
Q_castanea	Q_crassifolia	0.709799928
Q_crassifolia	Q_obtusata	0.702822815
Q_castanea	Q_rugosa	0.688255892
Q_laurina	Q_rugosa	0.681947675
P_leiophylla	P_michoacana	0.679708633
P_montezumae	P_pseudostrobus	0.67936622
Q_obtusata	Q_rugosa	0.677055444
P_oocarpa	P_pringlei	0.66958948
P_michoacana	P_montezumae	0.665640235
P_michoacana	P_pseudostrobus	0.663403472
P_douglasiana	P_montezumae	0.660494937
P_oocarpa	P_pseudostrobus	0.651868276
P_leiophylla	P_pringlei	0.649749308
Q_crassifolia	Q_rugosa	0.648898201
P_pseudostrobus	P_teocote	0.646646845
P_pringlei	P_teocote	0.641123832
P_pringlei	P_pseudostrobus	0.639137491
P_montezumae	P_teocote	0.634235662
P_montezumae	P_pringlei	0.630630744

Frecuencia		
TERMINO 1	TERMINO 2	NIVEL SIMILITUD
P_michoacana	P_michoacana	0.629459699
P_leiophylla	P_montezumae	0.629415024
P_leiophylla	P_teocote	0.62799682
P_montezumae	P_oocarpa	0.627877989
Pinus_montezumae	Pinus_montezumae	0.626935506
P_pseudostrobus	P_teocote	0.623625249
P_montezumae	P_teocote	0.617479841
P_douglasiana	P_pseudostrobus	0.616808459
P_pseudostrobus	P_pseudostrobus	0.613222363
P_leiophylla	P_oocarpa	0.611248771
P_leiophylla	P_pseudostrobus	0.610116913
P_michoacana	P_teocote	0.602078389
P_leiophylla	P_pseudostrobus	0.597452836
P_michoacana	P_pseudostrobus	0.597356133
P_michoacana	P_teocote	0.595818764
P_michoacana	P_pringlei	0.589238742
P_oocarpa	P_teocote	0.588467696
P_michoacana	P_oocarpa	0.585287088
P_leiophylla	P_michoacana	0.584505824
P_pringlei	P_teocote	0.582376514
P_michoacana	P_montezumae	0.581935697
Alnus_acuminata	Arbutus_spp	0.581318359
Pinus_spp	Quercus_spp	0.577562523
P_pseudostrobus	P_teocote	0.571040241
P_michoacana	P_pringlei	0.569958719
P_douglasiana	P_leiophylla	0.560403683
P_pringlei	P_pseudostrobus	0.556957993
P_oocarpa	P_pseudostrobus	0.552272651
P_leiophylla	P_teocote	0.551168486
P_michoacana	P_pseudostrobus	0.548128128
P_pseudostrobus	P_teocote	0.540620506
P_michoacana	P_teocote	0.535485231
Q_rugosa	Quercus_rugosa	0.533391581
P_teocote	P_teocote	0.532127643

Frecuencia		
TERMINO 1	TERMINO 2	NIVEL SIMILITUD
P_oocarpa	P_teocote	0.530336335
Q_castanea	Quercus_rugosa	0.527644853
P_douglasiana	P_oocarpa	0.525973954
P_michoacana	P_teocote	0.522202693
P_montezumae	P_teocote	0.520685266
Q_crassifolia	Quercus_rugosa	0.518262172
P_pringlei	P_teocote	0.517699489
P_douglasiana	P_teocote	0.513936217
P_michoacana	P_oocarpa	0.512195122
P_douglasiana	P_teocote	0.508181626
Q_laurina	Quercus_rugosa	0.505964426
Q_rugosa	Quercus_arizonica	0.502651402
P_michoacana	P_pseudostrobus	0.500541712

Tabla 26. Variantes Denominativas con Frecuencia - Evaluación #1

Finalmente, al aplicar el algoritmo de agrupamiento pudimos observar que aunque los niveles de similitud son bajos, agrupa las variantes identificadas y aquellos términos como "R. Mangle" y "Bixa Orellana" que no son variantes denominativas, se encuentran más alejados entre ellos.


Acrocomia_aculeata
R_mangle
Opuntia_ficus-indica
Simmondsia_chinensis
Opuntia_spp
Pinus_oocarpa
L_Divaricata
Crescentia_alata
Spondias_mombin
Bauhinia_divaricata
Diospyros_cuneata
Malmea_depressa
B_simaruba
Krugiodendron_ferreum
Cedrela_odorata
Guazuma_ulmifolia
Prosopis_laevigata
Acacia_farnesiana
Leucaena_esculenta
Erythrina_flabelliformis
Serjania_schiedeana
Gliricidia_sepium
B_bipinnata
Bursera_glabrifolia
Pithecellobium_dulce
Acacia_cochliacantha
Acacia_bilimekii
Lysiloma_acapulcensis
Acacia_pennatula
Bursera_simaruba
Eysenhardtia_polystachyc
Prosopis_laevigata
Haematoxylum_campechianum
Hyperbaena_winzerlingii
Cameraria_latifolia
Coccoloba_cozumelensis
Neomillspaughia_emarginata
Ipomoea_arborescens
S_montanus
Stenocereus_thurberi
Carica_papaya
H_albicans
Arbutus_xalapensis
Taxodium_mucronatum
Quercus_laurina
Clethra_mexicana

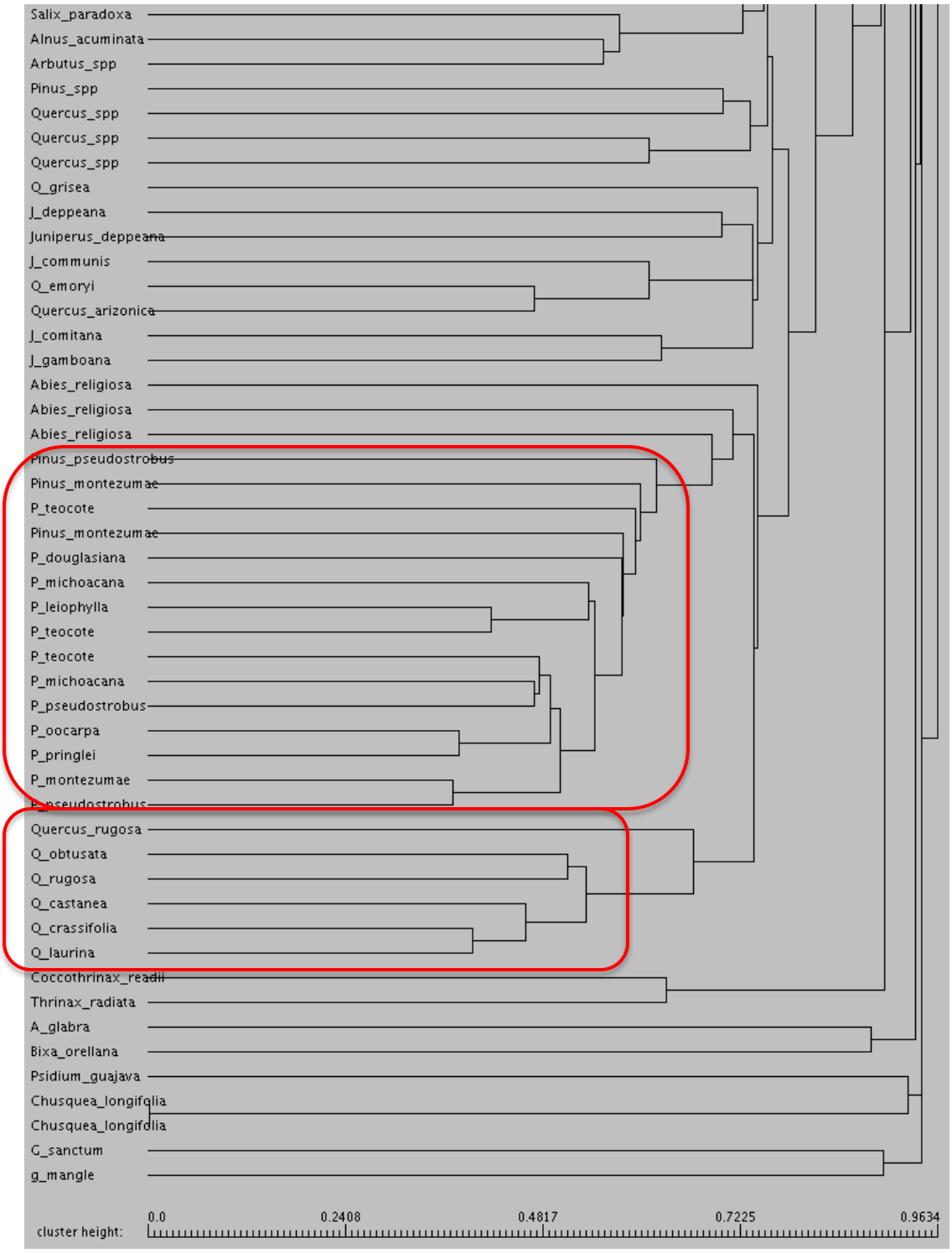


Figura 12. Dendograma - Evaluación #1

En esta primera evaluación pudimos observar que la metodología propuesta funciona, ya que únicamente de una lista de términos, en este caso nombres

científicos, pudimos obtener algunas variantes denominativas que, aunque el nivel de similitud entre los términos es bajo ya que no rebasa el 0.75, las variantes denominativas obtenidas en su mayoría son correctas.

Evaluación #2

Esta segunda evaluación es muy parecida a la anterior, debido a que ya que contábamos con los nombres científicos y nombres comunes de algunos elementos de la Biodiversidad de México, decidimos tomarlos por completo, a diferencia de la evaluación anterior donde únicamente hicimos uso de los nombres científicos, esperando obtener un mayor número de variantes denominativas correctas.

Seleccionamos de manera indiscriminada los términos, es decir, tanto nombres científicos como nombres comunes, en un listado, el cual fue la entrada para nuestro programa. Posteriormente seguimos con el mismo procedimiento que en los procesos anteriores, obtuvimos los contextos de cada término, la matriz de coocurrencia, los vectores semánticos y finalmente las distancias coseno y el nivel de similitud entre pares de términos, obteniendo los siguientes resultados:

Nivel de similitud	Resultados	
	Frecuencia	PMI
1.0 a 0.9	0	0
0.9 a 0.8	0	0
0.8 a 0.7	19	0
0.7 a 0.6	29	11
0.6 a 0.5	33	15
TOTAL	81	26

Tabla 27. Total de Variantes Denominativas - Evaluación #2

Como se puede observar en la tabla anterior, el número total de variantes denominativas obtenidas con cada una de las medidas de asociación utilizadas es diferente, obteniendo una mayor cantidad al utilizar únicamente la medida de

frecuencia. Además podemos ver que el nivel de similitud es muy bajo, se muestran a continuación las variantes denominativas obtenidas.

PMI		
TERMINO1	TERMINO2	NIVEL SIMILITUD
P_montezumae	P_pseudostrobus	0.697756243
Bursera_glabrifolia	copal_chino	0.695240022
Acacia_bilimekii	Tehuistle	0.64374624
ahuehuete	Sabino	0.627839654
Salix_paradoxa	huejote	0.623625598
P_oocarpa	P_pringlei	0.620628183
copal	copal_chino	0.620290006
Krugiodendron_ferreum	Chintok	0.614428072
B_bipinnata	copal_chino	0.612194439
Q_crassifolia	Q_laurina	0.604596021
Bursera_glabrifolia	Copal	0.60437277
P_leiophylla	P_teocote	0.583602705
Bixa_orellana	achiote	0.556436106
Acacia_bilimekii	cubata	0.549031135
P_michoacana	P_teocote	0.547023209
B_simaruba	chintok	0.542953244
Q_castanea	Q_crassifolia	0.539734712
P_michoacana	P_montezumae	0.539428754
Q_emoryi	Quercus_arizonica	0.529482192
P_pseudostrobus	P_teocote	0.522084877
Taxodium_mucronatum	sabino	0.516339991
Alnus_acuminata	huejote	0.515024346
P_leiophylla	P_michoacana	0.508105556
A_glabra	corcho	0.507057206
P_michoacana	P_pseudostrobus	0.504503358
P_montezumae	P_teocote	0.502064104

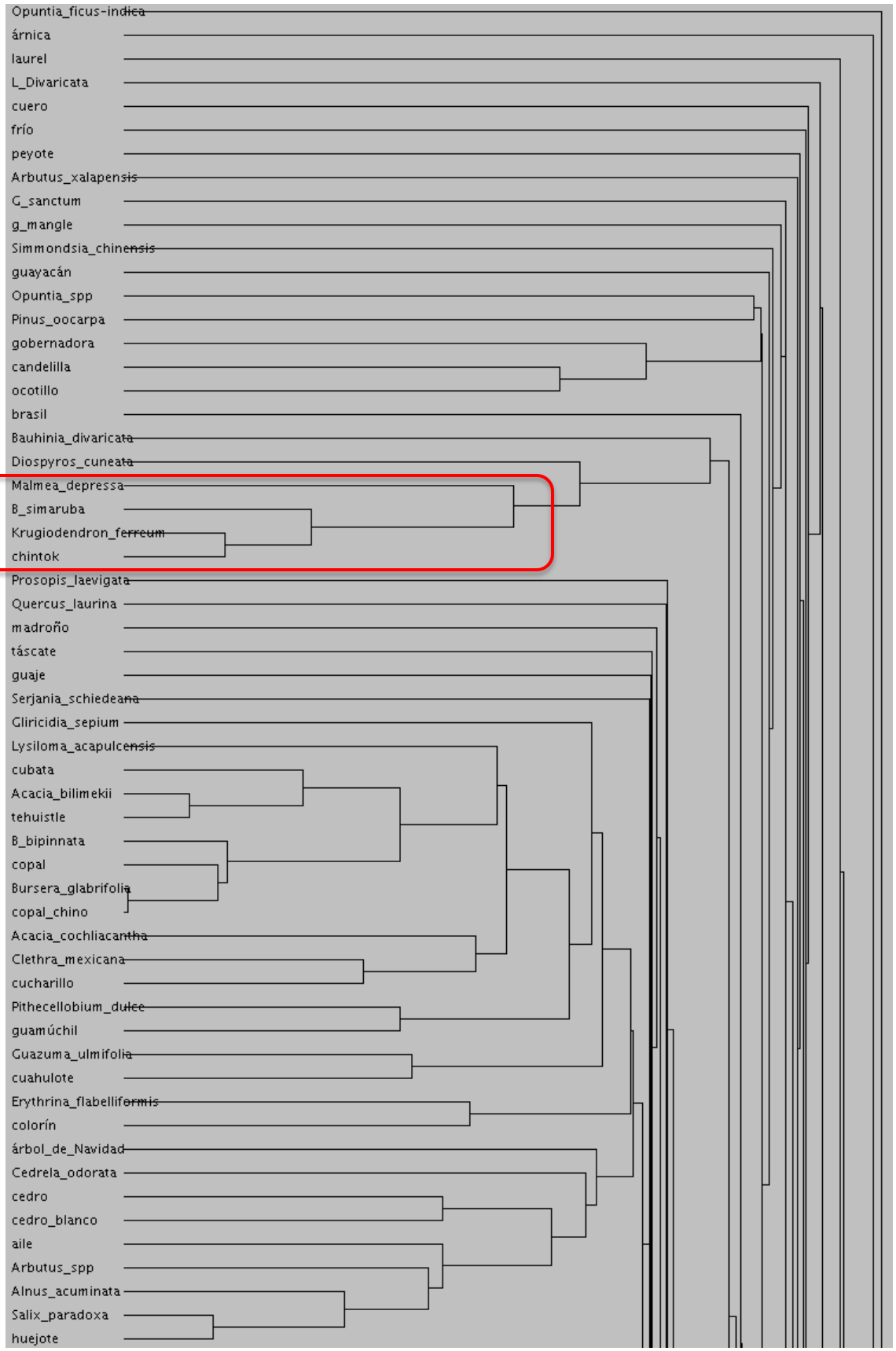
Tabla 28. Variantes Denominativas obtenidas con PMI - Evaluación #2

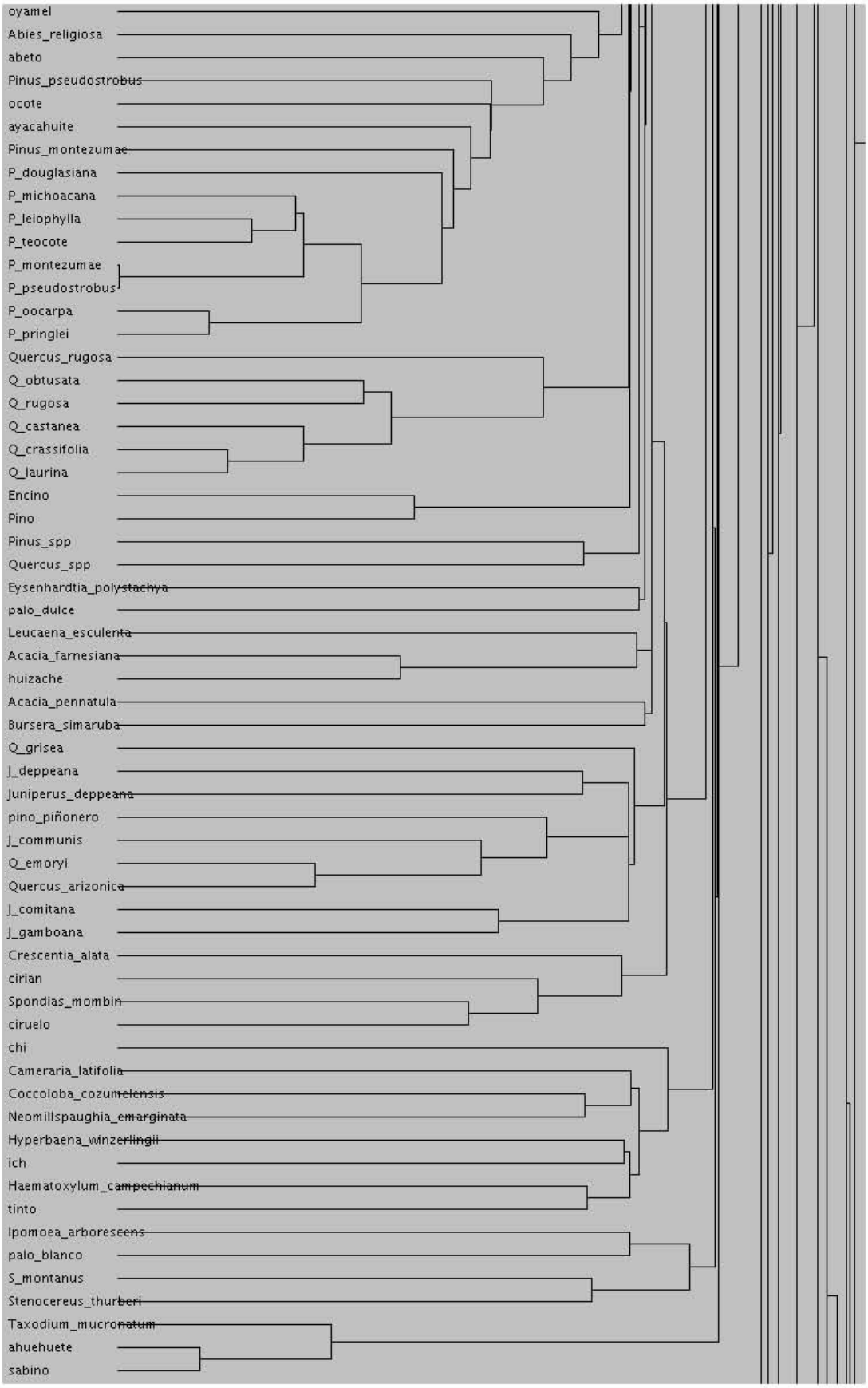
FRECUENCIA		
TERMINO1	TERMINO2	NIVEL SIMILITUD
P_montezumae	P_pseudostrobus	0.78062475
P_leiophylla	P_teocote	0.777123908
Encino	Pino	0.762462369
Q_castanea	Q_obtusata	0.756351406
Q_laurina	Q_obtusata	0.751789979
P_pseudostrobus	P_teocote	0.750064932
P_montezumae	P_teocote	0.730911345
ahuehuete	sabino	0.725615751
P_michoacana	P_teocote	0.725308724
Q_crassifolia	Q_laurina	0.722222222
Q_castanea	Q_laurina	0.719864255
P_pringlei	P_teocote	0.718626991
P_michoacana	P_pseudostrobus	0.714781199
Q_castanea	Q_crassifolia	0.709799928
Q_crassifolia	Q_obtusata	0.702822815
Acacia_bilimekii	tehuistle	0.7
Bursera_glabrifolia	copal_chino	0.7
Krugiodendron_ferreum	chintok	0.7
Salix_paradoxa	huejote	0.7
P_leiophylla	P_michoacana	0.697490581
P_michoacana	P_montezumae	0.693187821
Q_castanea	Q_rugosa	0.688255892
Q_laurina	Q_rugosa	0.681947675
Q_obtusata	Q_rugosa	0.677055444
P_leiophylla	P_pseudostrobus	0.67237264
P_oocarpa	P_pringlei	0.66958948
P_oocarpa	P_pseudostrobus	0.669455549
Pino	táscate	0.668813344
P_pringlei	P_pseudostrobus	0.66513304
P_douglasiana	P_montezumae	0.660494937
P_pseudostrobus	ayacahuite	0.65819311
R_mangle	mangle	0.657927204
P_leiophylla	P_pringlei	0.649749308
Q_crassifolia	Q_rugosa	0.648898201

FRECUENCIA		
TERMINO1	TERMINO2	NIVEL SIMILITUD
B_bipinnata	copal_chino	0.645497224
P_michoacana	P_pringlei	0.641417907
Encino	oyamel	0.635086357
P_montezumae	P_pringlei	0.630630744
P_leiophylla	P_montezumae	0.629415024
P_michoacana	ayacahuite	0.628413798
P_montezumae	P_oocarpa	0.627877989
Pino	oyamel	0.621658644
P_leiophylla	ayacahuite	0.620039685
P_teocote	ayacahuite	0.619580725
P_oocarpa	P_teocote	0.616657178
P_leiophylla	P_oocarpa	0.611248771
P_michoacana	P_oocarpa	0.609776697
P_montezumae	ayacahuite	0.603588272
Encino	táscate	0.59738781
P_douglasiana	P_pseudostrobus	0.597112393
Campeche	Grado	0.596840841
Pinus_montezumae	ayacahuite	0.591680153
Alnus_acuminata	Arbutus_spp	0.581318359
Arbutus_spp	huejote	0.581318359
Abies_religiosa	ayacahuite	0.578653565
Pinus_spp	Quercus_spp	0.576473351
P_douglasiana	P_teocote	0.566044561
P_douglasiana	P_leiophylla	0.560403683
Bursera_glabrifolia	Copal	0.559016994
Copal	copal_chino	0.559016994
Bixa_orellana	achiote	0.553509303
P_michoacana	Abeto	0.551410967
Jabalí	venado	0.546361797
Q_rugosa	Quercus_rugosa	0.533391581
Q_castanea	Quercus_rugosa	0.527644853
Thrinax_radiata	Chit	0.526401715
P_douglasiana	P_oocarpa	0.525973954
P_teocote	Ocote	0.523758208

FRECUENCIA		
TERMINO1	TERMINO2	NIVEL SIMILITUD
Oyamel	táscate	0.522843505
P_pseudostrobus	Ocote	0.519974579
Q_crassifolia	Quercus_rugosa	0.518262172
P_pseudostrobus	Pinus_montezumae	0.512250536
Corcho	palo_de_corcho	0.507833375
Q_laurina	Quercus_rugosa	0.505964426
Ayacahuite	Ocote	0.503355936
Copal	cubata	0.503115295
Q_rugosa	Quercus_arizonica	0.502651402
P_montezumae	Pinus_montezumae	0.502485881
Leucaena_esculenta	huizache	0.502244099
Acacia_bilimekii	cubata	0.5
Alnus_acuminata	huejote	0.5

Tabla 29. Variantes Denominativas obtenidas con Frecuencia - Evaluación #2


Opuntia_ficus-indica
árnica
laurel
L_Divaricata
cuero
frío
peyote
Arbutus_xalapensis
G_sanctum
g_mangle
Simmondsia_chinensis
guayacán
Opuntia_spp
Pinus_oocarpa
gobernadora
candelilla
ocotillo
brasil
Bauhinia_divaricata
Diospyros_cuneata
Malmea_depressa
B_simaruba
Krugiodendron_ferreum
chintok
Prosopis_laevigata
Quercus_laurina
madroño
táscate
guaje
Serjania_schiedeana
Gliricidia_sepium
Lysiloma_acapulcensis
cubata
Acacia_bilimekii
tehuistle
B_bipinnata
copal
Bursera_glabrifolia
copal_chino
Acacia_cochliacantha
Clethra_mexicana
cucharillo
Pithecellobium_dulce
guamúchil
Guazuma_ulmifolia
cuahulote
Erythrina_flabelliformis
colorín
árbol_de_Navidad
Cedrela_odorata
cedro
cedro_blanco
aile
Arbutus_spp
Alnus_acuminata
Salix_paradoxa
huejote


oyamel
Abies_religiosa
abeto
Pinus_pseudostrobus
ocote
ayacahuite
Pinus_montezumae
P_douglasiana
P_michoacana
P_leiophylla
P_teocote
P_montezumae
P_pseudostrobus
P_oocarpa
P_pringlei
Quercus_rugosa
Q_obtusata
Q_rugosa
Q_castanea
Q_crassifolia
Q_laurina
Encino
Pino
Pinus_spp
Quercus_spp
Eysenhardtia_polystachya
palo_dulce
Leucaena_esculenta
Acacia_farnesiana
huizache
Acacia_pennatula
Bursera_simaruba
Q_grisea
J_deppeana
Juniperus_deppeana
pino_piñonero
J_communis
Q_emoryi
Quercus_arizonica
J_comitana
J_gamboana
Crescentia_alata
cirian
Spondias_mombin
ciruelo
chi
Cameraria_latifolia
Coccoloba_cozumelensis
Neomillspaughia_emarginata
Hyperbaena_winzerlingii
ich
Haematoxylum_campechianum
tinto
Ipomoea_arborescens
palo_blanco
S_montanus
Stenocereus_thurberi
Taxodium_mucronatum
ahuehuete
sabino

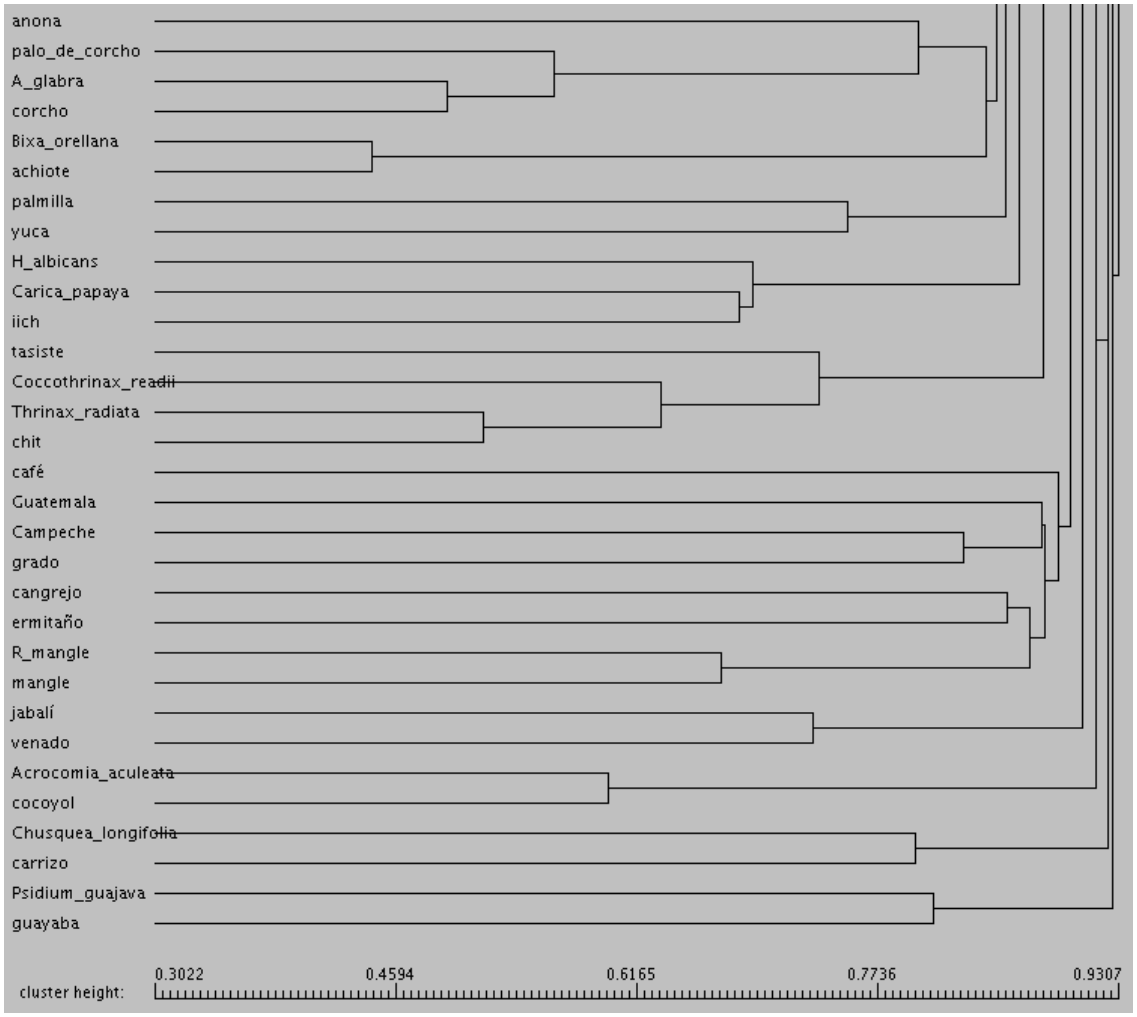


Figura 13. Dendograma - Evaluación #2

La figura 13 muestra el resultado de obtener las distancias coseno entre los vectores y la aplicación del algoritmo jerárquico de agrupamiento para obtener las variantes denominativas entre términos: los nombres comunes y los nombres científicos como “guayaba” y "Psidium guajava” o incluso entre nombre científicos o nombres comunes únicamente como “pino” y “cedro” o “Q amory” y “Quercus arizonica”

En esta segunda evaluación pudimos observar que al igual que la evaluación anterior, el nivel de similitud de las variantes denominativas obtenidas es bajo, ya que no rebasa el 0.79 para el uso de la frecuencia como medida de asociación, por lo que fue necesario realizar otro tipo de evaluaciones diferentes,

las cuales nos permitieron obtener mejores resultados y de esa manera concluir que la metodología propuesta funciona correctamente.

Evaluación #3

Una vez que realizamos las evaluaciones anteriores pudimos observar que el número de términos fue muy limitado, además que el nivel de similitud entre los términos de las variantes denominativas obtenidas fue muy bajo. Por lo que fue necesario utilizar otros términos diferentes, para ello hicimos uso de una muestra de los catálogos de autoridades taxonómicas de la CONABIO, en la que se encuentran 56,969 relaciones denominativas claramente identificadas por expertos del área.

Dicha base de datos incluye diferentes datos, entre ellos un identificador (IdCAT), el grupo al que pertenece dicha variante (Grupo), el nombre científico (Taxon) y el nombre común (NombreComun).

IdCAT	Grupo	Taxon	NombreComun
12576ANFIB	Anfibios	Acris blanchardi	rana grillo del noreste
12577ANFIB	Anfibios	Acris crepitans	Northern cricket frog
12577ANFIB	Anfibios	Acris crepitans	rana grillo norteña
12579ANFIB	Anfibios	Agalychnis callidryas	ninfa del bosque
20433ANGIO	Magnoliophyta	Bromelia pinguin	jocuiste
26279ANGIO	Magnoliophyta	Bromelia karatas	jocuiste

Tabla 30. Base de datos proporcionada por CONABIO

Como se puede observar en la tabla 31, existen nombres científicos (Acris crepitans) y Nombres comunes (jocuiste) repetidos, esto se debe a que un nombre común (jocuiste) puede identificarse con uno o varios nombres científicos (Bromelia pinguin y Bromelia karatas), así como un nombre científico (Acris crepitans) identifica a uno o más nombres comúnes (Northern cricket frog y rana grillo norteña).

Para esta evaluación hicimos uso de un listado de los nombres comunes y científicos de manera indiscriminada, obteniendo un total de 58,023 términos diferentes, con los cuales se aplicó, al igual que en las evaluación anteriores, la metodología propuesta para la identificación de variantes denominativas.

Al momento de realizar la evaluación observamos que dentro del contenido del corpus de Biodiversidad únicamente encontramos 3,818 términos de los 58,023 términos totales proporcionados por la CONABIO en la base de datos y utilizados para obtener las variantes denominativas, es decir, que el 93.42% de los términos no se encontraron en el corpus de Biodiversidad.

	Encontrados en el corpus	Proporcionados por CONABIO
Nombres científicos	1,896 (8.83%)	21,491
Nombres comunes	1,922 (5.26%)	36,532
Totales	**3,818 (6.58%)**	**58,023**

Tabla 31. Porcentaje de términos encontrados en corpus con respecto a los proporcionados por la CONABIO

Como lo indica la tabla anterior, los términos encontrados en el corpus fueron, 1,896 correspondientes a nombres científicos y 1,922 a nombres comunes, haciendo un total de 3,818 términos localizados dentro del corpus de Biodiversidad, los cuales fueron utilizados para encontrar las variantes denominativas entre ellos.

Una vez que aplicamos la metodología, obtuvimos 1,448 variantes denominativas cuyo nivel de similitud es mayor a 0.5, con la siguiente proporción:

- 629 variantes denominativas entre Nombre Común – Nombre Científico.
- 348 variantes denominativas entre Nombre Científico – Nombre Científico.
- 471 variantes denominativas entre Nombre Común – Nombre Común

PMI			
TERMINO 1	TERMINO 2	NIVEL SIMILITUD	TIPO DE VARIANTE DENOMINATIVA
Boa_constrictor	boa_constrictor	1	Científico-Común
Lacaena_bicolor	Lacaena_bicolor	1	Científico-Común
campana	flor_de_campana	1	Común-Común
chango	oreja_de_chango	1	Común-Común
orejona	papaya_orejona	1	Común-Común
pico_de_pato	rana_pico_de_pato	1	Común-Común
encino_laurelillo	laurelillo	1	Común-Común
chak_ach	urrac_chak_cha_ach	0.974349941	Común-Común
jiote	palo_jiote	0.96557774	Común-Común
chooch	chooch_kitam	0.951592452	Común-Común
camedor	palma_camedor	0.948463762	Común-Común
incienso	palo_de_incienso	0.94839073	Común-Común
Purpura_patula	purpura	0.936505495	Científico-Común
viuda	viuda_negra	0.934653496	Común-Común
cavador	sapo_cavador	0.934538345	Común-Común
Guazuma_ulmifolia	guazuma	0.93142082	Científico-Común
bejuquilla	Bejuquilla_verde	0.930110329	Común-Común
carpintero_imperial	imperial	0.927839151	Común-Común
cuatro_ojos	tlacuache_cuatro_ojos	0.926257012	Común-Común
Liquidambar_styraciflua	liquidambar	0.926110533	Científico-Común
duraznillo	nopal_duraznillo	0.925736417	Común-Común
culebra_ciempiés	culebra_ciempiés_de_Michoacán	0.924682676	Común-Común
achoque	achoque_de_Pátzcuaro	0.924540011	Común-Común
caballito	caballito_de_mar	0.924103021	Común-Común
almeja_gallito	gallito	0.923708221	Común-Común
Ostrea	Stereum_ostrea	0.922044292	Científico-Científico
barba_negra	colibrí_barba_negra	0.921880256	Común-Común
pino_piñonero	piñonero	0.920281637	Común-Común
oreja_de_palo	trompa_de_cochi	0.91976045	Común-Común
mesoplodonte	mesoplodonte_antillano	0.917283578	Común-Común
coquito	Coquito_de_aceite	0.91630758	Común-Común

candelero	candelero_americano	0.914673508	Común-Común
cheechem	sak_cheechem	0.913912059	Común-Común
Chrysobalanus_icaco	icaco	0.91270869	Científico-Común
Washingtonia_robusta	washingtonia	0.910130371	Científico-Común
Terrapene_carolina	carolina	0.909349177	Científico-Común
calabacita_italiana	italiana	0.906730747	Común-Común
calabacilla	calabacilla_loca	0.906430468	Común-Común
jaragua	zacate_jaragua	0.904737161	Común-Común

Tabla 32. Variantes Denominativas con PMI - Evaluación #3

En la tabla anterior mostramos los resultados obtenidos utilizando la medida de asociación PMI, únicamente aquellas variantes denominativas cuyo nivel de similitud fue mayor a 0.9, donde la mayoría de ellas resultó ser entre dos nombres comunes.

Una vez que obtuvimos los resultados pudimos observar en ellos variantes denominativas entre dos nombres científicos o entre dos nombres comunes, además de las variantes denominativas entre un nombre común y un nombre científico, las cuales pudieron ser validadas mediante la base de datos proporcionada por la CONABIO y así conocer si realmente corresponden a variantes denominativas entre ellas; Por ello comparamos las variantes denominativas obtenidas contra las relaciones entre un nombre científico y uno común, y observamos que de las 629 variantes obtenidas mediante la metodología planteada, **310 (49.28%) corresponden correctamente** con las proporcionadas por la CONABIO, es decir, menos de la mitad son correctas, lo cual es un índice muy bajo.

Termino 1	Termino 2	Distancia Coseno	Nivel de Similitud
campana	flor_de_campana	0	1
oreja_de_palo	trompa_de_cochi	0.08023955	0.91976045
Celtis_iguanaea	vainoro	0.40000509	0.599994906
Gila_minacae	tortuga_carey	0.98303423	0.016965773
Hura_polyandra	habillo	0.29945558	0.700544419

Tabla 33. Ejemplo de diversos niveles de similitud

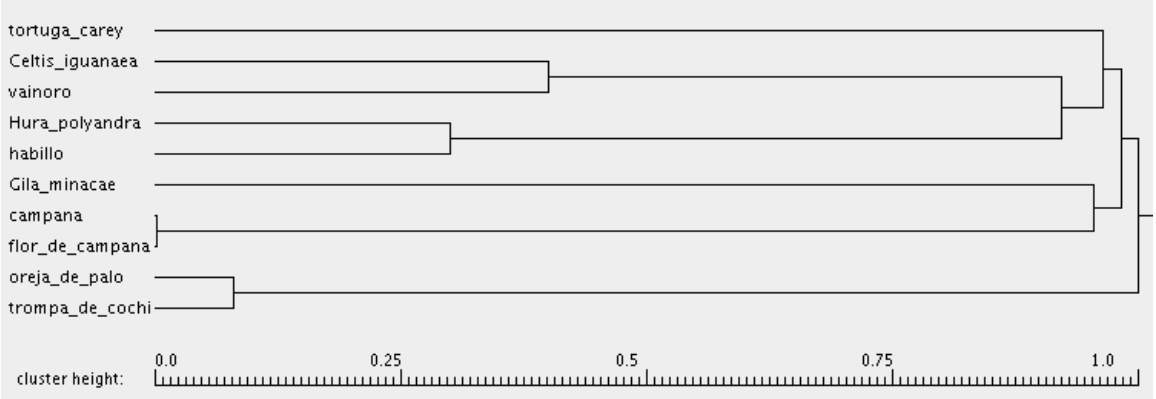


Figura 14. Dendograma ejemplo distancias coseno

Como ya mencionamos anteriormente, una vez que obtuvimos los vectores, es posible obtener el nivel de similitud con base en la distancia coseno entre cada par de vectores y aplicar un algoritmo de agrupamiento, en este caso, el jerárquico. En la Figura 14 observamos cómo se agrupan los términos de ejemplo con base en la distancia coseno calculada, así como el dendograma generado.

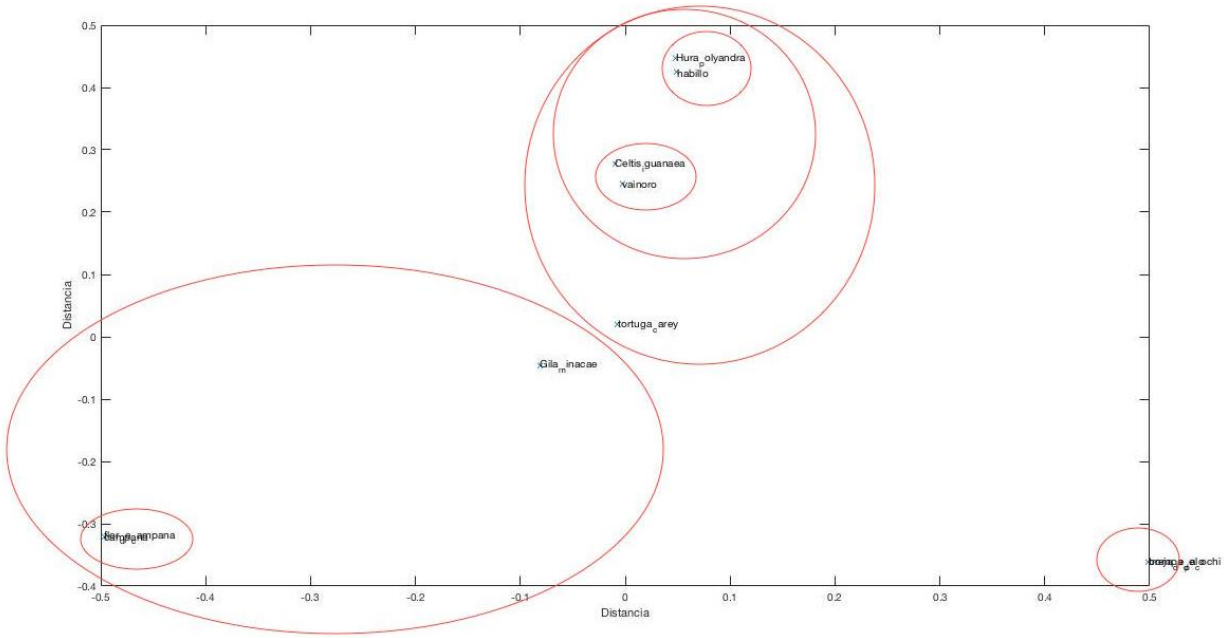


Figura 15. Gráfica mapa 2D de niveles similitud

En la figura 15 mostramos otra forma de representación gráfica de la similitud, mediante un mapa 2D donde podemos observar más claramente el nivel de

similitud entre los términos, así como la manera en que se realizan los agrupamientos con base en la distancia coseno obtenida entre cada par.

Para el caso de las variantes denominativas obtenidas entre dos nombres científicos, solamente fue posible conocer si pertenecen al mismo grupo, como es el caso de la variante denominativa obtenida entre los términos "Montastraea_faveolata" y "Porites_astreoides", los cuales pertenecen al Grupo "Cnidaria", pero no es posible conocer si dicha variante denominativa realmente es correcta. Por esta razón decidimos realizar una evaluación más, para conocer el nivel o grado de precisión al utilizar la metodología propuesta.

Finalmente, con la ayuda del equipo de minería de textos de coordinación de Ecoinformática de la CONABIO, realizamos una evaluación a los resultados obtenidos utilizando la "Biblioteca para la minería de textos de biodiversidad en español" (Barrios, Molina, Sierra-Alcocer, & Zenteno-Jimenez, 2015) para obtener información acerca de diversos términos. Dicha herramienta es de código libre y se puede obtener de manera gratuita en su sitio web, en la siguiente dirección electrónica: https://bitbucket.org/conabio_cmd/text-mining. Mediante esta biblioteca se pudo obtener información de cada uno de los términos obtenidos como parte de las variantes denominativas.

A través de dicha dirección electrónica es posible obtener un archivo en formato JSON, el cual contiene, entre otros datos, las categorías taxonómicas del término (Nombre Científico) solicitado, pudiendo así conocer hasta cuál nivel taxonómico son similares entre ellos.

[[{"id":1000001,"nombre_cientifico":"Animalia","arbol":"/1000001"
,"nombre_categoria_taxonomica":"Reino","nombre_autoridad":"Linnae
us,
1758","estatus":2},{"id":10000006,"nombre_cientifico":"Arthropoda

","arbol":"1000001/10000006","nombre_categoria_taxonomica":"phylum","nombre_autoridad":"von Siebold, 1848","estatus":2},{"id":10000007,"nombre_cientifico":"Mandibulata","arbol":"1000001/10000006/10000007","nombre_categoria_taxonomica":"subphylum","nombre_autoridad":"Snodgrass, 1938","estatus":2},{"id":10000008,"nombre_cientifico":"Atelocerata","arbol":"1000001/10000006/10000007/10000008","nombre_categoria_taxonomica":"infraphylum","nombre_autoridad":"Heymons, 1901","estatus":2},

Algunas de las categorías taxonómicas que se utilizan son:

1. Reino
2. división
3. subdivisión
4. clase
5. subclase
6. orden
7. familia
8. género
9. especie

Una vez que obtuvimos la información de cada uno de los términos involucrados en las variantes denominativas resultantes entre dos nombres científicos, la manera de comprobar si son correctas fue conocer su clasificación taxonómica y comprobar si es verdad que a mayor nivel de similitud, los términos deben coincidir en un nivel más bajo dentro de las categorías taxonómicas (misma especie, mismo género), de lo contrario deberán coincidir en un nivel más alto (mismo reino, misma división).

Realizamos la prueba con 6 pares de variantes denominativas entre dos términos científicos resultantes con diferentes niveles de similitud, para conocer hasta cuál nivel taxonómico coinciden, es decir, son semejantes. El método propuesto funcionó correctamente, ya que las primeras 5 variantes con mayor nivel de similitud, resultaron coincidir hasta la categoría taxonómica de especie, género o familia, mientras que la última variante con un nivel de similitud más bajo (0.589242633), únicamente coincidió con la categoría de orden.

TÉRMINO 1	TÉRMINO 2	Nivel de Similitud	Resultado
Ostrea	**Stereum ostrea**		
Reino = Protoctista división = Bacillariophyta clase = Bacillariophyceae subclase = Bacillariophycidae orden = Thalassiophysales familia = Catenulaceae género = Amphora especie = Amphora ostrearia	Reino = Fungi división = Basidiomycota subdivisión = Agaricomycotina clase = Agaricomycetes subclase = Incertae sedis orden = Russulales familia = Stereaceae género = Stereum especie = Stereum ostrea	0.922044292	Términos similares hasta la **misma especie**.
Reino = Animalia phylum = Mollusca clase = Gastropoda subclase = Caenogastropoda orden = Neotaenioglossa suborden = Neogastropoda superfamilia = Muricoidea familia = Muricidae subfamilia = Muricinae género = Calotrophon especie = Calotrophon ostrearum			
Reino = Fungi división = Basidiomycota subdivisión = Agaricomycotina clase = Agaricomycetes subclase = Agaricomycetidae orden = Agaricales familia = Pleurotaceae género = Pleurotus especie = Pleurotus ostreatus			
Reino = Fungi			

TÉRMINO 1	TÉRMINO 2	Nivel de Similitud	Resultado
división = Basidiomycota subdivisión = Agaricomycotina clase = Agaricomycetes subclase = Incertae sedis orden = Russulales familia = Stereaceae género = Stereum especie = Stereum ostrea			
Montastraea faveolata	**Porites streoides**		
Reino = Animalia phylum = Cnidaria clase = Anthozoa subclase = Hexacorallia orden = Scleractinia familia = Montastraeidae género = Montastraea especie = Montastraea faveolata	Reino = Animalia phylum = Cnidaria clase = Anthozoas subclase = Hexacorallia orden = Scleractinia familia = Poritidae género = Porites especie = Porites astreoides	0.882455856	Términos similares hasta el **mismo orden**.
Jabiru	**Jabiru_mycteria**		
Reino = Animalia phylum = Craniata subphylum = Vertebrata clase = Aves orden = Ciconiiformes familia = Ciconiidae género = Jabiru	Reino = Animalia phylum = Craniata subphylum = Vertebrata clase = Aves orden = Ciconiiformes familia = Ciconiidae género = Jabiru especie = Jabiru mycteria	0.878759837	Términos similares hasta el **mismo genero**.
Reino = Animalia phylum = Craniata subphylum = Vertebrata clase = Aves orden = Ciconiiformes familia = Ciconiidae género = Jabiru			

TÉRMINO 1	TÉRMINO 2	Nivel de Similitud	Resultado
especie = Jabiru mycteria			
Phyllospadix scouleri	**Phyllospadix torreyi**		
Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Phyllospadix especie = Phyllospadix scouleri	Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Phyllospadix especie = Phyllospadix torreyi	0.86775804	Términos similares hasta el **mismo genero**.
Phyllospadix scouleri	**Zostera marina**		
Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Phyllospadix especie = Phyllospadix scouleri	Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Zostera especie = Zostera marina	0.864675314	Términos similares hasta la **misma familia**.
	Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Zostera especie = Zostera marina forma = Zostera marina f. latifolia		

TÉRMINO 1	TÉRMINO 2	Nivel de Similitud	Resultado
	Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Zostera especie = Zostera marina forma = Zostera marina f. typica		
	Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Zostera especie = Zostera marina variedad = Zostera marina var. latifolia		
	Reino = Plantae división = Magnoliophyta clase = Liliopsida superorden = Alismatiflorae orden = Najadales familia = Zosteraceae género = Zostera especie = Zostera marina variedad = Zostera marina var. Stenophylla		
Lantana camara	**Podranea ricasoliana**		

TÉRMINO 1	TÉRMINO 2	Nivel de Similitud	Resultado
Reino = Plantae división = Magnoliophyta clase = Magnoliopsida subclase = Asteridae orden = Lamiales familia = Verbenaceae género = Lantana especie = Lantana camara	Reino = Plantae división = Magnoliophyta clase = Magnoliopsida subclase = Asteridae orden = Lamiales familia = Bignoniaceae género = Podranea especie = Podranea ricasoliana	0.589242633	Términos similares hasta el **misma orden**.

Tabla 34. Ejemplo clasificaciones taxonómicas con API de CONABIO

Como se puede observar en la tabla anterior, de acuerdo con la CONABIO, los nombres científicos pueden tener más de una clasificación taxonómica, como es el caso del término "`Ostrea`", por lo cual es necesario comparar cada una de ellas entre ambos términos de la variante denominativa, para conocer el nivel máximo de similitud entre ellos.

Posteriormente, realizamos un proceso de automatización para validar la totalidad de las variantes denominativas obtenidas. Este dicho proceso de automatización consistió en la elaboración de un algoritmo de comparación entre las clasificaciones taxonómicas de los términos y la elaboración de un programa en el lenguaje de programación Java obteniendo los siguientes resultados:

NIVEL TAXONOMICO	NUM. VARIANTES DENOMINATIVAS	RANGO SIMILITUD
Division	19	0.505942972 - 0.745373158
Phylum	7	0.593721855 - 0.861737085
Subphylum	2	0.614281027 - 0.630331278
Clase	41	0.500785503 - 0.690703679
Subclase	14	0.505349046 - 0.758563893
Grupo	1	0.624259693

NIVEL TAXONOMICO	NUM. VARIANTES DENOMINATIVAS	RANGO SIMILITUD
Superorden	1	0.558177155
Orden	68	0.500727824 - 0.882455856
Suborden	16	0.507476669 - 0.706841506
Infraorden	2	0.512422305 - 0.635912077
Superfamilia	4	0.527620787 - 0.574229598
Familia	59	0.500130143 - 0.864675314
Subfamilia	47	0.50300882 - 0.763647674
Tribu	9	0.501356504 - 0.764567053
Subtribu	1	0.507236706
Genero	47	0.503036936 - 0.86775804
Serie	5	0.504671957 - 0.640808391
Especie	2	0.878759837 - 0.922044292

Tabla 35. Totales y rangos de similitud de variantes denominativas con términos de CONABIO

En la tabla 35 mostramos los diferentes niveles taxonómicos, el número de variantes denominativas obtenidas cuyo máximo nivel taxonómico semejante corresponde con dicha categoría taxonómica, así como el rango de similitud entre los que se encuentran. Para los niveles taxonómicos de Especie, Familia y Género, los niveles de similitud máximos son altos, 0.922044292, 0.864675314 y 0.86775804 respectivamente y para los niveles taxonómicos de Clase y División sus niveles de similitud máximos son de 0.690703679 y 0.745373158 respectivamente, lo que muestra que la metodología propuesta en esta tesis funcionó correctamente para este tipo de términos.

Para las variaciones denominativas obtenidas entre un nombre científico y un nombre común, hicimos uso de la información de los nombres comunes relacionados a cada nombre científico, proporcionada también por la “Biblioteca para la minería de textos de biodiversidad en español” (Barrios, Molina, Sierra-Alcocer, & Zenteno-Jimenez, 2015) de CONABIO. Primeramente obtuvimos la información completa del nombre científico y posteriormente verificamos si dentro de los nombres comunes correspondientes a dicho nombre científico se encontraba el nombre común de la variante denominativa obtenida, por ejemplo:

Reino = Plantae
división = Magnoliophyta
clase = Magnoliopsida
subclase = Asteridae
orden = Asterales
familia = Asteraceae
subfamilia = Asteroideae
tribu = Heliantheae
género = Helianthus
especie = Helianthus annuus
Nombres Comunes:
-acahualli
-chimal-acatl
-chimalacate
-chimalatl
-chimálatl
-chimálitl
-chimalte
-flor de sol
-gigantón
-Girasol
-girasol
-lampote
-maíz de teja
-maíz de tejas
-maíz de Texas
-quisnaniqui-tonale
-xaricámata
-xuchipalli
-yendri

En el extracto anterior se puede observar cómo además de la clasificación taxonómica del nombre científico, se incluyen los nombres comunes relacionados a él. Uno de los resultados que obtuvimos fue la variación denominativa ("Helianthus annuus" y "girasol"), primero obtuvimos la información del término científico "Helianthus annuus" y pudimos observar

que dentro de los nombres comunes se encontraba "girasol", por lo que dicha variante denominativa es correcta.

A través del uso de la "Biblioteca para la minería de textos de biodiversidad en español" (Barrios, Molina, Sierra-Alcocer, & Zenteno-Jimenez, 2015) obtuvimos que 317 de las 629 variaciones denominativas entre un nombre científico y un nombre común fueron acertadas y en el resto de ellas no fue posible encontrar el nombre científico o no se encontró el nombre común entre los nombres comunes relacionados a dicho nombre científico en la API. Por lo cual tuvimos que verificarlas de manera manual, es decir, revisando los extractos del corpus en donde se encontraron ambos términos de la variante denominativa resultante para definir si era correcta o no.

De igual manera, las variantes denominativas resultantes entre dos nombres comunes, tuvieron que ser revisadas manualmente por nosotros, debido a que la API proporcionada por la CONABIO, no permite la búsqueda de información con base en el nombre común.

Tipo Variación Denominativa	Total	Correctos	No encontrados o incorrectos
Nombre Científico – Nombre Científico	348	345 **(99.13%)**	3
Nombre Común – Nombre Científico	629	396 **(62.95%)**	233
Nombre Común – Nombre Común	471	318 **(67.51%)**	153

Tabla 36. Total variaciones denominativas utilizando términos del experto en Biodiversidad

La tabla 36 muestra el número de variantes denominativas correctas e incorrectas o no encontradas (nombres científicos no encontrados por la API o variantes erróneas tras su revisión manual), donde podemos observar que la metodología propuesta funciona correctamente para la mayoría de las variantes denominativas resultantes, obteniendo porcentajes de entre el 62.95% y 99.13%.

7. Conclusiones

En esta tesis presentamos una metodología completa y detallada paso a paso para la obtención de variantes terminológicas dentro de dos diferentes áreas de especialidad, como lo es la Ingeniería de Software y la Biodiversidad, con lo que se cumplió el objetivo principal.

Además de haber cumplido con el objetivo principal de la tesis, realizamos diversas tareas y/o actividades para logar dicho objetivo, entre las que se encuentran:

- Investigación y uso de dos diferentes herramientas de Extracción Automática de Términos (Termext y Termostat), las cuales nos dieron los términos como materia prima para obtener las variantes denominativas entre ellos.
- Realizamos la obtención de los contextos de cada uno de los términos conforme a una ventana de contexto de un único tamaño (5 palabras a la derecha y 5 a la izquierda).
- Obtuvimos matrices de coocurrencia entre los términos y las palabras del corpus utilizando dos medidas de asociación entre ellas (Frecuencia y PMI). Evaluamos los resultados, concluyendo en la obtención de mejores resultados utilizando PMI como medida de asociación.
- Obtuvimos por cada término un vector semántico, con los cuales realizamos el cálculo de las distancias entre ellos utilizando la distancia coseno.
- Aplicamos un método de agrupamiento con base en las distancias coseno obtenidas, permitiéndonos analizar de mejor manera los resultados y el comportamiento de los términos.
- Realizamos el cálculo de los niveles de similitud entre cada par de términos, el cual va de 0 a 1, considerando únicamente para su análisis y evaluación aquellos cuyo nivel superó el 0.5 de similitud.
- Finalmente realizamos un análisis a los resultados obtenidos, así como una serie de evaluaciones haciendo uso de términos proporcionados por expertos en ambas áreas de especialidad.

Después de haber realizado el análisis de los resultados obtenidos y las evaluaciones a la metodología propuesta en esta tesis utilizando términos identificados por expertos en las áreas de especialidad, pudimos observar dos casos diferentes; Para el área de Ingeniería de Software, las variantes denominativas obtenidas, con base en nuestros conocimientos en el área, no fueron acertados, ya que incluso los términos indicados por los expertos no fueron obtenidos como variantes denominativas y algunos de ellos tampoco fueron obtenidos mediante la extracción automática con Termext y con Termostat. Para el área de Biodiversidad, el caso es diferente, las variantes denominativas obtenidas fueron correctas y validadas mediante una API que mantiene la información de dichos términos y en algunos de ellos fueron revisados de manera manual, pero la mayoría de los resultados corresponden a variantes denominativas dentro del área de especialidad.

Con base en las evaluaciones que realizamos, especialmente en el área de Biodiversidad, pudimos concluir que los resultados son favorables, aunque aún queda trabajo por realizar como otro tipo de pruebas y evaluaciones en otras áreas de especialidad, además de que es posible realizar modificaciones a la metodología propuesta para tratar de obtener mejores y más variantes denominativas como resultado.

Trabajo futuro

Si bien en este trabajo de tesis logramos los objetivos que se tenían planteados desde el inicio de la investigación, como son:

1. Obtener, analizar y representar las variantes denominativas dentro de la terminología utilizada en el área de especialidad de manera automática.
2. Conocer la terminología de áreas de especialidad, así como su similitud y variabilidad entre ellos.
3. Creación de corpus especializados de diferentes áreas de especialidad (Ingeniería de Software y Biodiversidad).
4. Mostrar que el modelo de distribución semántica y el análisis de variantes denominativas se puede aplicar sobre otras lenguas y dominios de especialidad.

Aun cuando logramos estos objetivos, quedan líneas de investigación abiertas, dudas por despejar y pruebas por realizar, además de que es posible ahondar mucho más en las áreas de Extracción Terminológica y Semántica Distribucional.

Los apartados siguientes son un ejemplo de la riqueza de explotación del tema de tesis en un futuro:

- Ahondar en la investigación dentro de la Extracción Terminológica automática dentro de las áreas de especialidad.
- Utilización y/o mejoramiento de diferentes herramientas y métodos de extracción automática de términos.
- Recopilación de corpus de diversas áreas de especialidad para su uso en diferentes proyectos de investigación.
- Evaluación de la metodología propuesta dentro de diferentes y diversas áreas de especialidad y evaluar su funcionamiento.
- Uso de diferentes distancias entre vectores semánticos, así como diferentes medidas de asociación entre los términos obtenidos y las palabras del corpus para la obtención de las variantes denominativas.
- Establecer pruebas variando los tamaños de la ventana de contexto establecida y evaluar las variantes denominativas obtenidas entre ellos.
- Realizar etiquetado POS con diferentes herramientas y obtener algunos patrones morfosintácticos que ocurren en el contexto de los términos.
- Al utilizar el modelo de distribución semántica, es posible aplicar esta metodología sobre otras lenguas, revisar los resultados y evaluar su funcionamiento.

Bibliografía

Abellán, R. (1999). Pasos para la creación y el uso de un corpus de inglés turístico. *Cuadernos de turismo*, 127-139.

Allen, J. (01 de 10 de 2014). The Unicode Standard / the Unicode Consortium. Mountain View, California, United States.

Barrios, J., Molina, A., Sierra-Alcocer, R., & Zenteno-Jimenez, E.-D. (2015). A Text Mining Library for Biodiversity Literature in Spanish. *International Journal of Computational Linguistics and Applications*, Vol. 6 No. 2.

Barrón Cedeño, L. A. (2008). *Manual para el extractor de términos TERMEXT.* Universidad Nacional Autonoma de México, Mexico, D.F.

Barrón, L. A., Sierra, G., & Villaseñor, E. (2006). C-value aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español. *In 7th Mexican International Conference on Computer Science (ENC 2006).* Los Alamitos: IEEE Computer Press.

Barrón, L. A., Sierra, G., Drouin, P., & Ananiadou, S. (2009). An Improved Automatic Term Recognition Method for Spanish. *Cicling*, 125-136.

Bourigault, D., Gonzalez-Mullier, I., & Gros, C. (1996). LEXTER, a Natural Language Processing tool for terminology extraction. *Proceedings of the 7th EURALEX International Congress*, (págs. 771-779).

Bowker, L. (1998). Variant terminology: frivolity or necessity. *EURALEX'98 Proceedings* (págs. 487-496). Liège: Université de Liège.

Burnard, L. (Enero de 2009). *The British National Corpus (BNC)*. Recuperado el 2 de Febrero de 2016, de http://www.natcorp.ox.ac.uk/corpus/index.xml

Cabré, M. T. (1993). *La terminología: Teoría, metodología y aplicaciones.* Barcelona, España: Antártida/Empúries.

Cabré, M. T. (1997). Standardization and interference in terminology. *The Changing Scene in World Languages: Issues and Challenges,*, (págs. 9, 49).

Cabré, M. T. (1999). *La terminología: Representación y comunicación.* Cataluña, España: Instituto Universitario de Lingüistica Aplicada.

Cabré, M., Estopà, R., & Vivaldi, J. (2001). Automatic term detection: A review of current systems. *Recent advances in computational terminology*, 53-88.

CAT, CONABIO. (Marzo de 2013). *Subcoordinación de Catálogos de Autoridades Taxonómicas.* Recuperado el 15 de Enero de 2016, de Catálogos de Autoridades Taxonómicas: http://www.biodiversidad.gob.mx/especies/CAT.html

CNCUB, CONABIO. (2012). *Comisión Nacional para el Conocimiento y Uso de la Biodiversidad*. Recuperado el 25 de Enero de 2016, de Biodiversidad Maxicana: http://www.biodiversidad.gob.mx/biodiversidad/que_es.html

Daille, B. (1996). ACABIT: une maquette d'aide à la construction automatique de banques terminologiques monolingues ou bilingues. *Lexicomatique et Dictionnairiques*, 123-136.

Dang, V., Xue, X., & Croft, W. (2009). Context-based quasi-synonym extraction. *idea*, 200.

David, S., & Plante , P. (1990). Le progiciel TERMINO: de la nécessité d'une analyse morphosyntaxique pour le dépouillement terminologique des textes. *Colloque International sur les Industries de la Langue: Perspectives des Années*, (págs. 71-88).

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *In Terminology*, 99-117.

Dubois, A. (2012). The distinction between introduction of a new nomen and subsequent use of a previously introduced nomen in zoological nomenclature. *Bionomina*, 57-80.

Enguehard, C., & Pantera, L. (1995). Automatic natural acquisition of a terminology. *Journal of quantitative linguistics*, 27-32.

Galicia-Haro, S. N., & Gelbukh, A. (2016). Assessing Context for Extraction of Near Synonyms from Product Reviews in Spanish. *Lecture Notes in ComputerScience. Volume 9924. En prensa*.

Gelbukh, A. (2007). Procesamiento de lenguaje natural y sus aplicaciones. *Komputer Sapiens*, 6-11.

Guevara, R. (2011). *Minería de textos en la red social Twitter.* México: Universidad Nacional Autónoma de México, Facultad de Ingeniería.

Harris, Z. (1954). Distributional structure. *Word, 10*, 146-162.

Hazem, A., & Daille, B. (2014). Semi-compositional method for synonym extraction of multi-word terms. *9th edition of the Language Resources and Evaluation Conference.* LREC.

Henriksson, A., Moen, H., Skeppstedt, M., Eklund, A.-M., Daudaravicius, V., & Hassel, M. (2012). Synonym Extraction of Medical Terms from Clinical Text Using Combinations of Word Space Models. 10-17.

IEEE. (2002). *Standard Glossary of Software Engineering Terminology.* IEEE Standards Board.

INECC. (03 de Julio de 2013). *¿Qué es la Biodiversidad?* Recuperado el 11 de Diciembre de 2015, de Instituto Nacional de Ecología y Cambio Climático: http://www.inecc.gob.mx/con-eco-biodiversidad

Jaramillo, A., Zapata, C., & Isaza, F. (2007). Una propuesta para la asistencia al proceso de interpretación de textos utilizando técnicas de procesamiento del lenguaje natural e ingeniería de software. *Avances en Sistemas e Informática*, 39-46.

Kao, A., & Poteet, S. (2007). *Natural language processing and text mining.* Springer Science & Business Media.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2), 211.

Landauer, T., Foltz, P., & Laham, D. (1988). An introduction to latent semantic analysis. *Discourse processes.*, 259-284.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28(2), 203-208.

Manning, C., Raghavan, P., & Schütze, H. (2009). *An Introduction to Information Retrieval.* Cambridge, England: Cambridge University Press.

Martin, J., & Jurafsky, D. (2000). *Speech and language processing.* International Edition.

Matsuo, Y., & Ishizuka, M. (2004). Keyword extraction from a single document using word co-occurrence statistical information. *International Journal on Artificial Intelligence Tools*, (págs. 157-169).

Morin, E., & Jacquemin, C. (2004). Automatic acquisition and expansion of hypernym links. *Computers and the Humanities.*, 363-396.

Nazar, R., & Renau, I. (2012). Agrupación semántica de sustantivos basada en similitud distribucional. Implicaciones lexicográficas. *n V Congreso Internacional de Lexicografía Hispánica.* Madrid, Spain.

Nenadié, G., Ananiadou, S., & McNaught, J. (2004). Enhancing automatic term recognition through recognition of variation. *In Proceedings of the 20th international conference on Computational Linguistics* (pág. 604). Association for Computational Linguistics.

Perez, C., & Moreno, A. (2009). Lingüística Computacional y Lingüística de Corpus. Potencialidades para la Investigación Textual. En *Teoría y literatura artística en la sociedad digital: construcción y aplicabilidad de colecciones textuales informatizadas.* TREA.

Pressman, R., & Troya, J. (1988). *Ingeniería del software.* McGraw Hill.

Reyes, A. (2002). *Hacia una obtención computarizada de términos. (Aplicación concreta al léxico de la física en el nivel bachillerato).* México: Universidad Nacional Autónoma de México.

Runeson, P., Alexandersson, M., & Nyholm , O. (2007). Detection of duplicate defect reports using natural language processing. *Software Engineering. 29th International Conference* (págs. 499-510). IEEE.

Sahlgren, M. (2006). *The Word-space model.* Gothenburg University: Doctoral dissertation.

Schutze , H. (1992). Dimensions of meaning. *Supercomputing '92* (págs. 787-796). IEEE.

Schütze, H. (1993). Word space. *Advances in neural information processing systems*, 5, 895-902.

Sierra, G. (2008). Diseño de corpus textuales para fines lingüísticos. *Proceedings of the IX Encuentro Internacional de Lingüística en el Noroeste* (págs. 445-462). México, D.F.: Universidad Nacional Autónoma de México.

Sierra, G. (2014). *Introducción a los corpus lingüisticos.* Distrito Federal, México: Instituto de Ingeniería.

Sierra, G., Medina, A., & Lázaro, J. (2009). Determinación de la terminología básica en sexualidad a partir de la web como corpus. *In A survey of corpus-based research*.

Sierra, G., Villaseñor, E., & Barrón, L. (2006). C-value aplicado a la extracción de términos multipalabra en documentos técnicos y científicos en español. *7th Mexican International Conference on Computer Science.* San Luis Potosí, México: Press, IEEE Computer.

Sommerville, I. (2005). *Ingeniería del software.* Pearson Educación.

Stefan, E. (2010). Distributional semantic models. *In Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Los Angeles: Association for Computational Linguistics.

Torruella , J., & Llisterri, J. (1999). Diseño de corpus textuales y orales. *Filología e informática. Nuevas tecnologías en los estudios filológicos* (págs. 45-77). liceu.uab.cat.

Uszkoreit, H. (03 de 05 de 2000). What is Computational Linguistics? Alemania.

Vilares, J., Cabrero, D., & Alonso, M. (2001). Applying Productive Derivational Morphology. *Computational linguistics and intelligent text processing*, 336.348.

Vivaldi, J., Cabrera-Diego, L., Sierra, G., & Pozzi, M. (2012). Using Wikipedia to Validate the Terminology found in a Corpus of Basic Textbooks. *LREC*, (págs. 3820-3827).

Wang, Y. (2007). *Software engineering foundations: A software science perspective.* CRC Press.

Widdows, D., & Cohen, T. (2010). The semantic vectors package: New algorithms and public tools for distributional semantics. *In Semantic Computing (ICSC) on IEEE Fourth International Conference* (págs. 9-15). IEEE.

Apéndices

Etiquetas utilizadas (tagset)

Para el etiquetado de partes de la oración se utilizan dos conjuntos diferentes de etiquetas, ya que para el uso de Termext y Termostat se hace uso de la herramienta TreeTagger como parte del proceso de extracción terminológica y el conjunto de etiquetas que utiliza son las siguientes:

ACRNM acronym (ISO, CEI)
ADJ Adjectives (mayores, mayor)
ADV Adverbs (muy, demasiado, cómo)
ALFP Plural letter of the alphabet (As/Aes, bes)
ALFS Singular letter of the alphabet (A, b)
ART Articles (un, las, la, unas)
BACKSLASH backslash (\)
CARD Cardinals
CC Coordinating conjunction (y, o)
CCAD Adversative coordinating conjunction (pero)
CCNEG Negative coordinating conjunction (ni)
CM comma (,)
CODE Alphanumeric code
COLON colon (:)
CQUE que (as conjunction)
CSUBF Subordinating conjunction that introduces finite clauses (apenas)
CSUBI Subordinating conjunction that introduces infinite clauses (al) CSUBX Subordinating conjunction underspecified for subord-type (aunque)
DASH dash (-)
DM Demonstrative pronouns (ésas, ése, esta)
DOTS POS tag for "..."
FO Formula
FS Full stop punctuation marks
INT Interrogative pronouns (quiénes, cuántas, cuánto)
ITJN Interjection (oh, ja)
LP left parenthesis ("(", "[")
NC Common nouns (mesas, mesa, libro, ordenador)
NEG Negation
NMEA measure noun (metros, litros)
NMON month name
NP Proper nouns
ORD Ordinals (primer, primeras, primera)
PAL Portmanteau word formed by a and el
PDEL Portmanteau word formed by de and el
PE Foreign word
PERCT percent sign (%)
PNC Unclassified word
PPC Clitic personal pronoun (le, les)
PPO Possessive pronouns (mi, su, sus)
PPX Clitics and personal pronouns (nos, me, nosotras, te, sí)
PREP Negative preposition (sin)
PREP Preposition
PREP/DEL Complex preposition "después del"
QT quotation symbol (" ' `)
QU Quantifiers (sendas, cada)
REL Relative pronouns (cuyas, cuyo)
RP right parenthesis (")", "]")
SE Se (as particle)
SEMICOLON semicolon (;)
SLASH slash (/) SYM Symbols
UMMX measure unit (MHz, km, mA)
VCLIger clitic gerund verb
VCLIinf clitic infinitive verb
VCLIfin clitic finite verb
VEadj Verb estar. Past participle
VEfin Verb estar. Finite
VEger Verb estar. Gerund

VEinf	Verb estar. Infinitive
VHadj	Verb haber. Past participle
VHfin	Verb haber. Finite
VHger	Verb haber. Gerund
VHinf	Verb haber. Infinitive
VLadj	Lexical verb. Past participle
VLfin	Lexical verb. Finite
VLger	Lexical verb. Gerund
VLinf	Lexical verb. Infinitive
VMadj	Modal verb. Past participle
VMfin	Modal verb. Finite
VMger	Modal verb. Gerund
VMinf	Modal verb. Infinitive
VSadj	Verb ser. Past participle
VSfin	Verb ser. Finite
VSger	Verb ser. Gerund
VSinf	Verb ser. Infinitive

También hicimos uso de la herramienta Freeling, la cual se basa en las etiquetas propuestas por el grupo EAGLES.

A continuación presentamos las etiquetas que el analizador morfológico utiliza para el castellano en formato de tabla y algunos ejemplos de cada categoría.

Para cada categoría presentamos los atributos, valores y códigos que puede tomar. Las tablas tienen el siguiente formato:

ETIQUETAS			
Posición	**Atributo**	**Valor**	**Código**
Columna 1	*Columna 2*	*Columna 3*	*Columna 4*

ADJETIVOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adjetivo	A
2	Tipo	Calificativo	Q
		Ordinal	O
3	Grado	Aumentativo	A
		Diminutivo	D
		Comparativo	C
		Superlativo	S
4	Género	Masculino	M
		Femenino	F
		Común	C
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Función	-	0
		Participio	P

ADVERBIOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adverbio	R
2	Tipo	General	G
		Negativo	N

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5-6	Clasificación semántica	Persona	SP
		Lugar	G0
		Organización	O0
		Otros	V0
7	Grado	Aumentativo	A
		Diminutivo	D

SIGNOS DE PUNTUACIÓN			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Puntuación	F

CONJUNCIONES			
Pos.	**Atributo**	**Valor**	**Código>**
1	Categoría	Conjunción	C
2	Tipo	Coordinada	C
		Subordinada	S

DETERMINANTES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Determinante	D
2	Tipo	Demostrativo	D
		Posesivo	P
		Interrogativo	T
		Exclamativo	E
		Indefinido	I
		Artículo	A
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Invariable	N
6	Poseedor	Singular	S
		Plural	P

INTERJECCIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Interjección	I

PRONOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Pronombre	P
2	Tipo	Personal	P
		Demostrativo	D
		Posesivo	X
		Indefinido	I
		Interrogativo	T
		Relativo	R
		Exclamativo	E
3	Persona	Primera	1
		Segunda	2
		Tercera	3
4	Género	Masculino	M
		Femenino	F
		Común	C
		Neutro	N
5	Número	Singular	S
		Plural	P
		Impersonal /Invariable	N
6	Caso	Nominativo	N
		Acusativo	A
		Dativo	D
		Oblicuo	O
7	Poseedor	Singular	S
		Plural	P
8	Politeness	Polite	P

FECHAS Y HORAS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Fecha/Hora	W

NUMERALES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Cifra	Z
2	Tipo	partitivo	d
		Moneda	m
		porcentaje	p
		unidad	u

PREPOSICIONES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Adposición	S
2	Tipo	Preposición	P
3	Forma	Simple	S
		Contraída	C
3	Género	Masculino	M
4	Número	Singular	S

VERBOS			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Verbo	V
2	Tipo	Principal	M
		Auxiliar	A
		Semiauxiliar	S
3	Modo	Indicativo	I
		Subjuntivo	S
		Imperativo	M
		Infinitivo	N
		Gerundio	G
		Participio	P
4	Tiempo	Presente	P
		Imperfecto	I
		Futuro	F
		Pasado	S
		Condicional	C
		-	0
5	Persona	Primera	1
		Segunda	2
		Tercera	3
6	Número	Singular	S
		Plural	P
7	Género	Masculino	M
		Femenino	F

Listado de palabras funcionales (Stoplist).

DE	TIENE	NADA	VUESTRO
LA	TAMBIÉN	MUCHOS	VUESTRA
QUE	ME	CUAL	VUESTROS
EL	HASTA	SEA	VUESTRAS
EN	HAY	POCO	ESOS
Y	DONDE	ELLA	ESAS
A	HAN	ESTAR	ESTOY
LOS	QUIEN	HABER	ESTÁS
DEL	ESTÁN	ESTAS	ESTÁIS
SE	DESDE	ESTABA	ESTÉ
LAS	TODO	ESTAMOS	ESTÉS
POR	NOS	ALGUNAS	ESTEMOS
UN	DURANTE	ALGO	ESTÉIS
PARA	TODOS	NOSOTROS	ESTÉN
CON	UNO	MI	ESTARÉ
NO	LES	MIS	ESTARÁS
UNA	NI	TÚ	ESTARÁ
SU	CONTRA	TE	ESTAREMOS
AL	OTROS	TI	ESTARÉIS
ES	FUERON	TU	ESTARÁN
LO	ESE	TUS	ESTARÍA
COMO	ESO	ELLAS	ESTARÍAS
MÁS	HABÍA	NOSOTRAS	ESTARÍAMOS
PERO	ANTE	VOSOTROS	ESTARÍAIS
SUS	ELLOS	VOSOTRAS	ESTARÍAN
LE	E	OS	ESTABAS
YA	ESTO	MÍO	ESTÁBAMOS
O	MÍ	MÍA	ESTABAIS
FUE	ANTES	MÍOS	ESTABAN
ESTE	ALGUNOS	MÍAS	ESTUVE
HA	QUÉ	TUYO	ESTUVISTE
SÍ	UNOS	TUYA	ESTUVO
PORQUE	YO	TUYOS	ESTUVIMOS
ESTA	OTRO	TUYAS	ESTUVISTEIS
SON	OTRAS	SUYO	ESTUVIERON
ENTRE	OTRA	SUYA	ESTUVIERA
ESTÁ	ÉL	SUYOS	ESTUVIERAS
CUANDO	TANTO	SUYAS	ESTUVIÉRAMOS
MUY	ESA	NUESTRO	ESTUVIERAIS
SIN	ESTOS	NUESTRA	ESTUVIERAN
SOBRE	MUCHO	NUESTROS	ESTUVIESE
SER	QUIENES	NUESTRAS	ESTUVIESES

ESTUVIÉSEMOS
ESTUVIESEIS
ESTUVIESEN
ESTANDO
ESTADO
ESTADA
ESTADOS
ESTADAS
ESTAD
HE
HAS
HEMOS
HABÉIS
HAYA
HAYAS
HAYAMOS
HAYÁIS
HAYAN
HABRÉ
HABRÁS
HABRÁ
HABREMOS
HABRÉIS
HABRÁN
HABRÍA
HABRÍAS
HABRÍAMOS
HABRÍAIS
HABRÍAN
HABÍAS
HABÍAMOS
HABÍAIS
HABÍAN
HUBE
HUBISTE
HUBO
HUBIMOS
HUBISTEIS
HUBIERON
HUBIERA
HUBIERAS
HUBIÉRAMOS
HUBIERAIS
HUBIERAN
HUBIESE
HUBIESES
HUBIÉSEMOS
HUBIESEIS
HUBIESEN
HABIENDO
HABIDO
HABIDA
HABIDOS
HABIDAS
SOY
ERES
SOMOS
SOIS
SEAS
SEAMOS
SEÁIS
SEAN
SERÉ
SERÁS
SERÁ
SEREMOS
SERÉIS
SERÁN
SERÍA
SERÍAS
SERÍAMOS
SERÍAIS
SERÍAN
ERA
ERAS
ÉRAMOS
ERAIS
ERAN
FUI
FUISTE
FUIMOS
FUISTEIS
FUERA
FUERAS
FUÉRAMOS
FUERAIS
FUERAN
FUESE
FUESES
FUÉSEMOS
FUESEIS
FUESEN
SIENDO
SIDO
TENGO
TIENES
TENEMOS
TENÉIS
TIENEN
TENGA
TENGAS
TENGAMOS
TENGÁIS
TENGAN
TENDRÉ
TENDRÁS
TENDRÁ
TENDREMOS
TENDRÉIS
TENDRÁN
TENDRÍA
TENDRÍAS
TENDRÍAMOS
TENDRÍAIS
TENDRÍAN
TENÍA
TENÍAS
TENÍAMOS
TENÍAIS
TENÍAN
TUVE
TUVISTE
TUVO
TUVIMOS
TUVISTEIS
TUVIERON
TUVIERA
TUVIERAS
TUVIÉRAMOS
TUVIERAIS
TUVIERAN
TUVIESE
TUVIESES
TUVIÉSEMOS
TUVIESEIS
TUVIESEN
TENIENDO
TENIDO
TENIDA
TENIDOS
TENIDAS
TENED
UN
UNA
UNAS
UNOS
UNO
SOBRE
TODO
TAMBIÉN
TRAS
OTRO
ALGÚN
ALGUNO
ALGUNA
ALGUNOS
ALGUNAS
ESTA
ESTAIS
ESTAN
COMO
EN
PARA
ATRAS
PORQUE
POR QUÉ
ANTE
ANTES
AMBOS
PERO
POR
PODER
PUEDE
PUEDO
PODEMOS
PODEIS
PUEDEN
HACER
HAGO
HACE

HACEMOS
HACEIS
HACEN
CADA
FIN
INCLUSO
PRIMERO
DESDE
CONSEGUIR
CONSIGO
CONSIGUE
CONSIGUES
CONSEGUIMOS
CONSIGUEN
IR
VOY
VA
VAMOS
VAIS
VAN
VAYA
GUENO
TENER
TENEIS
EL
LA
LO
LAS
LOS
SU

AQUI
MIO
TUYO
ELLOS
ELLAS
NOS
NOSOTROS
VOSOTROS
VOSOTRAS
SI
DENTRO
SOLO
SOLAMENTE
SABER
SABES
SABE
SABEMOS
SABEIS
SABEN
ULTIMO
LARGO
BASTANTE
HACES
MUCHOS
AQUELLOS
AQUELLAS
SUS
ENTONCES
TIEMPO
VERDAD

VERDADERO
VERDADERA
CIERTO
CIERTOS
CIERTA
CIERTAS
INTENTAR
INTENTO
INTENTA
INTENTAS
INTENTAMOS
INTENTAIS
INTENTAN
DOS
BAJO
ARRIBA
ENCIMA
USAR
USO
USAS
USA
USAMOS
USAIS
USAN
EMPLEAR
EMPLEO
EMPLEAS
EMPLEAN
AMPLEAMOS
EMPLEAIS

VALOR
MUY
ERAMOS
MODO
BIEN
CUAL
CUANDO
DONDE
MIENTRAS
QUIEN
CON
ENTRE
SIN
TRABAJO
TRABAJAR
TRABAJAS
TRABAJA
TRABAJAMOS
TRABAJAIS
TRABAJAN
PODRIA
PODRIAS
PODRIAMOS
PODRIAN
PODRIAIS
YO
AQUEL

Terminología de la Ingeniería de Software

A continuación mostramos el listado de los términos extraídos automáticamente y utilizados para obtener las variantes denominativas:

Termext

software
información
proyecto
sección
calidad
datos
organización
empresas
procesos
proceso
sistema
trabajo
desarrollo
parte
empresa
cliente
pruebas
tiempo
desarrollo de software
producto
personas
problemas
figura
proyectos
años
usuario
negocio
equipo
vez
diseño
código
forma
caso
usuarios
aplicación
uso
manera
herramientas
aplicaciones
organizaciones
méxico
objetivo
objetivos
día
problema
valor
mundo
modelo
artículo
ingeniería de software
tecnología
análisis
internet
conocimiento
modelos
cuenta
ciclo de vida
arquitectura
momento
autor
sistemas
servicios
nube
base
big data
proceso de pruebas
elementos
experiencia
etc.
resultados
ti
productos
tabla
clientes
requerimientos
bases de datos
certificaciones
gestión de proyectos
proyectos de explotación
seguridad
definición
necesidades
prueba
tecnologías
servicio
mercado
aspectos
http
ingeniería
características
necesidad
lado
tareas
contexto
persona
ejemplo
mayoría
capacidad
año
mejora
actividades
base de datos
paso
certificación
resultado
cosas
sg
contenido
lecturas adicionales y fuentes de información
bio
febrero
gestión
acceso
equipo de desarrollo
evaluación
arquitectura de software
relación
esfuerzo
solución
importancia
casos
cambios
cambio
ejecución
costo
cabo
programa
idea

nivel
beneficios
abril
acuerdo
mejora de procesos
tema
dispositivos
cmmi
futuro
dispositivos móviles
fin
innovación
recursos
tiempo real
país
metodología
plataforma
tarea
herramienta
com
lugar
puntos
riesgo
técnicas
capÍtulo
versión
infraestructura
requisitos
éxito
método
pruebas unitarias
proyecto de mejora
resumen
investigación
documentación
métodos
estudio
reto
punto
operación
moprosoft
productos de software
procesos de negocio
valores
práctica
certificaciones profesionales
usabilidad
personal
comportamiento
proceso de desarrollo
interacción
fuentes de información
adicionales
funcionalidad
volúmenes de datos
mantenimiento
veces
áreas
integración
actividad
estructura
autores
toma de decisiones
tendencias
redes sociales
comunicación
implementación
realidad
desarrolladores
web
creación
agosto-octubre
dominio
crecimiento
decisiones
gente
listado
acciones
código fuente
oportunidad
conjunto
administración
revista
presentación
responsabilidad
miembros
defectos
hardware
modelo de negocio
casos de uso
atributos de calidad
programas
mejora de procesos software
trabajos
estado
ati
soluciones
ideas
estrategia
impacto
países
participantes
número
confianza
partes
conceptos
continuación
publicado
comité editorial
patrones
mayo-julio
universidad
dispositivo
años de experiencia
componentes
línea
equipos de trabajo
análisis de datos
metodologías ágiles
revisión sistemática
punto de vista
adopción
luis daniel soto maldonado
españa
interés
control
riesgos
situación
tamaño
eventos
propuesta
nombre
grupo de calidad
ciclo de desarrollo
inicio
sentido
producto de software
casos de prueba
proyectos de software
proceso de revisión
noviembre
testing
productividad
errores
negocios
preguntas
alcance
colaboración
resto
metodologías
ayuda
reicis
diferencia

objetos
conocimientos
equipos
implantación
concepto
final
red
tecnologías de información
respuesta
columna
rol
ámbito
construcción
estándares
red de ingenierÍa
informática
soporte
pruebas de software
prueba de software
lenguajes de programación
número de reicis
desempeño
área
contribuciones
niveles
dirección
usuario final
lenguaje
principio
explotación de información
apoyo
lenguajes
documentos
descripción
referencia
enfoque
madurez
configuración
compañía
programación
tipo
campo
estrategias
evolución
días
costos
vida
relaciones
opción
página
competencia
salario promedio
evento
interfaz de usuario
software libre
pruebas de aceptación
proyecto de software
tipo de organización
disciplina
principios
hanna oktaba publicado
gunnar wolf publicado
plataformas
historia
error
java
escenarios
mejora continua
asignaturas
desarrollador
ocasiones
//www
&lt
visión
interfaz
atención
medida
ventajas
retos
pymes
elemento
pyme
rendimiento
temas
producción
conclusión
nivel mundial
ventaja
modelo relacional
palabras clave
sistemas de información
posibilidad
inversión
hora
factores
memoria
aplicaciones móviles
ejemplos
api
caso de uso
línea de productos
dólares
oráculo de prueba
gestión de procesos
software engineering institute
visualización de programas
diagrama de objetivos
puntos de vista
decisión
forma de trabajo
gradle
comunidad
modelo de proceso
pregunta
visualización
compañías
actualidad
privacidad
introducción
métodos ágiles
iso
conclusiones
participación
escenario
revista latinoamericana
modelo de referencia
meses
open source
eficiencia
analista
cultura
sitio web
objetivos organizacionales
proveedor
www

tendencia
operaciones
cómputo
inglés
fases
técnica
desarrollos
camino
clase
esquema
condiciones
norma
ej
salarios
capacidades
métricas
procesos de educción temprana

iso/iec
servidores
propósito
revisión
procesos de ingeniería
microsoft
carlos mario zapata jaramillo
normas
razón
patrones de diseño
estudios
arquitectura de información
datos de sensores
procesamiento de datos
conceptos de diseño
expectativas
proceso de gestión
ser
análisis de requisitos
programas de certificación
sector de ti
secretaría de economía
nivel de madurez
representación
productos de trabajo
fase de definición
programa de mejora
lista
palabras
entorno
entendimiento
objeto
selección
documento
medio
plazo
etapa
década
pasos
casa
lenguaje natural
modelo de procesos
import org
líder
estándar
cantidad
septiembre
compromiso
formas
espacio
sensores
julio
computadora
estimación
apps
equipo de trabajo
dependencias
octubre
mx
propuestas
reglas
usuarios finales
opciones
asignatura
papel
satisfacción
especificación

Termostat

prueba
diseño
modelo
usuario
dato
tabla
proceso
atributo
requerimiento
clase
ingeniería
interfaz
proyecto
negocio
herramienta
sección
arquitectura
figura
equipo
cliente
diagrama
patrón
interacción
técnica
componente
análisis
metodología
web
software
variable
funcionalidad
sistema
información
awareness
descripción
implementación
caso de prueba
dominio
caso de uso
calidad
característica
especificación
rol
algoritmo
lenguaje
webapp
módulo
perfil
estimación
evento
reporte
costo
ti
modelo de proceso
actividad
proyecto de explotación
soporte
código
validación
uso
escenario
valor
explotación de información
ciclo
plataforma
desarrolladores
certificación
computadora
ingeniero
publicado
nodo
defecto
comportamiento
base de dato
uml
configuración
nube
error
servidor
tarea
secuencia
standard
aplicación
medición
organización
conocimiento
procesamiento
impacto
archivo
nuestro
usabilidad
entorno
documentación
función
explotación
hardware
verificación
madurez
flujo
salida
curva
área
computación
webapps
usuario final
producto
fase
gerente
método
rango
informática
nuestra
repositorio
aseguramiento
universidad
fig
tamaño
semántica
http
objeto
proceso de desarrollo
ciclo de vida
cómputo
sensores
interfaz de usuario
actor
perfil de usuario
casasegura
código fuente
nuestros
elemento
sistema operativo
programación
sub-sección
www
miembro del equipo
abstracción
promedio
desarrollador
formato
subproceso
tecnología
capítulo
reingeniería
modelado
notación
conjunto

process
prototipo
complejidad
subsistema
operación
estructura
mejora de proceso
requisito
concepto
sitio web
metodología ágil
entrada
área de proceso
artefacto
entrega
conceptualización
guía
diseñador
rango de valor
lineamientos
red social
nivel
iteración
proceso de prueba
cloud
revisión
número
columna
identificación
cmmi
gestión de proyecto
habilidad
falla
video
desarrollo
colaborativo
esquema
paradigma
acs
mejora
analista
lenguaje de programación
planeación
marco conceptual
acoplamiento
versión
variable independiente
confiabilidad
pantalla
evaluación
caso de estudio
bio
dispositivo
caso
automatización
com
rendimiento
equipo de desarrollo
fuente
modelo de requerimiento
diagrama de secuencia
revisión sistemático
autor
prueba unitarias
oráculo
implantación
internet
testeo
reicis
ed
gestión
retroalimentación
problema de negocio
desempeño
estudio
patrón de diseño
consumo de bien
etapa
vinod
nivel de abstracción
número de caso
despliegue
matriz
colección
refinamiento
métricas
modelo conceptual
ítems
nivel de componente
simulación
api
definición
eur
chatbot
estructura de dato
sensor
comando
sei
navegación
educción
nivel de madurez
página
ontología
proceso de explotación
visualización
clasificación
problema de explotación
caja
resultado
profesionista
sintaxis
outsourcing
base
mantenibilidad
ibm
adaptación del proceso
proceso de gestión
número de atributo
enfoque
valor específico
pymes
model
probabilidad
entendimiento
tiempo real
dispositivo móvil
capa
etc
tipo
atributo de calidad
prueba de aceptación
robot
cantidad
objeto de contenido
lector
modelo de calidad
revisión técnico
fuente de dato
modelo de negocio
nuestras
mps
máquina virtual
producto de entrada
participante
página web
gps
flecha
etcétera
total de caso
diseño arquitectónico
producto del trabajo
espacio virtual
calidad del producto

gráfico
proveedor
mapa
taxonomía
epeu
proceso de negocio
grupal
arquitecto
programadores
atributo útil
edición
entrenamiento
representación
trabajo colaborativo
escalabilidad
cbms
diagrama de flujo
navegador
aporte
minería
ontologías
objeto de dato
factor
selección
maestría
equipo de trabajo
soa
elemento de entrada
aplicación web
subprocesos
listado
descubrimiento
iteraciones
programador
design
subconjunto
prueba funcional
variable experimental
producto de trabajo
red semántica
subcaracterísticas
sql
producto de salida
servicio web
operador de mutación
prueba de caja
método formal
systems
tipo de awareness
clase de análisis
sg

gerente de proyecto
bajo
línea de investigación
experimental
ambiente
desarrollo del proyecto
aprendizaje
mejora continuo
doug
almacenamiento
actividad estructural
valor nulo
moprosoft
máquina
investigador
scrum
colaborativa
prueba de integración
conjunto de tarea
sitio
mutación
regla
institute
miles
lista de verificación
repositorios
flujo de trabajo
noviembre
consultoría
derivación
fuente de información
apps
escenario de usuario
neurona
resumen
unidad de medida
disco
modelo de ciclo
itil
minería de dato
diagrama de actividad
cambio
planificación
mx
caja blanco
mis
lenguaje natural
categoría
fase de análisis
variabilidad
bloque

ubicación
tipo de prueba
facilidad
performance
dominio de aplicación
chatbots
apis
gua
factor crítico
modelo de descubrimiento
consultor
aplicación móvil
estrategia de prueba
parámetro
big
clase de diseño
objetivo de negocio
mapeo
conjunto de caso
glosario de término
contenido
capacitación
conjunto de dato
asignatura
glosario
requerimiento funcional
rango de valor específico
prueba de unidad
diagrama de estado
modelo de diseño
criticidad
técnica de análisis
función de seguridad
depuración
modelo de referencia
plan de prueba
agile
quienes
gestión de servicio
administración de proyecto
identificador
proceso de verificación
psp
línea de código
atributo clase
estilo
manejo
pdm
meta
tam
politécnica

equipo de proyecto
metodología de desarrollo
proyecto de tamaño
ítem
gr
adicción
alice
escenario real
ingeniería inverso
framework
descomposición
restricción
menú
selenium
estudio primario
calendarización
especificación formal
disparador
proyecto de tamaño pequeño
punto de función
ej
proceso de administración
or
red
variación
agregado
partición
ecuación
valor bajo
diagrama de clase
tiempo de respuesta
proyecto de mejora
número de modelo
nivel de capacidad
revista
búsqueda
mantenimiento
servicio
porcentaje
trabajo
oportunidad de negocio
dato de prueba
service
prueba de regresión
objetivo del negocio
tablero
administración

Variantes denominativas de la Ingeniería de Software (Termostat-PMI)

F-SECURE, F-SECURE, 1
CATEGORÍA_DE_DESOCUPACIÓN, DESOCUPACIÓN, 0.999765245
DEFECT_REMOVAL, DEFECT, 0.999729679
DISEÑO_POSMODERNO, POSMODERNO, 0.999147833
DEFECTOS_CERRADOS, TOTAL_DE_ANTIGÜEDAD, 0.999133576
FORMA|SOMBRERO, FORMA|TRONCO, 0.999077242
INGENIERÍA|INGENIERÍAS, INGENIERÍA, 0.998371272
MÉTODO_ÁGIL|MÉTODOS_ÁGILES, MÉTODOS_ÁGILES, 0.997240164
REALIZAR|REALIZARA|REALIZARTE|REALIZARLOS|REALIZAREMOS, REALIZAR, 0.996930495
OBJETIVO|OBJETIVOS|OBJETIVAS, OBJETIVO|OBJETIVOS, 0.996798287
CONCEPTOS|HERMANOS, CONCEPTOS, 0.996472709
EXPERTO|EXPERTA|EXPERTAS, EXPERTO|EXPERTA, 0.994251752
EMPRESAS_TRACTORAS, TRACTORAS, 0.991838354
DEFINIR|DEFINIRÁ|DEFINIRLA|DEFINIRLOS|DEFINIREMOS, DEFINIR, 0.991338934
ARCHIVAR|ARCHIVO, ARCHIVO, 0.991109347
AUTENTICADO_SOLO, USUARIO_AUTENTICADO, 0.990732894
MUESTRA|MUESTRAS, MUESTRA, 0.990406884
DETERMINAR|DETERMINARÁ|DETERMINEN, DETERMINAR, 0.989770074
CONTROLAR|CONTROLARÁ|CONTROLARON|CONTROLARLOS, CONTROLAR, 0.989424913
DESARROLLAR|DESARROLLARÁ|DESARROLLÉ|DESARROLLARLOS|DESARROLLAREMOS, DESARROLLAR, 0.989178585
MANUFACTURA|MANUFACTURAS, MANUFACTURA, 0.988674134
DISEÑO|DISEÑOS, DISEÑO, 0.988574694
DISEÑO_EXPERIMENTAL|DISEÑOS_EXPERIMENTALES, DISEÑO_EXPERIMENTAL, 0.988051065
DESARROLLO|DESARROLLOS, DESARROLLO, 0.987925626
MEJORAR|MEJORARÁ|MEJORARÍA|MEJORARLAS, MEJORAR, 0.985481548
ARCHIVO_DE_INTERFAZ_EXTERNO|ARCHIVOS_DE_INTERFAZ_EXTERNOS,
ARCHIVO_DE_INTERFAZ|ARCHIVOS_DE_INTERFAZ, 0.984801199
OR_FIRMS, OR_SETTINGS, 0.984072493
TOTAL|TOTALES, TOTAL, 0.983876515
PLANIFICACIÓN|PLANIFICACIONES, PLANIFICACIÓN, 0.983380767
CONCEPTO, SINÓNIMOS|CONCEPTO, 0.982715081
REVISAR|REVISARÁ|REVISARLA, REVISAR, 0.982084637
FORMALIZAR|FORMALIZÓ, FORMALIZAR, 0.981751412
DATOS_DISPONIBLES, DATOS_DISPONIBLE|DATOS_DISPONIBLES, 0.980879583
METODOLOGÍA_ÁGIL|METODOLOGÍAS_ÁGILES, METODOLOGÍAS_ÁGILES, 0.979307893
EXPERTO|EXPERTA, EXPERTO, 0.978703957
SEIS_SIGMA, SIGMA, 0.978380811
RESOLVER|RESOLVERÁ|RESOLVIERA|RESOLVIERON|RESUELVA, RESOLVER, 0.977899703
VALORES_NULOS, VALORES_NULO|VALORES_NULOS, 0.977742927
ERROR_CUADRÁTICO_MEDIO, ERROR_CUADRÁTICO, 0.976371018
SYS|WHERE, WHERE, 0.976282328
APLICAR|APLICAS|APLICABA|APLICARÍAN|APLIQUEN, APLICAR, 0.976253467
IMAGEN_DE_MÁQUINA_VIRTUAL|IMÁGENES_DE_MÁQUINA_VIRTUAL|IMÁGENES_DE_MÁQUINAS_VIRTUALES,
IMAGEN_DE_MÁQUINA|IMÁGENES_DE_MÁQUINA|IMÁGENES_DE_MÁQUINAS, 0.976065507
PRUEBA_DE_ARREGLO_ORTOGONAL, PRUEBA_DE_ARREGLO, 0.975561899

ENTORNO_DE_INTEGRACIÓN_CONTINUA|ENTORNOS_DE_INTEGRACIÓN_CONTINUA,
ENTORNO_DE_INTEGRACIÓN|ENTORNOS_DE_INTEGRACIÓN, 0.975442415
VALOR_MEDIO|VALORES_MEDIOS, VALOR_MEDIO|VALORES_MEDIO, 0.975375663
CATEGORÍA_DE_ACTIVIDAD_ECONÓMICA, CATEGORÍA_DE_ACTIVIDAD, 0.97524815
CASOS_CLASE, CASOS_NÚMERO, 0.975008634
ASIGNAR|ASIGNARÁ|ASIGNEN, ASIGNAR, 0.974581187
EXPERIMENTALES_INDEPENDIENTES, VARIABLES_EXPERIMENTALES_INDEPENDIENTES, 0.974127562
COMISIÓN_DE_ACTOS_ILEGALES, COMISIÓN_DE_ACTOS, 0.974074946
EXPERTO|EXPERTA|EXPERTAS, EXPERTO, 0.973079854
ACTIVIDAD_DE_MARCO_CONCEPTUAL|ACTIVIDADES_DE_MARCO_CONCEPTUAL,
ACTIVIDAD_DE_MARCO|ACTIVIDADES_DE_MARCO, 0.972302139
PRIMITIVA|PRIMITIVAS, PRIMITIVO|PRIMITIVA|PRIMITIVAS, 0.972229453
FALLAS|FALLE, FALLAS, 0.971513134
OFRECER|OFRECES|OFRECÍ|OFREZCO|OFRECIMOS, OFRECER, 0.971403841
ASPECTO_EXPONENCIAL_Y_CRECIENTE, ASPECTO_EXPONENCIAL, 0.970156248
MOVIMIENTO_DE_CULTURA_LIBRE, MOVIMIENTO_DE_CULTURA, 0.970110929
CONOCIMIENTO_DIFERENTE|CONOCIMIENTO_DIFERENTES, ÁREAS_DE_CONOCIMIENTO_DIFERENTES,
0.969114155
CONOCER|CONOCES|CONOCIERA|CONOCIERON|CONOCEREMOS|CONOCÍAMOS|, CONOCER, 0.968834601
VALOR_BAJO|VALORES_BAJOS, VALOR_BAJO, 0.967154446
NEURONALES, REDES_NEURONALES, 0.966847448
FORMA|FORMAS, FORMA|TRONCO, 0.966502042
FUENTE_TEMPORARIA, TEMPORARIA, 0.966226533
FORENSE|FORENSES, FORENSE, 0.96601294
OR_FIRMS, OR_TEAM, 0.965994584
FORMA|FORMAS, FORMA|SOMBRERO, 0.965957817
CICLO_DE_PROMOCIÓN_EXCESIVA, CICLO_DE_PROMOCIÓN, 0.965737522
CONCEPTO|PADRE, CONCEPTO, 0.965281007
CONTAR|CONTARA|CONTASE|CONTABAN|CUENTO|CUÉNTANOS|, CONTAR, 0.964172909
MEDICIÓN_DE_TAMAÑO_FUNCIONAL|MEDICIONES_DE_TAMAÑO_FUNCIONAL,
MEDICIÓN_DE_TAMAÑO|MEDICIONES_DE_TAMAÑO, 0.964044435
INTRUSIÓN, PRUEBA_DE_INTRUSIÓN|PRUEBAS_DE_INTRUSIÓN, 0.963302915
VERSIONES_DE_SISTEMAS_OPERATIVOS, VERSIONES_DE_SISTEMAS, 0.963249921
INFORMACIÓN_DE_TIPO_AWARENESS, INFORMACIÓN_DE_TIPO, 0.962929019
INTERACCIONES_EN_ESPACIOS_VIRTUALES, INTERACCIONES_EN_ESPACIOS, 0.962873112
VERIFICAR|VERIFICARÁ|VERIFICARLO|VERIFIQUE|VERIFIQUEN, VERIFICAR, 0.962383319
NEURONAL, RED_NEURONAL, 0.962383223
COMPORTAMIENTO|COMPORTAMIENTOS, COMPORTAMIENTO, 0.962145352
LISTA-DE-PARÁMETROS, NOMBRE-DEL-EVENTO, 0.962113632
AGREGAR|AGREGARÁ|AGREGARÍA|AGREGAMOS, AGREGAR, 0.961834148
MOVER|MOVERÁ|MOVIERA|MOVERNOS|MUEVAN, MOVER, 0.961504345
APLICACIÓN_DE_FUENTE_ABIERTA|APLICACIONES_DE_FUENTE_ABIERTA,
APLICACIÓN_DE_FUENTE|APLICACIONES_DE_FUENTE, 0.960849917
NÚCLEO_DE_OPENSTACK, PROYECTOS_NÚCLEO, 0.959982026
ACTIVIDAD_DE_ESPARCIMIENTO_EMOCIONAL|ACTIVIDADES_DE_ESPARCIMIENTO_EMOCIONAL,
ESPARCIMIENTO_EMOCIONAL, 0.959821618
PORCENTAJE_PROMEDIO, PROMEDIO_DE_DEBILIDADES, 0.958967682
PANEL_DE_CONTROL|PANELES_DE_CONTROL, PANEL|PANELES, 0.955719665
TIMBRE_FISCAL_DIGITAL, TIMBRE_FISCAL, 0.955431667
FIRMS, TEAM, 0.954917144
FAMILIA_MONOPARENTAL, MONOPARENTAL_NUCLEAR, 0.954682007

RECURSO_HUMANO|RECURSOS_HUMANOS, RECURSOS_HUMANOS, 0.954087159
FACTORES_DE_ÉXITO_CRUCIALES, ÉXITO_CRUCIALES, 0.954086102
EDUCCIÓN_TEMPRANA, TEMPRANA_DE_REQUISITOS, 0.954075219
ÁRBOL_DE_DESCOMPOSICIÓN_FUNCIONAL, ÁRBOL_DE_DESCOMPOSICIÓN, 0.953727599
VARIABLE_EXPERIMENTAL|VARIABLES_EXPERIMENTALES, VARIABLES_EXPERIMENTALES, 0.953345266
INFRAESTRUCTURA_TECNOLÓGICA|INFRAESTRUCTURAS_TECNOLÓGICAS,
INFRAESTRUCTURA_TECNOLÓGICA, 0.952975313
NORMA_UNE|NORMAS_UNE, NORMA_UNE, 0.952620221
IDEAL|IDEALES, IDEAL, 0.951732281
PROCESAMIENTO_DE_LENGUAJE_NATURAL, PROCESAMIENTO_DE_LENGUAJE, 0.95099155
OR_SETTINGS, OR_TEAM, 0.950335009
DISTANCIA_FK|DISTANCIAS_FK, FK, 0.950035126
INGENIERÍA_EN_SISTEMAS_COMPUTACIONALES, SISTEMAS_COMPUTACIONALES, 0.949987924
CONCEPTO|PADRE, SINÓNIMOS|CONCEPTO, 0.949924198
PROYECTO_DE_TAMAÑO_PEQUEÑO|PROYECTOS_DE_TAMAÑO_PEQUEÑO, TAMAÑO_PEQUEÑO,
0.949833413
TOMBOLA, TSV, 0.949825052
CONTADOR_DE_AES, MODO_CONTADOR, 0.949824854
CMMI_DE_PROCESOS, GESTIÓN_DE_CAPACIDAD, 0.949454122
SISTEMA_DE_TIEMPO_REAL|SISTEMAS_DE_TIEMPO_REAL, SISTEMA_DE_TIEMPO|SISTEMAS_DE_TIEMPO,
0.946726846
EXPLORACIÓN_DE_CONCEPTOS_INICIALES, EXPLORACIÓN_DE_CONCEPTOS, 0.946621539
RESERVORIO_DE_BLOQUES, RESERVORIO, 0.946168758
CERTIFICACIÓN_DE_EMPRESAS, PENDIENTE_DE_DEFINICIÓN, 0.945826844
ENFOQUE_DE_CUARTO_LIMPIO, ENFOQUE_DE_CUARTO, 0.944052323
MEDIO_DE_TRASPORTE, TRASPORTE, 0.943767845
ACCION-R, ACCION-T, 0.943440226
SELECCIONAR|SELECCIONARÁ|SELECCIONARÁN|SELECCIONAMOS, SELECCIONAR, 0.943426258
SAFETY, SEGURIDAD_SAFETY, 0.942598545
ARREGLO_ORTOGONAL, ORTOGONAL, 0.942478948
CARACTERÍSTICA|CARACTERÍSTICAS, CARACTERÍSTICO|CARACTERÍSTICOS|CARACTERÍSTICAS,
0.942437975
AHOGO_FINANCIERO, AHOGO, 0.941852748
COLATERALES, EFECTOS_COLATERALES, 0.941544613
UTILIDAD|UTILIDADES, UTILIDAD, 0.94036894
ARQUITECTURA|ARQUITECTURAS, ARQUITECTURA, 0.93938018
ESTUDIO_COMPARATIVO|ESTUDIOS_COMPARATIVOS, ESTUDIO_COMPARATIVO, 0.939232064
ENFOQUE_DE_INTEGRACIÓN_INCREMENTAL, INTEGRACIÓN_INCREMENTAL_CONVENCIONAL, 0.93775368
EXPLORACIÓN_INICIAL, REPORTE_DE_EXPLORACIÓN_INICIAL, 0.937365246
BASE_DE_DATOS_ALFANUMÉRICA|BASE_DE_DATOS_ALFANUMÉRICAS,
DATOS_ALFANUMÉRICA|DATOS_ALFANUMÉRICAS, 0.936624367
CLOUD-TO-CLOUD, SERVICIOS_DE_CONTINUIDAD, 0.936516785
DIAGNÓSTICO|DIAGNÓSTICOS, DIAGNÓSTICO, 0.936220587
DISEÑO_DE_CUARTO_LIMPIO, DISEÑO_DE_CUARTO, 0.935564278
POBLACIONAL, SITUACIÓN_POBLACIONAL, 0.935464721
COBERTURA_TIEMPO, SEG, 0.934979881
NORDESTE_ARGENTINO, NORDESTE, 0.934821055
CAMPO_VECTORIAL, VECTORIAL, 0.933831698
CONTEXTOS_DE_MASIVIDAD, MASIVIDAD, 0.933461597
EVALUACIÓN_POR_EVALUADORES, PENDIENTE_DE_DEFINICIÓN, 0.933447327
FILA_DESTINO, FILA_ORIGEN, 0.93333775

ACTOS_ILEGALES, COMISIÓN_DE_ACTOS_ILEGALES, 0.930527108
COMPARACIÓN|COMPARACIONES, COMPARACIÓN, 0.929585788
PROTOCOLO_DE_COMUNICACIÓN_INTERNA, PROTOCOLO_DE_DOCUMENTACIÓN_INTERNA, 0.929571093
RESEÑAR|RESEÑA, RESEÑA|RESEÑAS, 0.928930523
BASE_DE_DATOS_RELACIONAL|BASES_DE_DATOS_RELACIONAL, DATOS_RELACIONAL, 0.927989882
CANTIDAD_DE_TUPLAS_DISPONIBLES, CANTIDAD_DE_TUPLAS, 0.927419136
EMERGENCY|SITUATION, SITUATION, 0.926415112
OR_FIRMS, TEAM, 0.9258616
EXPLOTACIÓN_DE_INFORMACIÓN, EXPLOTACIÓN, 0.925755406
DIAGRAMA_ENTIDAD-RELACIÓN, ENTIDAD-RELACIÓN, 0.925595534
FIRMS, OR_SETTINGS, 0.92556818

Variantes denominativas de la Ingeniería de Software (Termext-PMI)

TECNOLOGÍAS_DE_INFORMACIÓN, TECNOLOGÍAS_DE_INFORMACIÓN, 1
CONJUNTO_DE_REGLAS_SINTÁCTICO-SEMÁNTICAS, REGLAS_SINTÁCTICO-SEMÁNTICAS, 0.966173986
ARQUITECTURA_DE_INFORMACIÓN_AUTOR, INFORMACIÓN_AUTOR, 0.965477056
OR_SETTINGS, OR_TEAM_OR_FIRMS, 0.959786489
FÁBRICA_DE_TELARES, PRODUCCIÓN_DE_MATERIA_PRIMA, 0.939283641
LÍDERES_DE_GRUPO, LÍDERES_SERVIDORES, 0.927649156
TASA_DE_ACEPTACIÓN, TASA_DE_RECHAZO, 0.910237048
FLUJO_DE_VALOR, MAPA_DE_FLUJO, 0.909542727
EXPORTADORES_DE_SERVICIOS, PAÍSES_EXPORTADORES, 0.907787424
CONTROLES_DE_FORMULARIOS, POSICIÓN_DE_CONTROLES, 0.904502606
FLEXIBILIDAD_DE_USO, MITIGACIÓN_DE_RIESGOS_AMBIENTALES, 0.894741424
GESTO_DISCRETO, TIPOS_DE_GESTOS, 0.885160007
ADMINISTRADOR_DE_BASE, CERTIFICADO_DE_ADMINISTRADOR, 0.88083403
GESTIÓN_DE_RECURSOS, GESTIÓN_DE_RECURSOS_HUMANOS, 0.874275016
LIDERAZGO_RESPONSABLE, TI_ESPECIALIZADO, 0.867647072
BACKLOG_DE_PROYECTOS, LISTA_PRIORIZADA_DE_PROYECTOS, 0.863096235
QUARTZSCHEDULER, SHUTDOWN_INFO, 0.860429577
ASESOR_DE_NEGOCIOS, TI_ESPECIALIZADO, 0.858100499
MEMORYSTREAM_MISTREAM, XMLWRITER_ESCRITORXML, 0.855874539
COMERCIO, COMERCIO_ELECTRÓNICO, 0.849112014
BASES, BASES_DE_DATOS, 0.841224756
CRITERIOS_DE_USABILIDAD, RECOMENDACIONES_VIGENTES, 0.839438458
CONECTORES_DE_REDES, CONECTORES_DE_TRANSPORTE, 0.837313689
EVENTOS_DE_INTERÉS_CIENTÍFICO, INTERÉS_CIENTÍFICO_Y_TÉCNICO, 0.833584987
MITIGACIÓN_DE_RIESGOS_ECONÓMICOS, SEGURIDAD_DE_USO, 0.828863362
ESQUEMA_DE_CONTRATACIÓN, RANGO_DE_EDAD, 0.826446277
TILEINMEDIATO, TOASTINMEDIATO, 0.824504489
INCOMPATIBILIDAD_DE_FONDO, PROJECT_PORTFOLIO_MANAGEMENT, 0.823255836
DEBILIDAD_PARTICULAR, GANZÚA_SOLO, 0.819971923
ADOPCIÓN_ESPAÑOLA, ADOPCIÓN_EUROPEA, 0.819447034
ESTRUCTURA_COMPATIBLE, ESTRUCTURA_DE_CMMI, 0.81643683
BOTTOM, LEFT, 0.816342543
CONSEJO_EDITORIAL, EDITORIAL_REICIS, 0.809614173
ACTIVIDADES_PRIMARIAS, CADENA_CUENTA, 0.80632744
ASESOR_DE_NEGOCIOS, LIDERAZGO_RESPONSABLE, 0.804340009
CUBETA, TOALLA, 0.801152625
DESARROLLO_DE_EMPRESAS, TECNOLOGÍA_OPERACIONAL_NECESARIA, 0.797914383

AMENAZAS_GENERALES, GANZÚA_SOLO, 0.794508186
PAGOS_DE_SERVICIOS, TIEMPO_AIRE, 0.7925805
COLABORATIVO, TRABAJO_COLABORATIVO, 0.791874031
CENTROS_DE_AYUDA, DOCUMENTACIÓN_DE_PRUEBA, 0.790189257
FÓRMULA_ANTERIOR, INTERVALO_VALOR, 0.790079941
ERRORES_DE_ESTÁNDARES, ERRORES_DE_PLATAFORMA, 0.788964757
PRUEBA_DE_HIPÓTESIS, VALIDACIÓN_DE_RESULTADOS, 0.779857454
CONSUMO_PÚBLICO, SIMULACIÓN_DE_USO, 0.778961118
DIRECCIÓN_DE_ORIGEN, OBJETO_CONCEPTUAL_O_FÍSICO, 0.778619088
ACUERDO_DE_LICENCIA, CONTRATO_DE_LICENCIA, 0.777180759
EFECTIVIDAD_DE_USO, EFICIENCIA_DE_USO, 0.775519481
ESPACIOS_COWORKING, ESPACIOS_LABORALES, 0.772744838
APLICACIONES_DE_TI, LÍDERES_DE_NEGOCIO, 0.772673063
CÓDIGO_DE_APP, EJEMPLO_DE_CÓDIGO, 0.772438407
ETAPA_DE_REQUISITOS_TARDÍOS, SUB_ETAPAS, 0.772343926
CONJUNTO_DE_BASES, DATOS_ALTERNATIVAS, 0.772026923
ADOPCIÓN_DE_SOFTWARE_LIBRE, FACTOR_DE_ADOPCIÓN, 0.77044616
FINALES, USUARIOS_FINALES, 0.769583404
FUNDADOR_DE_STARBUCKS, TRABAJO_DE_GENTE, 0.769426115
DE_SOFTWARE, SOFTWARE, 0.7688469
COACH_SOLO, PROCESO_DE_COACHING, 0.767057317
ANÁLISIS_SERIO, PROYECTOS_COSTOSOS, 0.764230174
BOTTOM, RIGHT, 0.764228843
SHUTDOWN_INFO, SHUTTING_DOWN, 0.763591135
INGENIEROS, INGENIEROS_DE_SOFTWARE, 0.761594408
LÍNEA_DE_MONTAJE, MÉTODO_EXISTENTE, 0.761579474
MÓDULO_DE_COBRO, TIPO_DE_DERECHO, 0.760956999
CICLO, CICLO_DE_VIDA, 0.75909071
ESTILO_AUTORITARIO, INSTRUCCIÓN_CLARA, 0.758538268
OPERACIÓN_DE_TI, ÁREAS_DE_INFRAESTRUCTURA, 0.75778875
LEFT, RIGHT, 0.757479345
ACTIVIDADES_PRIMARIAS, ACTIVIDADES_SECUNDARIAS, 0.756306839
SALAS_DE_JUEGOS, SALAS_DE_LECTURAS, 0.754939048
MARGE_DESARROLLO, RAMA_DESARROLLO, 0.753830402
BASE_DE_CÓDIGOS, CORPUS, 0.75242693
TEXTBLOCK, TEXTWRAPPING, 0.751491359
LÍDERES_DE_GRUPO, MOTIVADORES_EXTRÍNSECOS, 0.750361288
GESTOS_CONTINUOS, TIPOS_DE_GESTOS, 0.74860721
DOCUMENTACIÓN_TÉCNICA_CORPORATIVA, FUENTE_DE_DOCUMENTACIÓN, 0.747975181
PUNTO_DE_VISTA, VISTA, 0.743932481
CAPACIDAD_TÉCNICA, ORGANIZACIÓN_PERSONAL, 0.74365075
EXISTENCIA_DE_SELECCIÓN, MODO_DE_ENTRADA, 0.741702902
FALTA_DE_ACCESO, PROBLEMAS_DE_DESEMPEÑO, 0.741600441
ARQUITECTOS_DE_TI, ASOCIACIÓN_GLOBAL, 0.741454935
CERTIFICACIONES_EMPRESARIALES, CERTIFICACIONES_TÉCNICAS, 0.740540657
ESTIMACIÓN_EXISTENTES, MULTITUD_DE_MÉTODOS, 0.738985756
EQUIPO_INADECUADO, FALLAS_DE_EQUIPO, 0.738921105
ESQUEMATIZACIONES_SENCILLAS, MATRICES_DE_MOMENTOS, 0.738915383
GIT_CHECKOUT_DESARROLLO, RAMA_LLAMADA, 0.738440694
APOYO_DE_MARCOS_SPI, PRODUCTO_ISO, 0.738361219
CONDICIONES_HABITACIONALES, CONJUNTO_RESIDENCIAL, 0.737494113

DATOS_OPTIMIZADA, PLATAFORMA_DE_BASE, 0.737314016
LÍDERES_SERVIDORES, MOTIVADORES_EXTRÍNSECOS, 0.736741821
ENTORNOS_LABORALES, YANG, 0.736563151
GUID, PHONEID, 0.736439673
CONJUNTO_DE_DISCOS, RAIDØ, 0.736362756
APLICACIÓN_ANTERIOR, APLICACIÓN_FUTURA, 0.735686891
DONATIVOS_DE_PERSONAS, DONATIVOS_DE_USUARIOS, 0.734393198
ITERACIONES_DE_DISEÑO, SESIÓN_DE_CUESTIONAMIENTO, 0.733873486
MODELO_DE_RECIENTE_CREACIÓN, SER_MOPROSOFT, 0.732910117
PORCENTAJE_DE_PROFESIONISTAS, TI_TODAVÍA, 0.732593237
EQUIPOS_DE_LIDERAZGO, PENSAMIENTO_CREATIVO, 0.731881855
CDS, DIRECTOR_DE_PROYECTOS_INFORMÁTICOS, 0.73096233
IMPACTO_IMPORTANTE, MÉTODO_DE_TRABAJO, 0.730681746
MÁQUINAS, MÁQUINAS_VIRTUALES, 0.728222184
ALMACENAMIENTO_REMOTO, DIRECTORIO_LOCAL, 0.727156675
ESTÁNDAR_XML, FORMATO_DE_INTERCAMBIO, 0.725641749
PROBLEMAS_RECURRENTES_DE_DISEÑO, SOLUCIONES_CONCEPTUALES, 0.725271884
PROBLEMAS_DE_NEGOCIACIÓN, PROBLEMAS_ECONÓMICOS, 0.722423521
ÁRBOL_DE_OBJETIVOS, ÁRBOL_DE_PROBLEMAS, 0.722043327
REPRESENTACIÓN_ININTELIGIBLE, TEXTO_CLARO, 0.719992564
E-COMERCIO_EXTERIOR, E-RED_SOCIAL, 0.719055327
CONSTELACIÓN_DE_DESARROLLO, SUB-PRÁCTICAS, 0.717822034
ACCESO_FRECUENTE, ACCESO_VELOZ, 0.71754975
CONCEPCIÓN_TRADICIONAL, DISEÑO_ABIERTO, 0.71696178
LA_TECNOLOGÍA, TECNOLOGÍA, 0.716737518
EMPRESA_PUBLICADO, MUNDO_MÓVIL_PUBLICADO, 0.71598415
GESTO_CONTINUO, MÉTODO_SOLO, 0.715016446
CONSECUENCIAS_NOCIVAS, SISTEMA_MÚSCULO-ESQUELÉTICO, 0.714538047
APLICACIÓN_EMPRESARIAL, EJBS, 0.713733067
DESARROLLO_DE_INNOVACIÓN, GENERACIÓN_DE_EMPLEOS, 0.713529343
MICRO_AHORRO, MICRO_CRÉDITOS, 0.712970427
EQUIPO_TENIA, SIGNOS_DE_CORAJE, 0.712813837
CONDICIONES_DE_EXCEPCIÓN, ESCENARIO_DE_USO_IDEAL, 0.712626407
PESOS_MEXICANOS, TIPO_DE_CAMBIO, 0.711978403
CARGA_DE_PÁGINA, UBICACIÓN_VÁLIDA, 0.711354465
CDS, SEGENTE, 0.7113192
MANEJO_DE_XML, TIPO_DE_ASPECTOS, 0.71123307
RFS, RNFS, 0.709754191
FORMA_ESCALABLE, SUB-CONJUNTOS_DE_DATOS, 0.709601952
CADENAS_GENERALES_BÁSICAS_DE_BÚSQUEDA, OR_SETTINGS, 0.708598751
ENVÍO_DE_DINERO, PAGOS_DE_SERVICIOS, 0.707998038
AMBIENTES_ABIERTOS, TIPO_DE_CORPORACIONES, 0.707973248
CONCEPTUALIZACIÓN, PROCESO_DE_CONCEPTUALIZACIÓN, 0.707915994
ARQUITECTURAS_DE_REFERENCIA, DISEÑOS_COMPLETOS, 0.707638119
NARANJA, TONALIDADES_VIVAS_Y_BRILLANTES, 0.707394614
CONSUMO_DE_ENERGÍA, ENERGÍA_ELÉCTRICA, 0.706830638
ACCESO_MASIVO, DESARROLLO_DE_EMPRESAS, 0.706325552
LABOR_DE_ADMINISTRACIÓN, LABOR_DE_VENTA, 0.705627396
CAPACITADORA_EXCELENTE, ESTÁNDARES_DE_TI, 0.705601244
ASUNTO_ESPECÍFICO, EQUIPO_DE_PROYECTO_COMPLETO, 0.70498792
AGUINALDO, REPARTO_DE_UTILIDADES, 0.70472132

COSAS_BÁSICAS, TIPOS_DE_UVAS, 0.70336762
PREGUNTA_FÁCIL, RESPUESTA_DIFÍCIL, 0.702617748
EXISTENCIA_DE_SELECCIÓN, ORDEN_DE_TABULACIÓN, 0.702167972
CARPETA_REFERENCES, CLICK_DERECHO, 0.701608554
ERRORES_DE_ESTÁNDARES, PROBLEMAS_FÍSICOS, 0.701159779
ATAQUES_MALICIOSOS, TARJETAS_INTELIGENTES, 0.700990454
ABANICO_DE_SOLUCIONES, DESARROLLADOR_CUENTA, 0.700549078
ENTREGA_DE_SERVICIOS, SERVICIOS_DE_APLICACIONES, 0.700485742
DISFUNCIÓN_ERÉCTIL, INTENCIÓN_INICIAL, 0.699770132
TÉCNICAS_DE_KANBAN, ÁRBOL_DE_ALTO_DESEMPEÑO, 0.699662142
GESTOS_CONTINUOS, GESTOS_MULTITOUCH, 0.69905796

Terminología de la Biodiversidad

A continuación mostramos el listado de los términos extraídos automáticamente y utilizados para obtener las variantes denominativas:

Termext

especies
estado
conservación
méxico
parte
biodiversidad
años
estado de méxico
región
chiapas
estado de campeche
península de yucatán
país
suelo
recursos naturales
especie
entidad
figura
número de especies
yucatán
distribución
mayoría
cuadro
estado de aguascalientes
agua
uso
plantas
importancia
géneros
familias
superficie
población
aves
diversidad
año
vegetación
grupo
ecosistemas
poblaciones
áreas
campeche
norte
tipos de vegetación
estudio
información
zona
desarrollo
sur
hábitat
estudios
cuerpos de agua
caso
bosques
conocimiento
suelos
municipios
mundo
comunidades
área
familia
manejo
presencia
diversidad biológica
fauna
lugar
pastizales
protección
vez
aguascalientes
grupos
animales
organismos
zonas
forma
sierra madre
regiones
sitios
educación ambiental
diversidad de especies
agricultura
base
cambio de uso
msnm
tierra
producción
peces
estado de morelos
estado de chihuahua
situación
fauna silvestre
flora
c.
tipo de vegetación
manera
medio ambiente
golfo de méxico
reserva
territorio
mar
relación
árboles
diversidad genética
cuenca
bosque
sierra
flores
quintana roo
recursos
aprovechamiento
actividades
estados
áreas naturales
ríos
acciones
peligro de extinción
cuenta
nivel nacional
riqueza de especies
riesgo
pérdida
naturaleza
mamíferos
casos
estudio de caso
habitantes
ambiente
insectos
personas
hongos

reptiles
servicios ambientales
impacto
hojas
datos
cambios
crecimiento
actividad
riqueza
trabajo
estado de chiapas
década
centro
actividades humanas
localidades
abundancia
clima
cultivo
calidad
tiempo
análisis
total
cultivos
usos
altos de chiapas
ganadería
valor
estructura
desierto chihuahuense
resto
ecosistema
especies nativas
amenazas
resultados
pastizal
hombre
condiciones
especies endémicas
morelos
nivel
costa
incremento
especies de aves
características
incendios forestales
lado
especies de plantas
chihuahua
cabo
resultado
estado de yucatán
vida silvestre
partes
individuos
humedales
participación
enfermedades
selvas
factores
alimento
comunidad
disminución
sierra fría
erosión
nivel mundial
spp
fin
cambio climático
salud
países
acuerdo
materia orgánica
km2
importancia económica
vegetación secundaria
abejas
productos
selva baja
conjunto
extensión
registros
deforestación
maíz
género
número
estado de michoacán
mesófilo de montaña
temperatura
laguna de términos
contaminación
anfibios
momento
gobierno
capacidad
tema
estado de conservación
altura
sistema
ambientes
construcción
investigación
lugares
michoacán
necesidad
especies de plantas vasculares
control
sentido
petenes
origen
aguas
anp
medida
reducción
Áreas naturales protegidas
capítulo
año 2000
altitud
descripción
fecha
ciclo de vida
lagunas
selva baja caducifolia
tamaño
protección especial
elementos
ley
continuación
relieve
periodo
semarnat
actualidad
inegi
selva
madera
introducción de especies exóticas
sitio
paisaje
órdenes
frecuencia
bosque tropical caducifolio
biosfera
cm
agua dulce
estudios de diversidad genética
sociedad
rocas
cuerpo
superficie estatal
cap
formas
semillas
bosques de encino

mariposas
trabajos
explotación
décadas
pastos marinos
humedad
montañas
proceso
lluvias
mujeres
lagunas costeras
amenaza
mitad
producto
casas
día
aumento
ocasiones
milpa
asentamientos humanos
categoría de riesgo
condiciones ambientales
península
precipitación
rzedowski
programas
productores
composición
ciudad de aguascalientes
ganado
promedio
deterioro
zona norte
recurso
efectos
año 2005
cambio
especies vegetales
lagos
registro
servicios
hectáreas
madre de chiapas
plazo
papel
problema
cobertura
reproducción
vida
especies de peces
establecimiento
bosque de pino-encino
aspectos
siglo xx
conabio
sociedad civil
pinus
poblaciones silvestres
consecuencia
efecto
bosque de encino
epífitas
arbustos
frutos
norma oficial mexicana
problemas
destrucción
manglares
planta
fragmentación
comunidades vegetales
algas
meses
oeste
comunidades rurales
conclusiones
guatemala
objetivo
cactáceas
fuego
alimentación
terreno
matorrales
límites
terrenos
respecto
presas
especies de mamíferos
distrito federal
universidad
veces
campo
recursos genéticos
barrancas
interior
estrategia
tendencia
moluscos
caribe
interés
programa
pobladores
café
investigaciones
restauración
recursos forestales
río
raíces
instituciones
río conchos
continente americano
punto de vista económico
incendios
cuencas
pastos
arroyos
degradación
revisión
escala
selva lacandona
Área de protección
especies silvestres
esperanza de vida
mantenimiento
planeta
existencia
mercado
extracción
planicies
vertebrados
valores
nombre
diferencia
pastizales naturales
profundidad
matorral subtropical
verano
zona costera
cola blanca
siglo
atención
este
república mexicana
metros
esfuerzos de conservación
diario oficial
mayas
agricultura de riego
ejemplo
hábitats
vegetación acuática
especies invasoras
partes bajas

esfuerzos
formas de vida
comparación
proporción
vegetal
introducción
venado
economía
estrategias
líquenes
aplicación
murciélagos
grupo de plantas
superficie total
época
municipio
provincia
sector primario
climas
modificación
condición
densidad

colaboradores
contreras-macbeath et
pinos
guzmán
valor económico
patrones de distribución
valle de méxico
población humana
principios
línea de costa
nevado de toluca
procesos
importancia ecológica
instituto
biología
crustáceos
punto de vista
educación
ordenamiento ecológico
microorganismos
ejidos
altitudes

proyectos
norteamérica
gramíneas
águila real
mg/g ps
temperatura media anual
alteración
supervivencia
martínez
formación
historia
et
balsas
invertebrados
niveles
encinos
m.
pérdida de hábitat
bosques de pino-encino
selvas bajas caducifolias
ganadería extensiva
años de edad

Termostat

especie
conservación
vegetación
suelo
área
biodiversidad
municipio
manejo
ecosistema
bosque
selva
planta
sierra
ave
fauna
cuadro
figura
género
diversidad
hongo
superficie
aprovechamiento
agua
hábitat
río
mamífero

cuenca
entidad
pastizales
spp
flora
insecto
matorral
distribución
uso
área natural
recurso natural
reptil
estudio
tipo de vegetación
anp
árbol
colección
maya
península
ubicar
ambiente
población
sitio
anfibio
localidad
localizar

educación ambiental
vegetal
laguna
selva baja
msnm
sustentable
cuerpo de agua
número de especies
pino
zona
roca
lluvia
mexicano
pez
humedales
sp
familia
biosfera
fauna silvestre
promedio
reportar
habitar
extinción
servicio ambiental
venado
pastizal

extracción
deforestación
temperatura
uma
recurso
riqueza
alga
ejido
abundancia
presas
cultivo
restauración
porción
planeación
diversidad biológica
erosión
humedad
cacería
tortuga
precipitación
hoja
pasto
monitorear
encino
vertebrados
mariposa
tierra
abeja
especies nativas
diversidad genéticos
vida silvestre
cm
sustentabilidad
especies endémicas
arroyo
protegidas
cactáceas
milpa
caducifolia
nuestro
murciélago
ganadería
actividad
encontrar
petenes
peligro de extinción
estado
paisaje
leña
forestal
manglares

cuerpo
especie de planta
manantial
laguna de términos
lago
apéndice
cambio de uso
agua dulce
manglar
altitud
calakmul
parque
larva
conabio
variedad
encinos
microorganismos
pobladores
coníferas
predio
cola
arrecife
biología
impacto
materia orgánica
distribuir
mangle
quienes
serpiente
monte
riqueza de especies
estudio de caso
cerro
planicies
nutrientes
registrar
marina
diversidad de especies
camarón
madera
arbusto
cobertura
universidad autónoma
águila
temporada
lagartija
especies exóticas
estado de conservación
autónoma
hábitats
norte

comunidad vegetal
depresión
incendio
flor
recurso forestal
orquídeas
araña
ecología
pino-encino
dunas
moluscos
janos
helecho
jardín
coral
crustáceos
palma
rana
plaga
campesino
colecta
riego
maderables
especie de ave
característica
región
desierto
variación
altos
gramíneas
caracol
cola blanca
superficie estatal
corredor biológico
estrato
lomeríos
lerma
helmintos
ordenamiento ecológico
clima
reproducción
norte del estado
listado
equinodermos
desarrollo sustentable
selva mediana
patrón
sedimentos
meseta
aves acuáticas
tallo

instituto
actividad humana
ordenamiento
ladera
reporte
conchos
presencia
inegi
cenotes
densidad
cocodrilo
ornato
perro
vegetación secundarios
maíz
valle
marino
bosque de encino
barrancas
secretaría
generar
epífitas
venado cola
águila real
topografía
pastoreo
conocimiento
corredor
plantas vasculares
variabilidad
especie silvestre
chile
uso del suelo
roedor
norma oficial
sustrato
pradera
pérdida
puma
laguna costera
ora
hembra
montaña
década
nuestro país
anidación
registro
alteración
germoplasma
distribución geográfica
plan de manejo
salinidad
territorio estatal
colonia
bosque de pino
cascabel
acción de conservación
asentamiento humano
deterioro
categoría de riesgo
huevo
total
centro
bacteria
frijol
planta medicinal
mascota
altitudes
extensión
bosque tropical
suroeste
invertebrados
henequén
desierto chihuahuense
pronatura
pasto marino
cobertura vegetal
líquenes
individuo
vivero
pers
estrategia de conservación
pib
semilla
lagarto
ex
acuacultura
alacrán
orquídeas
pluma
ecoturismo
protección especial
organismo
color
población humana
palomillas
hongos comestibles
escarabajos
importancia ecológica
ciclo
condiciones ambientales
litoral
caducifolio
etcétera
especie de mamífero
caracterizar
área de protección
volcán
política pública
nido
producción
fertilidad
superficie total
río conchos
conanp
botánicos
var
cuenca del río
trucha
distrito federal
hectárea
ejemplar
recarga
población silvestre
programa de manejo
variar
degradación
subhúmedo
endemismos
sobrepastoreo
ecosistema acuático
sobreexplotación
nivel estatal
angiospermas
bosque mesófilo
selva alta
actividad agropecuaria
temperatura media
oeste
evento
actividad productiva
introducción de especies
estrategia estatal
semarnat
mexicanus
smo
duna
parte baja
café
inventario
residuos sólidos
conafor
ea

protozoos
depredadores
ciclo de vida
especie vegetal
gestión ambiental
carne
subcaducifolia
mascotas
colectas
parte alta
drenaje
especie invasora
acacia
mejoramiento
equilibrio ecológico
maya
habitante
calizas
aprovechamiento forestal
normatividad
albergar
esquema
presentar
tilapia
pecarí
herbarios
endemismo
aprovechamiento sustentable
eje neovolcánico
carpa
vegetación natural
reforestación
mosca
palo
nuestros
recreación
gradiente
miles
bromelias

costera
superficie del estado
arcilla
antropogénicas
cayo
noroeste
identificar
importancia económica
comunidad
usd
depresión central
álvarez
manto
bosque de pino-encino
diario oficial
sobresale
costa
pato
llano
vegetación acuática
herbario
cañada
apropiación
alimentar
pantano
mortalidad
interacción
especie del género
uso de suelo
cedro
playa
tala
capacitación
sequía
predomina
época
implementación
mosquito
virginianus

ecosistema natural
riqueza biológica
manchones
dominancia
descripción
nivel del mar
paloma
gusano
caza
corteza
pse
introducción de especies exóticas
calvillo
opiliones
cafetal
coyote
deterioro ambiental
tamaño
incendio forestal
sapo
manejo integral
especie de anfibio
biodiversidad del estado
taxones
artrópodos
sonora
grupo de plantas
época de lluvia
piel
com
universidad
tarahumara
especie de hongo
ictiofauna
área de distribución
tiburón
reserva
valor económico

Variantes denominativas de la Biodiversidad (Termostat-PMI)

MESETA_CENTRAL|MESETAS_CENTRALES, MESETA_CENTRAL, 1
VITRO, VITRO, 1
PIM_PIM, PIM, 0.99902876
ASENTAMIENTO_HUMANO|ASENTAMIENTOS_HUMANOS, ASENTAMIENTOS_HUMANOS, 0.992742529
LEÑA|LEÑAS, LEÑA, 0.990875142
TANGENTES_TRIFÁSICOS, TRIFÁSICOS, 0.990682757
ACTIVIDADES_DE_COLECTA_CIENTÍFICA, ACTIVIDADES_DE_COLECTA, 0.990610679
RECURSO_NATURAL|RECURSOS_NATURALES, RECURSOS_NATURALES, 0.989534075
RATÓN_DE_PATAS_BLANCAS|RATONES_DE_PATAS_BLANCAS, RATÓN_DE_PATAS|RATONES_DE_PATAS, 0.989184695
PARTE_CENTRAL|PARTES_CENTRALES, PARTE_CENTRAL, 0.987883749
BROWNORUM, LITHOBATES_BROWNORUM, 0.984676903
PRECIPITACIÓN_MEDIA_ANUAL, PRECIPITACIÓN_MEDIA, 0.981104292
BALUMILAL_TSELTAL, YAXCHI, 0.979938132
POBLACIÓN_URBANAPOBLACIÓN, URBANAPOBLACIÓN_URBANAPOBLACIÓN, 0.979285099
ZONAS_PROTECTORAS_FORESTALES, ZONAS_PROTECTORAS, 0.977832937
REGAR|RIEGAN|RIEGO, RIEGO|RIEGOS, 0.977805894
SCRIPTA, TRACHEMYS_SCRIPTA, 0.977713261
COLECTAR|COLECTA, COLECTA, 0.977427959
ESPECIE_INVASORA|ESPECIES_INVASORAS, ESPECIES_INVASORAS, 0.975565016
ACH_URRAC, CHAK_ACH, 0.974349941
EX_SITU, SITU, 0.973639088
AVICENNIA_GERMINANS, AVICENNIA, 0.971348781
HUMEDALES_DE_IMPORTANCIA_INTERNACIONAL, HUMEDALES_DE_IMPORTANCIA, 0.971296549
HÑÄ_HÑU, HÑÄ, 0.971228704
INCENDIO_FORESTAL|INCENDIOS_FORESTALES, INCENDIOS_FORESTALES, 0.970200806
TALAR|TALA, TALA|TALAS, 0.969993385
INSTITUTO_DE_HISTORIA_NATURAL, INSTITUTO_DE_HISTORIA, 0.965761687
CAMBIO_CLIMÁTICO|CAMBIOS_CLIMÁTICOS, CAMBIO_CLIMÁTICO, 0.964410752
BIÓXIDO_DE_CARBONO, BIÓXIDO, 0.963856514
ADN_MITOCONDRIAL, MITOCONDRIAL, 0.963729367
ESCUDOVOLCANES, TIPO_ESCUDOVOLCANES, 0.963021554
FRENTE_FRÍO|FRENTES_FRÍOS, FRENTES_FRÍOS, 0.962668078
PROCYON_LOTOR, PROCYON, 0.961102007
ARA_MACAO, MACAO, 0.959759064
WAKAX_TSELTAL, YOK_WAKAX, 0.959503855
RECURSOS_DE_USO_ARTESANAL, USO_ARTESANAL_IMPACTADOS, 0.958909191
PRECIPITACIÓN_TOTAL_ANUAL, PRECIPITACIÓN_TOTAL, 0.958628725
BALSAS, BALSA|BALSAS, 0.957679477
PALO|PALOS|PALILLOS, PALO, 0.957292722
GRAMINEUS, POOECETES_GRAMINEUS, 0.957245185
INSTITUTO_DE_EDUCACIÓN|INSTITUTOS_DE_EDUCACIÓN, INSTITUTOS_DE_EDUCACIÓN_SUPERIOR, 0.955912015
ARCULARIUS, POLYPORUS_ARCULARIUS, 0.955659836
MATORRAL_XERÓFILO, XERÓFILO, 0.95476792
RESERVA_FORESTAL|RESERVAS_FORESTALES, RESERVAS_FORESTALES, 0.954186342
MONITOREAR|MONITOREO, MONITOREO, 0.953836554
ASTROPHYTUM_MYRIOSTIGMA, MYRIOSTIGMA, 0.953628689
METOPIUM_BROWNEI, METOPIUM, 0.953581708

EDUCADORES_AMBIENTALES, EDUCADOR|EDUCADORES, 0.952397544
LITHOBATES_NEOVOLCANICUS, NEOVOLCANICUS, 0.951510092
REGIÓN_CENTRAL|REGIONES_CENTRAL, REGIÓN_CENTRAL, 0.950615912
FORMACIÓN_DE_RECURSOS_HUMANOS, FORMACIÓN_DE_RECURSOS, 0.949674853
SUBIMBRICATA, TILLANDSIA_ELONGATA, 0.949426045
SUELO_DE_TERRENOS_FORESTALES, SUELO_DE_TERRENOS, 0.949391548
TRUCHA|TRUCHAS, TRUCHA, 0.948604228
TERRAZAS, TERRAZA|TERRAZAS, 0.948458791
CANTHARELLUS_CIBARIUS, CIBARIUS, 0.947358379
PHANTERA_ONCA, PHANTERA, 0.94717443
POMARINUS, SALTEADOR, 0.947122882
PECARÍ_DE_LABIOS_BLANCOS, PECARÍ_DE_LABIOS, 0.946266968
HARPIA_HARPYJA, HARPYJA, 0.945096731
DOTACIÓN_DE_AGUA_POTABLE, DOTACIÓN_DE_AGUA, 0.945064402
AMMODRAMUS_SAVANNARUM, SAVANNARUM, 0.944776776
PTEROCEREUS_GAUMERI, PTEROCEREUS, 0.942188789
SITIO_DE_DISPOSICIÓN_FINAL|SITIOS_DE_DISPOSICIÓN_FINAL,
SITIO_DE_DISPOSICIÓN|SITIOS_DE_DISPOSICIÓN, 0.942062283
CUENCA_CERRADA|CUENCAS_CERRADAS, CUENCAS_CERRADAS, 0.941709346
INDIGO, LACTARIUS_INDIGO, 0.941386601
BOLSUDOS, RATONES_BOLSUDOS, 0.941092314
ESPECIES_DE_VIDA_SILVESTRE, ESPECIES_DE_VIDA, 0.940927945
MONOLINGÜE_INDÍGENA, POBLACIÓN_MONOLINGÜE, 0.940775088
PECES_ÓSEOS, ÓSEOS, 0.940418787
HYPOMYCES_LACTIFLUORUM, LACTIFLUORUM, 0.940389234
TROMPA|TROMPAS, TROMPA|TROMPO, 0.940142773
PAPITA_GÜERA, SOLANUM_SPP, 0.940093893
BOSQUE_MESÓFILO|BOSQUES_MESÓFILO, MESÓFILO, 0.940018862
LOPHOCAMBA_CIBRIANI, LOPHOCAMBA, 0.939732984
HÑU, HÑÄ_HÑU, 0.938849705
ESPECIES_DE_GANADO_MAYOR, GANADO_MAYOR_Y_MENOR, 0.93874173
ESPECIE_PRIORITARIA|ESPECIES_PRIORITARIAS, ESPECIES_PRIORITARIAS, 0.938667962
THRINAX_RADIATA, THRINAX, 0.93838672
FANEROGÁMICA, FLORA_FANEROGÁMICA, 0.938246297
COCHINITA_PIBIL, COCHINITA, 0.937482851
SURIANA_MARITIMA, SURIANA, 0.936545491
RANGIA_CUNEATA, RANGIA, 0.935991091
ÍNDICE_DE_RIESGO_NUTRICIONAL, ÍNDICE_DE_RIESGO, 0.935505339
PROGRAMA_DE_EDUCACIÓN_AMBIENTAL|PROGRAMAS_DE_EDUCACIÓN_AMBIENTAL,
PROGRAMA_DE_EDUCACIÓN|PROGRAMAS_DE_EDUCACIÓN, 0.935303907
PTEROGLOSSUS_TORQUATUS, PTEROGLOSSUS, 0.934845432
ORDENAMIENTO_TERRITORIAL|ORDENAMIENTOS_TERRITORIALES, ORDENAMIENTO_TERRITORIAL,
0.933714944
CONTENIDO_DE_MATERIA_ORGÁNICA, CONTENIDO_DE_MATERIA, 0.933038854
CAMPOSTOMA_ORNATUM, ORNATUM, 0.93162105
AVES_ICTIÓFAGAS, ICTIÓFAGAS, 0.931425131
LOTOR, PROCYON_LOTOR, 0.931372589
ANANAS_COMOSUS, COMOSUS, 0.927987115
CATEGORÍA_DE_ÁREA_NATURAL, CATEGORÍA_DE_ÁREA, 0.927269334
VEGETACIÓN_DE_DUNA_COSTERA, VEGETACIÓN_DE_DUNA, 0.926111983

HURÓN_DE_PATAS_NEGRAS|HURONES_DE_PATAS_NEGRAS, HURÓN_DE_PATAS|HURONES_DE_PATAS, 0.925502159
PECES_DE_AGUA_DULCE, PECES_DE_AGUA|PECES_DE_AGUAS, 0.925311881
ESTRADIOL, TESTOSTERONA, 0.924958131
MARIPOSA_MONARCA|MARIPOSAS_MONARCA, MONARCA|MONARCAS, 0.922090884
CICHLASOMA_GRAMMODES, GRAMMODES, 0.921732092
PROTECCIÓN_FORESTAL, ZONAS_DE_PROTECCIÓN_FORESTAL, 0.921584523
LABIOS_BLANCOS, LABIOS, 0.921020263
ROCAS_DE_ORIGEN_VOLCÁNICO, ROCAS_DE_ORIGEN, 0.920837539
PALO_TROMPA, TROMPA_DE_COCHI, 0.91976045
MORTANDAD_DE_AVES_ACUÁTICAS, MORTANDAD_DE_AVES, 0.914521527
HYLA_WALKERI, WALKERI, 0.914335888
ADMINISTRACIÓN_PÚBLICA_FEDERAL, PÚBLICA_FEDERAL, 0.913681094
CLADIUM_JAMAICENSE, JAMAICENSE, 0.913617911
HONGOS_MACROSCÓPICOS, MACROSCÓPICOS, 0.913149028
CONSTRUCCIÓN_DE_INFRAESTRUCTURA_HABITACIONAL, INFRAESTRUCTURA_HABITACIONAL_Y_COMERCIAL, 0.912354429
AGUA|AGUAS, AGUA, 0.912108216
COM_PERS, COM, 0.911716652
ENTEROLOBIUM_CYCLOCARPUM, ENTEROLOBIUM, 0.909749111
HÑU, HÑÄ, 0.908809389
MANTENIMIENTO_DE_ÁREAS_VERDES, MANTENIMIENTO_DE_ÁREAS, 0.907422417
CYCLOCARPUM, ENTEROLOBIUM_CYCLOCARPUM, 0.907155586
FLORIBUNDUM, GYMNOPODIUM_FLORIBUNDUM, 0.906602187
ARENA|ARENAS, ARENA, 0.906534144
PACHYCEREUS_PECTEN-ABORIGINUM, PECTEN-ABORIGINUM, 0.906024228
CAPSAICINA, CONTENIDO_DE_CAPSAICINA, 0.905720995
CYRTOPODIUM_MACROBULBON, CYRTOPODIUM, 0.903221914
ANP_DE_CARÁCTER_FEDERAL, ANP_DE_CARÁCTER, 0.902893627
ESCAMAS, ESCAMA|ESCAMAS, 0.90240384
CHUPASAVIA, SPHYRAPICUS, 0.902290397
SU-GRAN, Y-AHAU, 0.901949955
CARACTERÍSTICAS_GLÁNDULAS, GLÁNDULAS_REPULSIVAS, 0.901785348
DELICIOSUS, LACTARIUS_DELICIOSUS, 0.901402577
AVES_CANORAS, CANORAS, 0.900063649
CONDICIÓN_ADMINISTRATIVA_LEGAL, CONDICIÓN_ADMINISTRATIVA, 0.900026996
COCHI_OREJA, OREJA_DE_COCHI, 0.899908565
DESARROLLO_LOCAL_SUSTENTABLE, ESTRATEGIA_DE_DESARROLLO_LOCAL|ESTRATEGIAS_DE_DESARROLLO_LOCAL, 0.899034517
CHANGO_OREJA, PALO_TROMPA, 0.897718038
CEPEDIANUM, PEZ_CEBRA, 0.897624612
ALISIOS, VIENTOS_ALISIOS, 0.897569416
LORO|LORA, LORO, 0.897453225
MELIOSMA_DENTATA, MELIOSMA, 0.897447443
COCCOTHRINAX_READII, COCCOTHRINAX, 0.897438494
ESTUDIO_DE_DIVERSIDAD_GENÉTICA|ESTUDIOS_DE_DIVERSIDAD_GENÉTICA, ESTUDIO_DE_DIVERSIDAD|ESTUDIOS_DE_DIVERSIDAD, 0.897260698
BAHAMENSIS, MIMOSA_BAHAMENSIS, 0.8970123
RECURSO_FORESTAL|RECURSOS_FORESTALES, RECURSOS_FORESTALES, 0.895748359
PRODUCTO_FORESTAL|PRODUCTOS_FORESTALES, PRODUCTOS_FORESTALES, 0.895223258
GUARDALAGUA, VARIADOR, 0.895108533

HUILOTA, PALOMA_HUILOTA, 0.895074227
TANGENTES_TRIFÁSICOS, TANGENTES, 0.895069949
OREJAS_ROJAS, TORTUGA_DE_OREJAS_ROJAS|TORTUGAS_DE_OREJAS_ROJAS, 0.894735998
ALOUATTA_PIGRA, ALOUATTA, 0.894276289
CAMINO_DE_TERRACERÍA|CAMINOS_DE_TERRACERÍA, TERRACERÍA, 0.894025014
LEMNA_TRISULCA, TRISULCA, 0.893657087
CAMPECHIANUM, HAEMATOXYLUM_CAMPECHIANUM, 0.893589526
CHAGAS, ENFERMEDAD_DE_CHAGAS, 0.893464292
AULLADOR, MONO_AULLADOR, 0.892880304

Variantes denominativas de la Biodiversidad (Termext-PMI)

PROGRAMA_DE_PROTECCIÓN, PROGRAMA_DE_PROTECCIÓN, 1
ARCO_IRIS, TRUCHA_ARCO_IRIS, 0.986522501
ANAGNOSTIDIS, KOMÁREK, 0.985283176
CONSERVACIÓN_EX, CONSERVACIÓN_EX_SITU, 0.97833553
AGUA_PEQUEÑOS, CUERPOS_DE_AGUA_PEQUEÑOS, 0.9710916
AESCULIFOLIA, CEIBA_AESCULIFOLIA, 0.970599565
EDUCADORES, EDUCADORES_AMBIENTALES, 0.967909271
NATURALES_PROTEGIDAS, ÁREAS_NATURALES_PROTEGIDAS, 0.96581132
WIMMERIA, WIMMERIA_PERSICIFOLIA, 0.963351192
INSTITUTO_NACIONAL_DE_ESTADÍSTICA, NACIONAL_DE_ESTADÍSTICA, 0.962682097
LAGUNCULARIA, LAGUNCULARIA_RACEMOSA, 0.96191689
REGIONES_PRIORITARIAS_TERRESTRES, REGIONES_PRIORITARIAS_TERRESTRES_DE_CONSERVACIÓN, 0.960097327
PRECIPITACIÓN_TOTAL, PRECIPITACIÓN_TOTAL_ANUAL, 0.958628725
CHIPE_MANGLERO, MANGLERO, 0.957088867
ANIM, TERRÍCOLA_WUTZ_ANIM, 0.953570997
COORDINACIÓN_GENERAL, COORDINACIÓN_GENERAL_DE_ÁREAS, 0.952916419
HOEK, HOEK_ET, 0.952269402
BUCHLOE, BUCHLOE_DACTYLOIDES, 0.9513178
CASTRO-HERNÁNDEZ, CASTRO-HERNÁNDEZ_ET, 0.951100782
CENTROCESTUS_FORMOSANUS, METACERCARIAS_DE_CENTROCESTUS_FORMOSANUS, 0.948345884
BRITTON_ET_ROSE, ET_ROSE, 0.947424344
EQUINODERMOS_DE_AGUAS, EQUINODERMOS_DE_AGUAS_SOMERAS, 0.947217827
MATERIA_PRIMA, PRIMA, 0.94592026
ESTABLECIMIENTO_DE_ÁREAS_NATURALES, ESTABLECIMIENTO_DE_ÁREAS_NATURALES_PROTEGIDAS, 0.945756286
CREACIÓN_DE_ÁREAS_NATURALES, CREACIÓN_DE_ÁREAS_NATURALES_PROTEGIDAS, 0.944305377
NAL_TEL, TEL, 0.943270985
LIPPIA, LIPPIA_GRAVEOLENS, 0.942720215
PATAS_BLANCAS, RATONES_DE_PATAS_BLANCAS, 0.941494776
CORALILLO_MICRURUS_DISTANS, MICRURUS_DISTANS, 0.941399956
PRECIPITACIÓN_TOTAL_ANUAL, TOTAL_ANUAL, 0.940465727
LACANTUNIA, LACANTUNIA_ENIGMATICA, 0.940453916
HYPOMYCES_LACTIFLUORUM, LACTIFLUORUM, 0.940389234
HONGOS_MICORRIZÓGENOS, HONGOS_MICORRIZÓGENOS_ARBUSCULARES, 0.940333522
GOMPHUS, GOMPHUS_FLOCCOSUS, 0.939430605
NANNOTRIGONA, NANNOTRIGONA_PERILAMPOIDES, 0.93859562
DOMINGO_KESTÉ, SANTO_DOMINGO_KESTÉ, 0.937460151
CUARTAS, CUARTAS_PARTES, 0.936957728

INVENTARIO_NACIONAL, INVENTARIO_NACIONAL_FORESTAL, 0.936916927
SURIANA, SURIANA_MARITIMA, 0.936545491
BASASEACHI, CASCADA_DE_BASASEACHI, 0.936536863
PURPURA, PURPURA_PATULA, 0.936505495
ANTHRAX, EXOPROSOPA, 0.934282864
ROMO, ROMO_ET, 0.933832038
NORMA_OFICIAL, NORMA_OFICIAL_MEXICANA, 0.932665826
CANIS_LATRANS, LATRANS, 0.93224781
CAMPOSTOMA_ORNATUM, ORNATUM, 0.93162105
GUAZUMA, GUAZUMA_ULMIFOLIA, 0.93142082
LARREA, LARREA_TRIDENTATA, 0.931120348
COTIJA, DEPRESIÓN_DE_COTIJA, 0.929750046
ARCA_ZEBRA, ZEBRA, 0.929521598
MAYRAND, PAQUIN, 0.929265219
CATEGORÍA_DE_ÁREA, CATEGORÍA_DE_ÁREA_NATURAL, 0.927269334
VEGETACIÓN_DE_DUNA, VEGETACIÓN_DE_DUNA_COSTERA, 0.926111983
AMPHIODIA_GUILLERMOSOBERONI, GUILLERMOSOBERONI, 0.925859973
TIPO_CRÓNICO, TIPO_CRÓNICO_DEGENERATIVO, 0.922770409
ALACRANES_SINANTRÓPICOS, FAUNA_DE_ALACRANES_SINANTRÓPICOS, 0.922383028
SWALLOW, SWALLOW_ET, 0.922126225
SAN_CRISTÓBAL, UNIDAD_SAN_CRISTÓBAL, 0.922114503
MANCHA, MANCHA_URBANA, 0.922038035
VAINORO, VAINORO_CELTIS, 0.921177105
MONTE_NEGRO, SIERRA_MONTE_NEGRO, 0.921171541
EXOPROSOPA, LIGYRA, 0.920361832
ISCHADIUM_RECURVUM, RECURVUM, 0.919054066
CAYUELA, CAYUELA_ET, 0.918966906
JAPA, SABAL_JAPA, 0.918290478
PIPILO_FUSCUS, VIEJITA, 0.917907017
BRACHIDONTES_EXUSTUS, EXUSTUS, 0.917350108
CENTRAL_YUCATECA, PLANICIE_CENTRAL, 0.917294647
CONOCIMIENTO_ECOLÓGICO_TRADICIONAL, CONOCIMIENTO_ECOLÓGICO_TRADICIONAL_MAYA, 0.917188916
ESTUDIO_TÉCNICO_JUSTIFICATIVO, TÉCNICO_JUSTIFICATIVO, 0.914509933
AGREGADO, VALOR_AGREGADO, 0.914101358
TORNO, TORNO_TRADICIONAL, 0.913318175
COM, COM_PERS, 0.911716652
ESTUDIOS_DE_DIVERSIDAD, ESTUDIOS_DE_DIVERSIDAD_GENÉTICA, 0.910896908
CUCURBITAS, CUCURBITAS_SILVESTRES, 0.910497736
AMPHIODIA, AMPHIODIA_GUILLERMOSOBERONI, 0.910332909
GÉNEROS_DE_AVES, GÉNEROS_DE_MAMÍFEROS, 0.9100395
BOSQUE_TROPICAL_CADUCIFOLIO, CADUCIFOLIO, 0.909901254
MEDIANA_SUBCADUCIFOLIA, SELVA_MEDIANA_SUBCADUCIFOLIA, 0.908961592
PRIORITARIAS_DE_MÉXICO, REGIONES_PRIORITARIAS_DE_MÉXICO, 0.908960352
MICROCHLOA, MICROCHLOA_KUNTHII, 0.908770982
BATRACHOCHYTRIUM_DENDROBATIDIS, HONGO_BATRACHOCHYTRIUM_DENDROBATIDIS, 0.907908401
INSTITUTO_NACIONAL, INSTITUTO_NACIONAL_DE_ESTADÍSTICA, 0.907615975
NIVEL_ESTADO, NIVEL_PAÍS, 0.907295816
ALATUS, ISOGNOMON_ALATUS, 0.906739408
MUCH_LACANDÓN, OUDEMANSIELLA, 0.906394287
STORCH, WELSCH, 0.905948238

GONZÁLEZ-SOLÍS, VIDAL-MARTÍNEZ_ET, 0.90591349
CAPSAICINA, CONTENIDO_DE_CAPSAICINA, 0.905720995
MUNICIPIOS_DE_DZÁN, PLANTAS_DE_HUANO, 0.90565932
FORESTIERA_PHILLYREOIDES, PHILLYREOIDES, 0.903956248
BOSQUES_MESÓFILOS_DE_MONTAÑA, MESÓFILOS_DE_MONTAÑA, 0.903952548
CYRTOPODIUM, CYRTOPODIUM_MACROBULBON, 0.903221914
AGUA_EPICONTINENTALES, CUERPOS_DE_AGUA_EPICONTINENTALES, 0.90297653
NORMA_MEXICANA, NORMA_MEXICANA_DE_TOXICIDAD, 0.902565487
GUSANO_PELUDO, GUSANO_PELUDO_ESTIGMENE, 0.90152022
AVES_CANORAS, CANORAS, 0.900063649
CONDICIÓN_ADMINISTRATIVA, CONDICIÓN_ADMINISTRATIVA_LEGAL, 0.900026996
TREMBLAY_ET, WIDMER, 0.899993093
MORALES-PÉREZ, MORALES-PÉREZ_ET, 0.899315333
PANTHERA, PANTHERA_ONCA, 0.898944692
MARÍA, PALO_MARÍA, 0.898934163
PRECIPITACIÓN_TOTAL, TOTAL_ANUAL, 0.898707543
PROBLEMA_DE_SALUD, PROBLEMA_DE_SALUD_PÚBLICA, 0.898555554
TACANÁ, VOLCÁN_TACANÁ, 0.898213382
BRODO, BRODO_ET, 0.897729861
OREJA_DE_CHANGO_OREJA, OREJA_DE_PALO_TROMPA, 0.897718038
ALISIOS, VIENTOS_ALISIOS, 0.897569416
MELIOSMA, MELIOSMA_DENTATA, 0.897447443
SOCIETY, SOCIETY_FOR, 0.896962618
MATRIZ_DE_VEGETACIÓN, MATRIZ_DE_VEGETACIÓN_INUNDABLE, 0.896576672
LYCURUS, LYCURUS_PHLEOIDES, 0.896123324
TRANSMEXICANA, VOLCÁNICA_TRANSMEXICANA, 0.896082987
ENFERMEDADES_CARDIACAS, ENFERMEDADES_HEPÁTICAS, 0.895799675
PEPINOS, PEPINOS_DE_MAR, 0.895654192
PAG, PDF, 0.895314168
GUARDALAGUA, VARIADOR, 0.895108533
PIES_AMBULACRALES_SUCTORES, PIES_SUCTORES, 0.894867826
MATORRAL_DE_DUNA, MATORRAL_DE_DUNA_COSTERA, 0.894621495
SISTEMA_NACIONAL, SISTEMA_NACIONAL_DE_UNIDADES, 0.894212997
CHAGAS, ENFERMEDAD_DE_CHAGAS, 0.893464292
ISOGNOMON, ISOGNOMON_ALATUS, 0.893177785
BAJA_SUBPERENNIFOLIA, SELVA_BAJA_SUBPERENNIFOLIA, 0.892642749
MENDOZAFRANCO_ET, VIDAL-MARTÍNEZ_ET, 0.891412995
VALORES_MENORES, VALORES_MENORES_DE_DIVERSIDAD, 0.891101541
DEPÓSITOS_LACUSTRES_Y_FLUVIALES_POLIGENÉTICOS, VULCANÓGENO-SEDIMENTARIO, 0.889970126
DENDROPANAX, DENDROPANAX_ARBOREUS, 0.889877633
LEPIDANTHRAX, TOXOPHORA, 0.888275432
LATHROP, ZUILL, 0.888255243
ARBUSCULARES, HONGOS_MICORRIZÓGENOS_ARBUSCULARES, 0.887989752
ASCOMICETOS, BASIDIOMICETOS, 0.887869956
CYCADOPHYTINA, PINOPHYTINA, 0.887758847
HEATH, LONG, 0.886801365
INTRODUCCIÓN_DE_ESPECIES, INTRODUCCIÓN_DE_ESPECIES_EXÓTICAS, 0.886685669
RAMPHASTOS, RAMPHASTOS_SULFURATUS, 0.886258481
PDF, PDF_PAG, 0.885813872
TRYPANOSOMA, TRYPANOSOMA_CRUZI, 0.885748829
AAZPA, MISIÓN_DE_RECREACIÓN, 0.885633595

EN_QUINTANA_ROO, QUINTANA, 0.885218835
QUINTANA, QUINTANA_ROO, 0.885218835
ESCURRIMIENTO_MEDIO, ESCURRIMIENTO_MEDIO_ANUAL, 0.8839566
EREMOPHILA, EREMOPHILA_ALPESTRIS, 0.882341218
KAYOCH_LACANDÓN, USO_LIGNÍCOLA_COMPAÑERO, 0.882018327
OREJA_DE_CHANGO_OREJA, OREJA_DE_COCHI_OREJA, 0.881978688
FRUTO_COMPLETO, TEJIDO_PLACENTARIO, 0.881974691
CHANGO, OREJA_DE_CHANGO_OREJA, 0.880436182
DESCORTEZADORES, INSECTOS_DESCORTEZADORES, 0.879991077
COMUNIDADES_VEGETALES_LLAMADAS_PETENES, VEGETALES_ENDÉMICAS_LLAMADAS_PETENES, 0.879631065
COPALES, PAPELILLOS, 0.879425857
CHILE_DE_MONTE, CHILE_SILVESTRE, 0.879399873
ANGIOSPERMOFITOS, MAGNOLIOPHYTA, 0.879311074
NEACREOTHRICHUS, TOXOPHORA, 0.879008369

Contextos de evaluación para los términos de Ingeniería de Software.

Gestión-Administración

objetivo. Conclusiones El uso de Kanban, sumado con conceptos de gestión ágil, nos ha permitido aumentar la calidad de nuestros d
alcanzado el nivel 2 de capacidad (por ejemplo, en el proceso de gestión de proyectos), alcanzar el nivel 3 de capacidad es más s
bre 2011) Sección: Código Innovare La adopción de metodologías de gestión de procesos de negocio ha aumentado considerablemente en
a ampliación del modelo CMMI El modelo del SEI específico para la gestión de servicios de TI, CMMI for Services (CMMI-SVC) [13] su
específicos, para comprender el fin y llevar a cabo una correcta gestión de requerimientos basada en el entendimiento de las acti

tware. El proceso de software forma la base para el control de la administración de proyectos de software, y establece el contexto
y que aprovechar ahora el enfoque de diversos grupos como PMI con administración ágil, RUP con su versión para desarrollo ágil y d
cionales, estructurales, performance, automatización, así como la administración de las pruebas, etc. El desarrollo del Testing ha
los proyectos, están la inadecuada administración de proyectos y administración de requerimientos. El objetivo central del análi
iendo la evaluación de los datos. En este contexto, el sistema de administración de datos de sensores facilitará las tareas de ges

Computadora-Ordenador

an algunas de las métricas de diseño más comunes para software de computadora. Cada una puede proporcionarle comprensión mejorada
ivas [sin apoyo cuantitativo]”. Horst Zuse bles de un programa de computadora. Al usar conceptos similares a los propuestos en IEE
ropósito principal es ejercitar por completo el sistema basado en computadora. Aunque cada prueba tenga un propósito diferente, to
os sentimos hace tantos años, al tener la oportunidad de usar una computadora por primera vez, al darnos cuenta de que en realidad
alla o pasa de forma determinista. Las pruebas que dependen si la computadora está configurada correctamente, o cualquier otro tip

web como un método de evaluación de la seguridad de un sistema de ordenadores o una red mediante la simulación de un ataque. Una p
eben abordar el trabajo de pruebas es también variado4: o Grandes ordenadores: 20,59% o Estaciones de trabajo o PCs: 47,06% o Ento
la posibilidad de que varias personas trabajen juntas utilizando ordenadores y tecnología informática, facilitando el trabajo en

eniería y la informática aunque, siendo niñas, disfrutaban usando ordenadores incluso de forma avanzada. La falta de información s
“frikis” asociales y que consumen un exceso de horas frente a un ordenador. En este sentido, no se trata de atraer nuevas vocacio

Requerimientos-Requisitos

plir, pueden ser divididos en las siguientes cuatro categorías: • Requerimientos Funcionales: Referido a las reglas de negocio que
ba funcional Esta prueba verifica el correcto cumplimiento de los requerimientos del usuario. Prueba de operación El éxito de esta
mica, incluso si no se anticipan” [Bar06]. Por su naturaleza, los requerimientos emergentes conducen al cambio. ¿Cómo se controla
aración del producto tratando de dejar una liga clara de cómo los requerimientos iniciales concluyeron con los resultados finales.
den dividir en dos categorías: Requerimientos Funcionales (RFs) y Requerimientos No Funcionales (RNFs). De acuerdo a Wiegers [1],

s requisitos que capturan la mayoría de las facetas de cómo estos requisitos funcionales deben llevarse a cabo se les conoce como
sión y análisis de las metodologías para la educción temprana de requisitos de software que utilizan directamente objetivos y pro
tención de requisitos X X ENG.1.1 Proceso de análisis y diseño de requisitos de sistema X ENG.1.2 Proceso de análisis de requisito
so particular. La intención es mostrar que es posible suministrar requisitos de comportamiento con bajo coste de especificación, y
ntre el comportamiento mostrado por el programa bajo prueba y los requisitos de comportamiento suministrados. Son por tanto condic

Archivos-Ficheros

ivo enviado por actor CONTENIDO OBJETO Es el material, documento, archivo, artefactos de software que un actor envía a otro actor.
s correspondiente a la entrada de la aplicación a desarrollar. El archivo de salida será de igual formato que los que ingresan, pa
cada iteración) se utilizarán tres posibilidades distintas: • Un archivo de configuración en el cuál se especificarán ciertos par
iales clave para la pila. Rackspace contribuyó con su plataforma "Archivos en la Nube" (código) para alimentar la parte de almacen
le el hito y se ganan los puntos correspondientes. Al poner en un archivo una lista con los componentes e hitos logrados (puntos g

nivel de servidor y que tratar como CIs los distintos sistemas de ficheros, crea una complejidad innecesaria para la estructura de
ema, Fecha de Desarrollo, Descripción del Sistema, enumeración de ficheros (fuentes y objeto), tablas y de datos. Ingeniería Inver
un servidor y la configuración de cada programa, los sistemas de ficheros que hay, los discos que los soportan y mucha más inform
XSLT, SAX o DOM. Algunos de ellos son adecuados para operar sobre ficheros de datos XML, otros están adaptados a la consulta de re
ser suministrados al oráculo en forma de script XQuery, a modo de fichero de configuración. 1) Implementación relajada del program

Pruebas-Ensayos

sarrollo que no realice de manera más o menos exhaustiva y formal pruebas de software. Las pruebas de software1 se definen como "u
tes conceptos: haciendo mención a técnicas para realizar pruebas (pruebas de caja blanca y de caja negra), dando nombre a diferent
n cuenta decisiones de diseño mientras que cuando se escribe cada prueba unitaria, hay que tener en cuenta lo que hace el código i
nal y el Laboratorio de Ensayos de Plataformas, donde se realizan pruebas de desempeño y se asiste a la industria para resolver pr
de los casos de prueba formales del sistema. Por último, realizar pruebas de integración, hace referencia a ejecutar las pruebas d

funcionales de un producto de software utilizada en el Centro de Ensayos de Software. Las principales características de la estra
ales independientes, un 33,6% corresponde a consultoría, un 30% a ensayos de plataformas y un 3,4% a capacitación. • Un 90% son pr
EN-50128 propone principalmente las siguientes: ensayos formales, ensayos probabilísticos, análisis estático y análisis dinámico.
específicos o liberaciones de producción. Ensayos de auditoría. Un ensayo de auditoría establece información adicional acerca de cu
allazgos en forma escrita. 24.6 EL PRINCIPIO W5HH En un excelente ensayo acerca del proceso de software y los proyectos, Barry Boe

Plantillas-Formularios

sicos de OpenStack en un sistema de plantillas de un archivo. Las plantillas permiten la creación de la mayoría de los tipos de re
y construido en los años de experiencia de IBM, los patrones son plantillas que consisten en software y recursos de la máquina vi
algunas estrategias de prueba de software. Todas proporcionan una plantilla para la prueba y tienen las siguientes características
el paquete java.util.regex para, mediante su expansión, obtener "plantillas de prueba". Una plantilla de prueba consiste en una s
ecución de la revisión sistemática en [10] se ha desarrollado una plantilla del protocolo para la revisión. El objetivo es que sir

la Dirección Nacional de Protección de Datos (DNPDP) y que en el formulario de registro ha denunciado como destino de transferenc
el desarrollo. Los autores explican cómo los scripts, templates y formularios del PSP son ajustados, describiendo en particular lo
si es la primera vez, deberá registrarse previamente, llenando un formulario que corresponde a su perfil. Esta información permiti
s a mejorar. Definir la propuesta de los nuevos procesos. Diseñar formularios para realizar el proceso. Elaborar manuales de uso d
s y la semántica. Una de las funciones será dar validez vínculos, formularios, HTML dinámico, contenido de streaming, etc. Del mis