UNIVERSIDAD CATÓLICA DE SANTA MARIA

CIENCIAS E INGENIERÍAS FISICAS Y FORMALES

INGENIERÍA DE SISTEMAS

MODELO DE TEXT CLUSTERING PARA EL DESCUBRIMIENTO DE PATRONES EN TEXTOS TÉCNICOS CORTOS NO ESTRUCTURADOS

Tesis presentada por el Bachiller:

Llamosas Lazo Brayan Gilmer

Para optar por el Título Profesional de:

Ingeniero de Sistemas

AREQUIPA - PERÚ

2014

PRESENTACIÓN

Sra. Directora del Programa Profesional de Ingeniería de Sistemas.

Sres. Miembros del Jurado Examinador de Tesis.

De conformidad con la disposición del reglamento de Grados y Títulos del Programa Profesional de Ingeniería de Sistemas, remito a vuestra consideración el estudio de investigación titulado "Modelo de Text Clustering para el descubrimiento de patrones en textos técnicos cortos no estructurados", el mismo que al ser aprobado me permitiría optar por el Título Profesional de Ingeniero de Sistemas.

Arequipa, Noviembre de 2014

Brayan Gilmer Llamosas Lazo

AGRADECIMIENTOS

Un especial agradecimiento para el Mg. Christian López del Álamo, ya que sin su desinteresado apoyo y recomendaciones no habría podido llegar a los resultados obtenidos en este trabajo.

Así también un gran agradecimiento al Dr. Ángel Chirinos por su apoyo incondicional y su gran contribución para las pruebas de esta investigación.

A mis padres, ya que
sin su apoyo constante
no me hubiera sido
posible realizar mis
estudios superiores, ni
elaborar esta Tesis.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE FÓRMULAS

ÍNDICE DE ALGORITMOS Y TABLAS

ÍNDICE DE ALGORITMOS

ÍNDICE DE TABLAS

RESUMEN

En la actualidad, la generación y estructura de la información ha variado bastante respecto a cómo era hace una década; ahora podemos ver en la internet una información más variada, concisa, de varios temas, a varios niveles especialización y, en muchos casos, sin seguir ningún tipo de estructura, ya que la información tiende a ser clara y concisa para la mayor cantidad de gente posible, dejando de lado algunos formalismos (textos de escritura libre).

En la siguiente investigación se busca ahondar en las técnicas existentes actualmente en el estado del arte del Procesamiento de Lenguaje Natural, para el tratamiento de textos con las características explicadas anteriormente. El objetivo es adaptar y mejorar la aplicación de estas técnicas para este tipo de textos; así también se pretende que el modelo de tratamiento resultante de esta investigación sea aplicable para cualquier tipo de textos, sin importar dominio del conocimiento de estos.

ABSTRACT

At present, the generation and structure of information has quite varied regarding how was a decade ago; now we can see on the internet an information more varied, concise, on various topics, at several specialization levels, and in many cases, without following any type of structure, because information tends to be clear and concise for as many people as possible, leaving aside some formalisms (free writing texts).

The following investigation seeks to delve into the techniques currently available in the state of the art of the Natural Language Processing for the processing of texts with the characteristics discussed above. The aim is to adapt and improve the applying of these techniques to such texts; and it is also intended that the treatment model resulting from this research be applicable to any type of text, regardless of knowledge domain of these.

CAPÍTULO 1

PLANTEAMIENTO TEÓRICO

1.1. TÍTULO DESCRIPTIVO DEL PROYECTO:

Modelo de Text Clustering para el descubrimiento de patrones en textos técnicos cortos no estructurados.

1.2. DESCRIPCIÓN DEL PROBLEMA

1.2.1. Definición del Problema:

Con el paso de los años, la información tiende a almacenarse cada vez más en medios digitales y, de la mano con este cambio, la representación de la información es más propensa a ser reducida o resumida. Si bien el Procesamiento de Lenguaje Natural es capaz de lidiar con textos no estructurados, el análisis de textos de corta longitud se dificulta ya que, por su misma longitud limitada, la cantidad de características que se pueden extraer de ellos también se reduce. Un buen ejemplo de este caso es la explotación de información presente en las redes sociales, en los cuales la longitud de cada texto es bastante corta.

Otro problema presente en la actualidad es la cantidad de información que existe sobre temas específicos, ya sea en un nivel muy especializado o generalizado; el problema con este tipo de textos es que las palabras usadas en ellos son de carácter técnico y, por lo general, estas palabras representan las características más importantes para el descubrimiento de sus patrones; la dificultad con estas palabras es que no suelen seguir las reglas gramaticales tradicionales, además que su relevancia sobre el texto está altamente relacionada con su contexto. Un buen ejemplo de este caso combinado con el expuesto en el párrafo anterior son los diagnósticos que redactan los médicos, estos son de una longitud bastante corta, presentan un lenguaje altamente especializado y no presentan ningún formato formal, de poder analizar esta información automáticamente, los médicos podrían tener a su disposición información útil de manera más rápida.

1.2.2. Área Científica a la que corresponde el Problema:

Área: Inteligencia Artificial.

Línea: Procesamiento de Lenguaje Natural.

1.2.3. Tipo y Nivel de Investigación:

Tipo: Es de tipo aplicada, ya que en esta investigación se revisará el estado del arte para encontrar la mejor aplicación del conocimiento encontrado sobre textos con características específicas.

Nivel: Es de nivel experimental pues, después de analizar el estado del arte, se aplicarán los procedimientos más adecuados para resolver el problema específico, ensayando varias instancias de las variables dependientes de este (técnicas, algoritmos, fórmulas, umbrales, etc.); eliminando, limitando o recomendando la aplicación de cada uno de estos procedimientos.

1.2.4. Objetivos Generales y Específicos:

1.2.4.1. Objetivo General:

Proponer un modelo de Text Clustering que permita el descubrimiento de patrones característicos en textos técnicos cortos no estructurados sin importar el ámbito al cual estos estén referidos.

1.2.4.2. Objetivos Específicos:

- El modelo propuesto se basará en el estado del arte actual de todas las áreas afines al problema dentro del Procesamiento de Lenguaje Natural, tomando los aspectos más relevantes de cada uno.
- El modelo propuesto se ajustará especialmente para textos técnicos cortos no estructurados, es decir que los procesos tomados y las técnicas recomendadas serán las que mejores resultados den en la experimentación.
- El modelo propuesto no debe estar dirigido a textos de un ámbito en específico, este modelo debería ser aplicable a textos de cualquier ámbito con las características antes descritas.

- El modelo propuesto debe estar en base al uso de técnicas de Text Clustering no supervisadas.
- Optimizar una técnica de Text Clustering con la cual aumente la tasa de aciertos del nuevo modelo.

1.2.5. Justificación:

Actualmente el estado del arte para el campo del Text Clustering es abundante; sin embargo, con el paso de los años el modo de representación de la información va cambiando, en los últimos años la información que se puede encontrar en línea ha crecido exponencialmente, esto debido a que cada vez más personas pueden publicar la información que les parece relevante desde su punto de vista, ya sea una opinión, comentario, reseña, resumen, etc. Esto genera una gran cantidad de información no estructurada que se representa de manera diferente a la que se podía ver hace un par de décadas.

El modelo de Text Clustering que se desarrollará en esta investigación está orientado a lidiar con toda clase de textos que presenten principalmente las siguientes características:

- Escritura libre.
- Textos cortos.
- Escritura muy especializada.

Las dos primeras son altamente representativas para la información que se está generando últimamente; los comentarios en redes sociales, blogs, noticias, videos, etc.; toda información a ser publicada por una persona con la única intención de comunicar algo tiende a ser corta y sin controles gramaticales, una información cuya representación es muy distinta a una publicación formal como una noticia o una investigación. El análisis de esta información puede ayudar a determinar tendencias en el mercado sobre nuevos productos, opiniones sobre imagen empresarial, política, y en general todo lo que sea de interés público.

Si bien el modelo de Text Clustering que se presentará en este trabajo no podrá realizar un análisis tan preciso para ese tipo de información, está orientado a trabajar con textos de similares características, por lo que sería un buen punto de partida para saber cómo abordar, representar y tratar ese tipo de información.

La tercera característica "escritura muy especializada" delimita el universo de textos analizado anteriormente, a solo los que tratan de temas técnicos o científicos; bajo esta acotación, el modelo de Text Clustering que se presentará podría ser utilizado para organizar la información manualmente escrita sobre resúmenes de cualquier área de estudio. Quizás, una de las áreas más relevantes para la aplicación del modelo de Text Clustering que se propondrá es la medicina; su aplicación genera información que el médico puede utilizar para su asistencia. Por ejemplo, en el caso de estudio se evidencia el problema que los diagnósticos históricos de pacientes de una clínica no estén clasificados por la enfermedad que estos presentaron; si esta información estuviese disponible, un médico que esté analizando a un paciente cuyo diagnóstico no puede concluir con seguridad, podría consultar el histórico de todos los diagnósticos con síntomas similares a los que sufre el paciente y el estudio que se hizo de ellos, con esta información el especialista tendría oportunidad de comparar los síntomas comunes que su paciente presenta contra los diagnósticos encontrados y esto le ayudará a dar su diagnóstico final.

1.2.6. Alcances y Limitaciones:

1.2.6.1. Alcances:

- Maximizar el porcentaje de aciertos del modelo propuesto.
- Se hará un profundo análisis del pre-procesamiento para textos con las características de los estudios médicos para el Text Clustering.
- Se hará un profundo análisis de la representación de textos con las características de los estudios médicos para el Text Clustering.

1.2.6.2. Limitaciones:

- No se analizarán los algoritmos de clustering de la literatura, solo se utilizarán los que presenten mejores resultados en los antecedentes investigativos.
- La solución planteada no tomará en cuenta la optimización de la complejidad algorítmica que esta represente.

CAPÍTULO 2
MARCO TEÓRICO

2.1. IR (INFORMATION RETRIEVAL – RECUPERACIÓN DE INFORMACIÓN):

2.1.1. Introducción:

Según [1] "Information retrieval (IR) es encontrar material (usualmente documentos) de naturaleza no estructurada (usualmente textos) que satisface una necesidad de información obtenida de una gran colección de información". En [2] se define como "Un campo que está en la intersección de las ciencias de la información y las ciencias de la computación, IR se ocupa de la indexación y recuperación de información de recursos de información heterogéneos y en su mayoría textuales".

El campo de IR se ocupa de la organización automática de la información textual. Su objeto es proporcionar información relevante al usuario con una consulta simple hecha de palabras claves (como se ve en los motores de búsqueda), sin embargo, a medida que los textos se llevan a un nivel más especializado, su vocabulario y modo de escritura se torna más técnico, por lo que se requiere de nuevas técnicas que sean capaces de brindar una buena organización a estos textos especializados [2].

La IR se divide en 2 etapas básicas:

- Indexing (Indexación).
- Retrieval (Recuperación).

2.1.2. Indexing (Indexación):

La indexación está dirigida a la asignación de metadata (datos sobre los datos) a cada texto dentro de la base de datos, esta metadata es una representación que se le da al texto que le permite ser comparado más fácilmente con los demás (más adelante se referirá a esta metadata como vector característico). La principal preocupación en la indexación es buscar la mejor representación posible que permita una buena clasificación de los

textos en la base de datos. Los tipos de representación para la indexación pueden ser [2]:

- **Indexing Terms:** Un término de indexación puede estar compuesto de una o varias palabras, este término tiene una gran relevancia dentro del conjunto de textos a evaluar.
- **Indexing Attributes:** En algunos casos se da que la totalidad de los textos a tratar tienen una buena organización predefinida y dentro de ella tiene ciertas secciones que contienen palabras clave.

En la mayoría de los casos no se cuenta con Indexing Attributes (puesto que se generación es un arduo trabajo), pero de contarse con ellos es preferible llevar a cabo la indexación con ellos, de lo contrario se debe generar los mejores Indexing Terms [2].

Según [3] si se hace una medida de la frecuencia con la que aparecen todas las palabras (del inglés) en cualquier tipo de textos, se puede observar que las palabras con mayor frecuencia suelen ser palabras con poca significancia por ejemplo preposiciones, sin embargo observó que las palabras que presentaban una frecuencia media, en la medición, tienen mucha más probabilidad de representar al documento.

El proceso de indexación se puede dividir en las siguientes etapas:

- Pre-procesamiento.
- Definición de Thesaurus (Espacio Característico)
- Generación de vectores característicos.

2.1.2.1. Pre-procesamiento:

Para obtener una mayor eficiencia y eficacia de las técnicas de IR, se realizan una serie de adecuaciones a los textos para representarlos de una forma que sea más fácil de procesar. En figura 2.1 se esquematiza todas las posibles adecuaciones que se puede hacer a los textos.

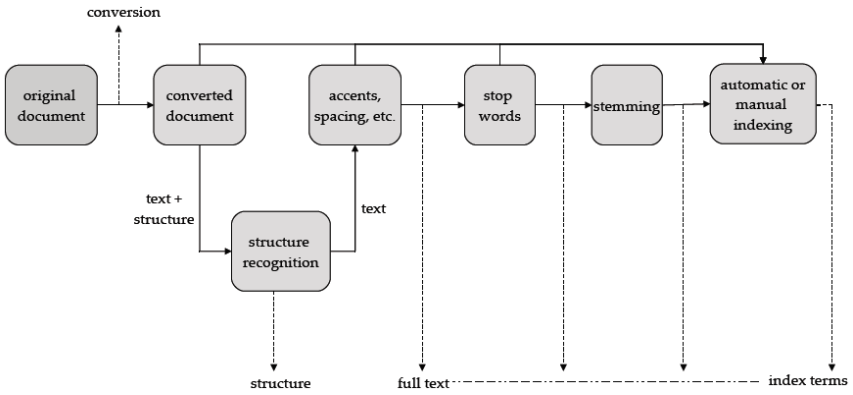

Figura 2.1: Posibles fases de pre-procesamiento del documento [4].

2.1.2.1.1. Convertir Formato:

Los textos pueden estar almacenados en formatos que no faciliten su manipulación, por ejemplo: PDF, Microsoft Word, HTML, XML. En este paso se debe llevar los textos al formato que más convenga para realizar las futuras operaciones sobre el texto [4].

2.1.2.1.2. Reconocimiento de la estructura:

Algunos documentos presentan una estructura preestablecida estándar, la cual puede ser útil para analizar su información para cada una de las partes de su estructura por separado, o solo para analizar la parte que es de nuestro interés. Este paso se encarga de encontrar la forma de identificar la estructura del texto [2] [4] [5].

2.1.2.1.3. Corrector de ortografía:

Este paso puede ser tanto manual como automático. Se trata de corregir los posibles errores al escribir ciertas palabras o en el modo de escritura que puedan alterar el verdadero sentido del texto. Este paso es muy recomendable de tomar cuando se trata con textos de escritura libre [2] [6].

2.1.2.1.4. Análisis Léxico:

Este paso trata de la identificación de palabras en el texto. Separarlo por espacios no es suficiente, se debe tener en cuenta números, guiones y signos de puntuación. La dificultad principal en este paso es determinar que si algún número, separación, combinación de números y caracteres son relevantes y podrían ser tomados como index terms, por ejemplo en un texto medico la palabra "B6", haciendo referencia a la vitamina, puede ser muy relevante [4].

2.1.2.1.5. Negative Dictionary:

También conocido como Stop Word List, este paso se trata de la eliminación de las palabras que suelen muy frecuentes en todos los textos, pero que no aportan mayor sentido semántico. Estas palabras se suelen utilizar para conectar ideas entre sí o como complementos, por ejemplo los artículos o las preposiciones. El Negative Dictonary no solo puede estar compuesto por estas palabras, sino también por palabras que se considere no aportarán valor para el dominio del conocimiento del que se traten los textos. La eliminación de los Stop Words no siempre es favorable dado que puede hacer perder sentido al texto, más aun en términos médicos, por ejemplo: Hepatitis **A**, Vitamina **A** [2] [4] [5] [6].

2.1.2.1.6. Stemming:

Este paso trata de reducir todas las palabras de los textos a su forma base, es decir su raíz o lexema. Para lograr esta tarea se necesita eliminar sus afijos (prefijos, sufijos e infijos). La aplicación o no de este paso depende mucho del dominio del conocimiento que se esté tratando y del lenguaje dado que los algoritmos de stemming no son 100% confiables, por ejemplo los términos técnicos suelen no seguir las reglas gramaticales tradicionales, por ejemplo un término medico puede reducirse incorrectamente: AIDS→AID [2] [4] [5] [6].

2.1.2.2. Thesaurus (Espacio Característico):

El thesaurus, también conocido como vocabulario controlado, es a lo que se hace referencia para realizar la indexación de los textos. Este vocabulario controlado está compuesto por los index terms que se definan para el conjunto de textos. En un modelo de IR Espacio Vectorial, el thesaurus vendría representando el espacio característico para los vectores característicos de cada texto [2] [4].

Algunos parámetros que se deben tomar en cuenta al momento de conformar los Indexing Terms son [2]:

- **Profundidad:** Grado de detalle o precisión del lenguaje utilizado. Esta precisión debe ir de acuerdo al conocimiento de los usuarios y a la información de la base de datos, no debe ser ni demasiado ni muy poco preciso para el usuario.
- **Amplitud:** Se refiere a la cantidad de términos que serán considerados dentro del diccionario. Generalmente se considera relevante un término principal dentro del texto. Mientras más exhaustiva sea más resultados tendrá, pero si es demasiado los resultados no serán tan buenos.

Los problemas más comunes, sobre todo en textos de carácter técnico como los médicos son [2]:

- **Sinónimos**
- **Polisemia:** Una palabra con varios significados.
- **Contenido:** No refleja el foco del artículo.
- **Contexto:** Las palabras en conjunto significan algo diferente.
- **Morfología:** Aparición de prefijos, sufijos, etc.
- **Granularidad:** Conceptos a distintos niveles jerárquicos.

Thesaurus más especializados toman en cuenta las relaciones entre términos (palabras o conjunto de palabras referidas a un concepto), los tipos de relación se dividen en [2]:

- **Jerárquicos:** Pueden ser más amplios o más específicos, ayuda a tener una vista de árbol de diccionario y una mejor búsqueda.
- **Sinónimos:** Utilizados en las búsquedas para una mejor performance.
- **Relacionados:** Términos que no son jerárquicos ni sinónimos, pero están relacionados de alguna forma. Pueden ser utilizados en la búsqueda.

Existe una variedad de thesaurus médicos públicos como el PubMed; uno de los más utilizados y completos es el MeSH (Medical Subject Heading), este nos provee de un buen diccionario organizado en forma de árbol de todos los términos médicos en general, implementa varios tipos de relaciones entre términos para una mejor referencia al momento de la búsqueda. La identificación de sus términos se hace mediante un código único relacionado con el tema del que trata. Los tipos de conceptos relacionados en el MeSH son los siguientes [2] [7]:

- **See related:** Cuando te lleva a otro título (que está estrechamente relacionado con el actual) que podría ser más apropiado para ese caso.
- **Consider also:** Cuando existe más de un significado para un término.
- **Sugerencias:** Annotation, relación de textos relacionados con el actual.

En las figuras 2.2 y 2.3 se pueden ver la estructura del diccionario MeSH:

Figura 2.2: Composición de árbol del diccionario MeSH [7].

MeSH Heading	Virus Diseases
Tree Number	C02
Annotation	GEN: prefer specifics; / drug ther: consider also ANTIVIRAL AGENTS
Scope Note	A general term for diseases produced by viruses.
Entry Term	Viral Diseases
See Also	Antiviral Agents
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH TM UR US VE VI
Entry Version	VIRUS DIS
Online Note	pre-explosion = VIRUS DISEASES (PX)
Date of Entry	19990101
Unique ID	D014777

Figura 2.3: Información de un término médico en el diccionario MeSH [7].

2.1.2.3. Generación de Vectores Característicos:

Teniendo ya el espacio característico definido (representado por el thesaurus) solo queda asignarle un peso a cada index term del espacio característico para formar el vector característico de cada texto. El peso que se le asigne a cada index term debe reflejar cuanto este contribuye con la semántica del documento, para esto existe una variedad de técnicas. A continuación mencionaremos las técnicas de pesado más utilizadas en IR [1] [2]:

2.1.2.3.1. Binario:

Toma en cuenta solo la presencia o ausencia del index term en el documento, si está presente le da un peso de 1 y si no lo está le da un peso de 0 [1] [2].

2.1.2.3.2. TF (Term Frequency):

Le asigna el peso al index term según su frecuencia en el documento, es decir que el peso del index term es igual al número de veces que este se presenta en el documento [1] [2]:

$$\mathrm{TF}(\mathrm{term}, \mathrm{document}) = \text{frequency of term in document} \quad (2.1)$$

2.1.2.3.3. TF – IDF (Term Frequency – Inverse Document Frequency):

Es la más utilizada en IR dado que esta técnica da un alto valor a las palabras que ocurren frecuentemente a en el documento, pero no en los demás. La técnica del TF-IDF tiene los siguientes componentes [1] [2] [5]:

- **TF (Term Frequency):** Análogo a la técnica explicada en el punto anterior.
- **IDF (Inverse Document Frequency):** Indica inversamente la frecuencia con que se presenta la palabra en todos los documentos de la base de datos, dará un menor valor mientras se encuentre en más documentos

Y se expresa bajo la siguiente fórmula (2.3) [1]:

$$\mathrm{IDF}(\mathrm{term}) = \log \frac{\text{number of documents in databse}}{\text{number of documents with term}} \quad (2.2)$$

$$\mathrm{WEIGHT}(\mathrm{term}, \mathrm{document}) = \mathrm{TF}(\mathrm{term}, \mathrm{document}) * \mathrm{IDF}(\mathrm{term}) \quad (2.3)$$

2.1.3. Retrieval (Recuperación por consulta):

Se refiere a como se obtendrá la respuesta con los documentos más relevantes para una consulta de usuario. En la sección anterior vimos el

proceso de indexación para un modelo de IR Espacio Vectorial por lo que para el momento de la consulta cada texto ya debe tener un vector característico asociado. Una vez que el usuario ingresa su consulta, se crea el vector característico de esta y se calcula la distancia a cada documento que se tenga en la base de datos, a menor distancia se considera que el documento es más relevante y es en base a esa distancia que se crea el ranking de documentos a ser recuperados para el usuario. Es importante elegir una distancia que se adecue al dominio del conocimiento de los textos y a las técnicas escogidas para crear el vector característico, de modo que optimice los resultados [2] [6] [8] [9].

2.2. TEXT CLUSTERING (TC):

2.2.1. Introducción:

En [10] se define como "El text clustering es uno de los problemas centrales en las áreas de text mining e information retrieval. La tarea de text clustering es agrupar textos similares".

El TC toma mucho del funcionamiento básico de IR debido a que ambos funcionan en base al contenido de los textos. El mayor parecido está en la indexación de los textos dado que también se basa en un espacio vectorial [5] [11].

A continuación se explicará las principales diferencias en el proceso de indexación de TC con respecto a IR que son: Pre-procesamiento y Reducción Dimensional. Posteriormente se dará una explicación de las técnicas de clustering.

2.2.2. Pre-procesamiento:

El pre-procesamiento en TC es bastante similar al de IR, sin embargo se considera un paso más debido a la complejidad extra que tiene el hacer clustering en comparación al retrieval de IR.

- **Podado:** Se refiere a la eliminación de las palabras muy poco frecuentas, a pesar que esta palabras tiene un buen poder de discriminación estas contribuirían a obtener clusters muy

pequeños (con pocos documentos). Bajo el mismo criterio, en algunas ocasiones, se eliminan las palabras con mucha frecuencia. El criterio de eliminación que se suele tomar es un umbral preestablecido basado en un porcentaje de aparición de la palabra en la totalidad de los textos (DF – Document Frequency), por ejemplo: se eliminan todas las palabras que aparezcan en menos del 5% de los textos y en más del 40% de los textos [11].

2.2.3. WordNet:

WordNet es un Thesaurus hecho de un lenguaje completo o varios. A diferencia de un Thesaurus, WordNet tiene información semántica que los Thesaurus no tienen, esta información está representada por agrupaciones que se dan a las palabras y relaciones entre estas como: sinónimos, antónimos, hiperónimos, hipónimos, etc. Las agrupaciones que provee son las siguientes [12] [13] [14]:

- **Categoría gramatical (Part of Speech - POS):** Sustantivo, verbo, adjetivo y adverbio. Cada palabra puede tener más de una categoría dependiendo de su contexto.
- **Dominio:** Se refiere al sentido que puede tomar la palabra dependiendo de su contexto, por lo que cada palabra puede tener más de un dominio. Algunos ejemplos de dominios son: Animales, plantas, medicina, economía, geometría, etc.

2.2.4. Reducción Dimensional:

Debido a que el vector característico de los documentos está hecho en base a los index terms seleccionados (thesaurus), este vector suele tener una cantidad de dimensiones bastante grande. A diferencia de IR, un vector característico con una cantidad de dimensiones bastante grande conlleva a una gran caída en la performance de los algoritmos de clustering, es por ello que se aplican técnicas para reducir la cantidad de dimensiones de los vectores sin que ellos pierdan el valor semántico que representa a los textos (en algunas ocasiones mejorándolo); esto además tiene el beneficio de evitar el overfitting o sobre especialización del documento [5] [10] [15] [16] [17].

Las técnicas de reducción dimensional en TC se pueden dividir en 2 [5] [10] [15] [16] [17]:

2.2.4.1. Term Selection:

También llamado Term Space Reduction (TSR). Puede aumentar el porcentaje de aciertos hasta en un 5%. Se filtran los index terms, manteniendo solamente los que reciban un mayor peso aplicándoles una función de medida de importancia. Los index terms seleccionados serán totalmente comprensibles para leer dado que son un subconjunto de los index terms originales [5] [10] [16] [17].

Las funciones de Term Selection se pueden dividir en 2 grupos:

- **No supervisadas:** Se pueden aplicar directamente.
- **Supervisadas:** Estas funciones depende del conocimiento de las categorías destino por lo que estas funciones se pueden aplicar para Text Categorization pero no para Text Clustering.

A continuación se describirán las funciones de medida de importancia no supervisadas más relevantes según [5] [10] [16] [17]:

2.2.4.1.1. Document Frequency (DF):

Una alternativa simple y efectiva, consiste en tomar como indicador la cantidad de documentos en la que se presenta el index term. Se puede reducir la dimensión en un factor de 10 sin perder efectividad, y en un factor de 100 casi sin perder efectividad, otros autores eliminan los términos que se repiten en 3 o 5 documentos o menos. Esto no contradice la ley de IR que dice que los términos medianamente y poco frecuentes son los más relevantes, dado que en el cuerpo la mayoría de los términos son los muy poco frecuentes, los cuales serán los eliminados, mientras que las términos medianamente y poco frecuentes serán preservados [5].

2.2.4.1.2. Term Strength (TS):

Este método estima la importancia del index term basándose en que tanto el index term se presenta en textos relacionados. La idea es tener un conjunto de entrenamiento de documentos del cual se analice cada par de documento con el objetivo de determinar si ese par está relacionado basándose en sus vectores característicos usando una distancia (por ejemplo coseno), para esto se debe tener un valor umbral como máximo de distancia de dos textos relacionados. Ya con todos los pares de documentos relacionados en el conjunto de entrenamiento, la importancia de un index term se mide por cuan informativa es para cada par de documentos relacionados, por ejemplo: para un documento "x" y un documento "y" relacionados, el Term Strength "s(t)" del index term "t" se define por la siguiente probabilidad [16] [17]:

$$s(t) = P(t \in y | t \in x) \tag{2.4}$$

Entonces el Term Strength de un index term para todos los documentos (calculado solo en base al conjunto de entrenamiento) se determina por la siguiente fórmula [16]:

$$s(t) = \frac{\text{Number of pairs in which } t \text{ occurs in both}}{\text{Number of pairs in which } t \text{ occurs in the first of the pair}} \tag{2.5}$$

Dado que se necesita de un conjunto de entrenamiento para poder determinar el Term Strength este método no es 100% no supervisado, por lo que se puede considerar como semi-supervisado.

2.2.4.1.3. Entropy-based Ranking (En):

En este método la importancia del index term se mide en base a la reducción de entropía después de eliminarse del espacio característico. La fórmula que determina la entropía "E(t)" de un

index term "t" para una "n" cantidad de documentos es la siguiente [10] [16]:

$$E(t) = -\sum_{i=1}^{n}\sum_{j=1}^{n}(S_{ij}\cdot\log(S_{ij}) + (1-S_{ij})\cdot\log(1-S_{ij})) \quad (2.6)$$

Donde S_{ij} es la similitud entre los documentos "i" y "j" después de que "t" es removido, S_{ij} se define por la siguiente fórmula [16]:

$$S_{ij} = 2^{-\frac{dist(i,j)}{\overline{dist}}} \quad (2.7)$$

Donde "dist(i, j)" es la distancia entre los documentos "i" y "j" después de que "t" es removido, y "$\overline{dist}$" es el promedio de diferencia de todos los documentos después de que "t" es removido. El mayor problema de este método es la complejidad computacional que representa [10] [16].

2.2.4.1.4. Term Contribution:

Este método toma en cuenta el peso que se le da al index term y mide su importancia en base a la contribución del index term para la medida de similitud entre los textos. Este método fue propuesto con el objetivo de solucionar un problema del DF, los index terms con un DF muy alto tienden a tener una distribución uniforme entre todos los clusters generados, por lo que estos no contribuyen con los algoritmos de clustering. Considerando que la similitud entre dos documentos se representa por la siguiente fórmula [10] [16]:

$$sim(d_i, d_j) = \sum_{t} f(t, d_i) \times f(t, d_j) \quad (2.8)$$

Donde "f(t, d_i)" es el peso DF-ITF del index term "t" en el documento d_i. La contribución de un index term en la similitud de todos los documentos está dada por la siguiente fórmula [10]:

$$TC(t) = \sum_{i,j \cap i \neq j} f(t,d_i) \times f(t,d_j) \tag{2.9}$$

Como el método anterior, un gran problema del Term Contribution es su complejidad computacional [10] [16].

2.2.4.2. Term Extraction:

Transforma el conjunto de index terms en otro conjunto de index terms artificiales bajo la premisa que los index terms solos por separado están expuestos a problemas de polisemia, homonimia, sinónimos, etc. Los index terms extraídos podrían no ser comprensibles para leer dado que estos resultan de la combinación de los index terms originales o de la aplicación de un método de transformación sobre todos ellos. Algunos de los métodos de Term Extraction más utilizados son: Term Clustering, Latent Semantic Index (LSI) y Principal Components Analysis (PCA). A continuación se describirá dos métodos de Term Clustering [5] [10] [11] [15] [16] [18]:

- **N-Gramas:** Son una secuencia de letras o palabras del texto. La idea es que los index terms tienen un mayor valor semántico cuando se les toma como un conjunto de letras o palabras en el orden que están en el texto. "N" representa el número de letras o palabras consecutivos que contendrá cada grama, entonces para obtener el nuevo espacio característico se buscan todas las combinatorias de N-gramas en todos los textos, posteriormente se seleccionan los N-gramas más frecuentes en todos los textos para representar el espacio característico, para esto se puede determinar un umbral para la frecuencia mínima o para escoger un número determinado de

N-Gramas que sean los más frecuentes [15] [19] [20] [21] [22].

- **Secuencias Frecuentes Maximales (SFM):** Sigue la misma idea de agrupar las palabras consecutivas de los N-Gramas, sin embargo en este método no hay un "N" predefinido. Este método tiene las siguientes reglas [23] [24]:
    - Se tiene un umbral predefinido que representa el número mínimo DF del index term en todos los documentos.
    - Las secuencias pueden ser de cualquier tamaño.
    - Ninguna secuencia debe contener a otra.

  En otras palabras, una secuencia de tamaño "n" no puede contener ninguna secuencia de tamaña "n - 1" o inferior; entonces, suponiendo que se va buscando secuencias de menor a mayor tamaño empezando por un tamaño igual a 2, cuando se busquen las secuencias de tamaño 3, se tendrán que dejar de considerar como parte del SFM a todas las secuencias de tamaño 2 que estén contenidas dentro de las secuencias de tamaño 3, procediendo análogamente hasta un tamaño para el cual ya no se encuentren secuencias que superen o igualen el umbral [23] [24].

  Se pone el siguiente ejemplo, considérese las oraciones en la figura 2.4 como el conjunto de textos [24]:

1. MILES DE MUERTOS SE ENCONTRARON EN LA REGION AFGANA
2. EN LA REGION DEL CAIRO MURIERON MUCHAS PERSONAS A RAIZ DEL TEMBLOR
3. EN MEXICO MILES DE MUERTOS QUEDARON POR LAS CALLES
4. EN TOTAL MURIERON MUCHAS PERSONAS POR EL INCENDIO
5. SE DERRUMBARON MILLONES DE CASAS POR EL SISMO
6. EN LAS CERRANIAS SE DERRUMBARON MILLONES DE CASAS
7. MILES DE HOGARES FUERON AFECTADOS EN ARGENTINA
8. MUCHAS PERSONAS QUEDARON INCOMUNICADAS
9. AUNQUE MUCHAS PERSONAS FUERON DAMNIFICADAS
10. EN COZUMEL MILES DE HOGARES FUERON AFECTADOS POR LA INUNDACION

Figura 2.4: Conjunto de oraciones referentes al tema "desastres naturales" [24].

Se obtiene las SFM con un umbral de 2 obteniendo el resultado mostrado en la figura 2.5 [24]:

```
2 SFM de tamaño 1
        [2] QUEDARON
        [2] LAS
1 SFM de tamaño 2
        [2] POR EL
3 SFM de tamaño 3
        [2] MILES DE MUERTOS
        [2] EN LA REGION
        [2] MURIERON MUCHAS PERSONAS
2 SFM de tamaño 5
        [2] MILES DE HOGARES FUERON AFECTADOS
        [2] SE DERRUMBARON MILLONES DE CASAS
```

Figura 2.5: SFM obtenidas [24].

Como se puede ver la secuencia "muchas personas" no aparece en las SFM dado que, a pesar de ser una secuencia frecuente, no es maximal, ya que la SFM para ella es "murieron muchas personas" [24].

2.2.5. Clustering:

2.2.5.1. Clustering en textos pequeños:

Este tipo de textos se deben afrontar de manera diferente a los normales debido a que su escasez de palabras hace que los vectores característicos de estos textos no contengan muchos valores, lo que afecta a su funcionamiento [25].

Tomando en cuenta que, por lo general, el pesado de las palabras se hace con la técnica de TF-IDF, "En el caso de los textos cortos, ya que el term frequency de la mayoría de los index terms está limitado a los documentos (mayormente 1, raramente 2 o 3) el vector característico con TF-IDF se podría reducir a un vector con IDF. Podría ser suficiente reprsentar el documentos como un vector de 1/0 dependiendo de la presencia/ausencia del index term" [25].

2.2.5.2. Bisecting K-means:

En [26] se define como: "Es una extensión directa del algoritmo básico de K-means y se basa en la siguiente idea: para obtener K

clusters, separa todo el conjunto de puntos en 2 clusters, selecciona uno de esos clusters para separar, y continuar de la misma forma hasta obtener K clusters". El algoritmo del Bisecting K-means se describe en [26] con el seudocódigo mostrado en la figura 2.6.

```
1: Initialize the list of clusters to contain the cluster consisting of all points.
2: repeat
3:    Remove a cluster from the list of clusters.
4:    {Perform several "trial" bisections of the chosen cluster.}
5:    for i = 1 to number of trials do
6:       Bisect the selected cluster using basic K-means.
7:    end for
8:    Select the two clusters from the bisection with the lowest total SSE.
9:    Add these two clusters to the list of clusters.
10: until Until the list of clusters contains K clusters.
```

Figura 2.6: Algoritmo de Bisecting K-means [26].

Al momento de separar 1 cluster en 2, se realizan varias pruebas de separación con el algoritmo de K-means, la prueba que resulte con mayor similitud total de clusters producidos es la que se toma. El criterio que se toma para determinar cuál será el siguiente cluster a separar es analizando cual tiene la menor similitud total de cluster [26] [27].

2.2.6. Distancias:

2.2.6.1. Similitud Coseno:

Es una de las medidas de similitud más populares para documentos de texto tal como aplicaciones de IR o TC. Una propiedad importante de la similitud coseno es su independencia de la longitud del documento. Para 2 documentos P y Q su similitud coseno se calcula por la fórmula (2.10) [8] [9] [28]:

$$s_{Cos} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sqrt{\sum_{i=1}^{d} P_i^2}\sqrt{\sum_{i=1}^{d} Q_i^2}} \qquad (2.10)$$

La medida de la similitud está en el rango del 0 al 1, siendo 1 totalmente idénticos y 0 totalmente diferentes. Su correspondiente medida de distancia se calcula por la fórmula (2.11) [9]:

$$D_{Cos} = 1 - S_{Cos} \tag{2.11}$$

2.2.6.2. Coeficiente Jaccard:

También conocido como coeficiente Tanimoto. Mide la similitud como la intersección dividida por la unión de los objetos. Para documentos de texto, el coeficiente Jaccard compara la suma de los pesos de los index terms compartidos por ambos documentos con la suma de los pesos de los index terms que solo están presentes en 1 de los 2 documentos. Para 2 documentos P y Q su coeficiente Jaccard se calcula por la fórmula (2.12) [8] [9] [28]:

$$S_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} \tag{2.12}$$

La medida de la similitud está en el rango del 0 al 1, siendo 1 totalmente idénticos y 0 totalmente diferentes. Su correspondiente medida de distancia se calcula por la fórmula (2.13) [9]:

$$D_{Jac} = 1 - S_{Jac} \tag{2.13}$$

2.2.7. Evaluación:

Para las fórmulas que definen las medidas en las siguientes 2 subsecciones se describe los que representan los siguientes símbolos:

x = Número de clusters

c_i = Cluster (vector característico promedio o centroide)

$n_i^{(j)}$ = Número de documentos total, en el cluster i o en clase j según corresponda

$d_i^{(j)}$ = Documento en cluster i o en clase j según corresponda

k = Número de clases

$m_i^{(j)}$ = Numero de textos en cluster i y en clase j

sim = Medida de similitud utilizada (Ejm: coseno, Jaccard)

2.2.7.1. Medidas Internas:

Se refiere a las medidas que se pueden realizar para evaluar la calidad de los clusters solo basándose en clusters del resultado, si el resultado presenta buenas medidas internas se interpreta que se ha hecho una buena representación de los textos [29].

2.2.7.1.1. Intra Similitud:

Mide la calidad de cada cluster por separado en términos de cuan similares son cada uno de los textos del cluster (cohesión), mientras más grande sea la medida de similitud resultante, mejor calidad de cluster representa. Se mide bajo la fórmula (2.14) [29]:

$$S_{Intra} = \sum_{i=1}^{x} \sum_{j=1}^{n} sim(d_j, c_i) \tag{2.14}$$

2.2.7.1.2. Inter Similitud:

Mide la calidad de todos los clusters en conjunto en términos de que tan separados están todos los clusters entre sí. Mientras más pequeña sea la medida de similitud, mejor calidad de clusters representa. Se mide bajo la fórmula (2.15) [29]:

$$S_{Inter} = \sum_{i=1}^{x} \sum_{j=i+1}^{x} sim(c_i, c_j) \tag{2.15}$$

2.2.7.2. Medidas Externas:

Se refiere a las medidas que se pueden realizar para evaluar la calidad de los clusters comparándolos con una clasificación manual de los textos hecha por un experto, si el resultado presenta buenas medidas externas se interpreta que la representación de los textos ayuda a conseguir resultado útiles y relevantes [29].

En la descripción de las medidas externas que se explicarán a continuación, se hará mención a los siguientes dos términos de medida de performance de TC [29]:

- **Precision:** Compara un cluster con una clase evaluando la cantidad de coincidencia contra la cantidad de documentos del cluster. Se representa bajo la fórmula (2.16) [29]:

$$p_i^{(j)} = \frac{m_i^{(j)}}{n_i} \tag{2.16}$$

- **Recall:** Compara un cluster con una clase evaluando la cantidad de coincidencia contra la cantidad de documentos de la clase. Se representa bajo la fórmula (2.17) [29]:

$$r_i^{(j)} = \frac{m_i^{(j)}}{n^{(j)}} \tag{2.17}$$

2.2.7.2.1. Puridad:

La puridad de un cluster representa la máxima medida de precision de un cluster. La puridad de un solo cluster se define bajo la fórmula (2.18) [29]:

$$pur_i = \max_j \left\{ p_i^{(j)} \right\} \tag{2.18}$$

La puridad de todo el conjunto de clusters representa el porcentaje de acierto de cada cluster con su correspondiente clase. Se define bajo la fórmula (2.19) [29]:

$$pur = \sum_{i=1}^{x} \frac{n_i}{n} pur_i = \frac{\sum_{i=1}^{x} \max_j \left\{ m_i^{(j)} \right\}}{n} \tag{2.19}$$

La medida de la evaluación de la puridad está en el rango del 0 al 1, siendo 1 la mejor calificación y o la peor [29].

2.2.7.2.2. Entropía:

La entropía toma en consideración todas las clases en consideración. La entropía de un cluster se define bajo la fórmula (2.20) normalizada en el rango de 0 a 1 [29]:

$$entr_i = \frac{-\sum_{j=1}^{n} p_i^{(j)} \log\left(p_i^{(j)}\right)}{\log(x)} \tag{2.20}$$

La entropía de todo el conjunto de clusters se define bajo la fórmula (2.21) [29]:

$$entr = \frac{\sum_{i=1}^{x} n_i entr_i}{n} \tag{2.21}$$

La medida de la evaluación de la entropía está en el rango del 0 al 1, siendo 0 la mejor calificación y 1 la peor [29].

CAPÍTULO 3
PROPUESTA DE MODELO DE TEXT CLUSTERING PARA TEXTOS TÉCNICOS CORTOS NO ESTRUCTURADOS

El modelo propuesto está dirigido al clustering de textos técnicos cortos no estructurados sin importar el ámbito al cual estén referidos, es decir que si se quisiera cambiar el contexto de los textos técnicos, bastaría con cambiar los parámetros adecuados para indicar a que contexto están referidos estos nuevos textos para poder aplicarse sin perder performance. A continuación se describe más específicamente las características de los textos que se desean clusterizar:

- Escritura libre: Los textos no están sujetos a estándares formalmente impuestos ni correctores de ortografía.
- Textos Cortos: En el caso de estudio no superan las 150 palabras en su mayoría. Los textos son escritos muy rápidamente y apuntando a transmitir solamente lo esencial, en su mayoría no superan los 2 párrafos ni las 200 palabras en total. En estos textos prima la presencia de términos técnicos sobre los stop words, siendo no muy legibles para personas no expertas en el área.
- Escritura muy especializada (técnica): Los textos hacen referencia a áreas del conocimiento muy específicas, por lo que utilizan un lenguaje muy técnico, es decir que en estos textos prima la presencia de términos técnicos sobre los stop words, por lo que solamente puede ser entendidos a su cabalidad por un especialista en el área.

Bajo esta perspectiva, se propone un modelo de TC que se compone de las mejores prácticas encontradas en el estado del arte para este tipo de textos; además, se propone una mejora a la técnica de reducción dimensional SFM orientada a mejorar sus resultados en textos de escritura muy especializada [30].

3.1. MODELO PROPUESTO:

Dentro de la literatura podemos encontrar un modelo básico de aplicación TC; sin embargo, existen muchas variaciones de este dependiendo del entorno en el cual es utilizado. Para el modelo propuesto de TC se incluyeron los procesos

que demostraron rendir mejores resultados en el clustering de textos con las características descritas anteriormente. En la figura 3.1 se puede ver todo el modelo de TC propuesto.

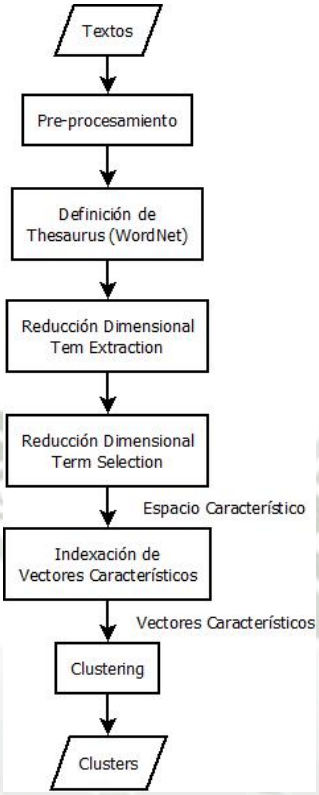

Figura 3.1: Modelo propuesto de Text Clustering [Fuente: propia].

3.1.1. Pre-procesamiento:

Se lleva a cabo en varias tareas las cuales están representadas en la figura 3.2 y serán descritas a continuación.

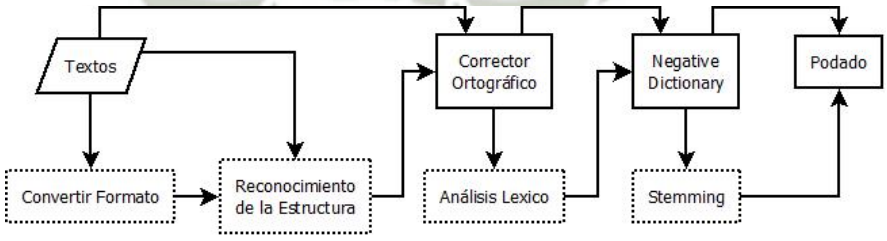

Figura 3.2: Pre-procesamiento de textos. Las tareas en recuadros punteados son opcionales [Fuente: propia].

3.1.1.1. Convertir Formato (opcional):

Esta tarea no siempre es aplicable dado que solo es necesario si los textos que se quieren clusterizar se encuentran en un formato que no pueda ser directamente procesado por el lenguaje de programación que se escoja, por ejemplo: PDF, Microsoft Word, HTML, etc., son formatos que además del contenido del documento guardan información extra o están codificados. El objetivo de esta tarea es extraer el contenido del texto junto con su estructura (si la tiene). Frameworks open source como iText y POI pueden ser utilizados para tratar documentos PDF y Microsoft Office respectivamente.

3.1.1.2. Reconocimiento de la Estructura (opcional):

Esta tarea es opcional, ya que no siempre se presentan textos que tengan una estructura formalizada. Solamente, en esos casos, se aplica esta tarea y solamente, en los casos que el diferenciar toda la estructura del texto apoye en los resultados. Por ejemplo, si se estuvieran clusterizando periódicos y el criterio principal por el que se quiere clusterizar es por los anuncios de empleos que estos tienen sería de gran ayuda poder identificar la estructura de los periódicos para poder centrar el análisis en la sección de empleos.

3.1.1.3. Corrector Ortográfico:

Esta tarea representa un gran problema por la frecuente presencia de los términos técnicos, los cuales no suelen seguir las reglas gramaticales tradicionales sobre las que se basan los algoritmos de corrección de ortografía. En caso no se encuentre algún algoritmo que haga una corrección adecuada de la ortografía de los textos, se propone la aplicación del corrector ortográfico semiautomático que se expone a continuación.

Partiendo de la premisa que un mismo error ortográfico no se debería presentarse muy frecuentemente se presenta un algoritmo de corrección semiautomática en el algoritmo 3.1.

```
Algoritmo 3.1 Corrector Ortográfico Semiautomático.
Input: Lista de documentos, umbral.
Output: Documentos corregidos.
Comentario frecuencia → arreglo de frecuencias para las palabras en los textos
frecuencia := {}
Para cada documento a ser corregido
    Para cada palabra en el documento
        frecuencia[palabra]++
    Siguiente palabra
Siguiente documento
Para cada palabra en frecuencia[]
    Si frecuencia[palabra] <= umbral
        *Revisar si palabra está bien escrita, considerar su contexto.
    Fin Si
Siguiente palabra
```

Algoritmo 3.1: Corrector ortográfico semiautomático, (*) proceso manual [Fuente: propia].

Seguidamente se explicará paso a paso el algoritmo 3.1:

- Se cuentan las repeticiones de todas las palabras en todos los textos.
- Se toman solo las palabras cuyas repeticiones sean menores o iguales a un umbral (se sugiere un valor de 5).
- Se revisa y corrige manualmente todas las palabras mal escritas, si se tiene dudas revisar el contexto de la palabra.

Esta forma de corrector ortográfico genera gran carga de trabajo manual. Sin embargo, la aplicación de esta tarea se corresponde con una mejora en los resultados, el cual depende de la cantidad de errores ortográficos que los textos presenten. Otro punto a favor al realizar esta forma de corrector ortográfico es el conocimiento que uno toma al analizar muchos de los textos, lo cual puede ayudar en los siguientes procesos del modelo.

3.1.1.4. Análisis Léxico (opcional):

Esta tarea es opcional dado que trata de identificar si palabras con caracteres no alfabéticas son relevantes para permanecer en el texto; el poder evaluar esto requiere de conocimiento del área de los textos.

Al estar tratando con varios términos técnicos, se recomienda dejar de lado esta tarea, dado que posteriormente, si los términos no alfabéticos son o no relevantes para los textos, los procesos posteriores se encargarán de tomarlos o no en cuenta para su representación según corresponda. Por ejemplo, el termino B6 puede ser no relevante para muchos textos, pero en textos médicos representa una vitamina, lo que puede llegar a ser muy relevante.

3.1.1.5. Negative Dictionary:

Se debe tener cuidado con la fuente de donde se obtiene el Negative Dictionary, se recomienda obtenerlo de la página oficial de una institución reconocida e idealmente, que sea un Negative Dictionary especial para el área del conocimiento de los textos, el cual no contenga Stop Words que pueda estar contenido en términos que resulten relevantes para los textos.

3.1.1.6. Stemming (opcional):

Esta tarea representa un problema dado que los términos técnicos no sigues las reglas gramaticales tradicionales sobre las cuales los algoritmos de stemming se basan. Si no se encuentra un algoritmo de stemming que se ajuste al dominio del conocimiento de los textos, se recomienda no aplicar esta tarea.

3.1.1.7. Podado:

Se recomienda su utilización para la eliminación de las palabras muy poco frecuentes usando un umbral menor al número promedio esperado de documentos en los clusters resultantes.

3.1.2. Definición del Thesaurus:

En este proceso no se definirá el Thesaurus como el espacio característico final; el objetivo principal es dejar solamente las palabras que se consideren relevantes para la representación del texto, para esto se eliminan todas las palabras que se considere puedan ser perjudiciales para el criterio por el cual se quieran clusterizar los textos. Por ejemplo,

siguiendo con el ejemplo de los anuncios de empleos en los periódicos, suponiendo que el criterio por el cual se quiere clusterizar los periódicos es por el tipo de empleo que estos ofrecen, para este motivo palabras referentes a las carreras profesionales o los conocimientos que se necesitan son muy relevantes; sin embargo, palabras como los años de experiencia o la empresa que está solicitando trabajadores dan otro criterio de clustering que puede competir contra el criterio del tipo de empleo que es el criterio que nos interesa, por lo que resulta conveniente eliminar ese tipo de palabras de nuestro Thesaurus, sobre todo si presentan una frecuencia suficiente como para identificar un cluster (frecuencia mediana).

Se recomienda altamente usar un WordNet para filtrar las palabras que realmente representan el criterio por el cual queremos clusterizar los textos; para esto se debe escoger los dominios convenientes y, si cabe el caso, las categorías gramaticales adecuadas, y finalmente filtrar únicamente las palabras que cumplan estos requisitos. Si existiese algún thesaurus especializado en el tema de los textos que se quiere clusterizar y brinda la accesibilidad necesaria a sus datos para hacer un filtrado automático, sería preferible utilizar ese thesaurus.

3.1.3. Reducción Dimensional por Term Extraction:

Este proceso identifica todos los index terms posibles basándose solo en las palabras presentes en el thesaurus definido en el proceso anterior. Lo que hacen las técnicas de Term Extraction es extraer nuevos términos a partir de las letras o palabras presentes en el texto, luego eligen los términos relevantes bajo algún criterio, y posteriormente seleccionan los más relevantes usando una técnica de Term Selection (la cual es TF en la mayoría de los casos). Lo que se está haciendo aquí es separar este proceso en 2 para dar la flexibilidad de utilizar la combinatoria de Term Extraction y Term Selection que mejor se adapte al caso que se esté tratando.

3.1.4. Reducción Dimensional por Term Selection:

Como se explicó en el proceso anterior, este proceso es el encargado de seleccionar los index terms más relevantes de los identificados por la técnica de Term Extraction. Si bien es cierto que la técnica más ampliamente usada es TF, existen otras técnicas que pueden dar mejores resultados bajo ciertos contextos, por lo que generalmente este proceso se podría obviar. Para el caso de Text Categorization existen una mejor y más amplia variedad de técnicas. Basándose en los antecedentes investigativos [5] [10] [17] se recomienda usar:

- **Document Frequency (DF):** Es un índice de 0 a 1 que representa el porcentaje de documentos en los que se encuentra el index term.

$$\mathrm{DF}(\mathrm{term}) = \frac{\text{number of documents with term}}{\text{number of ducuments in database}} \tag{3.1}$$

- **Term Contribution:** Este método toma en cuenta el peso que se le da al index term y mide su importancia en base a la contribución del index term para la medida de similitud entre los textos. Si se utiliza un pesado binario el resultado de esta técnica será igual a la de un DF. Se define bajo la fórmula (3.2):

$$TermContr(t) = \sum_{i=1}^{d} \sum_{j=i+1}^{d} sim(t, d_i, d_j) \tag{3.2}$$

Dónde:

d = Número de documentos a clusterizar.

t = Index term a evaluar.

sim() = Medida de similitud para t entre los documentos d_i y d_j.

3.1.5. Indexación de Vectores Característicos:

La asignación de pesos a cada index term tiene un enfoque diferente dado que los textos son de muy corta extensión, lo que limita la cantidad de index terms que estos puedan contener, así como la posibilidad que alguno de ellos se presente más de 1 vez. Basándose en lo dicho y en el

estudio hecho en [25], además de otros antecedentes investigativos, se recomienda la utilización de:

- **Pesado Binario:** Se pondera con 1 si es index term está presente en el documento o con 0 si no está presente.
- **TF-IDF:** Da un alto valor a las palabras que ocurren frecuentemente a en el documento, pero no en los demás.

3.1.6. Clustering:

Nuevamente, la corta extensión de los documento hace que se tenga un enfoque diferente de este proceso. Al ser textos cortos los vectores característicos tendrán muy pocas dimensiones con un valor distinto a 0. Basándose en los antecedentes investigativos [25] [27] se sugiere la utilización de los siguientes algoritmos de clustering:

- **Bisecting K-means:** Utiliza el algoritmo tradicional del K-means iterativamente, empieza separando el conjunto de textos de 2 clusters, luego escoge el de menos calidad y lo vuelve a separar, se repite el proceso hasta tener el número deseado de clusters. Se define bajo el algoritmo mostrado en la figura 3.3.

```
1: Initialize the list of clusters to contain the cluster consisting of all points.
2: repeat
3:    Remove a cluster from the list of clusters.
4:    {Perform several "trial" bisections of the chosen cluster.}
5:    for i = 1 to number of trials do
6:       Bisect the selected cluster using basic K-means.
7:    end for
8:    Select the two clusters from the bisection with the lowest total SSE.
9:    Add these two clusters to the list of clusters.
10: until Until the list of clusters contains K clusters.
```

Figura 3.3: Algoritmo de Bisecting K-means [26].

Usando las siguientes medidas de distancia:

- **Similitud Coseno:** Para 2 documentos P y Q su similitud coseno se calcula por la fórmula (3.3):

$$s_{Cos} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sqrt{\sum_{i=1}^{d} P_i^2} \sqrt{\sum_{i=1}^{d} Q_i^2}} \quad (3.3)$$

La medida de la similitud está en el rango del 0 al 1, siendo 1 totalmente idénticos y 0 totalmente diferentes. Su correspondiente medida de distancia se la por la fórmula (3.4):

$$D_{Cos} = 1 - S_{Cos} \quad (3.4)$$

- Coeficiente Jaccard: Para 2 documentos P y Q su coeficiente Jaccard se calcula por la fórmula (3.5):

$$s_{Jac} = \frac{\sum_{i=1}^{d} P_i Q_i}{\sum_{i=1}^{d} P_i^2 + \sum_{i=1}^{d} Q_i^2 - \sum_{i=1}^{d} P_i Q_i} \quad (3.5)$$

La medida de la similitud está en el rango del 0 al 1, siendo 1 totalmente idénticos y 0 totalmente diferentes. Su correspondiente medida de distancia se la por la fórmula (3.6):

$$D_{Jac} = 1 - S_{Jac} \quad (3.6)$$

3.2. TÉCNICA DE RD PROPUESTA: N-GRAMAS HÍBRIDO:

Esta técnica está basada en la técnica de N-Gramas la cual agrupa distintos grupos de palabras consecutivas (frases) según su frecuencia en los textos a clusterizar a los cuales a partir de ahora se denotarán con la letra T; la extensión de la frase está determinada por N. Una forma de utilización de los N-gramas también incluye generar términos o frases con distintos valores de N, generalmente desde 1 hasta un numero dado por el usuario y luego se aplica algún criterio de selección como el recoger los N-Gramas más frecuentes generados con todos los valores de N; sin embargo estos criterios de selección

no toman en cuenta la relación semántica absorbente que puede tener un N-Grama respecto a otro cuyo N sea menor que el primero, lo cual se puede apreciar en la figura 3.4, los 2 Bi-Gramas mostrados están contenidos dentro del Tri-Grama.

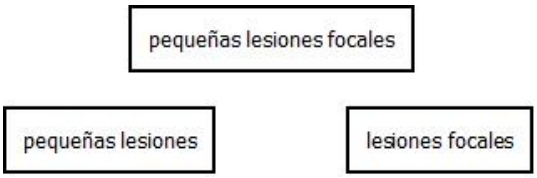

Figura 3.4: Ejemplo de 1 Tri-Grama y sus 2 Bi-Gramas correspondientes [Fuente: propia].

Las Secuencias Frecuentes Maximales (SFM) toman en cuenta el criterio de absorción, sin embargo dejan atrás a muchos N-gramas que pueden ser muy relevantes, por ejemplo en la figura 3.5, el SFM ignora completamente los Bi-Gramas “tc cerebral” y “cerebral alteraciones”; sin embargo, estos Bi-Gramas por sí mismo pueden resultar bastante relevantes para los textos.

Figura 3.5: Ejemplo de Secuencias Frecuentes Maximales. Entre paréntesis esta la frecuencia de cada N-Grama [Fuente: propia].

El objetivo de los N-Gramas Híbridos es el de tener un criterio de selección de N-Gramas que tome en cuenta las relaciones semánticas entre los N-Gramas generados con distintos valores para N, tomando los que resulten más relevantes para el clustering de T.

3.2.1. Relaciones Semánticas:

En principio, el objetivo de los N-Gramas es el de extraer grupos de palabras frecuentes de T dado que las palabras tienen un mejor valor, semánticamente hablando, si es que se las agrupa; sin embargo, se debe recordar que estos N-Gramas también deben tener la propiedad de

identificar lo mejor que puedan a cada texto, lo cual es responsabilidad del criterio de selección de N-Gramas.

Figura 3.6: N-Grama padre y sus 2 N-Gramas hijos [Fuente: propia].

Como se puede ver en la figura 3.6, todo "N-Grama padre" está compuesto por 2 "N-Gramas hijos" con un N menor en uno que su padre, y tomando esto en cuenta se pueden tomar 2 posturas ante esta relación:

- Que los 2 N-Gramas hijos no tengan el valor semántico suficiente por sí mismos, por lo que tendrían que ser absorbidos por el N-Grama padre, puesto que es más probable que tenga un mayor valor semántico por tener una mayor cantidad de palabras.
- Que los 2 N-Gramas hijos sean autosuficientes, en cuyo caso habría que evaluar el valor semántico del N-Grama padre y tomando en cuenta el valor semántico del N-Grama padre y sus 2 N-Gramas hijos por separado, decidir cuales puede resultar relevantes para el clustering de T.

3.2.2. Absorción en Cadena:

El objetivo de las absorciones es el de generar N-Gramas que representen tópicos lo más específicos posibles, pero sin perder su relevancia para el clustering de T, por esta razón que el orden de las absorciones se puede ver como una absorción en cadena desde los valores de N más bajos hasta los más altos, dado que cuando se están absorbiendo N-Gramas que pertenecen a un solo tópico, esta absorción parará cuando el N-Grama padre ya no represente a ese tópico, sino que ya estaría representando a un tópico mucho más específico; sin embargo, ese N-Grama padre luego puede seguir absorbiéndose con sus semejantes; si estos tienen suficiente relevancia también se les tomará al momento de la selección de los N-Gramas Híbridos más relevantes. En la figura 3.7 se puede ver un ejemplo

de lo que sería una absorción en cadena desde un valor de N=2 hasta un valor de N=4.

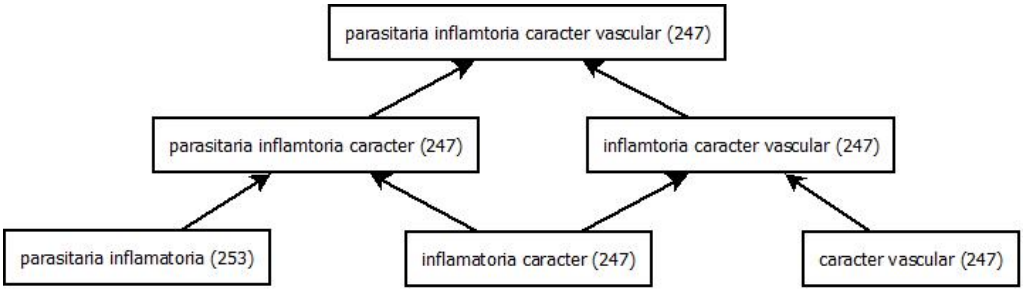

Figura 3.7: Escenario de Absorción en Cadena [Fuente: propia].

3.2.3. Criterios de Absorción:

Como se vio en el punto anterior, la absorción de N-Gramas depende de la relevancia semántica de cada N-Grama en relación a la de su padre e hijos, para la técnica desarrollada la determinación de la relevancia semántica será directamente proporcional a la frecuencia de cada N-Grama en T. Se puede afirmar que:

$$freq(N - Grama\ Padre) \leq freq(N - Grama\ Hijo) \qquad (3.7)$$

Dado que todas las ocurrencias en T que tenga el N-Grama padre, también las tendrán sus N-Gramas hijos, además cada N-Grama hijo tiene la posibilidad de tener más de un N-Grama padre; de este modo se deduce que, un N-Grama hijo tendrá como mínimo un N-Grama padre, en cuyo caso sus frecuencias serían iguales, si el N-Grama hijo tiene más de un padre entonces la frecuencia de este puede ser superior a la de sus N-Gramas padres si estos no están relacionados entre sí. En la figura 3.7 podemos ver un ejemplo de lo que se acaba de describir, el Bi-Grama "inflamatoria carácter" tiene 2 Tri-Gramas padres, sin embargo estos están relacionados entre sí ya que representan a un mismo Tetra-Grama que es su padre, es por esto que la frecuencia del Bi-Grama "inflamatoria carácter" es igual a la de sus padres a pesar de tener más de un padre. Dicho esto podemos ver los 2 siguientes casos:

- **Igualdad de frecuencia del padre y sus hijos:** Cuando se presenta este caso se puede afirmar que las ocurrencias del padre en T son las mismas que presentan sus hijos en T, por lo que en esencia, lo que representan en T es lo mismo; sin embargo, por el hecho que el N-Grama padre tiene un mayor valor semántico por tener más palabras en él, en este caso el N-Grama padre absorbe a sus 2 N-Gramas hijos, esto es poner la frecuencia de los N-Gramas hijos en 0. En la figura 3.8 se puede ver un ejemplo.

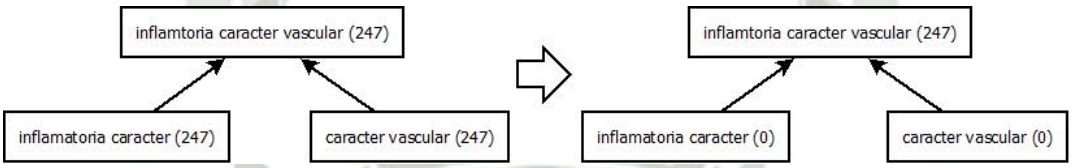

Figura 3.8: Absorción total de 2 N-Gramas hijos por su N-Grama padre [Fuente: propia].

- **Hijos con una mayor frecuencia que sus padres:** Cuando se presenta esto los N-Gramas hijos tienen ocurrencias en T que no involucran a su padre, por lo que se puede deducir que abarcan más de un tópico determinado dado que tienen más de un padre. En estos casos se resta la frecuencia del padre de la frecuencia de sus hijos, esto con el motivo de que finalizado el proceso la frecuencia de cada N-Grama represente lo que es por sí mismo, y no represente también a sus N-Gramas padres. En caso que la frecuencia del padre no supere el umbral definido, este no resta su frecuencia de la de sus hijos dado que no será escogido como un N-Grama hibrido relevante. En la figura 3.9 se puede ver un ejemplo.

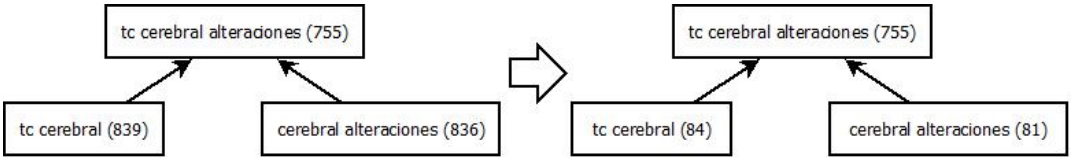

Figura 3.9: Absorción parcial de 2 N-Gramas hijos por su N-Grama padre [Fuente: propia].

Cabe resaltar que al momento de obtener los N-Gramas para todos los valores de N que se determine, se aplica un filtro de frecuencia por el umbral definido, de modo que si, en el proceso, un N-Grama no tiene padres es porque ninguno de ellos supero el umbral y por tanto no se considera relevante.

3.2.4. Hijo Común de N-Gramas Hermanos:

Dos N-Gramas son hermanos si estos descienden del mismo padre, estos 2 N-Gramas tienen una particularidad que hay que cuidar: El segundo hijo del primero hermano es el mismo N-Grama que el primer hijo del segundo hermano, en la figura 3.10 se puede ver un ejemplo del caso planteado.

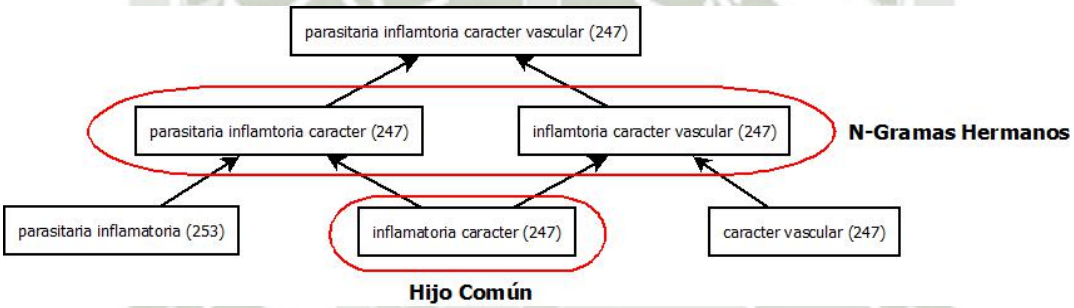

Figura 3.10: Hijo Común de N-Gramas Hermanos [Fuente: propia].

Teniendo esto en consideración, al momento de hacer la absorción en cadena de abajo hacia arriba, el hijo en común seria absorbido por sus 2 N-Gramas padres hermanos; sin embargo existen 2 problemas con esto:

- Después de ser absorbido por su primer padre, el hijo común podría quedarse con menos frecuencia que su segundo padre, lo cual sería una contradicción a la afirmación antes hecha sobre la relación de frecuencia de N-Gramas padres e hijos en la fórmula (3.7).
- Los N-Gramas hermanos tienen un padre en común, por lo que si los 2 absorben a su hijo en común, en realidad se le estaría sustrayendo 2 veces al hijo en común el valor semántico de su N-Grama abuelo, dejándolo con menos valor semántico (frecuencia) del que realmente debería tener.

En la figura 3.11 se puede ver un ejemplo de los problemas antes descritos.

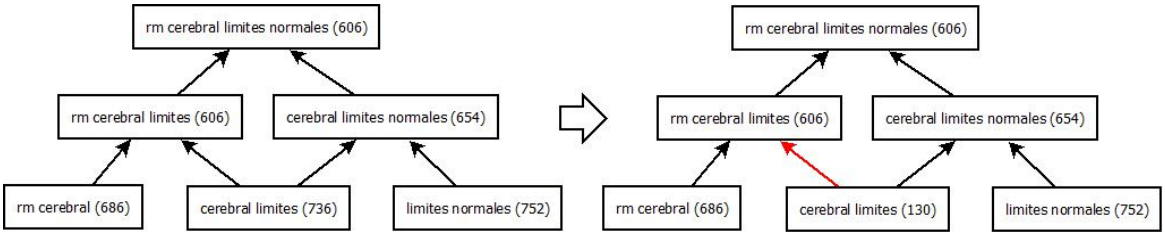

Figura 3.11: Hijo común se queda con menor frecuencia que su segundo padre después de ser absorbido por su primer padre [Fuente: propia].

El problema de sustraer más de una vez, a los hijos en común, el valor semántico de su N-Grama abuelo, puede ocurrir también con sus ancestros remotos.

Para solucionar estos problemas se abordará de forma diferente la absorción en cadena, de arriba hacia abajo, es decir desde el valor de N más alto hasta el más bajo; para esto cada N-Grama absorberá a todos sus descendientes, en otras palabras su frecuencia será sustraída de la de todos ellos. Haciendo esto se elimina su representación semántica de todos sus hijos, eliminando el problema de la doble sustracción semántica para el hijo en común. Ahora que se hace una absorción para toda la descendencia, se debe evaluar a cada N-Grama antes de realizar su absorción en cadena puesto que este puede quedarse con una frecuencia por debajo del umbral como consecuencia de la absorción de sus ancestros, si es así no será considerado como un N-Grama hibrido relevante ni hará su absorción en cadena ya que no es suficientemente representativo como para absorber a sus descendientes. En la figura 3.12 se presenta un ejemplo de absorción en cadena de arriba hacia abajo.

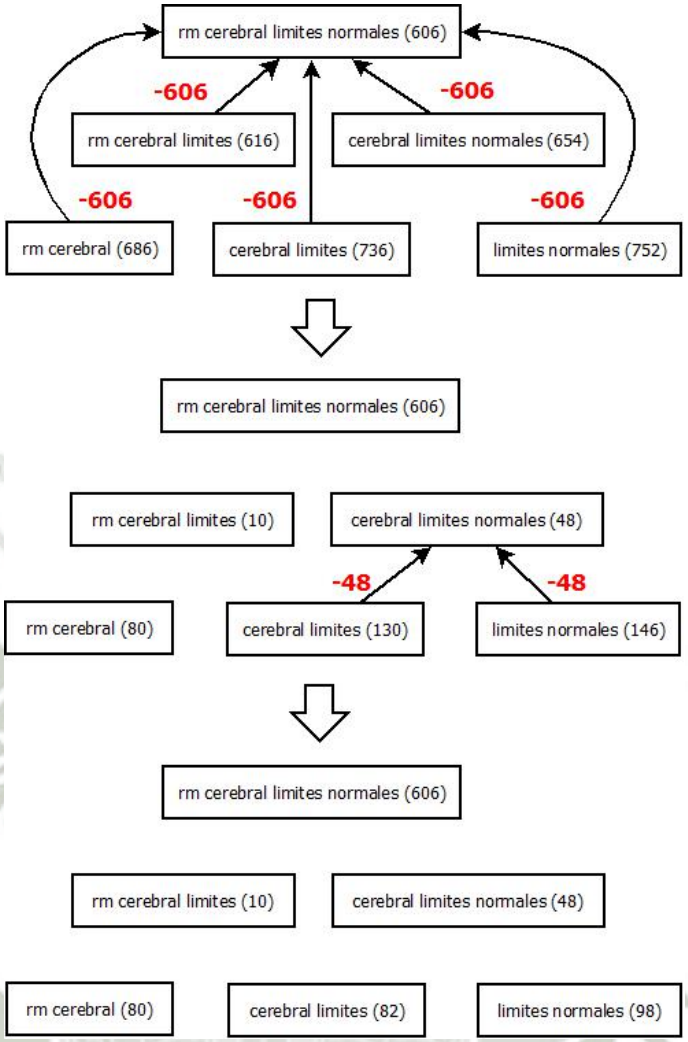

Figura 3.12: Ejemplo de absorción en cadena de arriba hacia abajo con un umbral de 40. El Tri-Grama "rm cerebral límites" no hace absorción en cadena porque se queda con una frecuencia menor al umbral [Fuente: propia].

3.2.5. Algoritmo:

Finalmente se plantea el algoritmo 3.2 tomando en cuenta todas las acotaciones hechas anteriormente, posteriormente se explicará el algoritmo.

```
Algoritmo 3.2 N-Gramas Híbridos.
Input Lista de textos, thesaurus, maxLong, umbral.
Output N-Gramas Híbridos.
Comentario Obteniendo todos los n_gramas
Comentario n_gramas → arreglo de frecuencias de n-gramas normales
Comentario n → longitud de grama
n_gramas := {}
n := 0
Repetir
    n++
    n_gramas[n - 1] = obtener_ngramas(n, umbral)
Mientras ngramas[n - 1].length > 0 Y n < maxLong
n--
Comentario Absorción en cadena
x = n
Mientras x >= 1
    Para cada n_grama en n_gramas[x - 1]
        Comentario Selección de n_gramas hibridos relevantes
        Si n_grama.frecuencia < umbral
            Siguiente n_grama (continue)
        Fin Si
        n_grama es relevante
        Para cada n_grama_des descendiente de n_grama
            n_grama_des.frecuncia -= n_grama.frecuencia
        Siguiente n_grama_des descendiente
    Siguiente n_grama
    x--
Repetir
```

Algoritmo 3.2: Obtención de N-Gramas Híbridos [Fuente: propia].

En primer lugar se obtienen todos los N-Gramas que superen o igualen un umbral de frecuencia, se buscan N-Gramas hasta un valor de N que ya no genere N-Gramas que superen o igualen el umbral puesto.

Posteriormente se realiza una absorción en cadena de abajo hacia arriba, tal como se explicó en el punto anterior; y finalmente se seleccionan todos los N-Gramas híbridos que superen o igualen en umbral definido, los cuales luego pasarán al proceso de reducción dimensional por Term Selection, si se considera oportuno.

CAPÍTULO 4
CASO DE ESTUDIO

4.1. INTRODUCCIÓN:

La clínica SEDIMED, especialista en diagnósticos por imágenes, posee una gran base de datos de estudios médicos basados en imágenes, y cada una de ellas tiene su correspondiente diagnóstico dada por un médico o médicos especialistas. A pesar que los diagnósticos de cada estudio ya han sido dados, no tienen sus estudios clasificados bajo ningún criterio. Utilizando el modelo de Text Clustering y la técnica de N-Gramas hibrido propuesta en el capítulo anterior, se clusterizarán los estudios médicos por la enfermedad que estos presenten, basándose en los diagnósticos de cada uno. Solamente se utilizarán los diagnósticos referentes a resonancias y tomografías encefálicas, y por consiguiente se clusterizarán los estudios médicos por enfermedades encefálicas.

El conjunto de textos tomados de la clínica SEDIMED para realizar la experimentación consiste en 4961 diagnósticos, los cuales están almacenados en una base de datos Oracle propiedad de SEDIMED.

4.2. APLICACIÓN DEL MODELO DE TEXT CLUSTERING:

A continuación se describirá la aplicación de cada una de los proceso del modelo propuesto para el caso de estudio. En la figura 4.1 se puede ver esta información de manera resumida.

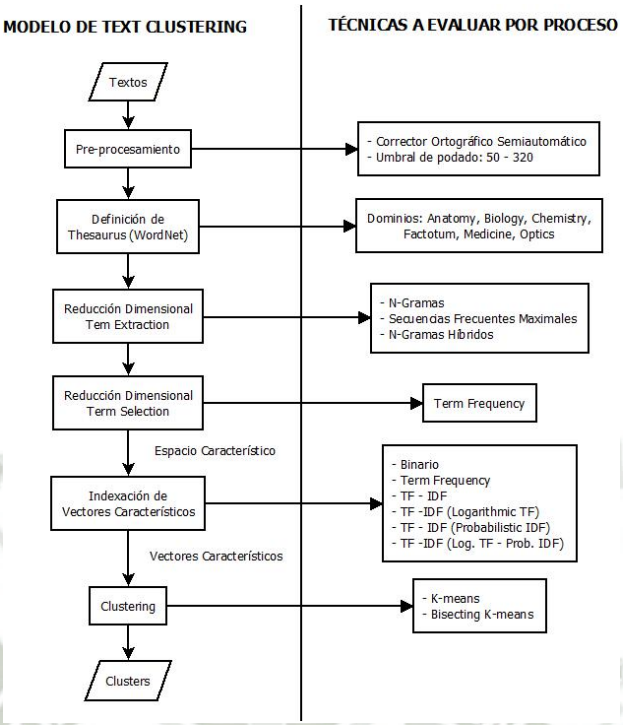

Figura 4.1: Aplicación del modelo de Text Clustering propuesto para la evaluación del caso de estudio [Fuente: propia].

4.2.1. Pre-procesamiento:

4.2.1.1. Convertir Formato:

Se extrajo los diagnósticos médicos de la base de datos Oracle de la clínica a un formato de texto para facilitar su tratamiento.

4.2.1.2. Reconocimiento de la Estructura:

A pesar de ser textos de escritura libre, los textos tienen una escritura muy similar, dado que todos sus autores son médicos de la misma clínica. Se pudo identificar que los diagnósticos presentan categorías como: técnica, hallazgos y conclusión; sin embargo, las únicas que se presentan en todos los diagnósticos indefectiblemente son técnica y conclusión.

Cada diagnóstico se separó en 2 partes, la técnica y la conclusión dada. El criterio por el cual se podía identificar estas 2 categorías, el mismo que se usó para separarlas, es la presencia de la palabra "TECNICA" y "CONCLUSION" en el texto de cada conclusión,

después de la presencia de estas palabras seguía su correspondiente categoría. A partir de ahora en adelante, todo el procesamiento se hará solamente a la categoría de conclusión.

4.2.1.3. Corrector Ortográfico:

Se aplicó el corrector semiautomático descrito en el capítulo anterior. Se revisaron las palabras con una frecuencia menor o igual a 4, no se identificó ninguna palabra con una frecuencia de 4 mal escrita. A continuación se enumeran las principales dificultades que se encontraron para la identificación y corrección de las palabras mal escritas, así como otras observaciones que se consideran importantes:

- Las palabras más frecuentes son también las más frecuentemente mal escritas, por ejemplo: cerebral, desviación, inflamación, envejecimiento, etc.
- Uso indistinto de siglas o sus nombres completos, por ejemplo: ACM (Arteria Cerebral Media), ACV (Accidente Cerebro Vascular), DCV (Desorden Cerebro Vascular), etc.
- Uso inapropiado de abreviaciones personalizadas: enfermedad → enf, izquierdo → izq, derecho → der.
- Palabras que pueden ir tanto juntas como separas, por ejemplo: parieto occipital → parietooccipital, postquirúrgica → post quirúrgica, postraumática → post traumática.

4.2.1.4. Análisis Léxico:

No se realizó el análisis léxico puesto que muchas de las palabras no alfabéticas pueden resultar relevantes para el texto, pero esta evaluación necesitaría el apoyo de un experto en el área como apoyo. Como se mencionó en el capítulo anterior, se dejará este paso de lado y si los términos no alfabéticos son o no relevantes para los textos, los procesos posteriores se encargaran de tomarlos o no en cuenta para su representación según corresponda.

4.2.1.5. Negative Dictionary:

Se utilizó el Negative Dictionary publicado en [31]; este hace referencia a un stop word list utilizado por el sistema de IR SMART (System for the Mechanical Analysis and Retrieval of Text), proyecto que fue liderado por Gerard Salton.

4.2.1.6. Stemming:

Se utilizó el algoritmo de Porter para realizar el stemming, si los resultados obtenidos no son consistentes respecto a la forma base que deberían tener los términos médicos, se obviará este paso.

4.2.1.7. Podado:

Se eliminaron todas las palabras cuya frecuencia es menor al umbral definido. Todas las palabras que superan este umbral son consideradas para la definición del thesaurus en el siguiente procedimiento. Los umbrales a probar irán desde 50 hasta 320 aumentando de 10 en 10.

4.2.2. Definición de Thesaurus (WordNet):

Se utilizaron los WordNet de [32] [33] probando todas las combinatorias posibles de las categorías gramaticales: sustantivo, verbo, adjetivo y adverbio; y todas las combinatorias posibles de los siguientes dominios:

- Anatomy.
- Factotum.
- Medicine.
- Biology.

4.2.3. Reducción Dimensional por Term Extraction:

Se probarán las siguientes técnicas para una longitud de grama desde 1 hasta 10:

- N-Gramas.
- Secuencias Frecuentes Maximales.
- N-Gramas híbridos.

4.2.4. Reducción Dimensional por Term Selection:

Se probarán las siguientes técnicas:

- Term Frequency.

4.2.5. Indexación de Vectores Característicos:

Se probarán las siguientes técnicas de pesado:

- Binario.
- Term Frequency.
- TF-IDF 1.
- TF-IDF 2 (Logarithmic TF).
- TF-IDF 3 (Probabilistic IDF).
- TF-IDF 4 (Logarithmic TF - Probabilistic IDF)

4.2.6. Clustering:

Se probarán las siguientes técnicas de clustering:

- K-means.
- Bisecting K-means.

Y se utilizarán las siguientes medidas de distancia:

- Similitud Coseno.
- Coeficiente Jaccard.

CAPÍTULO 5
PRUEBAS Y RESULTADOS:

Con motivo de poder mejorar el proceso de validación de las pruebas, se pidió a un médico experto, ajeno a la clínica SEDIMED, que clasifique los estudios médicos por la enfermedad que estos representan; para esto se le brindo un software para que pueda hacer esta clasificación mucho más fácilmente, teniendo como resultado 18 clusters, y para cada uno de estos se les dio un tópico o título que representa su enfermedad o anomalía. En la tabla 5.1 se muestran los tópicos o títulos que se le dio a cada cluster y la cantidad de textos que tiene.

Tabla 5.1: Clusters resultantes de la clasificación hecha por el médico experto [Fuente: propia].

Título	Cantidad de Textos
Demencia senil	413
Esclerosis	4
RM normal	726
Leuco encefalopatía	11
Enfermedad desmielinizante	365
Patología quística	238
Tomografía normal	948
TEC	287
Encefalomalacia	182
Procesos inflamatorios	64
Proceso expansivo	574
Hidrocefalia	136
Enfermedad vascular cerebral	368
Sinusitis	169
Mastoiditis	32
Edema cerebral	237
Hipertrofia de cornetes	167
Miscelánea	40

A continuación se presentarán una serie de pruebas y resultados comparativos con el objetivo de encontrar la técnica o valor de umbral óptimo para cada uno de los procesos de modelo. Todas las pruebas irán orientadas principalmente a hacer una

comparación entre todas las técnicas de Term Extraction para poder evaluar la performance de la técnica propuesta, N-Gramas Híbrido.

Para cada una de las combinatorias de técnicas/umbrales se realizaron 10 pruebas de clustering para obtener resultados más confiables; estas pruebas de clustering solamente se hicieron con K-means como técnica de clustering y distancia coseno ya que, al hacer pruebas preliminares, se notó una performance muy baja con la técnica de Bisecting K-means y que al utilizar el coeficiente Jaccard habían muchos casos en los que K-means no llegaba a converger en un resultado con 18 clusters.

5.1. CORRECTOR ORTOGRÁFICO SEMIAUTOMÁTICO:

Una vez aplicado el corrector ortográfico semiautomático, la distribución de las palabras respecto a su frecuencia cambio considerablemente. En la figura 5.1 se muestra como fue variando esta distribución según como se iba corrigiendo todas las palabras con 1, 2 y 3 ocurrencias en todos los textos.

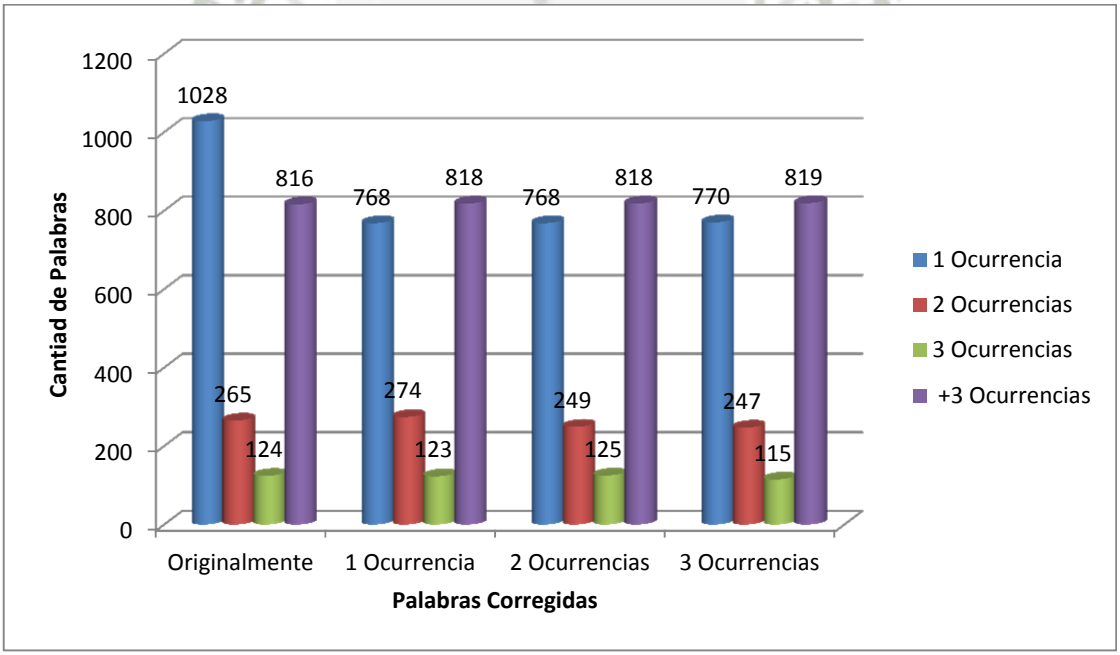

Figura 5.1: Variación de la distribución de todas las palabras por cantidad de ocurrencias en los textos según se fueron corrigiendo [Fuente: propia].

Finalmente tras la corrección, se redujeron 282 palabras del total encontrado al inicio. En la tabla 5.2 podemos ver la cantidad de palabras reducidas en cada

una de las etapas de la corrección y el porcentaje sobre el total que este representa.

Tabla 5.2: Representatividad de cada una de las etapas sobre toda la corrección completa [Fuente: propia].

Etapa – Al corregir palabras con	N° de Palabras Corregidas	Porcentaje del Total
1 ocurrencia	250	88.65%
2 ocurrencias	23	8.16%
3 ocurrencias	9	3.19%
Total	**282**	**100%**

5.2. STEMMING:

Se decidió no realizar el proceso de stemming a los diagnósticos médicos ya que en las pruebas realizadas con el algoritmo de Porter, no se pudo llevar de manera adecuada muchos términos médicos a su forma base.

5.3. PODADO:

Se hicieron pruebas con un valor de umbral desde 50 hasta 320 con un aumento de 10, este mismo umbral se utiliza para todas las demás etapas que requieran de un umbral. En la figura 5.2 se muestran los resultados de las pruebas hechas agrupadas por las técnicas de Term Extraction y los umbrales.

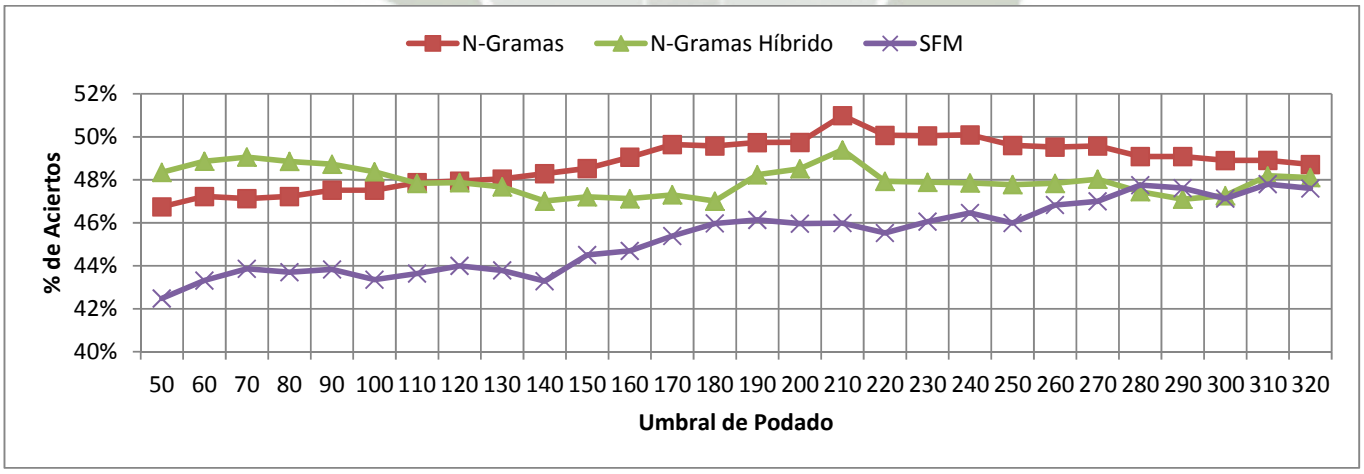

Figura 5.2: Resultados agrupados por las técnicas de Term Extraction y los umbrales [Fuente: propia].

5.4. DEFINICIÓN DE THESAURUS (WORDNET):

Se utilizaron todas las combinatorias posibles con las categorías gramaticales de WordNet (sustantivos, verbos, adjetivos y adverbios) para las pruebas; sin embargo, al hacer un análisis preliminar se observó que todos los adverbios que superaban el umbral mínimo, estaban contenidos dentro de las demás categorías, y que la cantidad de verbos era muy pequeña para representar una combinatoria por sí misma, por lo que la categoría de adverbios no es considerada para las combinatorias y la categoría de verbos estará presente, pero no se considerará como un elemento para las combinatorias; con estas observaciones las combinatorias probadas fueron las siguientes:

- Sustantivos.
- Adjetivos.
- Sustantivos, adjetivos y verbos.

A continuación, en la figura 5.3, se muestran los resultados de las pruebas hechas agrupadas por las técnicas de Term Extraction y las Categorías Gramaticales de WordNet.

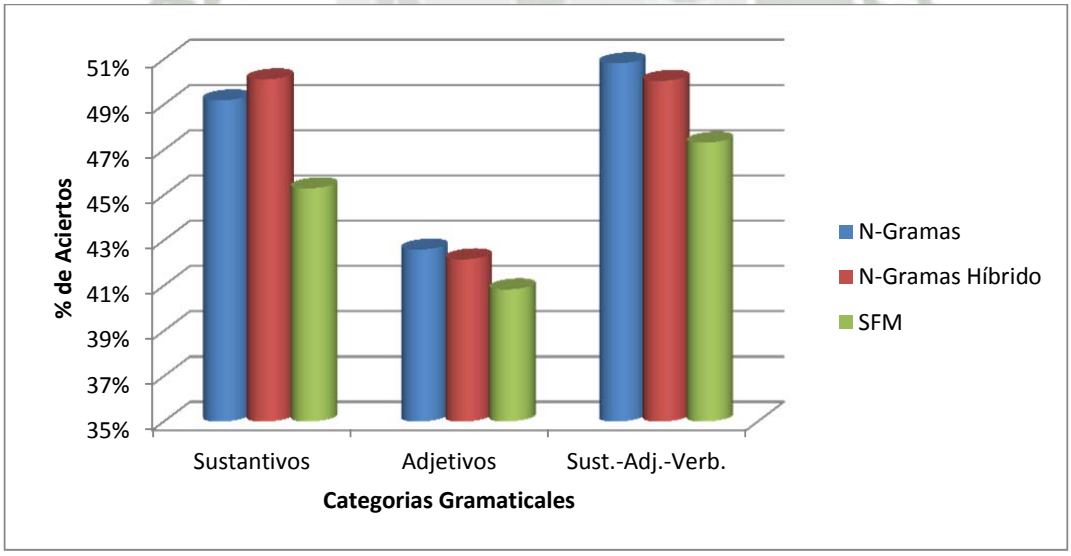

Figura 5.3: Resultados agrupados por las técnicas de Term Extraction y las Categorías Gramaticales de WordNet [Fuente: propia].

Además, se utilizaron todas las combinatorias posibles de los dominios que resultaban relevantes para el caso de estudio, estos son:

- Anatomy.
- Biology.
- Chemistry.
- Factotum.
- Medicine.
- Optics.

Al igual que con las Categorías Gramaticales, se hizo un análisis preliminar de los dominios con las palabras que superan el umbral minino, en este se observó que los dominios de Chemistry y Optics están completamente contenidos dentro de los demás dominios, y que solo una pequeña parte del dominio de Biology no está contenido dentro de los demás dominios; por estas razones los dominios de Chemistry y Optics no serán consideradas para las combinatorias, y el dominio de Biology estará presente, pero no se considerará como un elemento para las combinatorias; con estas observaciones las combinatorias probadas fueron las siguientes:

- Anatomy, biology.
- Factotum, biology.
- Medicine, biology.
- Anatomy, factotum, biology.
- Anatomy, medicine, biology.
- Factotum, medicine, biology.
- Anatomy, factotum, medicine, biology.

A continuación, en la figura 5.4, se muestran los resultados de las pruebas hechas agrupadas por las técnicas de Term Extraction y los dominios de WordNet.

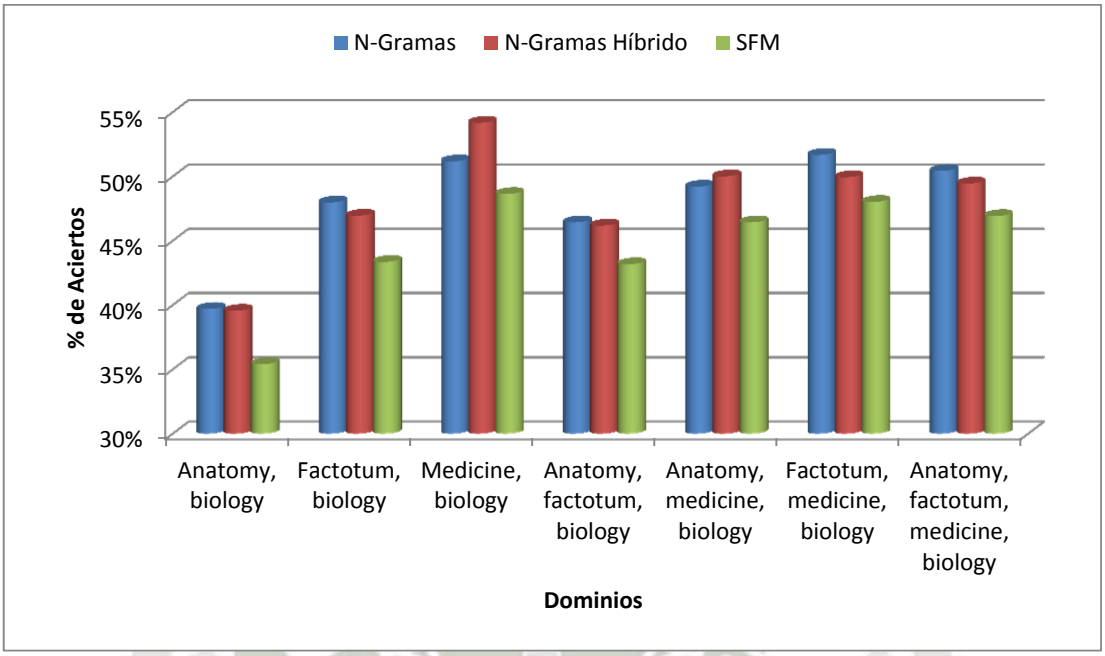

Figura 5.4: Resultados agrupados por las técnicas de Term Extraction y los Dominios de WordNet [Fuente: propia].

5.5. REDUCCIÓN DIMENSIONAL POR TERM EXTRACTION:

Se probaron reducciones dimensionales con una longitud de grama desde 1 hasta 10. Los N-Gramas Híbridos y las SFM partieron de una longitud de 2 ya que, por las características de los algoritmos, con una longitud de 1 harían exactamente lo mismo que los N-Gramas; al mismo tiempo, los N-Gramas solamente se probaron hasta una longitud de 2 ya que, con una longitud mayor a esa, no se generaban suficientes términos (características) para que el K-means pueda converger.

A continuación, en la figura 5.5, se muestran los resultados de las pruebas hechas agrupadas por las técnicas de Term Extraction y las longitudes de gramas.

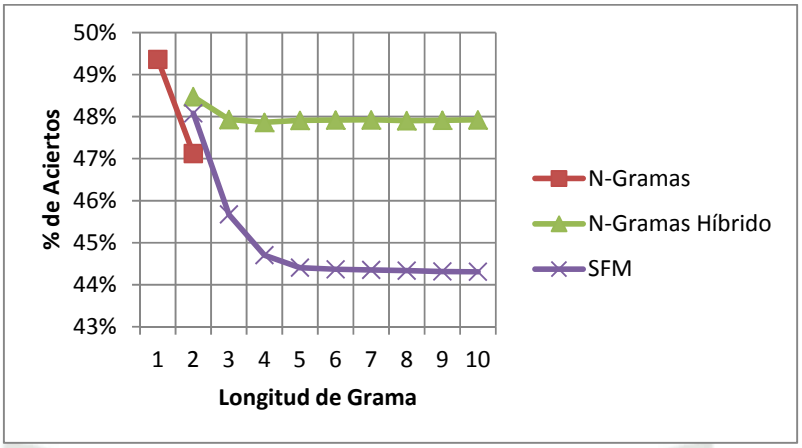

Figura 5.5: Resultados agrupados por las técnicas de Term Extraction y las longitudes de gramas [Fuente: propia].

5.6. INDEXACIÓN DE VECTORES CARACTERÍSTICOS:

Se probarán las siguientes técnicas de pesado:

- Binario.
- Term Frequency.
- TF-IDF 1.
- TF-IDF 2 (Logarithmic TF).
- TF-IDF 3 (Probabilistic IDF).
- TF-IDF 4 (Logarithmic TF - Probabilistic IDF).

A continuación, en la figura 5.6, se muestran los resultados de las pruebas hechas agrupadas por las técnicas de Term Extraction y las técnicas de pesado.

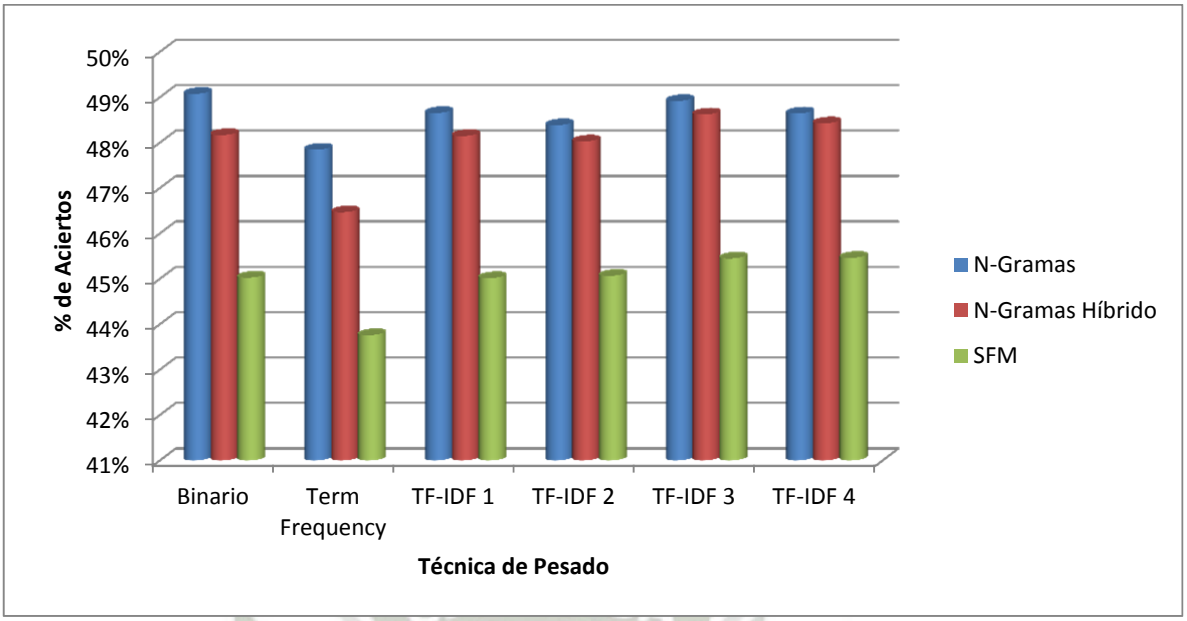

Figura 5.6: Resultados agrupados por las técnicas de Term Extraction y las técnicas de pesado [Fuente: propia].

5.7. RESULTADOS GENERALES:

En total se realizaron 545700 pruebas para un total de 54570 combinatorias probadas entre las distintas técnicas y umbrales. En la tabla 5.3 se muestran los resultados agrupados solamente por la técnica de Term Extraction:

Tabla 5.3: Resultados agrupados por las técnicas de Term Extraction [Fuente: propia].

TÉCNICA	% DE ACIERTOS
N-GRAMAS	48,5855%
N-GRAMAS HÍBRIDO	47,9747%
SFM	44,9635%

A continuación, en la tabla 5.4, se muestran las 50 mejores combinatorias promediando las 10 pruebas que se le hizo a cada combinatoria:

Tabla 5.4: 50 mejores resultados de las 54570 combinatorias probadas [Fuente: propia].

TERM EXTRACTION	PODADO	CATEGORIAS GRAMATICALES	DOMINIOS	LONG. GRAMA	PESADO	% DE ACIERTOS
N-GRAMAS	60	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 3	61,2235%
N-GRAMAS	60	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 1	60,9010%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 4	60,5362%
N-GRAMAS	70	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 4	60,3810%
N-GRAMAS	80	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 2	60,3346%
N-GRAMAS	60	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 2	60,2399%
N-GRAMAS	70	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 1	60,1028%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Anat., fact., med., bio.	2	TF-IDF 3	60,0806%
N-GRAMAS	60	Sust.-Adj.-Verb.	Anat., fact., med., bio.	1	TF-IDF 3	60,0363%
N-GRAMAS	70	Sust.-Adj.-Verb.	Anat., fact., med., bio.	1	TF-IDF 3	59,9153%
N-GRAMAS	50	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 2	59,8629%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 2	59,8327%
N-GRAMAS	70	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 3	59,8166%
N-GRAMAS	50	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 4	59,7561%
N-GRAMAS HÍBRIDOS	70	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 2	59,7541%
N-GRAMAS HÍBRIDOS	70	Sust.-Adj.-Verb.	Anat., fact., med., bio.	2	TF-IDF 4	59,6835%
N-GRAMAS HÍBRIDOS	70	Sust.-Adj.-Verb.	Anat., fact., med., bio.	2	TF-IDF 2	59,6714%
N-GRAMAS HÍBRIDOS	70	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 4	59,6553%
N-GRAMAS	60	Sust.-Adj.-Verb.	Med., bio.	1	TF-IDF 2	59,6533%
N-GRAMAS HÍBRIDOS	70	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 3	59,6170%
N-GRAMAS	70	Sust.-Adj.-Verb.	Anat., fact., med., bio.	1	TF-IDF 1	59,5888%
N-GRAMAS HÍBRIDOS	60	Sust.-Adj.-Verb.	Anat., fact., med., bio.	2	TF-IDF 3	59,5606%
N-GRAMAS	70	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 2	59,5525%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 3	59,5364%
N-GRAMAS	60	Sust.-Adj.-Verb.	Anat., fact., med., bio.	1	TF-IDF 2	59,5223%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 1	59,4638%
N-GRAMAS HÍBRIDOS	60	Sust.-Adj.-Verb.	Anat., fact., med., bio.	2	TF-IDF 2	59,4316%
N-GRAMAS	60	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 4	59,4175%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Anat., fact., med., bio.	2	TF-IDF 4	59,3812%
N-GRAMAS	60	Sust.-Adj.-Verb.	Med., bio.	1	TF-IDF 3	59,3671%
N-GRAMAS HÍBRIDOS	60	Sustantivos	Anat., fact., med., bio.	3	TF-IDF 3	59,3086%
N-GRAMAS	70	Sust.-Adj.-Verb.	Anat., med., bio.	1	TF-IDF 3	59,3026%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Fact., med., bio.	3	TF-IDF 4	59,2764%
N-GRAMAS	50	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 3	59,2038%
N-GRAMAS HÍBRIDOS	60	Sust.-Adj.-Verb.	Fact., med., bio.	3	TF-IDF 4	59,1836%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Anat., fact., med., bio.	3	TF-IDF 4	59,1836%
N-GRAMAS HÍBRIDOS	60	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 1	59,1796%
N-GRAMAS	80	Sust.-Adj.-Verb.	Anat., fact., med., bio.	1	TF-IDF 1	59,1655%
SFM	80	Sust.-Adj.-Verb.	Fact., med., bio.	2	Term Freq.	59,1594%
N-GRAMAS HÍBRIDOS	70	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 1	59,1554%
SFM	80	Sust.-Adj.-Verb.	Fact., med., bio.	2	Binario	59,1111%
SFM	80	Sust.-Adj.-Verb.	Fact., med., bio.	2	TF-IDF 3	59,0647%
N-GRAMAS HÍBRIDOS	50	Sustantivos	Anat., fact., med., bio.	2	TF-IDF 2	59,0566%
N-GRAMAS	50	Sust.-Adj.-Verb.	Anat., med., bio.	1	TF-IDF 4	59,0425%
N-GRAMAS	50	Sust.-Adj.-Verb.	Anat., fact., med., bio.	1	TF-IDF 1	59,0345%
N-GRAMAS HÍBRIDOS	80	Sustantivos	Anat., fact., med., bio.	4	TF-IDF 2	59,0163%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Anat., med., bio.	2	TF-IDF 3	59,0123%
N-GRAMAS HÍBRIDOS	50	Sust.-Adj.-Verb.	Anat., fact., med., bio.	2	TF-IDF 2	58,9921%
N-GRAMAS	50	Sust.-Adj.-Verb.	Fact., med., bio.	1	TF-IDF 1	58,9861%
N-GRAMAS	50	Sust.-Adj.-Verb.	Anat., fact., med., bio.	1	TF-IDF 3	58,9821%

CONCLUSIONES

- Por los resultados obtenidos, se puede dar como conclusión final que la técnica y el modelo propuestos son adecuados para textos técnicos no estructurados sin importar el ámbito al cual estos estén referidos; sin embargo, el hecho de que los textos con los que se han realizado las pruebas fueron de longitud muy corta, no contribuyó en el rendimiento de la técnica y el modelo propuesto.
- La diferencia tan acentuada entre la densidad de elementos de los clusters clasificados por el médico experto, se identificó como un factor muy fuerte en la reducción de la tasa de aciertos en la experimentación.
- La concepción del modelo propuesto, se basa en el estudio y análisis de las áreas más relevantes del Procesamiento de Lenguaje Natural tales como: Information Retrieval, Information Extraction y Text Clustering, así como en investigaciones de estudios similares.
- Por lo resultados obtenidos se puede resaltar que, si bien existen estudios de Text Clustering para textos cortos (aunque escasos), las técnicas sugeridas para el caso de estudio no brindaron buenos resultados. A continuación se describen los hallazgos más resaltantes para cada uno de los procesos del modelo propuesto:
    - Casi la totalidad de errores ortográficos en textos de escritura libre se encuentran dentro de las palabras que se presentan como máximo 3 veces en todos los textos, y la mayor cantidad de estos errores están en las palabras que solo se presentan una vez en todos los textos, esto debido a la poca probabilidad de equivocarse al escribir una misma palabra de la misma forma más de una vez.
    - Basándose en la observación que “Las palabras más frecuentes son también las más frecuentemente mal escritas”, se concluye que la corrección ortográfica es un proceso que no se debería ignorar del modelo, y que en caso de no existir un corrector totalmente automático, la propuesta de corrector semiautomático es una buena opción.
    - A pesar de ser textos de escritura libre, el estilo de escritura de los expertos en el área suele tener patrones muy similares.

- Si bien los umbrales de podado más altos promediaron mejores resultados, los mejores resultados individuales fueron con los umbrales más bajos; esto debido a que la clasificación que hizo el médico experto dio como resultado clusters con cantidades de elementos muy disparejas; en consecuencia, las características necesarias para clusterizar de manera adecuada los grupos más pequeños, no tendrán una frecuencia muy alta; dicho esto se puede concluir que los clusters con mayor cantidad de elementos son más fáciles de identificar ya que sus patrones comunes son muy difíciles de despreciar, razón por la cual, en promedio, los umbrales más altos de podado dieron mejores resultados en promedio; sin embargo, con un umbral bajo de podado y las técnicas complementarias adecuadas para escoger bien las características comunes de los clusters pequeños (que además no perjudique la identificación de los clusters grandes), se puede obtener mejores resultados en este tipo de casos, como se observó dentro de los mejores resultados.
- La utilización de WordNet para la definición del Thesaurus probó tener resultados positivos, dentro de esto se puede resaltar que:
  - Dentro de las Categorías Gramaticales, los sustantivos representan las características más importantes dentro de los patrones comunes; sin embargo, la utilización de todas las Categorías Gramaticales también puede traer buenos resultados.
  - Dentro de los Dominios que se definan como relevantes en el contexto de los textos, no necesariamente la utilización de todos da un mejor resultado. En los resultados se puede ver que, a pesar de que el dominio de "anatomía" parezca ser un criterio importante, su inclusión no representa necesariamente un alza en los resultados, notando inclusive que la sola utilización de este dominio da resultados bajos.
- Respecto a la longitud de los gramas, se observó una performance altamente superior de las longitudes más bajas, esto era algo de esperarse por los antecedentes que hay en el estado del arte; sin embargo, se esperaba un comportamiento diferente en este caso al

tratarse de textos de tipo técnico, ya que los patrones a buscar deberían ser termino técnicos compuestos por 2 o más palabras; esto pudo haberse visto afectado por las características de los clusters mencionadas anteriormente, y debido a la corta longitud de los textos.

- Entre las técnicas de pesado para la indexación, se nota un equilibrio entre la técnica binaria y todas las variaciones del TF-IDF cuando se promedian todos los resultados; sin embargo, dentro de los mejores, resultados resalta ampliamente la presencia del TF-IDF, lo que muestra que el pesado binario, no necesariamente, sea suficiente para la representación de textos cortos.
- Dentro del clustering se mostraron 2 bajas importantes en técnicas que se esperaban tuvieran mejores resultados basándose en los antecedentes investigativos:
  - El Bisecting K-means mostró resultados muy inferiores a los K-means; esto pudo deberse a las características de los clusters explicadas anteriormente.
  - El coeficiente Jaccard no lograba converger en un resultado en muchas de las pruebas realizadas, esto debido a que la longitud de algunos de los textos es muy corta y los mismo no tenían ninguna representación dentro del espacio característico resultante, por lo que se obtenía un vector con todas sus características en 0, y como consecuencia se tiene que el coeficiente Jaccard, por sus características matemáticas, tiene dificultades en ser procesado.

- Gracias a la flexibilidad que proporciona la utilización de WordNet, el modelo propuesto puede ser reutilizable en el clustering de textos de cualquier ámbito solamente indicando los dominios relevantes para el ámbito de los textos de entre los dominios que provee WordNet.
- Si bien los N-Gramas Híbridos presentaron resultados similares a los N-Gramas normales, estos probaron ser mejores que la técnica en la que se basaron (SFM), debido a que al observar los resultados agrupados en cada una de las técnicas y umbrales, se nota una superioridad clara en los resultados de los N-Gramas Híbridos sobre los SFM. Un resultado muy

interesante es el de la tasa de aciertos para diferentes longitudes de Grama (Figura 5.5), en ella se nota una mejor tasa de acierto de los N-Gramas Híbridos sobre las demás técnicas a partir de una longitud de 2, además de que la tasa de aciertos no disminuye al aumentar la longitud como si lo hacen las otras técnicas; bajo estas observaciones se puede concluir que la generación de nuevas características de los N-Gramas Híbridos es superior ya que no desprecia características relevantes de longitud de Grama corta que no son tomadas en cuenta por las otras técnicas. Se considera que los N-Gramas Híbridos pueden tener un mejor desempeño que los N-Gramas en textos técnicos con una longitud mayor a los utilizados en el caso de estudio.

RECOMEDACIONES Y TRABAJOS FUTUROS

- Se recomienda hacer siempre una corrección ortográfica a los textos a clusterizar, ya sea con una técnica automática o semi-automática.
- Mejorar los algoritmos de stemming para palabras de carácter técnico.
- Buscar una mejora en la representación y búsqueda de patrones en textos de corta longitud.
- Buscar alternativas en el proceso de clustering que ayuden a lidiar con la corta duración de los textos versus un espacio característico de gran tamaño.
- Buscar una mejora en la representación y búsqueda de patrones para textos en los que los clusters resultantes tengan un número de elementos muy desbalanceado.
- Probar el desempeño de los N-Gramas Híbridos en textos con otras características y dominios, así como corroborar si estos darían buenos resultados en textos técnicos de longitud mediana o larga.

BIBLIOGRAFÍA Y REFERENCIAS

[1] MANNING, Christopher D., RAGHAVAN, Prabhakar, SCHÜTZE, Hinrich. Introduction to Information Retrieval. Nueva York, Cambridge University Press, 2008.

[2] HERSH, William. Information Retrieval: A Health and Biomedical Perspective. 3ª Ed. Nueva York, Springer, 2008.

[3] LUHN, Hans Peter. A statistical approach to mechanized encoding and searching of literary information. IBM Journal of Research and Development, 1: 309-317, 1957.

[4] SPAT, Stephan. Prototype of a Medical Information Retrieval System for Electronic Patient Records. Diploma Thesis. Graz, Austria. Graz University of Technology, Institute of Information Systems and Computer Media (IICM), 2007. 117.

[5] SEBASTIANI, Fabrizio. Machine learning in automated text categorization. ACM Computing Surveys, 34(1): 1-47, Marzo 2002.

[6] SOYSAL, Ergin, CICEKLI, Ilyas, BAYKAL, Nazife. Design and evaluation of an ontology based information extraction system for radiological reports. Computers in Biology and Medicine, 40(1): 900-911, Noviembre-Diciembre 2010.

[7] NIH: U.S. National Library of Medicine. MeSH: Medical Subject Headings. [en línea] <http://www.nlm.nih.gov/mesh/MBrowser.html> [consulta: 10 octubre 2014].

[8] SRUTHI, K. y VENKATESHWAR REDDY, B.. Document Clustering on Various Similarity Measures. International Journal of Advanced Research in Computer Science and Software Engineering, 3(8): 1269-1273, Agosto 2013.

[9] HUANG, Anna. Similarity measures for text document clustering. Proceedings of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC2008, Christchurch, New Zealand, 49-56, Abril 2008.

[10] LIU, Tao, LIU, Shengping, CHEN, Zheng, MA, Wei-Ying. An evaluation on feature selection for text clustering. En: Proceedings of the 20th International Conference on Machine Learning, Washington DC, 488-495, Agosto 2003.

[11] ANDREWS, Nicholas y FOX, Edward. Recent Developments in Document Clustering. Department of Computer Science, Virginia Tech, Virginia, Octubre 2007.

[12] SEDDING, Julian y KAZAKOV, Dimitar. WordNet-based text document clustering. En: Proceedings of the 3rd Workshop on Robust Methods in Analysis of Natural Language Data, Pensilvania, 104-113, 2004.

[13] ELBERRICHI, Zakaria, RAHMOUN, Abdelattif, AMINE BENTAALAH, Mohamed. Using WordNet for Text Categorization. The International Arab Journal of Information Technology. 5(1): 16-24, Enero 2008.

[14] Universidad de Princeton. WordNet. [en línea] <http://wordnet.princeton.edu/> [consulta: 10 octubre 2014].

[15] LI, Y. H. y JAIN, Anil K. Classification of text documents. The Computer Journal, 41(8): 537-546, 1998.

[16] AGGARWAL, Charu y ZHAI, ChengXiang. A Survey of Text Clustering Algorithms. En su: Mining Text Data. Nueva York, Springer, 2012. 77-128.

[17] YANG, Yiming y PEDERSEN, Jan. A Comparative Study on Feature Selection in Text Categorization. En: Proceedings of the Fourteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 412-420. 1997.

[18] BERRY, Michael y CASTELLANOS, Malu. Survey of Text Mining II: Clustering, Classification, and Retrieval. Londres, Springer, 2008.

[19] FÜRNKRANZ, Johannes. A Study Using n-gram Features for Text Categorization. Austrian Research Institute for Artificial Intelligence, 1998.

[20] RAHMOUN, Abdellatif y ELBERRICHI, Zakaria. Experimenting N-Grams in Text Categorization. The Internacional Arab Journal of Information Technology, 4(4): 377-385, Octubre 2007.

[21] TAN, Chade-Meng, WANG, Yuan-Fang, LEE, Chan-Do. The Use of Bigrams to Enhance Text Categorization. Inf. Process. Manage, 38(4): 529-546, Julio 2002.

[22] Vincent Yip, Mutlu Mete, Umit Topaloglu, Sinan Kockara. Concept discovery for pathology reports using an n-gram model. AMIA Summit on Translational Bioinformatics, 43-47, Marzo 2010.

[23] GARCÍA BLASCO, Sandra. Extracción de secuencias maximales de una colección de textos. Tesis (Bachiller en Ingeniería Técnica en Informática de Gestión). Valencia, España. Universidad Politécnica de Valencia. 2009. 47.

[24] ORTA PALACIOS, Claudia Patricia. Métodos Basados en Patrones Léxicos para la Extracción de Información. Tesis (Maestría en Ciencias en la Especialidad de Ciencias Computacionales). Puebla, Mexico. 2008. 126.

[25] RANGREJ, Aniket, KULKARNI, Sayali, TENDULKAR, Ashish. Comparative study of clustering techniques for short text documents. En: Proceedings of the 20th international conference companion on World wide web, Nueva York, ACM, 2011, 111-112.

[26] TAN, Pang-Ning, STEINBACH, Michael, KUMAR, Vipin. Cluster Analysis: Basic Concepts and Algorithms. En su: Introduction to Data Mining. Boston, Addison-Wesley Longman Publishing Co., Inc., 2005. 487-568.

[27] STEINBACH, Michael, KARYPIS, George, KUMAR, Vipin. A comparison of document clustering techniques. En: KDD Workshop on Text Mining. Minnesota, 2000.

[28] CHA, Sung-Hyuk. Comprehensive Survey on Distance Similarity Measures between Probability Density Functions. International Journal of Mathematical Models and Methods in Applied Sciences, 1(4): 300-307, 2007.

[29] ROSELL, Magnus. Introduction to Information Retrieval and Text Clustering. Estocolmo, KTH CSC, 2006.

[30] LLAMOSAS LAZO, Brayan, LÓPEZ DEL ÁLAMO, Cristian, BELTRÁN CASTAÑÓN, César. Nuevo modelo para encontrar patrones en diagnósticos médicos de escritura libre. En: IV Simposio Peruano de Inteligencia Artificial SPIA2011. Arequipa, Diciembre 2011.

[31] Universidad de Neuchâtel. Stop Word List (Negative Dictionary) [en línea] <http://members.unine.ch/jacques.savoy/clef/spanishSmart.txt> [consulta: 10 octubre 2014].

[32] Universidad Politécnica de Cataluña, Center for Language and Speech Technologies and Applications (TALP). WordNet. [en línea] <http://adimen.si.ehu.es/web/MCR> [consulta: 10 octubre 2014].

[33] FBK – Fondazione Bruno Kessler. MultiWordNet. [en línea] <http://multiwordnet.fbk.eu/english/home.php> [consulta: 10 octubre 2014].

[34] LIDDY, Elizabeth. Natural Language Processing. En: DRAKE, Miriam. Encyclopedia of Library and Information Science. 2ª Ed. Nueva York, Marcel Dekker, 2003. 2126-2136.

[35] KAO, Anne y POTEET, Stephen. Natural Language Processing and Text Mining. Washington, Springer 2006.

[36] BETTS, Tom, MILOSAVLJEVIC, Maria, OBERLANDER, Jon. Advances in Information Retrieval. Roma, Springer, 2007. 295-306.

APÉNDICE A
GLOSARIO DE TERMINOS

- **Information Retrieval (IR):** Es un campo de intersección entre las ciencias de la información y las ciencias de la computación, la IR se ocupa de la indexación y recuperación de información de fuentes heterogéneas, en su mayoría textos.
- **Information Extraction (IE):** Esta área está dedicada a la extracción de información relevante de los textos. Esta información es extraída a partir de las relaciones semánticas que se encuentran a las cuales se les dan una interpretación de acuerdo al contexto.
- **Text Clustering (TC):** Es un área perteneciente al Text Mining. Trata de todo el procedimiento de clustering de textos, desde su pre-procesamiento hasta la obtención de los clusters.
- **Negative Dictionary – Stop Words:** Los Stop Words son palabras con un valor semántico muy pobre. Se caracterizan por tener una alta frecuencia en todo tipo de textos. El Negative Dictionary es un conjunto de Stop Words que se toman para textos de un entorno específico.
- **Stemming:** Es el proceso por el cual se llevan las palabras a su forma raíz o lexema suprimiendo todos sus afijos (prefijos, sufijos e infijos) dependiendo del lenguaje del que se trate.
- **Thesaurus:** Es un diccionario de palabras o términos relevantes para un tipo de textos especifico.
- **Vector Característico:** Un modo de representación de un algo del mundo real. Esta representación se hace creando un vector de números para cada objeto de la clase de objetos que estemos representando, esto facilita su comparación y clasificación automática.
- **Espacio Característico:** Es lo que representa un vector característico en cada una de las posiciones de su vector, por ejemplo: para un texto las posiciones pueden representar la cantidad de repeticiones de las palabras presentes en un thesaurus predefinido, para una imagen pueden representar la cantidad de pixeles de un color determinado.

- **Clustering:** Es la clasificación de objetos sin tener conocimiento a priori de las clases de objetos que existen. Esta clasificación agrupa los elementos por su mayor parecido entre sí.

APÉNDICE B

NLP (NATURAL LANGUAGE PROCESSING – PROCESAMIENTO DE LENGUAJE NATURAL)

Empecemos por ver dos definiciones de lo que es NLP, en [34] se define como "El Procesamiento de Lenguaje Natural es una amplia rama teórica de técnicas computacionales para el análisis y la representación de textos de origen natural en uno más niveles de análisis lingüístico con el propósito de lograr el procesamiento del lenguaje humano para una serie de tareas o aplicaciones". Según [35] es "NLP es el intento de extraer una representación con sentido más completo de un texto libre."

Como se puede ver, las definiciones abarcan una gran cantidad de temas, tratando de generalizar su definición para que esta sea flexible por la gran cantidad de campos que se pueden desagregar de ella [34].

B.1. Objetivos:

Existen distintas opiniones sobre los objetivos concretos que persigue el NLP, pero de nuevo, esto varía mucho por la gran envergadura que tiene. Para [34] los objetivos son los siguientes:

- Parafrasear un texto.
- Traducir un texto de un lenguaje a otro.
- Responder preguntas sobre el contenido de un texto.
- Realizar inferencias de un texto.

B.2. Niveles:

Se refiere a los niveles del procesamiento del lenguaje natural o procesamiento lingüístico y se utilizan para un mejor entendimiento de como el NLP desarrolla y basa sus técnicas y teorías en el proceso lingüístico en si [34]:

- **Fonología:** Se refiere a la interpretación de los sonidos del habla dentro ya través de las palabras. Hay, de hecho, tres tipos de reglas que se utilizan en el análisis fonológico [34]:

  - **Reglas fonéticas:** para los sonidos dentro de las palabras.
  - **Reglas fonológicas:** para las variaciones de la pronunciación cuando las palabras se hablan juntas.
  - **Reglas prosódicas:** por fluctuación en acento y la entonación a través de una oración.

- **Léxico:** Se refiere a la interpretación de cada una de las palabras por separado. En este nivel se puede definir un "lexicón" el cual es como un diccionario en el que se definen todos los elementos de una oración (verbo, sustantivo, adjetivo, etc.) que puede representar una palabra; este lexicón puede ser tan complejo como se quiera, puede adoptar información semántica de las palabras (cuantos argumentos puede tomar la palabra, sus limitaciones). Con la ayuda del lexicón se puede deducir que elemento de una oración puede representar una palabra según su contexto, en caso que la palabra pueda representar más de un elemento [34].
- **Sintáctico:** Analiza las palabras como parte de la oración, el orden en el que se presentan y la relación de dependencia estructural que pueden guardar un conjunto de palabras. En los siguientes 2 ejemplos se podrá apreciar la importancia del orden de las palabras a pesar de contener las mismas palabras [34]:
  - El perro muerde al niño.
  - El niño muerde al perro.
- **Semántico:** Se centra en determinar el significado de las palabras y con ellas determinar el posible significado de una oración basándose en la relación entre el significado de las palaras de la oración. Un desafío en este punto es el WSD (Word Sense Disambiguation – Desambiguación del sentido de la palabra), cuando una palabra tiene más de un significado se debe determinar cuál de ellos es el correcto para la correcta interpretación de la oración [34].
- **Discurso:** Analiza propiedades de texto entero que ayudan a encontrar su significado haciendo conexiones entre componentes de las oraciones. Los dos procesos de discurso más comunes son [34]:

  - Resolución de la Anáfora: Reemplazar elementos de la oración, como los pronombres, por las entidades a las que hace referencia.
  - Reconocimiento de la estructura del discurso/texto: Determina la función de las oraciones en el texto para así poder determinar que estructura tiene este, por ejemplo un periódico tendrá una estructura similar a la siguiente: historia principal, eventos anteriores, citas, etc.

- **Pragmático:** Este nivel trata de cómo interpretar un texto que utiliza el contexto como referencia, el objetivo del texto para expresar algo, que utiliza muchas anáforas; en otras palabras, como entender lo que el texto no dice directamente, como entender lo que dice suponiendo que el lector va comprendiendo lo que dice anteriormente [34].

Los sistemas actuales de NLP tienden a ir a los niveles más bajos y esto se debe a las siguientes razones [34]:

- Las aplicaciones no requieren de la interpretación de los niveles más altos.
- Los niveles más bajos han sido investigados e implementados mucho más.
- Los niveles más bajos tratan con unidades de análisis más pequeñas (morfemas, palabras y oraciones) que ya tienen reglas establecidas; por otro lado los niveles más altos tratan con el texto entero y con conocimiento en general que no tienen una buena base de reglas para su tratamiento.

B.3. Aplicaciones:

Las aplicaciones más frecuentes de NLP son las siguientes [34]:

- **Information Retrieval (IR):** A pesar de ser una aplicación de tanta relevancia, no muchas implementaciones importantes utilizan NLP.
- **Information Extraction (IE):** Un área de aplicación más reciente que se centra en el reconocimiento, etiquetado y obtención de una representación estructurada del texto. Esa estructura puede ser utilizada

por otras aplicaciones como question-answering, visualization y data mining.

- **Question-Answering:** En contraste al IR que te da una lista de documentos potencialmente relevantes ante una consulta, el question-answering te da únicamente un texto con la respuesta.
- **Sumarization:** Los niveles más altos del NLP, en particular el discurso, pueden ayudar a resumir textos.
- **Machine Translation:** Quizás la aplicación más antigua de todas las NLP, que va por todos los niveles del NLP, desde los más bajos con traducciones por palabra hasta aplicaciones con los más altos niveles.
- **Dialogue Systems:** Probablemente la aplicación omnipresente en el futuro para aplicaciones que interactúen con el usuario. Se cree que el aprovechamiento de todos los niveles del NLP ofrece la potencialidad de un verdadero dialogo entre dispositivo y humano.

APÉNDICE C

MODELOS DE IR

Existen tres modelos "clásicos" de IR [4]:

- El modelo booleano (Boolean Model).
- El modelo del espacio vectorial (Vector Space Model).
- El modelo probabilístico (Probabilistic Model).

"En estos modelo, las keywords (también llamadas index terms) describen el contenido del documento. Por esta razón, es fundamental que los index terms describan el contenido específico del documento. Por lo tanto, las palabras que aparecen en casi todos los documentos no tienen importancia para el proceso de IR, mientras que las palabras infrecuentes son significantes. Por consiguiente, se le asigna un peso numérico a cada index term. Este peso muestra que tan bien un index term k_i describe el contenido semántico de un documento d_j." [4]. En la figura C.1 se muestra el proceso base para los modelos de IR.

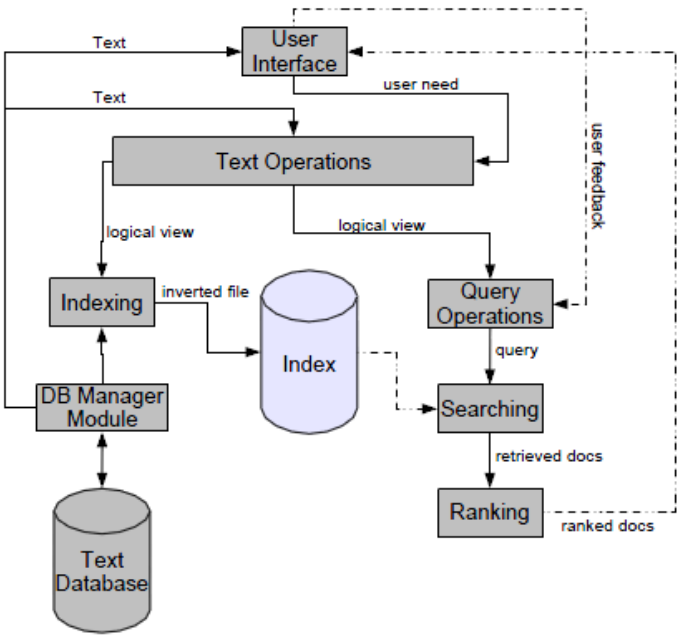

Figura C.1: Ilustración del proceso de IR [4].

Para el éxito de un modelo de IR, los siguientes puntos son necesarios [4]:

- Extraer los index terms más relevantes de los documentos.
- Asignar un peso significante a los index terms.
- Calcular la similitud entre los documentos y la consulta del usuario.

C.1. Set-theoretic Model:

Está basado en el modelo booleano. Los pesos de los index terms son booleanos (0 o 1), asimismo la respuesta a si un documento es relevante o no para una consulta de usuario es booleana (si es relevante o no es relevante). Sus principales ventajas y desventajas se muestran en la tabla C.1.

Tabla C.1: Ventajas y desventajas del modelo booleano [4].

Ventajas	Desventajas
- Fácil de entender - Exacto - El lenguaje de consulta es expresivo	- No hay coincidencias parciales - Los documentos recuperados no están rankeados - Una representación de "bag-of-words" podría no considerar la semántica de los documentos

C.2. Modelo Algebraico:

Está basado en el modelo del espacio vectorial el cual es el modelo más común en IR. Los index terms son representados en un vector, y utilizando este vector se puede usar una distancia métrica para calcular la similitud de un documento con una consulta [4].

Los pesos asignados a los index terms son números reales entre 0 y 1, para la asignación de estos pesos en necesario escoger un método, por ejemplo el esquema tf-idf, el cual da un alto peso a index terms que sean frecuentes en documentos relevantes y que sean poco frecuentes en documentos no relevantes. La similitud entre el vector del documento y de la consulta se puede calcular con una distancia algebraica, por ejemplo "ángulo del coseno" [4]. En la tabla C.2 se puede ver las ventajas y desventajas de este modelo.

Tabla C.2: Ventajas y desventajas del modelo espacio vectorial [4].

Ventajas	Desventajas
- Fácil de entender - Coincidencias parciales, documentos ordenados por ranking - Usa esquemas de pesados de index terms	- Esfuerzo para calcular la similitud - Una representación de "bag-of-words" podría no considerar la semántica de los documentos

C.3. Modelo Probabilístico:

Este modelo se basa en las probabilidades que cada documento sea relevante o no para una consulta y necesita de una retroalimentación por parte del usuario para refinar esta probabilidad [4].

Los pesos de los index terms son booleanos (0 o 1). Al inicio se determina arbitrariamente la probabilidad de que documentos son relevantes o no para una consulta. En base a eso se define una función de probabilidad de que un documento sea relevante para una consulta, y con esta función se aplica la siguiente fórmula [4]:

$$sim(d_j, q) \sim \sum_{i=1}^{t} \omega_{i,q} \times \omega_{i,j} \times (\log \frac{P(k_i|R)}{1 - P(k_i|R)} + \log \frac{1 - P(k_i|\overline{R})}{P(k_i|\overline{R})}) \quad \text{(C.1)}$$

Dónde:

t: cantidad de index terms.

d_j: Documento a evaluar.

q: Consulta.

R: Conjunto de documentos relevantes.

$\bar{R}$: Conjunto de documentos no relevantes.

$w_{i,q}$: Peso de index term i para la consulta.

$w_{i,j}$: Peso de index term i para el documento a evaluar.

$P(K_i|R)$: Probabilidad de que el index term Ki esté presente en un documento de R.

$P(K_i|\bar{R})$: Probabilidad de que el index term Ki no esté presente en un documento de R.

Se utiliza la fórmula anterior para calcular la similitud de los documentos con las consultas, de este modo obtenemos un ranking de los documentos más relevantes, los cuales son devueltos al usuario para que indique si los resultados son correctos. Con esta retroalimentación, de ser necesario, se modifica la función de probabilidad para que los resultados se ajusten a lo que se deben obtener [4]. En la tabla C.3 se muestran las ventajas y desventajas de este modelo.

Tabla C.3: Ventajas y desventajas del modelo probabilístico [4].

Ventajas	Desventajas
- Documentos rankeados por su relevancia	- Estimación inicial de los documentos relevantes y no relevantes - Modelo binario (pesos binarios) - Se asume que los index terms son independientes - Una representación de “bag-of-words” podría no considerar la semántica de los documentos

C.4. Modelo basado en Ontologías:

Este modelo utiliza una ontología basada en el dominio del conocimiento de los textos que se estén tratando para la recuperación de información importante, ya teniendo la información semántica más relevante de cada texto esta se utiliza como los index terms para aplicar y un método de pesado y su posterior cálculo de distancia con la consulta como se hace en el Modelo Algebraico [4]. En la tabla C.4 se muestran las ventajas y desventajas de este modelo.

Tabla C.4: Ventajas y desventajas del modelo basado en ontologías [4].

Ventajas	Desventajas
- Documentos rankeados por su relevancia - Se considera la semántica de los documentos - Este modelo supera a los modelos "clásicos" de la IR si se tiene una base de conocimiento adecuada.	- Se necesita un gran esfuerzo para construir la base de conocimiento

APÉNDICE D

IE (INFORMATION EXTRACTION – EXTRANCCIÓN DE INFORMACIÓN)

Según [6] IE es: "Una sub-disciplina del NLP enfocada en la identificación de datos específicos y relaciones dentro de textos no estructurados, la extracción de los valores relevantes, y su transformación en códigos estandarizados y/o información estructurada". Podemos encontrar una definición similar en [36]: "Es un proceso por el cual nosotros podemos identificar estructura dentro de un texto de lenguaje natural no estructurado".

Los textos médicos suelen ser más sencillos de procesar, desde el punto de vista gramático, debido a su naturaleza [6]:

- Como otros textos técnicos, utilizan un subconjunto limitado de la lengua con un número limitado de tipos de información.
- Terminología relativamente no ambigua.
- Presenta patrones predecibles.

IE tiene dos tareas básicas [6]:

- **Reconocimiento de entidades (NER – Named Entity Recognition):** Trata de identificar los límites del texto que representan entidades en un texto, por ejemplo: buscar todas las vitaminas que aparecen en el texto [6] [36].
- **Extracción de relaciones:** Trata de identificar las relaciones entre las entidades [6].

IE puede usar ontologías (exhaustivo y riguroso esquema conceptual de un dominio dado, contiene entidades relevantes y sus relaciones; estrechamente relacionado con los vocabularios fijos) como un esquema de estructura genérica de los textos a examinar para el reconocimiento de entidades y extracción de relaciones debido a que ha probado ser mucho más eficiente que su reconocimiento solo por los ítems del texto [6].

Mientras menos dominio del conocimiento se abarque y menos información gramatical se maneje, el sistema será más certero. Existen 2 enfoques en la IE [6]:

- **Método supervisado:** (Knowledge Engineering Approach): Las reglas son dadas por un experto en el área; la efectividad del sistema se verá limitada por el dominio que tenga el experto, aunque suele tener una mejor performance que el método no supervisado. Las principales desventajas son la necesidad del experto y la dificultad del sistema para adaptarse a un nuevo dominio.
- **Método no supervisado ó semi-supervisado (Automatic training Approach):** Crea un método de clasificación basado en un conjunto de entrenamiento. Las salidas del sistema, después de haber sido entrenado, pueden ser revisadas para su corrección; esta corrección resulta costosa dado que es necesaria la presencia de un experto y la modificación de reglas y plantillas (templates).

En [6] se realizó un el desarrollo de una ontología para el agrupamiento de diagnósticos médicos estomacales en turco, el proceso de dicha implementación de IE es como se muestra en la figura D.1.

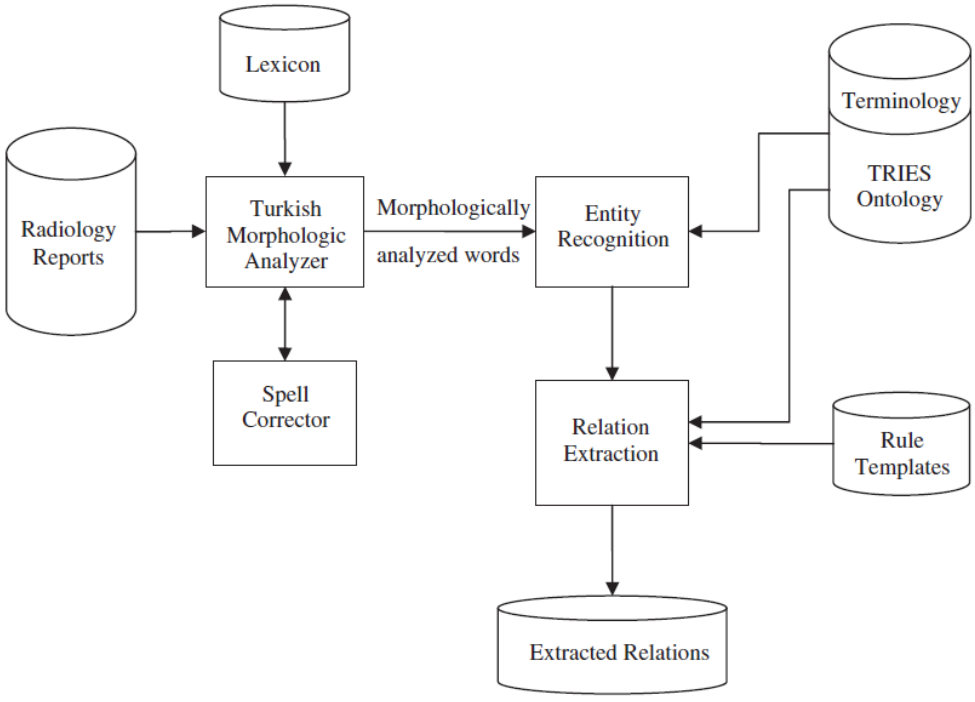

Figura D.1: Componentes del sistema de IE radiológico Turcos (TRIES) [6].

Todas las palabras de los reportes son analizadas por el "analizador morfológico", el cual con la ayuda del "lexicón", que contiene todas las raíces o lexemas de las posibles palabras que se presenten en este tipo de reportes, trata de encontrar el lexema de cada palabra y sus posibles morfemas. De no encontrar la forma base (lexema) en el "lexicón", se le envía al "spell corrector" para corregir un posible error de tipeo; luego se envía la palabra corregida al "lexicón" [6].

El "entity recognition" (NER – primera tarea básica del IE) se encarga de reconocer las secuencias de lexemas, sin sus morfemas, que tengan significado en conjunto, clasificándolas como conceptos ontológicos, atributos o valores de atributos. El resto de los morfemas se fusionan y se unen al término recién identificado (como modificadores) para su posterior análisis en el "rule extraction". Las entidades que pueden ser consideradas términos están contenidas en la "terminology" [6].

Después se pasan las oraciones por *el* "relation extraction" (segunda tarea básica del IE) para verificarlas mediante el "rule templates" el cual saca su información semántica en forma de relaciones. El "rule templates" se diseña en base a los "elementos conceptuales ontológicos" para hacer reglas más flexibles, además se incluyen expresiones regulares del lenguaje [6]. En la figura D.2 vemos el proceso de IE que se acaba de explicar aplicado a un pequeño texto.

Application of TRIES to a sample sentence. (POSS3SG: possessive suffix for 3rd singular person, NESS: -ness suffix, COP: copula).

Text
Karaciğer vertikal uzunluğu 14 cm'dir.
The height of liver is 14 cm.
Morphological analysis
Karaciğer vertikal uzun+NESS+POSS3SG 14 cm+COP
Liver vertical tall+NESS+POSS3SG 14 cm+COP
Named entity recognition
[Karaciğer] [vertikal uzun+NESS] +POSS3SG [14 cm] +COP
[entity:Liver] [attribute:height]+POSS3SG [value:NUMERIC: 14 cm] +COP
Relation extraction—rule to be matched, and rule constraints to be satisfied:
〈VisibleStructure O〉〈O:Attribute A〉 +POSS3SG 〈O:A:Value V〉 +COP
obj_has_attribute(Object, Attribute) – (Liver, height)
obj_attribute_accept_value(Object, Attribute, Value) – (Liver, height, 14 cm)
Extracted relation
Liver.height = 14 cm

Figura D.2: Aplicación de TRIES en una oración de ejemplo [6].

En el árbol de entidades las ontologías implementan 2 relaciones. La primera, que es el esqueleto de la ontología del árbol, es una relación tipo "es un" que se refiere a los atributos que están relacionados a cada entidad, está estrechamente relacionado con el modelo de información de la información extraída. La segunda relación es de padre a hijo como se ve en la figura D.3, en esta se hace una generalización de modo que modelo de información de la ontología se hace más simple, además cumple un rol importante en la validación de las limitaciones de las reglas [6].

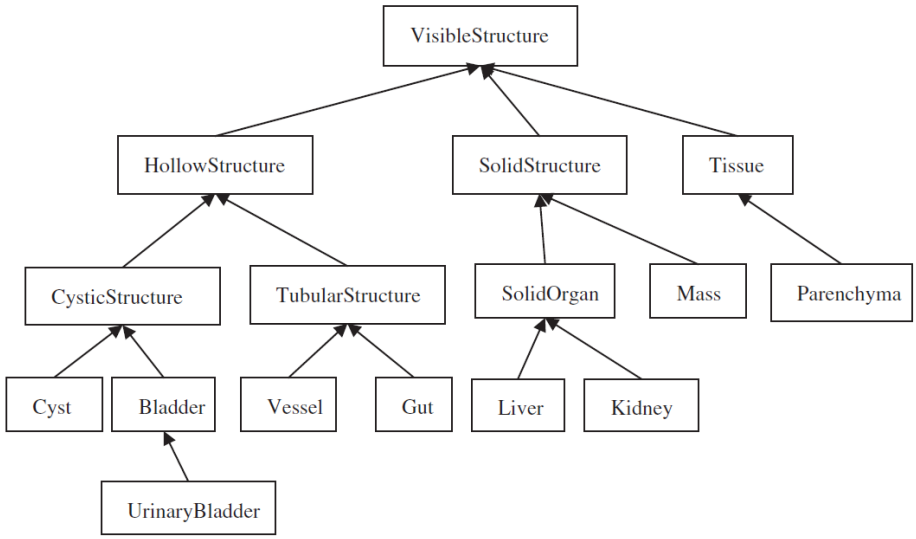

Figura D.3: Fragmento de la ontología TRIES diseñada usando Protegé. VisibleStructure es el padre de todas las otras entidades [6].