UNIVERSIDAD DE CONCEPCIÓN

FACULTAD DE INGENIERÍA

DEPARTAMENTO DE INGENIERÍA ELÉCTRICA

Profesor Patrocinante:

Dr. Rosa L. Figueroa I.

Informe de Memoria de Título
para optar al título de:

Ingeniero Civil Biomédico.

Clasificación de Artículos Científicos

Concepción, Marzo del 2015

Vanessa Del Rosario Andrade Alvarado

UNIVERSIDAD DE CONCEPCIÓN
Facultad de Ingeniería
Departamento de Ingeniería Eléctrica

Profesor Patrocinante:
Dr. Rosa L. Figueroa I.

Clasificación de Artículos Científicos

Vanessa Del Rosario Andrade Alvarado

Informe de Memoria de Título
para optar al título de

Ingeniero Civil Biomédico

Marzo 2015

Resumen

El objetivo del presente estudio es diseñar un método de identificación y extracción de tópicos que utilice como datos de entrada los abstract presentes en los documentos científicos adquiridos de una base de datos de BioMed Central. En este documento se describen las distintas etapas de este método que van desde el proceso de filtración de los datos hasta la implementación del clasificador no supervisado para su posterior evaluación.

El problema de clasificación presente en éste estudio abarca principalmente el exceso de información de la base de datos de Biomed Central, y a la vez los variados temas que abarcan el área de investigación científica, tal como estudios en medicina, cardiología, endocrinología, otorrinolaringología, broncopulmonar, etc. A partir de estos documentos y bajo la utilización de distintos métodos del procesamiento del lenguaje natural, se busca identificar tópicos para conocer de este modo el nivel de información que contiene la base de datos analizada. Las clases contempladas para este proceso de clasificación no supervisada corresponden a las clases “lung”, “kidney-heart” y “lung-kidney-heart”.

La metodología utilizada contempla el pre-procesamiento de los textos presentes en cada uno de los documentos en estudio, donde se utilizaron herramientas del procesamiento del lenguaje natural para poder normalizar y segmentar cada uno de los abstract científicos, con el objetivo de crear dos diccionarios de palabras claves. El primer diccionario se creó utilizando la biblioteca Topia de Python y el segundo utilizando reconocimiento de entidades (NER). Ambos diccionarios utilizados como entrada en la implementación del modelo de extracción de tópicos.

El clasificador utilizado en este estudio es Latent Dirichlet Allocation, el cual bajo la utilización de la biblioteca de Gensim de Python será el encargado de realizar la clasificación no supervisada de las clases mencionadas.

Los resultados obtenidos en la extracción de tópicos para cada clase seleccionada que contenían información acerca de pulmón, riñón y corazón nos dicen que para la clase “lung”, utilizando ambos diccionarios, los temas en estudio tienen algún tipo de información acerca de patologías, síntomas, casos de estudios en relación con el pulmón. Luego, para la clase “kidney-heart”, al igual que la clase anterior, los temas en estudio hacen mención al riñón y corazón, al

contener cada tópico términos como hipertensión, acute kidney injury (aki), coronary heart disease (chd), entre otras. Finalmente, la clase “lung-kidney-heart” entrega como resultados tópicos que tienen algún tipo de información acerca de pulmón, riñón, corazón. Adicionalmente, se descubren nuevos tópicos que hacen alusión a temas relacionados al sistema nervioso.

La evaluación del modelo se hizo de forma manual, tomando para cada clase 500 documentos. La medida de evaluación fue la precisión. Los resultados obtenidos por el modelo de clasificación no supervisada, Latent Dirichlet Allocation obtiene un 72% y 55.2% de precisión para la clase “lung”, un 75.8% y 57% de precisión para la clase “kidney-heart”, un 40.8% y 33.6% de precisión para la clase “lung-kidney-heart”. Los valores de precisión entregan el porcentaje de casos correctamente clasificados del total de 500 documentos evaluados según lo determinado por el programador. Es importante destacar que la mayor cantidad de diferencias en los resultados obtenidos en la clasificación están asociadas principalmente al tipo de segmentación utilizada en la creación de los dos diccionarios.

Finalmente, se tiene un modelo que da como resultado una serie de documentos clasificados a través de tópicos encontrados, esta aplicación informará al usuario de qué respuestas se asemejan más a los documentos seleccionados por el usuario.

La fe es una esperanza en aquello que no se ve y que es verdadero
Alma 32:27

Agradecimientos

Con esta Memoria de Título se da término a una etapa fructífera, llena de esfuerzo, dedicación y, por sobre todo, rebosadas de aprendizajes en lo académico y personal, la etapa universitaria.

Agradezco a todos los que formaron parte de ella, a Dios y a mi familia, en especial a mis padres por confiar en mí, a cada uno de mis hermanos, por su apoyo incondicional.

A las personas que hicieron que la ausencia de la familia por estudiar lejos de casa se volviera más fácil y alegre, a cada uno de los amigos.

A los académicos, a mi profesor guía.

A todos gracias.

Tabla de Contenidos

LISTA DE TABLAS ..... **IX**

LISTA DE FIGURAS ..... **X**

ABREVIACIONES ..... **XI**

CAPÍTULO 1. INTRODUCCIÓN **1**

1.1. INTRODUCCIÓN GENERAL 1
1.2. OBJETIVOS 2
1.2.1 Objetivo general *2*
1.2.2 Objetivos específicos *2*
1.3. ALCANCES Y LIMITACIONES 3
1.4. TEMARIO 3

CAPÍTULO 2. ESTADO DEL ARTE Y MARCO TEÓRICO **5**

2.1. ESTADO DEL ARTE 5
2.1.1 Introducción *5*
2.1.2 Modelación de Tópicos *5*
2.2. MARCO TEÓRICO 10
2.2.1 INTRODUCCIÓN 10
2.2.2 Aprendizaje (Supervisado VS no supervisado) *10*
2.2.3 Extracción de características *12*
2.2.4 Modelo de Tópicos *14*

CAPÍTULO 3. MATERIALES Y MÉTODOS **18**

3.1. MATERIALES 18
3.1.1 INTRODUCCIÓN. 18
3.1.2 BASE DE DATOS 18
3.2. METODOLOGÍA 20
3.2.1 INTRODUCCIÓN. 20
3.2.2 SELECCIÓN DE ARTÍCULOS DE INTERÉS. 20
3.2.3 PRE-PROCESAMIENTO DEL TEXTO. 21
3.2.3.1 Eliminación de palabras vacías. *22*
3.2.3.2 Normalización del texto. *22*
3.2.3.3 Segmentación. *22*
3.2.3.4 Lematización. *22*
3.2.3.5 Palabras claves. *23*
3.2.4 IDENTIFICADORES DE ENTIDADES NOMBRADAS (NER) 24
3.2.5 Modelado de Tópicos *25*
3.2.6 Evaluación del Modelo *28*

CAPÍTULO 4. RESULTADOS **31**

4.1. IMPLEMENTACIÓN 31
4.1.1 INTRODUCCIÓN 31
4.1.2 ADQUISICIÓN DE LA INFORMACIÓN 31
4.1.3 DESARROLLO DEL ALGORITMO 32
4.2. RESULTADOS 35
4.2.1 INTRODUCCIÓN 35
4.2.2 MODELADO DE TÓPICOS POR CLASES 36
4.2.2.1 CLASE "LUNG" 36
4.2.2.2 CLASE "KIDNEY -HEART" 41
4.2.2.3 CLASE "LUNG-KIDNEY-HEART" 47

CAPÍTULO 5. CONCLUSIONES **53**

5.1. CONCLUSIÓN 53

5.2. TRABAJO FUTURO. 55
ANEXO. CONJUNTO DE ETIQUETAS PENN TREEBANK [EXTRAÍDO DE [18]] 58
BIBLIOGRAFÍA 56
ANEXO…… 58

Lista de Tablas

Tabla 1: Contenido de base de datos 19
Tabla 2: Matriz de confusión para las distintas clases a modelar 30
Tabla 3: Clases a modelar 32
Tabla 4: Términos identificados en el análisis de la modelación de tópicos para los documentos de la clase "lung" 37
Tabla 5: Términos identificados en el análisis de la modelación de tópicos para los documentos de la clase "lung" 38
Tabla 6: Evaluación clase "lung" 40
Tabla 7: Evaluación clase "lung" 41
Tabla 8: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase "kidney-heart" 42
Tabla 9: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase "kidney-heart". 44
Tabla 10: Evaluación clase "kidney-heart" 45
Tabla 11: Evaluación clase "kidney-heart" 46
Tabla 12: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase "lung-kidney-heart" 48
Tabla 13: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase "lung-kidney-heart". 49
Tabla 14: Evaluación clase "lung-kidney-heart" 51
Tabla 15: Evaluación clase "lung-kidney-heart" 52

Lista de Figuras

Figura 1: Modelo LDA [extraído de [10]].....16
Figura 2: Esquema de uso del LDA 17
Figura 3: Diagrama EER. [Extraído de MySQL Workbench] 19
Figura 4: Arreglo asociativo.....26
Figura 5: Representación de una lista de listas.....27
Figura 6: Asociación de las palabras en el texto [extraído de [5]] 28
Figura 7 : Grado de información de los Abstract 32
Figura 8: Etiquetado POS en una oración. 34
Figura 9: Entidad gramatical reconocida.....34
Figura 10: Resultados modelo LDA, usando el diccionario n°1 para los documentos que pertenecen a la clase “lung”.....36
Figura 11: Resultados modelo LDA, usando el diccionario n°2 para los documentos que pertenecen a la clase “lung”.....38
Figura 12: Resultados modelo LDA.....40
Figura 13: Resultados modelo LDA.....40
Figura 14: Resultados modelo LDA, usando el diccionario n°1 para los documentos que pertenecen a la clase “kidney-heart” 42
Figura 15: Resultados modelo LDA, usando el diccionario n°2 para los documentos que pertenecen a la clase “kidney-heart” 43
Figura 16: Resultados modelo LDA.....45
Figura 17: Resultados modelo LDA.....46
Figura 18: Resultados modelo LDA, usando el diccionario n°1 para los documentos que pertenecen a la clase “lung-kidney-heart” 47
Figura 19: Resultados modelo LDA, usando el diccionario n°2 para los documentos que pertenecen a la clase “lung-kidney-heart” 49
Figura 20: Resultados modelo LDA.....51
Figura 21: Resultados modelo LDA.....52

Abreviaciones

Mayúsculas

TM	: Topic Model.
BMC	: BioMed Central
IE	: Information Extraction
IR	: Information Retrieval
NLP	: Natural Languaje Processing
LDA	: Latent Dirichlet Allocation
LSI	: Latent Semantic Indexing
LSA	: Latent Semantic Analysis
SVD	: Singular Value Decomposition
HMM	: Hidden Markov Model
SVM	: Support Vector Machine
POS	: Part-Of-Speech
NER	: Named Entity Recognition
NLTK	: Natural Language Toolkit
FDA	: Food and Drug Administration
ADN	: Acido Desoxirribonucleico

Minúsculas

egfr	: epidermal growth factor receptor
chd	: coronary heart disease
aki	: acute kidney injury
cpb	: cardiopulmonary bypass
nsclc	: non-small-cell lung cancer
ci	: confidence interval
chd	: coronary heart disease
snp	: single nucleotide polymorphism

Capítulo 1. Introducción

1.1. Introducción general

Con la literatura biomédica aumentando a un ritmo de varios de miles de documentos por semana es imposible para los investigadores estar al día con todas las publicaciones. Es por esto que para poder abarcar el creciente número de documentos, se han desarrollado métodos informáticos de clasificación automática de documentos para diversas tareas que se desarrollan en la sociedad contemporánea, en especial en el área del aprendizaje e investigación, los cuales tienen como objetivo seleccionar artículos pertinentes a un tema específico de grandes corpus y extraer información de interés para el investigador. [1]

El procesado y clasificación automático de textos es un área de investigación formada por diversas disciplinas. Estas incluyen Recuperación de la Información (IR), que se ocupa de encontrar documentos que satisfagan una determinada necesidad de información o consulta dentro de una gran base de datos de documentos; Procesamiento Natural del Lenguaje (NLP), que es una disciplina que abarca todos las técnicas de procesamiento automático tanto de lenguaje escrito como hablado; la Extracción de la Información (IE), que puede ser considerada un campo de NLP y está centrada en encontrar entidades explícitas y hechos dentro de un texto no estructurado. Finalmente, la Minería de Texto es el proceso de analizar el lenguaje natural escrito para descubrir información o conocimientos que son comúnmente difíciles de recuperar. [2]

La clasificación de información biomédica se convierte en uno de los desafíos más grandes debido al creciente número de artículos biomédicos que se presentan divididos en múltiples subgrupos. En vista de esta situación es que muchos investigadores han tratado de encontrar más maneras aplicables para la clasificación de la literatura biomédica con el fin de ayudar a los usuarios a encontrar los artículos relevantes en la web. Del mismo modo es importante tener en cuenta las técnicas de minería de textos supervisadas y no supervisadas, las cuales tienen diferentes propósitos. Para el caso de las técnicas de minerías de datos no supervisadas, también conocidas con el nombre de técnicas de descubrimiento del conocimiento, se utilizan para la detección de patrones ocultos en bases de datos de gran tamaño. [3] [4] Dentro de este tipo de clasificación se encuentra el modelado de tópicos, el cual es utilizado con el objetivo de caracterizar documentos dentro de un texto. El modelado de tópicos permite entregar información clara, correcta y coherente de los

textos.

En este contexto se propone un método de extracción de información biológica a partir de la literatura biomédica basado en la extracción de tópicos enriquecidos en términos encontrados en publicaciones de las bases de datos de BioMed Central. Dicha información es de gran importancia ya que los resultados de las distintas técnicas de minería de textos han de ser interpretados recurriendo a la información que contienen. Las técnicas informáticas que procesan literatura biomédica son útiles para facilitar el acceso a textos relevantes a biólogos, bioinformáticas e incluso a anotadores de bases de datos.

1.2. Objetivos

1.2.1 Objetivo general

Diseñar un método para descubrir tópicos de un conjunto de revistas científicas que permita la identificación y extracción de características de forma no supervisada.

1.2.2 Objetivos específicos

- Identificar artículos científicos de interés de la base de datos total
- Generar clases de los documentos de investigación propuestos.
- Realizar un pre-procesamiento del texto e identificación de entidad nombrada.
- Construir dos diccionarios de palabras únicas, que sean representativos en el texto.
- Descubrir tópicos a través de clasificación no supervisada.
- Obtener resultados a través de la modelación de tópicos.
- Analizar cada uno de los tópicos entregados por el modelo.

1.3. Alcances y Limitaciones

La información de las publicaciones científicas de BioMed Central se encuentra almacenada en una base de datos. Dicha base de información contiene información extensa y variada. Las distintas técnicas informáticas en la extracción de información han de ser, en muchos casos, interpretadas, recurriendo a la información que contienen de forma manual.

Debido a la creciente acumulación de información biomédica en la base de datos, se hace necesario el desarrollo de una aplicación que automatice la información en subgrupos de interés. Uno de los temas que más interesa en el área científica es precisamente el de clasificar la información de acuerdo a los procesos y funciones biológicas. Sabiendo de qué procesos biológicos habla cada documento, se podrá establecer asociaciones o clases en los documentos y finalmente en los tópicos encontrados. Dicho de otra manera, es posible caracterizar las clases a través de su función y del nivel de información que se encuentra en el estudio de la base de datos y, más en concreto, en la clasificación de los documentos.

1.4. Temario

Esta memoria se divide en 5 capítulos que se detallan a continuación:

Capitulo 1, introducción: En este capítulo se realiza una breve introducción al trabajo. Se comentan cuales son los objetivos que se pretenden alcanzar con la realización de este proyecto. También se especifican los alcances y limitaciones que se presentan durante el transcurso de este trabajo.

Capitulo 2, Estado del arte: En este capítulo se hace referencia al estado del arte, donde se hace mención a los trabajos investigados en el área de la clasificación no supervisada y extracción de tópicos. Además este capítulo contiene el marco teórico, donde se encuentran las definiciones de algunos términos que son indispensables para la comprensión de muchos puntos de la memoria, como lo son: tipos de aprendizaje, clasificadores, modelación de tópicos, entre otros.

Capitulo3, Materiales y métodos. En este capítulo se describen los materiales a utilizar, como es la base de datos con los artículos científicos a clasificar y las herramientas claves en el manejo de la información. También se describen las metodologías propuestas para la extracción de información

Capitulo 4, Resultados. En este capítulo se describen cuales son los requisitos de implementación que el modelo debe seguir. Además se presentan y explican cuáles son los datos a utilizar, así como donde se pueden obtener. Finalmente, en este capítulo se muestran cuáles son los resultados del modelado de tópicos realizados sobre cada una de las clases seleccionadas. Se presentan una serie de tablas y gráficos con la que la comprensión de los resultados se hace más sencilla.

Capitulo 5, Conclusiones. En este capítulo se presentan las conclusiones a las que se puede llegar con los resultados que se han obtenido y se comentan cuáles pueden ser los métodos que mejor se implementan durante el desarrollo del trabajo.

Capítulo 2. Estado del Arte y Marco teórico

2.1. Estado del arte

2.1.1 Introducción

En esta sección se da a conocer el estado del arte en relación al aprendizaje no supervisado en las técnicas de recuperación de información, y modelación de tópicos. La recuperación de información implica la búsqueda de diversos campos de investigación de modelos y ejemplos que se han utilizado para solucionar problemas basados en el aprendizaje no supervisado así como las aplicaciones prácticas que se han abordado mediante este. Además, dentro de esta revisión se incluyen las métricas más comúnmente utilizadas para determinar el rendimiento de las propuestas en modelación de tópicos, así como los conjuntos de datos que se utilizan con más frecuencia en la extracción de tópicos. Es por esto que para tener una visión global de la extracción de tópicos en artículos de revistas científicas de BioMed Central, se realiza una revisión bibliográfica sobre las investigaciones recientes y los resultados que se han obtenido. Todo esto permite evaluar que técnicas y métodos han tenido mejor desempeño en la clasificación de textos para ser implementados en este trabajo.

2.1.2 Modelación de Tópicos

La clasificación de artículos de interés busca encontrar tópicos. El desafío se encuentra en la determinación de estos y la asignación a la clase correcta. [3]

El precursor de los modelos de tópicos es David Blei, el cual en [5] describe de manera detallada los modelos de tópicos y las aplicaciones de estos. En [5], se define un tópico como el conjunto de elementos que pueden representar una temática presente en una colección de documentos sin pérdida de información estadística, es decir, en el cálculo de probabilidades que den información global de toda la colección de datos. Un modelo de tópicos tiene como objetivo identificar las relaciones latentes entre documentos pertenecientes a una colección, con el fin de dar una descripción sucinta de ésta sin perder información desde el punto de vista estadístico. Por

ejemplo, si existe una colección de documentos textuales que abarca múltiples temas, un tópico es un conjunto de palabras que logra describir estadísticamente uno de estos temas.

Se propone un método de aprendizaje no supervisado para el descubrimiento de tópicos e interacciones de los mismos. En este tipo de aprendizaje, los "rótulos" de clase son desconocidos, por ello se desea agrupar el conjunto de datos de acuerdo a ciertas similaridades. El aprendizaje no supervisado se divide en dos grupos: probabilístico como el bayesiano, LDA, entre otros y el no probabilístico que mide distancias, entropías o métricas.

A continuación se describen una serie de investigaciones referentes al modelado de tópicos, basándose en el procesamiento del lenguaje natural, recuperación de información, extracción de información y minería de datos.

En [6], los autores Gordon y Dumais, utilizan el modelo Indexación Semántica Latente (LSI), para descubrir la relación entre el aceite de pescado y la enfermedad de Raynaud usando la base de datos Medline. El modelo LSI es una técnica estadística que permite estimar la estructura latente entre las palabras. Igual que el modelo LSA utiliza un valor singular de descomposición que segmenta una gran matriz de datos de asociación de término-documento y permite construir un "espacio semántico" en el que se asocian entre sí términos y documentos.

En [7], el método Latent Semantic Analysis (LSA) da a conocer que existe una estructura latente en el uso de las palabras obstaculizado por la variabilidad en la selección de palabras, es decir, las palabras del mismo campo semántico suelen aparecer juntas o en similares contextos. Finalmente, se tiene un espacio semántico-vectorial en el que están representados todos los términos de los documentos. Estos términos aparecen y serán los que acotarán el uso de unos términos en concurrencia con otros y la probabilidad en que otros términos sean utilizados en estos mismos documentos. De esta manera se puede realizar clustering de los documentos y con los cuales encontrar similaridad entre la información evaluada. Los datos son pre-procesados usando una técnica de álgebra lineal conocida como Singular Value Decomposition (SVD), que por medio de un algoritmo recursivo, descompondrá la matriz que representa los términos y los documentos en que ocurren en dos matrices que representan vectores de términos y documentos y una matriz diagonal en los que se encuentran en orden descendente los valores singulares que representan las relaciones que mantienen ambas matrices de vectores singulares.

En [8], el análisis de tópicos está basado en reglas basadas en los modelos ocultos de

Markov (HMM). Las etiquetas basadas en modelos de Markov estiman la probabilidad de que una secuencia de etiquetas pueda ser asignada a una secuencia de palabras. Se entrena el sistema con el set de datos con el fin de estimar los parámetros del modelo utilizado.

Los autores Rodríguez y Bautista en [9], utilizan los modelos ocultos de Markov para el análisis de patrones espaciales en transectos de vegetación con datos de presencia-ausencia, es decir, tres estados ocultos; un primer estado, correspondiente a los arbustos, que denominan manchas, pequeños claros entre manchas y claros, donde la sucesión de los dos primeros estados constituyen agrupaciones de manchas, separadas entre sí por los claros. La mayor utilidad de los HMM como herramientas de análisis se basa en la posibilidad de estimar un modelo a partir de una serie de datos.

Por otro lado, se tiene el modelo Latent Dirichlet Allocation (LDA), el cual es un modelo probabilístico generativo. La idea básica es que los documentos a clasificar se representan como mezclas aleatorias de tópicos ocultos, donde cada tópico se caracteriza por una distribución sobre palabras. La distribución de categorías tiene una distribución a priori de Dirichlet. Este último modelo es el más utilizado en los trabajos investigados:

- Los autores Ruiz y Campos en [11], presentan un estudio en el que miden el desempeño de un algoritmo que combina kernels no lineales, concurrencias de rasgos craneales de forma, un método de selección de variables y un algoritmo estándar de reducción de dimensionalidad para caracterizar y clasificar malformaciones causadas por craneosinostosis primaria. La técnica de clasificación utilizada en este trabajo es que una clase particular de formas craneales pueda representarse por medio de patrones característicos. Para ello el modelo de reducción de dimensionalidad que se implementa está basada en el algoritmo LDA. El LDA se utiliza como un modelo generativo que muestrea rasgos craneales de forma a partir de una mezcla de tópicos geométricos. Los resultados de aquel estudio sugieren que la combinación de los cálculos de descriptores de forma (de los cráneos) y la técnica LDA resultan en bajas tasas de error de clasificación.
- En el trabajo de Bisgin, Liu, Fang, Xu, y Tong en [12], se busca determinar la relación de los medicamentos con los farmacos aprobados por la FDA. De esta manera se asocia cada fármaco al tópico más probable. El modelo utilizado en la extracción de tópicos es el LDA.
- El autor W. Hu en [13], emplea el aprendizaje no supervisado en el estudio de 2 libros

bíblicos (proverbios y salmos) debido a la asociación compartida de información. Los capítulos de cada uno de los dos libros se agrupan por contenido. El modelo que determina la asociación entre los libros es el algoritmo LDA, la extracción de tópicos de cada documento identificado por el modelo se utiliza para definir una correlación y medir la similitud entre ambos libros.

- El autor S. Rodríguez en [14], en su trabajo que lleva por nombre *"Estudio de técnicas no supervisadas para descubrir tópicos en videos deportivos"* da a conocer herramientas que analizan videos deportivos mediante el uso de técnicas de visión por computador y reconocimiento de patrones. Actualmente, para mejorar los resultados deportivos, los clubes disponen en sus plantillas de personas encargadas de analizar los partidos, tanto los propios como los de los equipos rivales, para encontrar patrones en su forma de juego y, de esta manera, estudiar la mejor forma de obtener ventajas competitivas. Las herramientas que se implementan en la recuperación de información, una vez evaluado su correcto funcionamiento del algoritmo, el siguiente paso es probar el modelo sobre videos deportivos y ver si es capaz de descomponerlos en tópicos. En la fase de descubrimiento de tópicos se utiliza el método de aprendizaje no supervisado LDA, para descubrir actividades e interacciones en lugares muy concurridos.

- Los autores Dueñas y Velásquez en [15], presentan una metodología alternativa para detectar tendencias en la Web a través del uso de técnicas de recuperación de la información, modelamiento de tópicos y minado de opiniones. Dado un conjunto de sitios Web, se extraen los tópicos que se mencionan en los textos recuperados y posteriormente se acude a las redes sociales para obtener la opinión por parte de sus usuarios en relación a estos. Se utiliza el LDA para la extracción de tópicos de importancia a partir de las opiniones consignadas en blogs y sitios de noticias.

- Seiter J., Amft O., Rossi M., y Tröster G. en [16], comparan el resultado de tres modelos no supervisados (LDA, n-gram TM, correlated TM (CTM)), en tres conjuntos de datos de actividades públicas para obtener directrices para la selección de parámetros TM dependiendo de las propiedades del conjunto de datos. En el trabajo se determinan que la principal limitación del modelo de unigramas es que supone que todos los documentos son sólo colecciones de palabras homogéneas, es decir, todos los documentos presentan un solo tema. Los resultados experimentales sobre el conjunto de datos seleccionados han

demostrado que la propuesta de LDA en el descubrimiento de información resulta ser menos sensible al ruido.

El método LDA fue desarrollado inicialmente para modelar conjuntos de datos discretos, aunque sobre todo los documentos textuales.

En la tarea de modelar un documento, el LDA se maneja mejor que el LSI y una mezcla de modelos de unigramas. El LSI sobre ajusta las probabilidades de la modelación de documentos para determinar los tópicos en un nuevo documento. Como era de esperar, el gran avance del LDA con el LSI fue que fácilmente se asignan probabilidades a un documento nuevo. Su aplicación en clasificación de documentos determina que el modelo LDA puede ser útil como un algoritmo de filtrado para la función de selección de tópicos.

El trabajo de S. Rodríguez en [14] fue más allá de los documentos de un texto simple. El estudio de técnicas no supervisadas para descubrir tópicos en videos, pretende implementar un método con el que se puedan extraer una serie de acciones comunes en videos con la mínima supervisión humana posible. En lugar de tener un documento de texto, tiene un usuario, y en vez de palabras sueltas, tiene videos elegidas por el usuario. El conjunto de datos se evalúa utilizando un estimador, una y otra vez y los resultados fueron que se desempeñaron mejor con el modelo LDA que con LSI y una mezcla de unigramas [1].

El LDA es un modelo robusto y genérico que es fácilmente extensible más allá de los datos empíricos de un pequeño conjunto discutido. Numerosos artículos han sido publicados sobre la aplicación del LDA a una amplia gama de áreas. Se ha aplicado a las tareas que van desde la detección del fraude en las telecomunicaciones a la detección de errores en el código fuente. A pesar de la amplia gama de aplicaciones, LDA no se ha aplicado a resumen automático de documentos, aunque la posibilidad es bastante factible.

El modelo LDA ha demostrado ser un método que permite asignar una clase o categoría a un objeto para generar aprendizaje, generando una buena separabilidad entre clases y una buena cohesión dentro de una misma clase. Es por esto, que en esta memoria se usará un tipo de aprendizaje no supervisado en la clasificación, porque no requiere conocimiento a priori de los artículos científicos. Los tópicos a descubrir pueden ser utilizados con éxito para la agrupación e identificación de información relevante dentro de la misma categoría científica.

(1) Un unigrama puede ser visto como una ventana de largo n=1 que se pone sobre el texto , es decir, abarca una palabra

Finalmente, el objetivo es demostrar cómo modelar tópicos, con una técnica de aprendizaje no supervisado, el cual puede contribuir como un nuevo espacio para el estudio de grandes corpus científicos. El modelado de tópicos puede agrupar y clasificar la información en función de sus características comunes sin necesidad de un conocimiento a priori, y por lo tanto tiene el potencial para amplias aplicaciones en la investigación biomédica, en particular para los documentos de BioMed Central.

2.2. Marco Teórico

2.2.1 Introducción

En la siguiente sección se da a conocer las técnicas de recuperación de información en la modelación de tópicos en textos y finalmente se explica el algoritmo de extracción de tópicos LDA. Como herramienta fundamental en el proceso de gestión del conocimiento se hace una comparación en el tipo de aprendizaje a seguir, eligiendo entre uno supervisado o no supervisado, con el fin de definir el tipo de clasificador a desarrollar durante el modelado y extracción de tópicos.

Las técnicas de recuperación de información son el pre-procesamiento del texto, Etiquetado POS, e Identificadores de Entidades Nombradas (NER). Es por esto, que los datos que se trabajan deben cumplir ciertos estándares y para ello es necesario aplicar estas técnicas de minería de textos, para evitar finalmente, errores en la clasificación.

2.2.2 Aprendizaje (Supervisado VS no supervisado)

Un aspecto importante que influye en el aprendizaje es el grado de supervisión. En algunos casos, un experto en el dominio proporciona al aprendiz retroalimentación acerca de lo que es apropiado para su aprendizaje. En otros casos, a diferencia del aprendizaje supervisado, ésta retroalimentación está ausente, dando lugar al aprendizaje no-supervisado. Y en otros casos, se combinan el aprendizaje supervisado y el no supervisado, dando lugar al aprendizaje semi-supervisado.

En minería de datos las técnicas supervisadas y no supervisadas tienen diferentes propósitos.

Las primeras se utilizan para construir modelos que serán utilizados para realizar predicciones mientras que las no supervisadas, o algoritmos de descubrimiento del conocimiento, se usan generalmente para la extracción de información útil a partir de grandes volúmenes de datos [4].

El aprendizaje supervisado es aquel en donde se intenta aprender de ejemplos como si estos fueran un maestro. Se asume que cada uno de estos ejemplos incluye características o atributos que especifican o definen a qué categoría o clase pertenece, de un conjunto de categorías o clases predefinidas, de esta manera cada ejemplo se asocia con su clase. Este tipo de aprendizaje es llamado supervisado por la presencia de los ejemplos para guiar el proceso de aprendizaje. Al conjunto de ejemplos del cual se intenta aprender se le llama conjunto de entrenamiento. Ejemplo de este tipo de aprendizaje son los modelos bayesianos, support vector machine (SVM), entre otros. Para saber si el método supervisado esta correcto, se define un *Gold Standar*d que posteriormente será utilizado como *corpus de texto* para el entrenamiento del modelo. Tras la creación del *Gold Standard* comienzan las distintas pruebas con los modelos de clasificación, donde se definen los datos correctamente clasificados (TP), los datos erróneos (FP), los datos que son verdaderos pero el clasificador lo toma como erróneos y viceversa. Con estos valores se evalúa el modelo a través del cálculo de las medidas de evaluación, recall, error de medición, precisión, entre otros.

El aprendizaje no supervisado no presupone ningún conocimiento previo sobre lo que se quiere aprender. Es decir, tiene la ventaja de que no es necesario que al aprendiz se le muestren todos los ejemplos existentes. La desventaja es que a pesar de lo anterior, sí es necesaria una gran cantidad de ejemplos para el entrenamiento. En el aprendizaje no supervisado, los ejemplos sólo incluyen los atributos, es decir, no se encuentran asociados a una clase. Dos ejemplos clásicos muy simples de aprendizaje no supervisado son clustering y la reducción de dimensionalidad. Para este caso la tarea se enfoca en descubrir patrones comunes entre los datos, que permitan separar los ejemplos en clases o tópicos. De éstas se podrán extraer caracterizaciones, o permitirán predecir características, o deducir relaciones útiles, a lo que se denomina como agrupación (clustering). Algunos de los algoritmos más comunes son: métodos probabilísticos (LDA, LSI, LSA), K- means, etc, siendo este último el más utilizado en el área de clasificación de textos. [17]

En el aprendizaje no supervisado se tiene un gold standard donde no siempre es posible realizar una evaluación de los datos, ya que no se sabe a priori el contenido de la información. Es por esto, que el desempeño del modelo no supervisado es evaluado a posteriori mediante revisión manual de un experto.

2.2.3 Extracción de características

El primer paso para realizar la tarea de clasificación de textos utilizando técnicas de aprendizaje computacional consiste en obtener los atributos que describan el texto a clasificar, así como transformarlos a una representación adecuada para ser utilizados por los algoritmos de aprendizaje. A este paso previo se le llama extracción de características y se divide comúnmente en tres etapas: Pre-procesamiento, Etiquetado POS, Identificadores de Entidades Nombradas (NER).

- **Pre-procesamiento del texto**

El Pre-procesamiento es la etapa más importante en cualquier actividad de minería de textos y el suceso y calidad de la minería depende directamente de la eficiencia de esta fase.

Durante la etapa de pre procesamiento de los documentos los documentos son analizados, eliminando aquellos elementos que se consideran no necesarios para algunas tareas. El pre procesamiento tiene procesos tales como: eliminar aquellos elementos que no contienen información relevante para el modelado de tópicos, así como la eliminación de palabras vacías, las cuales corresponden a los pronombres, preposiciones, conjunciones, artículos, etc. Dentro del pre procesamiento también se encuentra la normalización, segmentación y lematización de las palabras.

- **Etiquetado Part of Speech (POS)**

El etiquetado gramatical, también conocido como, POS tagging, es el proceso de asignar una etiqueta gramatical a cada una de las palabras de un texto según su categoría léxica. Un tagger o etiquetador es un programa que realiza este proceso automáticamente. La mayoría de los etiquetadores actuales utilizan modelos estadísticos que se generan a partir de un texto anotado previamente (corpus de entrenamiento).

El etiquetado gramatical muestra, hasta cierto punto, la estructura de un documento, brindando una gran cantidad de información sobre una palabra y sus vecinas. Una etiqueta gramatical puede ofrecer información relacionada con la pronunciación, el reconocimiento de sustantivos, adjetivos, tipo de derivación y/o inflexión.

El etiquetado gramatical es realizado manualmente por lingüistas o automáticamente por programas conocidos como etiquetadores gramaticales. La mayoría de las implementaciones actuales de estos programas están basadas en el aprendizaje; toman un corpus anotado correctamente con el cual se entrenan y luego emplean el conocimiento adquirido para etiquetar el corpus objetivo. En esa primera etapa conocida como entrenamiento, el etiquetador gramatical obtiene y preserva información sobre cada palabra, su etiqueta asignada y su contexto. Posteriormente dado un corpus objetivo como entrada el etiquetador determina una etiqueta para cada palabra.

Las etiquetas gramaticales pueden ser divididas en dos grandes categorías: clases cerradas y clases abiertas. Las clases cerradas son aquellas que tienen miembros relativamente fijos. Por ejemplo, las preposiciones son una clase cerrada porque hay un conjunto cerrado de ellas, es decir que son un grupo de palabras que raramente varía ya que raramente se agregan nuevas preposiciones. En contraste, los sustantivos y los verbos son clases abiertas ya que continuamente se introducen y eliminan nuevos verbos y sustantivos al lenguaje. Es probable que cualquier hablante o corpus tenga una clase abierta de palabras diferente, pero todos los hablantes de un lenguaje y corpora suficientemente grandes, seguramente van a compartir el conjunto de clases de palabras cerradas. Las clases de palabras cerradas también son generalmente palabras funcionales, por ejemplo: tienden a ser muy cortas, ocurrir frecuentemente y generalmente tienen usos estructurales en gramática. [18]

Dependiendo del tipo de problema que se quiera resolver, el contar con etiquetas POS dentro de un documento puede resultar de gran utilidad.

- **Identificadores de Entidades Nombradas (NER)**

La identificación de entidades nombradas en los textos es la identificación de algunos tipos predefinidos de entidades, tal como personas, organizaciones o lugares. Una entidad con nombre es un sintagma nominal. Un sintagma es una combinación de palabras que desempeña alguna función sintáctica dentro de una oración. Un sintagma nominal consiste en un sustantivo y un pronombre. Este puede denotar un objeto específico, tal como una persona, organización o ciudad. En el ámbito biomédico, estas entidades son por lo general genes, proteínas, enfermedades, líneas celulares, etc. [19]

La extracción de entidades es utilizada por sistemas que necesitan conocer de quien y en qué momento se está hablando para extraer información de los textos que se procesan. Esta técnica representa un paso previo para otras tareas de minería de texto, tales como la recuperación de información en documentos de artículos científicos que se encuentran relacionados en su temática. Es por esto que esta técnica incluye la identificación exacta de la mención en el texto, es decir, de sus límites, que se definen por la posición del primer y último caracteres que delimitan las palabras que lo componen.

El NER es útil en sistemas cuya funcionalidad consiste en filtrar información no deseada, o en sistemas que requieren llevar información no estructurada a un formato estructurado, así por ejemplo, construir una base de datos a partir de información identificada por un reconocedor NER. Otro ejemplo puede ser aquel que filtra contenido en documentos biomédicos, con la finalidad de que dicha información sea utilizada como referente a patologías o fármacos. [20]

2.2.4 Modelo de Tópicos

El modelado de tópicos es una herramienta potente que se desarrolla inicialmente con el objetivo de caracterizar documentos, es decir, identificando las relaciones latentes entre documentos. Este modelo generativo es usado para explicar observaciones multinomiales de aprendizaje no supervisado.

El objetivo del modelado de tópicos es colocar, de forma automática, documentos dentro de un número fijo de categorías (temas o clases) predefinidas, en función de su contenido. El procedimiento del modelado es dependiente del tipo de aprendizaje.

El autor Blei D. (2012), define un tópico como el conjunto de elementos que pueden representar una temática presente en una colección de documentos sin pérdida de información estadística. [5]

Entre los modelos de tópicos existentes, se encuentran el modelo Latent Dirichlet Allocation y el modelo CTM (Correlated Topic Model)

Estos modelos de tópicos se cimentan sobre las siguientes definiciones:

Una **palabra** w es la unidad elemental de un documento textual y se define como un elemento de un vocabulario indexado V. Para efectos de estos modelos, para representar una palabra se hace uso de vectores unitarios en donde la n-ésima palabra de V se representa con un vector de largo $|V|$ en el cual sólo su componente n-ésima es igual a 1.

Un **documento** es un arreglo de palabras descrito como $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)$, donde w_n es la n-ésima palabra de este.

Un **corpus** es una colección de documentos descrita como $D = (w_1, w_2, \ldots, w_M)$,

Un **tópico** es una distribución de probabilidad sobre un vocabulario V fijo. Por ejemplo, el tópico *político* está descrito por palabras como partido, diputado, senado, ley de manera frecuente y palabras como guerra, marcador, gol con probabilidad casi nula.

A continuación se describe el modelo LDA, utilizado en la modelación de tópicos.

2.2.4.1 Latent Dirichlet Allocation (LDA).

El método LDA es un modelo probabilístico generativo que permite explicar las similitudes en un conjunto de textos, utilizando grupos no observados. Los conjuntos de observaciones son las palabras recogidas en los documentos, donde cada documento es una mezcla de un pequeño número de tópicos. La creación de cada palabra se atribuye a uno de los tópicos del documento.

El modelo LDA trabaja bajo el supuesto de que los tópicos presentes en la colección de documentos que se está analizando no necesariamente están relacionados y por consiguiente no dependen entre ellos. Para extraer la estructura de tópicos presente en una colección, este modelo hace uso de un modelo estadístico de generación de documentos, tópicos y palabras a lo largo del tiempo que abarque ésta. [5]

Para cada documento presente en un corpus, se realiza los siguientes procesos:

1. Se define una distribución aleatoria para la presencia de los tópicos en el corpus y una distribución para la presencia de las palabras para cada tópico que se desea encontrar.
2. Luego, por cada palabra presente en el documento bajo análisis se debe:

    a) Escoger un tópico aleatoriamente haciendo uso de la distribución generada en el paso anterior.

    b) Escoger una palabra del documento aleatoriamente a partir de la distribución del vocabulario en relación al tópico escogido.

Para determinar la estructura de tópicos existente luego del proceso de generación, es necesario calcular las distribuciones condicionales entre los tópicos y sus documentos, el cual es un

problema, debido a que la cantidad de estructuras que pueden representar una colección de documentos crece exponencialmente en relación a la cantidad de documentos y palabras presente en esta. Este proceso es descrito a continuación, y es representado en la figura 1.

1. Se decide el número de palabras N que tendrá el documento M (Por ejemplo de acuerdo a una distribución Poisson (φ)).
2. Escoger θ ~ Dirichlet (α). Se elige una mezcla de tópicos para el documento (de acuerdo a una distribución Dirichlet sobre un conjunto fijo de K tópicos).
3. Por cada una de las N palabras, generar w_i

a. Escoger un tópico z_d ~ Multinomial(θ)

b. Elegir una palabra w_i para p(wi|zn, $\boldsymbol{\beta}$)

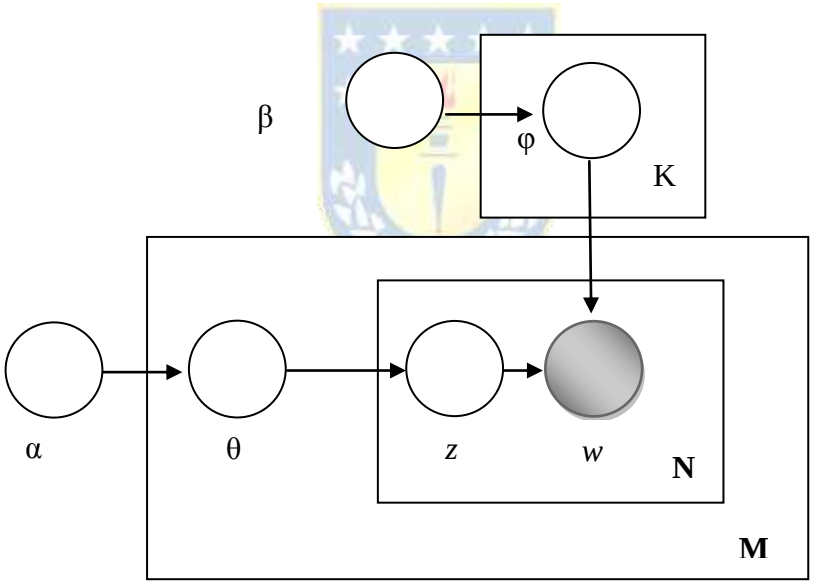

Figura 1: Modelo LDA [extraído de [10]]

En el esquema de la figura 2, se puede ver el algoritmo LDA, donde las entradas corresponden a: "k" que representa el número de tópicos, "diccionario" que es el conjunto de todas las palabras posibles, "matriz de datos" que contiene el número de veces que se ha repetido cada una de las palabras (del diccionario) por cada uno de los documentos. Luego las salidas

corresponden a: "P_{wz}" que es la probabilidad que tiene cada una de las palabras de formar parte de un tópico, y "P_{zd}" que es la probabilidad que tiene un documento de formar parte de un tópico.

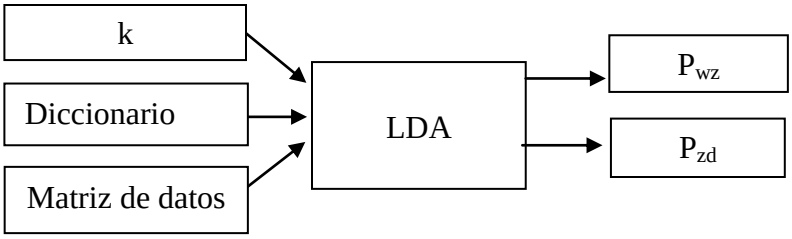

Figura 2: Esquema de uso del LDA

Capítulo 3. Materiales y Métodos

3.1. Materiales

3.1.1 Introducción.

En este apartado se da a conocer los materiales, como fuente de información en el estudio de la modelación y extracción de tópicos.

Se cuenta con una base de datos de artículos de revistas científicas en MySQL organizadas en tablas. Cada tabla contiene información entregada por cada artículo, por ejemplo autores, títulos, contenido, abstract.

Los abstract de la base de datos en MySQL serán los utilizados en el proceso de extracción de información. Las demás tablas serán usadas como apoyo en el proceso de extracción.

Las herramientas que se utilizarán en el trabajo de recuperación de la información son aplicaciones de MySQL y bibliotecas de Python como lenguaje de programación.

3.1.2 Base de datos

Los artículos de revista científica extraídos de BioMed Central se encuentran organizados en una base de datos de MySQL, permitiendo de esta manera descubrir tópicos y tendencias semiautomáticas del enorme corpus.

La base de datos cuenta con 119.339 artículos, que incluyen campos de información como Abstract, Autor, Classification, Keyword y Section (figura 3). Se observa que no todos los artículos contienen la misma información, y con ello se determina que un mismo artículo pueda no contener todos los campos de información. Ver tabla 1

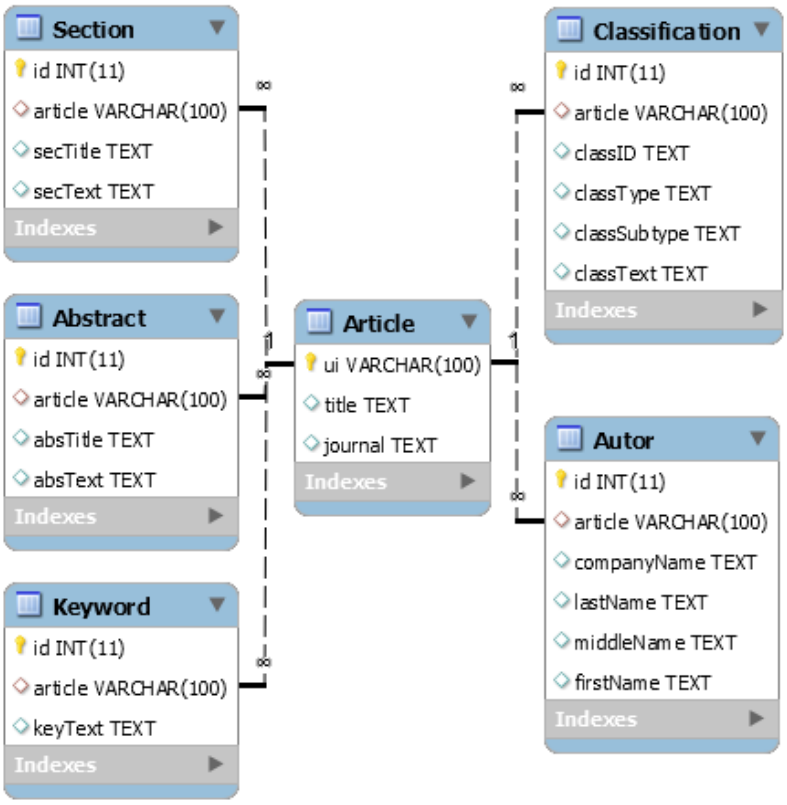

Figura 3: Diagrama EER. [Extraído de MySQL Workbench]

Tabla 1: Contenido de base de datos

Article	Abstract	Autor	Classification	Keyword	Section
(Total=119.339)	**102.946**	**660.111**	**27.368**	**97.294**	**720.473**

Los temas de información de esta base de datos, incluyen títulos generales de BMC Biology y BMC Medicine, además de revistas especializadas (por ejemplo, BMC Bioinformatics, Malaria Journal) que se centran en disciplinas particulares.

3.2. Metodología

3.2.1 Introducción.

En esta sección se detallan los métodos y algoritmos que han sido propuestos para el análisis y extracción de tópicos a partir de la literatura biomédica e información biológica de la base de datos existente.

La metodología que se emplea para el desarrollo de esta memoria consta de cinco etapas principales; selección de artículos de interés, pre–procesamiento del texto, identificadores de entidades nombradas, clasificación de los documentos y evaluación del clasificador implementado.

Para el desarrollo de la primera fase, se seleccionan artículos del ámbito biomédico necesarios para la adquisición de la información pertinente a la construcción de clasificador no supervisado.

Seguidamente se realiza un pre–procesamiento del texto, con el fin de obtener palabras o términos claves en un texto en particular. De forma paralela a este método, se realiza la extracción de entidades nombradas, donde las dos principales tareas son; detectar y clasificar entidades nombradas en fragmentos de textos.

Luego, se implementa el modelo de clasificación no supervisado, con el objetivo de extraer tópicos de interés biomédico.

Finalmente, se hace una evaluación del modelo LDA con el fin de determinar su desempeño en la modelación de tópicos en los artículos de las revistas científicas.

3.2.2 Selección de artículos de interés.

Esta metodología se centra en diseñar e implementar una aplicación en la cual se selecciona una serie de documentos científicos en idioma inglés. La aplicación extrae los conceptos que mejor definan a estos documentos.

Debido a la magnitud del corpus y a la diversidad de información contenida en él, se seleccionan los abstract de la base de datos, con el fin de saber a priori la información entregada por los artículos de revista científica. Se decide trabajar solo con los abstract de la base de datos porque

se encuentran en inglés y además porque las bibliotecas de Python a utilizar están diseñadas para tal idioma. Sin embargo, se hubiese podido escoger otro campo de información de la base de datos como los keyword pero estos contienen términos abreviados y en diferentes lenguajes. Otro motivo por el cual se decide trabajar con los abstract es que no todos los artículos presentan los mismos campos de información lo que reduce la posibilidad de definir a priori un gold standard representativo.

Los documentos que se deciden analizar hacen alusión al pulmón, corazón y riñón, por ser temas conocidos dentro del ámbito biomédico y por el cual se puede tener certeza de los conceptos entregados por los mismos. Luego, los abstract que contienen información relacionada a cada uno de estos temas, son agrupados en clases que posteriormente serán modelados mediante el modelo LDA.

La técnica utilizada en la selección de las clases es a través del uso de palabras claves como filtro dentro de los textos. Se tiene que para la clase pulmón se consulta en Workbench "SELECT * FROM Abstract WHERE absText LIKE '%lung%'". Luego la clase “pulmón” está asociada a todos los abstract que contengan la palabra “lung” y con ello se sabrá a priori que la información entregada en el texto hace mención a tal temática.

Finalmente, en la recuperación de información, los tópicos extraídos contienen conceptos relacionados y representativos de la clase modelada.

3.2.3 Pre-procesamiento del texto.

El Pre-procesamiento abarca todas aquellas técnicas de análisis de datos que permiten mejorar la calidad de un conjunto de datos de modo que las técnicas de extracción de conocimiento puedan obtener mayor y mejor información.

Estos tratamientos conllevan a analizar las palabras del lenguaje y conseguir obtener la información relevante y filtrar la que no tiene significado semántico. Dentro del pre-procesado de texto, se encuentran la eliminación de palabras vacías, lematización, normalización y segmentación del texto. Con los pasos anteriores cada documento de la clase seleccionada se convierte a una representación compacta adecuada para el algoritmo de aprendizaje no supervisado, LDA.

3.2.3.1 Eliminación de palabras vacías.

Las palabras vacías dentro de los artículos científicos son palabras que son muy frecuentes y que por lo general no contienen información importante para la extracción de información, por ejemplo: pronombres, preposiciones, conjunciones, artículos, etc.

Estas palabras deben ser excluidas del texto porque se mejora la eficiencia y efectividad de los sistemas de minería de texto. Además de agregar ruido y aumentar el tamaño de la representación, son palabras que pueden confundir ciertas aplicaciones como IR.

En el reconocimiento de entidades gramaticales esta técnica no es aplicable, ya que las palabras vacías juegan un rol fundamental en el texto, dependiendo del tipo de entidad a encontrar.

3.2.3.2 Normalización del texto.

El proceso de normalización es muy importante en el pre procesamiento de un texto, ya que impide que el programa tome como palabras distintas las palabras que son sintácticamente iguales. Para el caso de este trabajo sólo se pasó todo el texto a minúsculas. Lo anterior no afecta al resultado esperado dentro del contexto del corpus extraído.

3.2.3.3 Segmentación.

La segmentación es el proceso de separar el texto en unidades, son los denominados tokens. Una correcta tokenización del texto pasa por la separación de la puntuación, paréntesis, guiones y otros. En el caso de los artículos científicos, la aparición de estos tokens no siempre es trivial, como en el siguiente recorte: “flight"); "set"”. Se debe indicar donde terminan las oraciones o frases, en el caso de las mayúsculas indican el inicio de una oración, etc.

En general, no resulta complejo implementar estas operaciones de texto. Sin embargo, deben estudiarse las distintas excepciones con cuidado ya que pueden provocar un importante impacto en el momento de la recuperación de información.

3.2.3.4 Lematización.

Por lematización nos referimos al proceso de remover los sufijos para reducir una palabra a

su lema o raíz. Por ejemplo de los tópicos extraídos se encuentran raíces similares en el contenido de la palabra, dando paso a la agrupación por clases. Por ejemplo, representar"analysis","analyzer" o"analyzing" mediante"analy"

Frecuentemente, una palabra no aparece exactamente en un documento, pero sí alguna variante gramatical de la misma como plural, gerundios, sufijos de tiempo verbal, etc. Este problema puede resolverse con la sustitución de las palabras por su raíz (stem).

Un stem es la porción de una palabra que resulta de la eliminación de sus afijos (prefijos y sufijos). Los stems permiten reducir el tamaño de la estructura de indexación ya que el número de términos índice se reduce.

El algoritmo utilizado en eliminación de sufijos es el algoritmo de Porter [21]. Este algoritmo usa una lista para la detección de sufijos. La técnica se basa en aplicar una serie de reglas a los sufijos de las palabras del texto.

Sin embargo al aplicar lematización se pueden generar dos tipos de errores:

- Infraradicación (understemming): Es cuando dos palabras que deberían ser mapeadas a la misma raíz se mapean a diferentes raices. Dejan muchas letras de tal forma que con la misma raíz se mapean a diferentes raíces. Por ejemplo "alumnus" a "alumnu", "alumni" a "alumni", "alumna" a "alumna".
- Sobrerradicación (overstemming): Es cuando dos palabras con diferentes raíces son mapeadas a la misma raíz, es decir, sacan muchas letras de tal forma que palabras con diferentes significados se mapean a la misma raíz. Por ejemplo "comunismo", "comunal", "comunes" a "común". "universal", "university", "universo" a "univers"

3.2.3.5 Palabras claves.

Las palabras claves del texto son las obtenidas después de segmentar o tokenizar el texto en estudio. Estas entregan información estadísticas (mayor frecuencia, palabras más largas, etc.)

Una herramienta para deducir que términos, o conceptos, son los más relevantes de un texto es la bibilioteca de Python llamada topia.termextract, [22] esta implementa un algoritmo de procesamiento de lenguaje natural basado en Parts-Of-Speech, es decir, marca las palabras mediante el análisis sintáctico de las palabras individualmente y el análisis semántico dentro de las frases para aprender de qué habla cada frase y contabilizar cuantas veces se habla de cada concepto.

Esta biblioteca de Python cuenta con un diccionario de inglés así como de las reglas de

formación de palabras (tiempos verbales, plurales, etc) y las reglas de estructuración de frases en inglés para poder deducir qué tipo de palabras forman cada frase aunque no todas estén en el diccionario.

El algoritmo utilizado entrega palabras claves que pueden ser simples o compuestas, y además éstas pueden ser propias o comunes. Por este motivo toma de manera relevante las palabras claves que son compuestas antes de las simples y los propios por delante de los comunes.

3.2.4 Identificadores de Entidades Nombradas (NER)

El algoritmo encargado del reconocimiento de entidades con nombre (NER) tiene como objetivo delimitar en un texto aquellas frases simples que responden de forma directa a preguntas del tipo ¿quién?, ¿dónde?, ¿cuándo? o ¿cuánto?

Los sistemas NER pueden ser interpretados como un problema de clasificación en el que dado un texto representado como una secuencia de palabras (o tokens) $\hat{w} = w_1 \ldots w_T$ se desea asociar a cada palabra w_i una etiqueta t_i que determina el tipo de entidad que es. [23]

El etiquetado se realiza mediante la técnica POS, identificando cada palabra con su etiqueta correspondiente en función del contexto de la frase. A cada una de las palabras de un texto se le asigna su categoría gramatical (sustantivos, verbos, artículos, etc.). Este proceso se puede realizar de acuerdo con la definición de la palabra o considerando el contexto en que aparece dentro de algún texto. El algoritmo encargado de detectar las entidades nombrada, detecta el comienzo y el final de las mismas, viendo si cada palabra pertenece o no a la entidad nombrada. El conjunto de etiquetas Penn Treebank, se puede ver en el Anexo.

Una vez diseñado el algoritmo reconocimiento de entidades es posible la implementación en el modelo de aprendizaje no supervisado, LDA. Este modelo es aplicado a las clases previamente seleccionadas y el objetivo es capturar el conocimiento implícito es dicho conjunto de datos con la idea de extraer tópicos.

Una de las metas que se ha planteado en este trabajo es investigar qué grado de éxito puede ser alcanzado por el algoritmo NER en la base del modelo LDA.

3.2.5 Modelado de Tópicos

3.2.5.1 Introducción

En esta subunidad se explica la metodología de funcionamiento del algoritmo LDA donde a través de algoritmos de procesamiento de lenguaje natural, se extraen las palabras claves de texto que identifiquen los temas sobre los que trata los documentos. Por lo tanto, hay que decidir que conceptos son más importantes que otros, esto es lo que define finalmente el diccionario y matriz de datos que son clave en un buen modelado de tópicos.

La idea básica del modelo LDA es que los documentos en las clases a modelar se representan como una mezcla al azar sobre temas latentes y cada tema se caracteriza por una distribución de las palabras en los documentos.

3.2.5.2 Modelo LDA

El método LDA es un modelo generativo, lo que significa que trata de describir cómo se crea un documento. Se trata de un modelo probabilístico, ya que dice que un documento se crea mediante la selección de los temas y las palabras de acuerdo a las representaciones probabilísticas del texto natural. Por ejemplo, las palabras que se utilizan para escribir un párrafo se refieren a un subtema de un documento como un todo. Las palabras reales que se usan y lo componen son elegidas en base a ese tema. La probabilidad inherente en los modelos de selección de cada palabra se deriva del hecho de que el lenguaje natural nos permite utilizar múltiples palabras diferentes para expresar la misma idea. Expresar esa idea en el modelo LDA, sirve para crear un documento sin un corpus, lo que se podría determinar cómo una distribución de temas. Para cada palabra del documento que se está generando, se escoge un tema de una distribución de Dirichlet de temas. A partir de ese tema, se coge una palabra elegida al azar basada en otra distribución de probabilidad condicionada en ese tema. Esto se repite hasta que el documento se ha generado.

La idea básica que está detrás del modelo de un corpus con una distribución Dirichlet sobre temas es que los documentos tienen varios temas y estos se superpondrán. Por ejemplo, dentro del corpus de documentos de artículos científicos, se encuentran temas relacionados a biología y medicina. Es probable que haya algunas palabras que se utilizan con más frecuencia cuando se habla de biología, tales como: corazón, pulmón y riñón. El modelo LDA ve esto como un todo y elige los

temas a partir de ahí.

Para extraer los tópicos presentes en las distintas clases seleccionados, este modelo hace uso de un diccionario y una matriz de datos como entrada. A continuación se explican ambos:

1. Cuando los textos son pre–procesados se genera el diccionario de palabras que indexa palabras únicas y los identifica con un identificador o Id.

   El diccionario son todas las palabras que se tendrán en cuenta para tratar nuevos textos en la ejecución del modelo. Es decir, las palabras del diccionario serán todas las palabras que conozca el sistema para deducir información de los textos.

   Para las palabras que aparecen con menos frecuencia en los documentos únicos, pero son comunes en muchos documentos diferentes probablemente es indicativo de que existe un tema común entre los documentos.

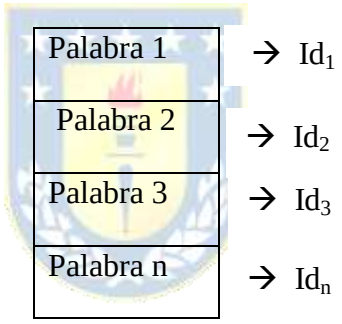

Figura 4: Arreglo asociativo

La figura 4, da a conocer el diccionario creado como un arreglo asociativo de palabras del texto, sin repeticiones, y asignándole una clave única o Id.

2. Luego, la "matriz de datos" es creada a partir del diccionario anterior, el cual contendrá todas las palabras de los textos de los documentos por su identificador seguido del número de apariciones de la palabra en el texto.

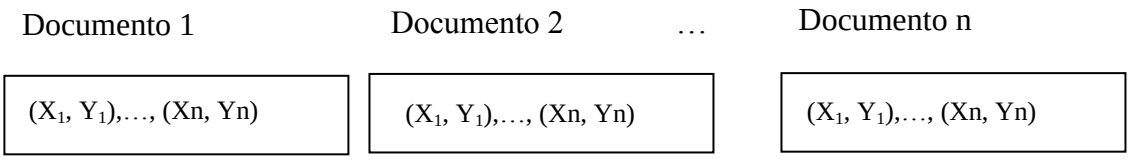

Figura 5: Representación de una lista de listas

La matriz de datos es representada como una lista de listas sin repetición donde cada palabra está formada por un identificador numérico único (X) y su número de apariciones en el texto (Y) (ver figura 5). Es importante que la matriz de datos sea lo más representativa posible del problema que se va a abordar ya que si el diccionario contiene unas palabras y los documentos a tratar no contienen dichas palabras el diccionario no sirve porque no hay identificador para esas palabras.

Considerando ambas entradas descritas, el modulo de Python será el encargado de descubrir los tópicos para cada clase, a través de la biblioteca Gensim [22]. Esta biblioteca ofrece soporte para guardar en memoria de forma optimizada toda esta cantidad de información y realizar los cálculos necesarios para acabar vectorizando cualquier texto y compararlos.

La salida del modelo LDA estará compuesta por dos matrices, donde la asociación de las palabras en el o los documentos está dada por las siguientes probabilidades.

- La matriz P_{wz}, que contiene tantas filas como palabras existen y tantas columnas como tópicos se haya indicado en el algoritmo. Con lo cual en la posición (i,j) de esta matriz se encontrará la probabilidad que tiene la palabra "i" de formar parte del tópico "j".
- La matriz P_{zd}, contiene una fila por cada uno de los documentos existentes y tantas columnas como tópicos se haya indicado en el algoritmo. Con lo que la posición (i, j) de la matriz indicará la probabilidad que tiene el documento "i" de formar parte del tópico "j".

Ambas matrices son el resultado de la asociación de las palabras en un texto, como se muestra en la figura 6.

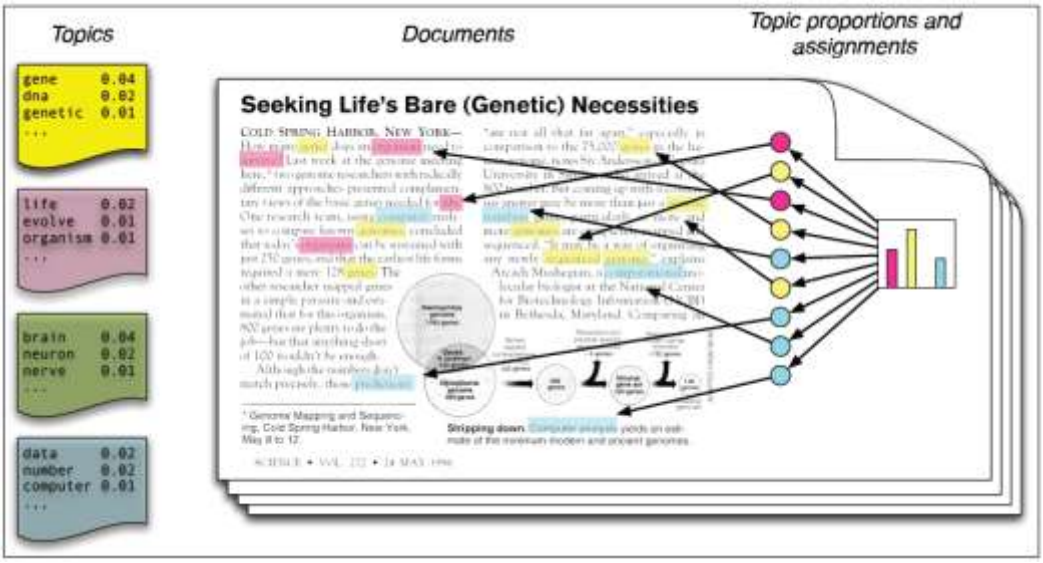

Figura 6: Asociación de las palabras en un texto [extraído de [5]]

Cada documento está caracterizado como una bolsa de palabras, es decir, que el orden no importa. El uso de una palabra es ser parte de un tema y comunica la misma información sin importar dónde se encuentre en el documento. Finalmente el modelo se encarga de encontrar la probabilidad de una palabra en el diccionario dado cada uno de los tópicos seleccionados y descubrir la probabilidad de un tópico dados cada uno de los documentos.

3.2.6 Evaluación del Modelo

Cuando se habla de aprendizaje supervisado, se tiene un conjunto de entrenamiento y un conjunto de prueba para probar el clasificador. Sin embargo, para la clasificación no supervisada, se tiene un conjunto de elementos descritos por un conjunto de características, sin conocer a que clase pertenece cada uno de ellos. Es por esto que para evaluar el clasificador se toman 500 abstract de cada clase, y se hace una revisión manual de cada uno de abstract seleccionados con el fin de hacer una comparación del resultado entregado por el modelo LDA y la información original que entrega el abstract, con esto se determina el desempeño que tiene el modelo en la clasificación de la información. En este caso por razones de tiempo y presupuesto es el mismo autor de este informe quien revisa manualmente los artículos.

Una medida de evaluación es la precisión del clasificador. La cual es definida como la probabilidad de que un documento etiquetado con la clase i corresponda realmente a esa clase, es decir, mide el número de términos correctamente reconocidos respecto al total de términos predichos, sean estos verdaderos o falsos. [24] La precisión representa el nivel de confianza del clasificador dado que una clase especifica ha sido predicha. La precisión mide el número de documentos correctamente clasificados respecto del total de los clasificados y viene establecido por la siguiente fórmula:

$$precisión = \frac{TP}{TP + FP} \tag{3.1}$$

Donde:

- TP (True positives): Representan el total de documentos que han sido correctamente clasificados y pertenecen a la clase.
- FP (False positives): Número de documentos clasificados en la clase siendo estos erróneos.
- FN (False negatives): Número de documentos que pertenecen a la clase y que no han sido clasificados como tal.
- TN (True negatives): Aquellos documentos que no pertenecen a clase y no han sido asignados en ella.

Una alternativa de medir la bondad del clasificador o tasa de error es la matriz de confusión. La matriz de confusión permite visualizar mediante una tabla de contingencia la distribución de errores cometidos por un clasificador.

Esta matriz de confusión para el caso de dos clases tiene la siguiente apariencia, como se puede apreciar en la Tabla 2.

Tabla 2: Matriz de confusión para las distintas clases a modelar [Extraído de [25]]

Predicción		
	Clase 1 = Pertenece	Clase 2 = No Pertenece
Clase 1 = Pertenece	TP	FN
Clase 2 = No Pertenece	FP	TN

Capítulo 4. Resultados

4.1. Implementación

4.1.1 Introducción

El proceso de implementación del modelo se desarrolla en dos etapas bien definidas. La primera de ellas se centra en la adquisición de información y creación de una estructura de datos que contenga la información más representativa. Para ello se accede a la base de datos de MySQL y se lleva a cabo el procesamiento del texto.

La segunda etapa tiene que ver con el desarrollo del algoritmo de extracción de tópicos, así como la presentación de los resultados.

4.1.2 Adquisición de la información

En esta fase se recupera toda la información de BioMed Central. Para ello nos servimos de la base de datos de MySQL Articles. En esta primera aproximación, se centra únicamente en los Abstract de la base de datos dedicada a la investigación.

Para el modelado de tópicos se utilizan los Abstract de los artículos científicos que contengan varios párrafos de información (ver figura 7), por ejemplo Background, Objectives, Methods, Results, Conclusions, Discussion o combinación de ellos. De esta forma se asegura que la información a estudiar sea más completa.

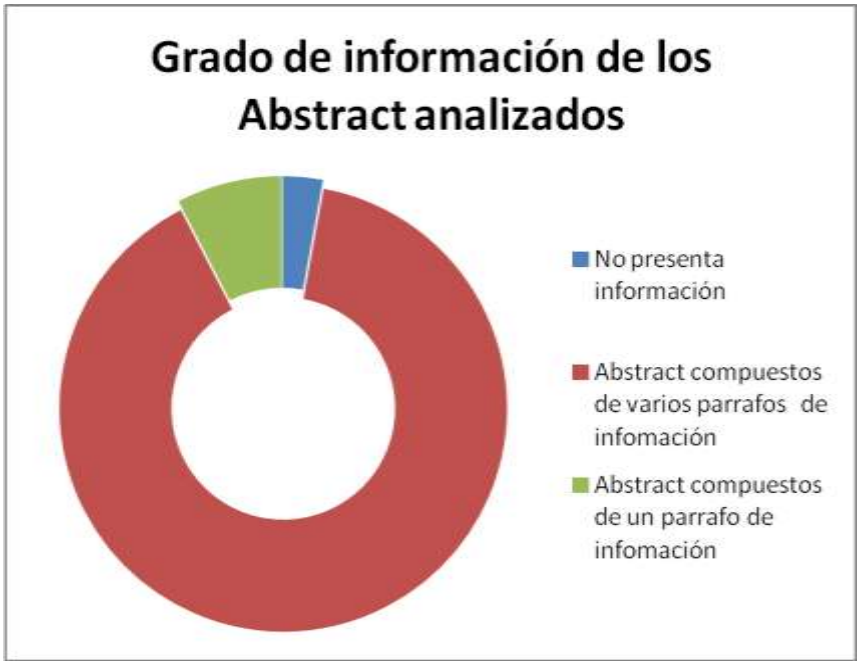

Figura 7: Grado de información de los Abstract

A través de la información seleccionada se busca encontrar similaridad entre los temas, creando 3 clases representativas. Estos serán los modelados finalmente. Ver tabla 3.

Tabla 3: Clases a modelar

Clase 1	**Clase 2**	**Clase 3**
"lung"	"kidney-heart"	"lung-kidney-heart"
3.554 Abstract	3.903 Abstract	7.132 Abstract

4.1.3 Desarrollo del algoritmo

Se decide crear 2 diccionarios para efectos de funcionamiento del modelo LDA, como se mencionó anteriormente el diccionario es clave en el modelado de tópicos.

Diccionario N°1.

Una vez seleccionada la clase se requiere de una herramienta para deducir que términos, o conceptos, son los más relevantes en un texto. Para ello se utiliza la biblioteca de Python llamada topia.termextract [22] que implementa un algoritmo de procesamiento de lenguaje natural basado en POS Tagging, es decir, marca las palabras individualmente a través del análisis sintáctico y análisis semántico dentro de las frases para aprender de qué habla cada frase y contabilizar cuantas veces se habla de cada concepto. Para esto, el algoritmo de Topia cuenta tanto con un diccionario de inglés así como de las reglas de formación de palabras (tiempos verbales, plurales, etc) y reglas de estructuración de frases en inglés para poder deducir qué tipo de palabras forman cada frase aunque no todas estén en el diccionario.

Diccionario N°2.

El diccionario 2 se crea a partir del reconocimiento de entidades con nombre (NER), en el cual se persigue delimitar el texto por oraciones. Por cada oración se hace un pre procesamiento del texto, donde este es normalizado, segmentado, eliminando los símbolos de exclamación, números, etc, en general todo aquello que genere ruido. Seguido de esto, las palabras son etiquetadas usando POS Tagging, rotulando las palabras del texto seleccionado, agregando la categoría gramatical (sustantivos, verbos, artículos, etc.) a cada una ellas por oración correspondiente.

Las herramientas utilizadas en el etiquetado POS son las proporcionadas por NLTK [22], de esta herramienta se escoge como modalidad de que la etiqueta gramatical aparezca al lado de la palabra asociada. Figura 8.

```
('recommendations', 'NNS'),
('arising', 'VBG'),
('from', 'IN'),
('evaluation', 'NN'),
('of', 'IN'),
('the', 'DT'),
('program', 'NN'),
('were', 'VBD'),
('used', 'VBN'),
('in', 'IN'),
('subsequent', 'NN'),
('stages', 'NNS'),
('of', 'IN'),
('implementation', 'NN')
```

Figura 8: Etiquetado POS en una oración.

El módulo NER reconoce la etiqueta gramatical por oración e identifica la entidad a seleccionar, pueden ser nombres propios, organizaciones, fechas y lugares o combinación de ellos. Luego la entidad gramatical a identificar es:

"NP: {<DT>?<JJ>*<NN|NNS>+}"

La entidad anterior indica que por cada oración se reconocerá el Articulo (DT) como "Opcional", Adjetivos (JJ), desde 0 a muchos y Sustantivo (NN), ya sea en singular o plural, considerando 1 o más.

Para cada oración del texto se genera un árbol, donde los nodos "NP" corresponden a las entidades reconocidas. Figura 9. Finalmente, estas serán las utilizadas como diccionario en el modelo de tópicos.

Figura 9: entidad gramatical reconocida

Por cada diccionario descrito anteriormente, se genera una matriz de datos distinta, de esta manera se obtendrán en los resultados tópicos que dependen directamente del diccionario creado.

La herramienta utilizada en la extracción de tópicos es Gensim [22], la cual es una biblioteca escrita en Python, para el aprendizaje no supervisado de los artículos científicos, contiene algoritmos que incluyen el análisis semántico latente (LSA, LSI) y la asignación latente de Dirichlet (LDA), este último utilizado en la clasificación de la base de datos de BioMed Central.

4.2. Resultados

4.2.1 Introducción

Una vez determinado el tipo de algoritmo, diccionario y cantidad de tópicos, el siguiente paso es determinar los tópicos o conjunto de palabas asociados que se quieren analizar en el campo de la clasificación.

La cantidad de tópicos a extraer depende de la información entregada, es por esto que una información variada provoca que un tópico descubierto por observación sea dividido en dos tópicos más pequeños pero altamente relacionados. Este suceso ocurre en todo dominio que se quiera analizar, sin embargo, a medida que el dominio bajo análisis es más amplio, la cantidad óptima de tópicos por periodo aumenta. Por lo tanto, es necesario ajustar el modelo LDA dependiendo del dominio bajo análisis. Es por esto, que se decide extraer 10 tópicos porque si bien hay una mayor fragmentación de la información al determinar más tópicos, sucede lo contrario si se decide tomar tópicos pequeños independientes que son absorbidos por ellos cuando la cantidad de tópicos por clases disminuye.

Los resultados se mostrarán en gráficos y tablas que se presentan en la sección 4.2.2, donde se busca descubrir: Los diferentes tópicos que se utilizan en cada uno de los documentos; La forma en la que están combinados varios tópicos en un mismo documento; Las palabras o términos que pertenecen a cada tópico; y finalmente la clasificación de los documentos no visualizados en estos tópicos.

4.2.2 Modelado de tópicos por clases

4.2.2.1 Clase “lung”

Como primera modelación de tópicos, se utiliza la clase “lung”, la cual contiene 3.554 abstract, a este set de documentos se construye el modelo LDA para la extracción de tópicos, teniendo como parámetros los diccionarios de palabras descrito en la sección 4.1.3

La información modelada en esta clase hace referencia a documentos que contienen la palabra “lung”, de ello se obtienen subtemas de la información en la clasificación final.

Los gráficos 10 y 11 presentan los resultados de la modelación de tópicos de la clase “lung”. Luego, para analizar la información entregada por los gráficos, dichos tópicos pueden ser categorizados como tendencias, se escoge un umbral prevalente, superior a la media, para ver la separabilidad entre tópicos.

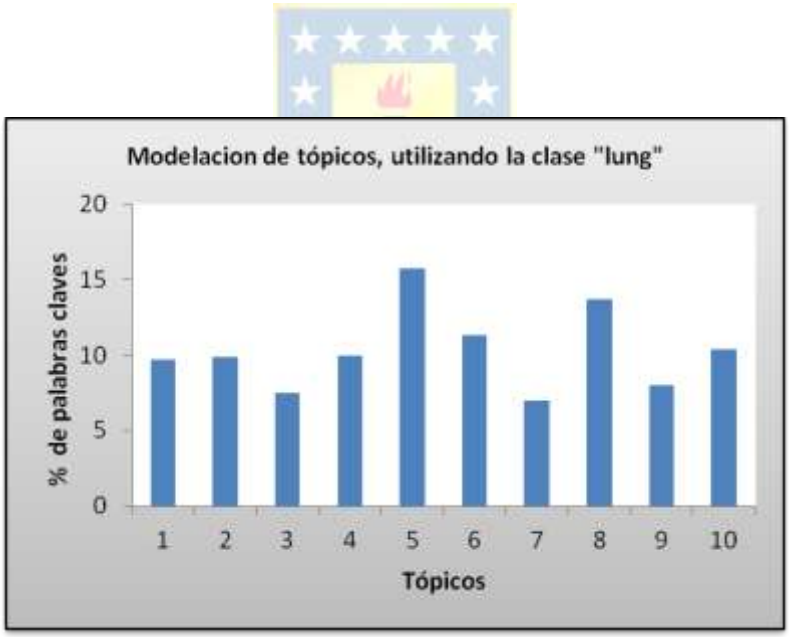

Figura 10: Resultados modelo LDA, usando el diccionario n°1 para los documentos que pertenecen a la clase “lung”

Del grafico anterior (figura 10), se toma un umbral mayor a 8.5 correspondiente al porcentaje de palabras claves que prevalecen dentro de cada tópico. Los tópicos superiores a este umbral son 1, 2, 4, 5, 6, 8, y 10.

El diccionario para la clase “lung” creado a través de la biblioteca Topia de Python, cuenta con 71.169 términos de palabras claves asociados a la temática modelada. Los términos asociados a cada tópico se muestran en la tabla 4.

Tabla 4: Términos identificados en el análisis de la modelación de tópicos para los documentos de la clase “lung”

Tópico 1	lung -group -cell -protein -cancer -injury -carcinoma -patient -tumor -case
Tópico 2	lung -peep -patient -group -cell -particle -effect -tumor -expression
Tópico 4	patient -lung -cancer -disease -cell -group -tumor - mice
Tópico 5	cell -lung -cancer -expression -tumor -gene -mice -patient -infection
Tópico 6	patient -cell -expression -lung -disease -nsclc -cd -level
Tópico 8	cancer -lung -patient -study -gene -expression -lungcancer -cell -egfr
Tópico 10	gene -cell -tumor -lung -expression -patient -cancer -study -group

De acuerdo a la tabla 4, los tópicos 1, 2, 4, 5, 6, 8, y 10 son los más representativos en la modelación, por presentar un mayor porcentaje de términos de palabras asociados a cada tópico. Cada tópico entregado contiene una temática diferente en la clasificación. Es por esto que cada palabra clave tiene relación al tópico modelado.

Analizando de forma general cada tópico y sus términos asociados, se puede determinar que el tópico 1 hace referencia a patologías que involucran al pulmón, el tópico 2 entrega información de los síntomas o efectos de las enfermedades del pulmón en los pacientes, el tópico 4, hace mención en el estudio de ratas sobre patologías asociados al cáncer de pulmón, el tópico 5, da a conocer la relación en investigaciones con ratas y sus aplicaciones en humanos, el tópico 6, entrega información sobre el desarrollo y progresión del cáncer de pulmón (nsclc), el tópico 8, da a conocer los estudios en el tratamiento de la enfermedad, como epidermal growth factor receptor (egfr), el tópico 10, hace mención de la expresión génica de cuyas patologías que involucran el pulmón en su estudio.

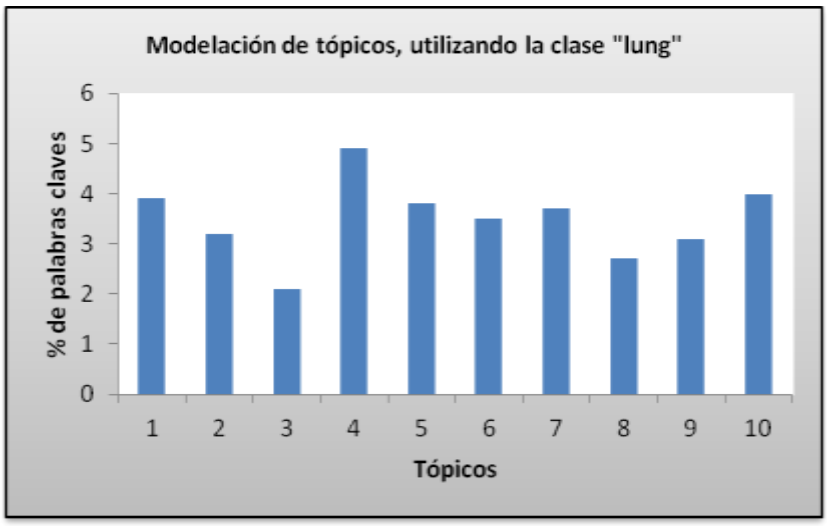

Figura 11: Resultados modelo LDA, usando el diccionario n°2 para los documentos que pertenecen a la clase "lung"

En la figura 11 se muestran los resultados de la modelación de tópicos utilizando el diccionario creado a través de la entidad gramatical descrita en la sección 4.1.3. El diccionario contiene 93.680 términos de palabras claves asociados a la clase "lung".

En la distribución del porcentaje de palabras claves se elige un umbral mayor a la media para determinar los términos más representativos en cada tópico. Los tópicos superiores a esta media corresponden a 1, 2, 4, 5, 7, y 10. Aquellos tópicos contienen palabras claves que son más influyentes que los demás.

Tabla 5: Términos identificados en el análisis de la modelación de tópicos para los documentos de la clase "lung"

Tópico 1	**patients -cells -expression -this study -hours -genes -smokers -lung cancer -lung injury**
Tópico 2	patients -mice -this study -genes -days -the expression -the lung -expression -the lungs
Tópico 4	patients -cells -expression -this study -mice -treatment -days -levels -vitro -vivo
Tópico 5	patients -years -lung cancer -cancer -this study -expression -age –women -genes -the lung

Tópico 7	patients -this study -case presentation -lung cancer -treatment -the patient -months -cases -mortality
Tópico 10	patients -asthma -months -mice -this study -infection -years -weeks -children -the patients

La tabla 5 muestra los resultados de la modelación de tópicos. Cada tópico contiene términos o palabras claves que determinan el subtema de la clase principal. Por ejemplo el tópico 1, hace mención al fumador y la relación con el cáncer de pulmón, el tópico 2 da a conocer estudios en ratas sobre expresiones en el pulmón, el tópico 4 entrega información acerca de la relación en investigaciones con ratas y sus aplicaciones en humanos, el tópico 5, hace mención a la expresión génica del cáncer de pulmón, el tópico 7 hace mención al cáncer de pulmón como caso de estudio, y el tópico 10 entrega información acerca de estudios del asma y su relación en ratas y posibles infecciones en niños.

- **Evaluación del modelo utilizando la clase "lung"**

El primer paso para determinar el desempeño del modelo LDA es encontrar la probabilidad más alta de acierto de un documento al tópico correspondiente. La respuesta correcta será dada por el número de resultados correctos e incorrectos que han sido obtenidos de forma manual. Luego, la precisión a obtener es el promedio de etiquetas acertadas frente a las predichas por el LDA.

Las figuras 12 y 13 muestran la modelación de tópicos de los documentos pertenecientes a la clase "lung". Cada documento es clasificado determinando la mayor probabilidad de acierto al tópico correspondiente.

La tabla 6 y 7 muestran el valor de precisión en función del número de resultados correctos e incorrectos que han sido obtenidos de forma manual.

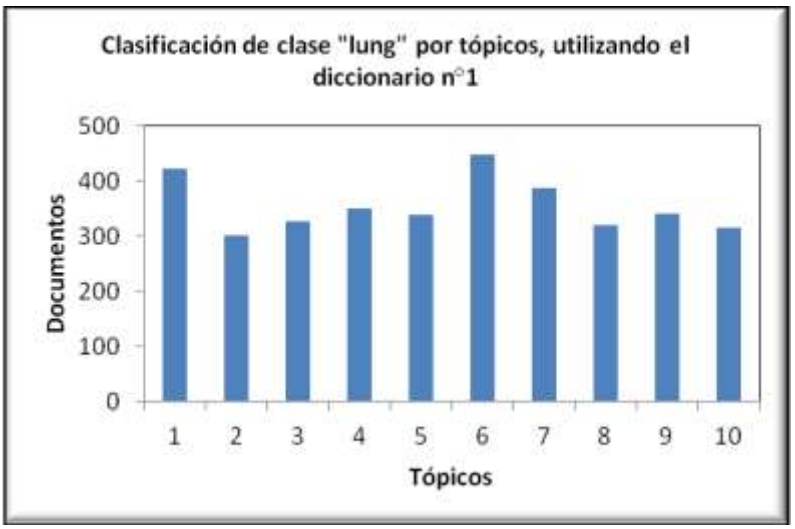

Figura 12: Resultados modelo LDA

Tabla 6: Evaluación clase "lung"

Clase	TP	FP	Precisión
Lung	360	140	0.72

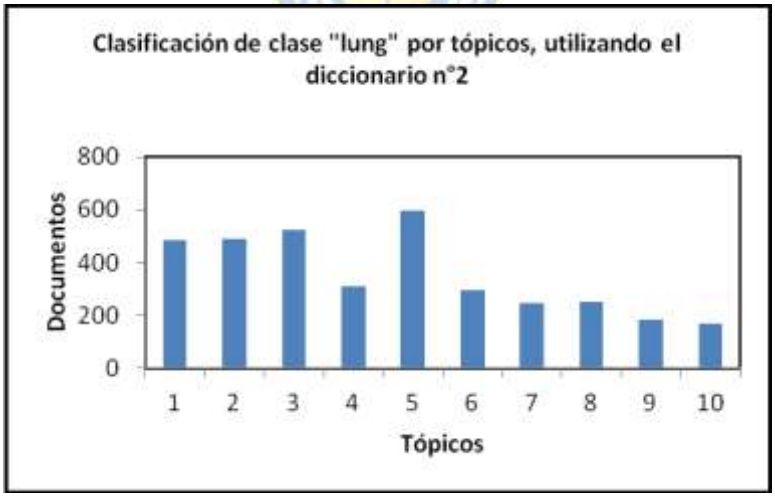

Figura 13: Resultados modelo LDA

Tabla 7: Evaluación clase "lung"

Clase	TP	FP	Precisión
Lung	**261**	**239**	**0.552**

- **Conclusiones de la modelación de tópicos utilizando la clase "lung"**

Los resultados de la modelación de tópicos para la clase "lung", usando los diccionarios descritos en la sección 4.1.3, muestran que los términos únicos del modelo LDA usando el reconocimiento de una entidad gramatical contienen más términos únicos en comparación con el diccionario creado por Topia. Sin embargo, el trabajar con entidades gramaticales en la extracción de información implica que por cada tópico, los términos identificados entreguen más información.

En las tablas 4 y 5 se muestran los resultados de la modelación de tópicos, en los cuales los términos identificados en el análisis de la modelación de tópicos son similares. Esto indica que los documentos científicos presentan como temática principal la palabra "lung".

En la evaluación se da a conocer que la cantidad de documentos acertados influye de manera significativa en la precisión del modelo implementado. Un sistema puede acertar siempre la categoría de cada documento porque clasifica pocos documentos, es decir, es un sistema muy preciso pero poco exhaustivo.

4.2.2.2 Clase "kidney -heart"

En esta sección se da a conocer la modelación de tópicos de la clase "kidney-heart", la cual contiene un total de 3.903 de documentos asociados a temas de riñón y corazón.

En las tablas 8 y 9 se muestran los resultados de la modelación para cada diccionario descrito en la sección 4.1.3. A partir del cual se analiza la separabilidad de la información, identificando los términos o palabras claves pertenecientes a cada tópico en base a probabilidades.

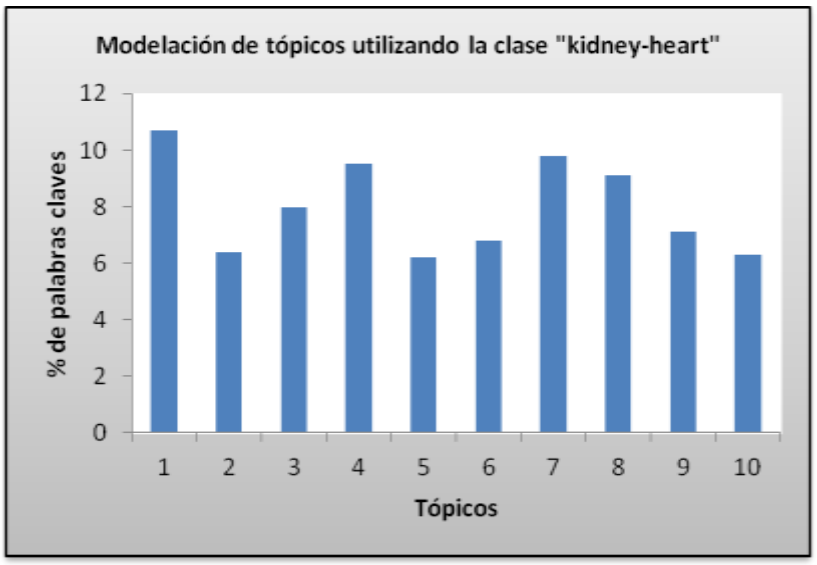

Figura 14: Resultados modelo LDA, usando el diccionario n°1 para los documentos que pertenecen a la clase "kidney-heart"

En la figura 14 se puede ver la distribución del porcentaje de palabras claves perteneciente a cada tópico como resultado de la clasificación no supervisada de la clase "kidney-heart". El diccionario de palabras utilizado por el modelo LDA es generado por Topia. Para esta clase el diccionario contiene 79.018 términos de palabras asociados a los documentos ya definidos. Luego, en la tabla 8 se da a conocer los términos asociados a cada tópico.

Tabla 8: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase "kidney-heart"

Tópico 1	**patient -group -aki -blood -pressure -control -study -kidney**
Tópico 2	patient -kidney -study -heart -mice -diabetes -protein -p -failure -disease
Tópico 3	cell -group -chd -patient -risk -level -protein -rat -
Tópico 4	patient -study -risk -group -treatment -health -factor -ci -disease
Tópico 5	cell -p -heart -expression -rat -gene -infection -protein
Tópico 6	cell -p -patient -tumor -kidney -tissue -day -expression -gene
Tópico 7	patient -heart -disease -rate -cancer -failure - -group -activity -mortality
Tópico 8	patient -study -disease -risk -heart -cell -gene
Tópico 9	gene -patient -model -level -risk -il -disease -expression -factor
Tópico 10	patient -level -group -heart -study -p -gene -syndrome

De acuerdo a la tabla 8 cada tópico contiene términos que lo definen. Analizando la información de dichos tópicos, los términos o palabras claves hacen referencia a los temas de corazón y riñón por la clase modelada. Por ejemplo, el tópico 1, hace referencia al riñón por contener aki. El tópico 2, tiene relación a ambos temas, corazón y riñón en estudios que involucran ratas. El tópico 3, indica relación al corazón por contener el término chd (coronary heart disease). El tópico 4, hace mención a estudios que involucran tratamientos para la salud. El tópico 5, indica relación al corazón en estudio de ratas. El tópico 6, informa acerca de patologías del riñón. El tópico 7, da a conocer información de insuficiencia cardiaca. El tópico 8, entrega información de estudios en pacientes con enfermedades al corazón. El tópico 9, hace mención del factor genético en patologías. Finalmente, el tópico 10, hace referencia al corazón.

La figura 15, muestra la distribución del porcentaje de términos por tópico como resultado de la clasificación de la clase “kidney-heart”. En esta modelación se utiliza el diccionario de palabras creado a través del reconocimiento de una entidad gramatical, descrito en la sección 4.1.3. El diccionario contiene 108.154 entidades reconocidas.

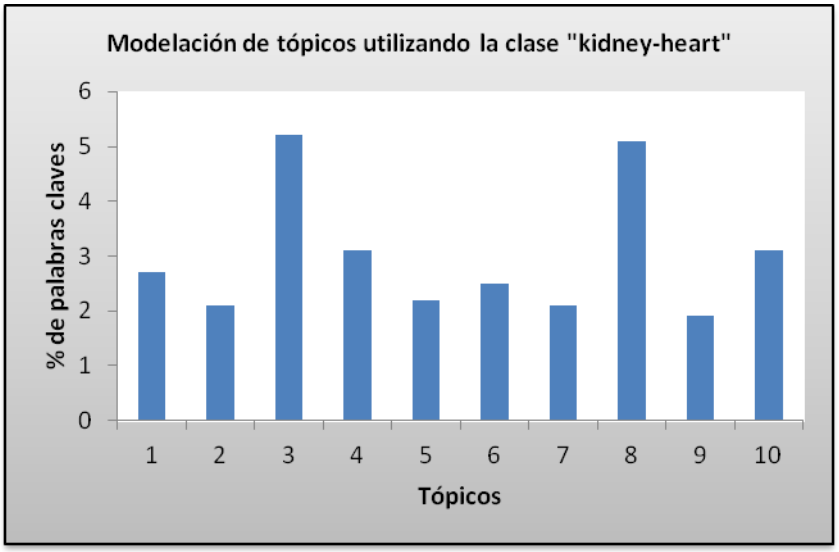

Figura 15: Resultados modelo LDA, usando el diccionario n°2 para los documentos que pertenecen a la clase “kidney-heart”

Tabla 9: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase "kidney-heart"

Tópico 1	**patients -this study –aki-years -hours -months -mortality -age -genes**
Tópico 2	patients -this study -weeks -the study -levels -hours -treatment -kidney -the effects
Tópico 3	patients -this study -risk -years -diabetes -hypertension -death -the presence -days -genes
Tópico 4	patients -aki -weeks -this study -cpb -baseline -children -heart failure -mortality
Tópico 5	patients -mortality -this study -treatment -years -women -the heart -risk -pigs
Tópico 6	patients -cells -mice -this study -years -expression -the study -genes -participants -levels
Tópico 7	patients -cells -expression -this study -proteins -genes -the heart -minutes -the present study
Tópico 8	patients -heart -age -years -this study -mortality -the patients –levels
Tópico 9	patients -this study -case presentation -days -years -subjects -mortality -hours –death
Tópico 10	patients -this study -women -years -men -days -subjects -mice –group

En la tabla 9 se da a conocer los términos o palabras claves asociados a cada tópico clasificado. En la clasificación no supervisada se utiliza el reconocimiento de una entidad gramatical (sección 4.1.3) para la creación del diccionario de palabras como entrada al modelo LDA. Los temas asociados a cada tópico definidos por el conjunto de palabras claves, hacen mención a: Tópico 1: Patologías asociadas al riñón como la lesión renal aguda (aki). El tópico 2: Estudio y tratamientos de enfermedades del riñón. El tópico 3: Patologías de origen genético como lo es la hipertensión y diabetes. El tópico 4: insuficiencia cardiaca y su relación lesiones renales. El tópico 5: Estudio y tratamientos de enfermedades del corazón. El tópico 6: Estudios y expresión génica utilizando ratas. El tópico 7: Estudio y expresión génica del corazón. El tópico 8: Mortalidad de pacientes con

enfermedad asociada al corazón. El tópico 9: Mortalidad como efecto de enfermedades en casos de estudio. El tópico 10: Estudio en ratones que involucran un diagnostico en pacientes.

- **Evaluación del modelo utilizando la clase “kidney-heart”**

Se determina el desempeño del modelo LDA encontrando la probabilidad más alta de acierto de un documento dado el tópico al cual corresponda. La respuesta correcta será dada por el número de resultados correctos e incorrectos que han sido obtenidos de forma manual. Luego, la precisión a obtener es el promedio de etiquetas acertadas frente a las predichas por el LDA.

Las figuras 16 y 17 muestran la distribución de los documentos pertenecientes a la clase “kidney-heart”. Cada documento es clasificado determinando la mayor probabilidad de acierto al tópico correspondiente.

Las tablas 10 y 11 muestran el valor de precisión en función del número de resultados correctos e incorrectos que han sido obtenidos de forma manual.

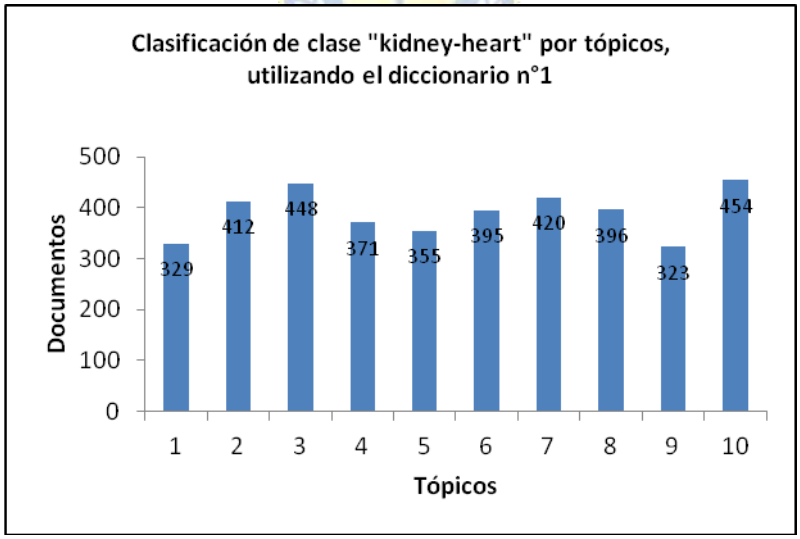

Figura 16: Resultados modelo LDA

Tabla 10: Evaluación clase “kidney-heart”

Clase	TP	FP	Precisión

kidney-heart	**379**	**121**	**0.758**

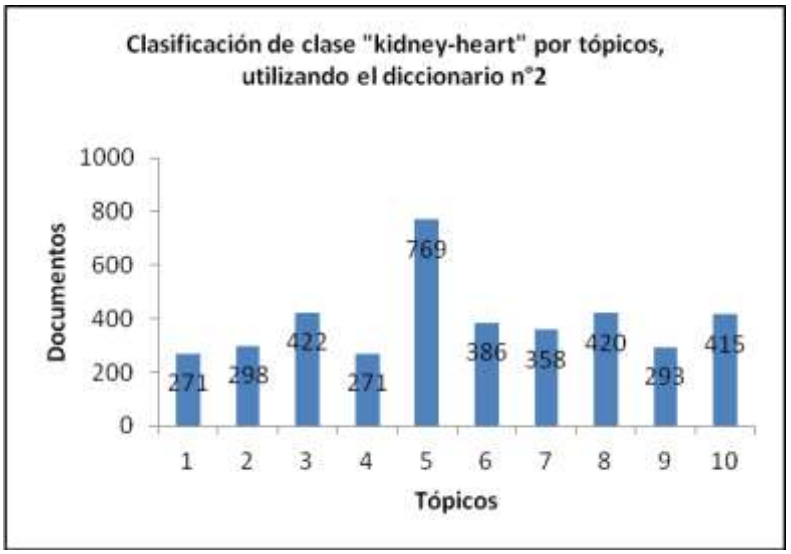

Figura 17: Resultados modelo LDA

Tabla 11: Evaluación clase "kidney-heart"

Clase	TP	FP	Precisión
kidney-heart	**285**	**215**	**0.57**

- **Conclusiones de la modelación de tópicos utilizando la clase "kidney-heart"**

Analizando los resultados obtenidos por el modelo, se puede decir que los tópicos extraídos muestran ciertas diferencias de la temática modelada (riñón y corazón).

En los gráficos 14 y 15, existen tópicos que son más influyentes que otros. En el caso del gráfico 14, los tópicos más representativos son los etiquetados como 1, 4, 7, y 8, mayor a 8 % de palabras claves por cada tópico. En el gráfico 15, los tópicos más influyentes son 3, 4, 8, y 10.

El diccionario asociado a cada clase modelada es primordial, ya que de este dependen los términos que se entreguen a cada tópico obtenido. El diccionario creado a través de la biblioteca Topia tiene mejor resultado en la clasificación de la clase "kidney-heart", por contener menos términos y además, porque cada término es una palabra clave fundamental en el documento.

Cabe hacer hincapié en que cada ejecución del modelo depende del número de documentos y del número de palabras claves que este posea en el diccionario. El diccionario creado a través del reconocimiento de una entidad gramatical contiene más términos claves, por lo que lo hace más complejo en el manejo de palabras claves asociados a cada tópico.

4.2.2.3 Clase “lung-kidney-heart”

Dada la dificultad de encontrar un corpus de documentos representativo, donde los temas sean conocidos y completos, para poder evaluar los resultados del modelo LDA, se recurre a la clasificación de la clase “lung-kidney-heart”. En esta clase prevalecen temas asociados al pulmón, riñón, corazón y documentos de temas desconocidos, los resultados son evaluados a partir de la temática conocida de la clase.

La clase “lung-kidney-heart” contiene 18.901 documentos, de los cuales 7.132 son temas conocidos (pulmón, riñón y corazón). A este set de documentos se aplica la metodología descrita en la sección 4.1.3, donde el diccionario de palabras claves corresponde a unas de las entradas del modelo LDA en la clasificación no supervisada.

Los tópicos extraídos constan de un conjunto de términos de palabras claves. Analizando las palabras asociadas a cada tópico, se considera enriquecida si aparece relacionada a, como mínimo, un número determinado de palabras.

Los gráficos 18 y 19 muestran la distribución de las palabras claves por cada tópico clasificado por el modelo LDA.

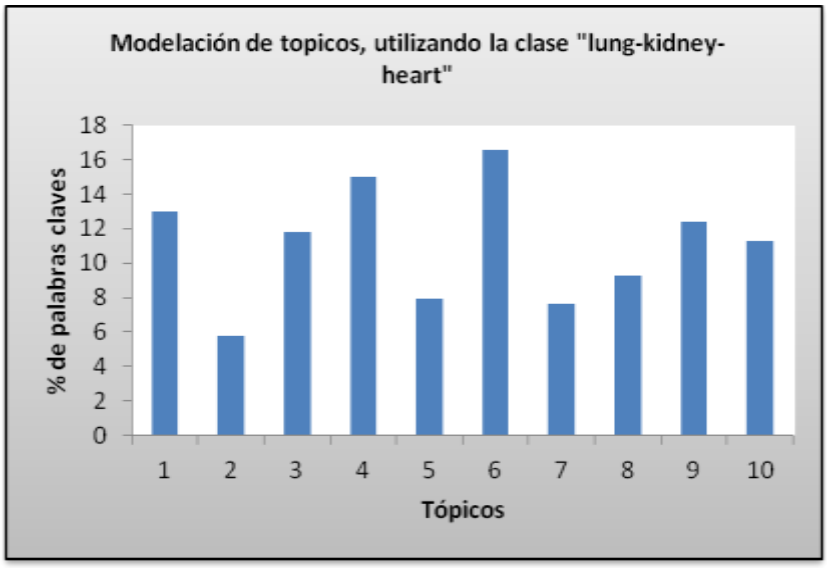

Figura 18: Resultados modelo LDA, usando el diccionario n°1 para los documentos que pertenecen a la clase "lung-kidney-heart"

La tabla 12 muestra los resultados de la modelación de tópicos de la clase "lung-kidney-heart". El diccionario utilizado por el modelo LDA es creado a través de la librería Topia descrita en la sección 4.1.3, este diccionario contiene 326.389 palabras claves.

Tabla 12: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase "lung-kidney-heart"

Tópico 1	**gene- sequence-genome-region-dna -family -snp -analysis -methylation**
Tópico 2	cell -population -map -chromosome-model -egfr -recombination -species -heart
Tópico 3	gene -expression -network -promoter -species -strain -study
Tópico 4	cancer -lung -breast -tumor -expression -cell -level -effect - il
Tópico 5	gene -ventilation -peep -icu -plant -cell -carcinoma -shock -lung -expression
Tópico 6	patient -group -risk -study -pressure -aki -mortality -factor -disease
Tópico 7	ml -kidney -expression -brca -group -receptor -stress
Tópico 8	mutation -patient -disease -test -ci -trial -study -association -rate
Tópico 9	cell -protein -infection -disease -interaction -hospital -function -airway
Tópico 10	gene -virus -method -strain -genome -study -sample -snp -cluster -association

El objetivo de la clasificación no supervisada de los documentos pertenecientes a la clase "lung-kidney-heart" es determinar tópicos asociados a los temas "lung", "kidney" y "heart" que son temas conocidos dentro de esta clase. Sin embargo a este sets de documentos, se agregaron temas desconocidos.

De acuerdo a la tabla 12 los términos asociados a cada tópico entregan información de los documentos de la clase descrita. Al analizar cada uno de ellos, los tópicos 2, 4, 5, 6, 7, y 9 hacen referencias a temas relacionados al pulmón, riñón y corazón. El tópico 2, hace mención al corazón.

El tópico 4, entrega información del pulmón y patologías asociadas a este. El tópico 5, hace mención al pulmón, considerando síntomas. El tópico 6, tiene relación al riñón. El tópico 7, hace mención al riñón y a la mutación BRCA asociado al cáncer. El tópico 9, hace referencia al pulmón, por contener el término “airway”.

La figura 19 muestra los resultados de la modelación de la clase“lung-kidney-heart” utilizando como entrada el diccionario de reconocimiento de entidad gramatical (sección 4.1.3) al modelo LDA. El diccionario creado a través de una entidad gramatical, busca en el texto pre procesado los términos que hacen referencia a la gramática buscada. Este contiene un total de 438.763 identidades reconocidas. Cada tópico descrito contiene una distribución de términos o palabras claves asociados que lo identifican.

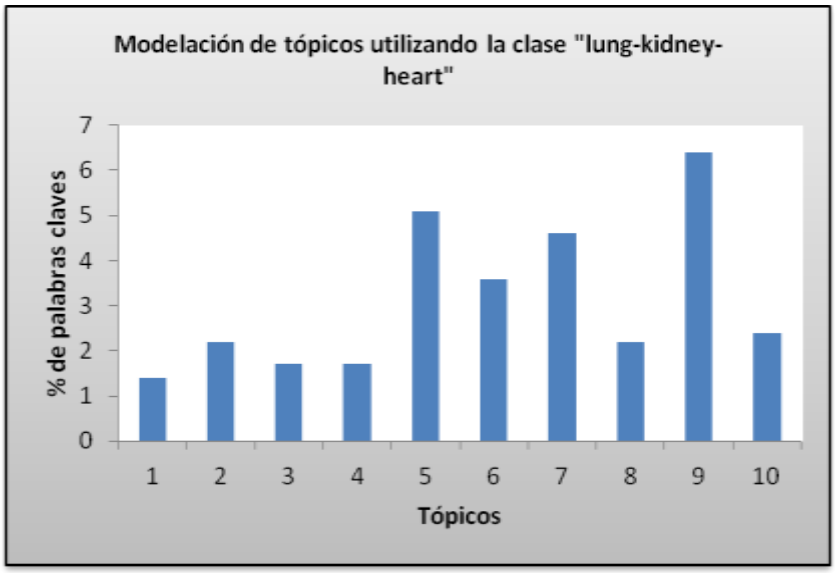

Figura 19: Resultados modelo LDA, usando el diccionario n°2 para los documentos que pertenecen a la clase “lung-kidney-heart”

Tabla 13: Términos identificados en el análisis de modelado de tópicos para los documentos de la clase “lung-kidney-heart”

Tópico 1	**species -the evolution -populations -viruses -this study -phylogenetic analysis -schizophrenia -high risk -animals -radiation**

Tópico 2	mutations -individuals -regions -this study -genetic variation-populations -confidence interval -variation -mutation - polymorphism
Tópico 3	this study -strains -peep -years -individuals -patients -isolates -dogs -acute respiratory distress syndrome -chromosome
Tópico 4	chemotherapy -snps -males -a case -females -this study -this review -cells -ad -mice
Tópico 5	genes -proteins -gene expression -expression -gene -the genome -analysis -the genes -function -the expression
Tópico 6	patients -days -children -ill patients -case presentation -this study -oxygenation -years -the diagnosis -mechanical ventilation
Tópico 7	patients -years -hours -this study -age -lung injury -baseline -months -groups
Tópico 8	patients -this study -chromosome -markers -tumors -rare variants -cm -individuals -linkage -survivors
Tópico 9	patients -expression -cells -this study -women -controls -treatment -mice -risk
Tópico 10	aki -cells -genes -mutations -expression -acute kidney injury aki -vitro -loss -methylation -a patient

Los términos que se asocian a cada tópico entregan información de la temática principal que presenta cada uno de ellos. En la tabla 13 se muestran los términos o palabras claves, donde la clasificación modelada utiliza el diccionario de reconocimiento de entidad gramatical descrito en la sección 4.1.3.

Al analizar cada tópico se puede determinar que los tópicos 3, 6, 7, y 10 hacen mención a los temas conocidos de la clase modelada. El tópico 3, hace mención a patologías del pulmonar. El tópico 6, hace referencia a funcionalidades del pulmón. El tópico 7, hace mención a estudios relacionado al pulmón. El tópico 10 da a conocer patologías o enfermedades que se asocian al riñón. Los demás tópicos descritos hacen mención a temas desconocidos.

- **Evaluación del modelo utilizando la clase “lung-kidney-heart”**

Debido a que la clase analizada es muy grande, no es posible evaluar uno a uno los documentos, y ver cuáles son los documentos más relevantes. Es por ello que se analizan de manera manual un total de 500 artículos. Este conjunto de documentos son los encargados de decir en último término si son relevantes o no. Este método asume que la gran mayoría de los documentos relevantes son encontrados, y los no recuperados pueden considerarse como no relevantes.

Las figuras 20 y 21 muestran la distribución de los documentos pertenecientes a la clase "lung-kidney-heart". Cada documento es clasificado determinando la mayor probabilidad de acierto al tópico correspondiente.

La tabla 14 y 15 muestra el valor de precisión en función del número de resultados correctos e incorrectos que han sido obtenidos de forma manual.

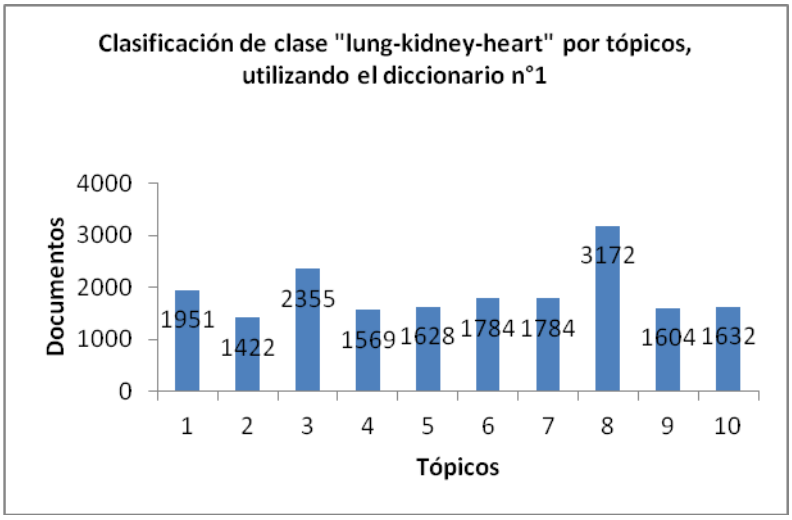

Figura 20: Resultados modelo LDA.

Tabla 14: Evaluación clase "lung-kidney-heart"

Clase	TP	FP	Precisión
lung-kidney-heart	**204**	**296**	**0.408**

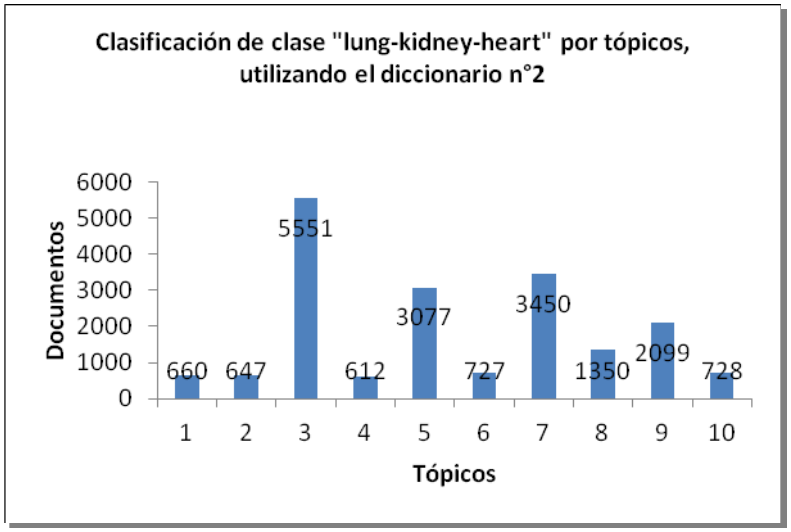

Figura 21: Resultados modelo LDA.

Tabla 15: Evaluación clase "lung-kidney-heart"

Clase	TP	FP	Precisión
lung-kidney-heart	**168**	**332**	**0.336**

- **Conclusiones de la modelación de tópicos utilizando la clase "lung-kidney-heart"**

Los resultados de la modelación de la clase "lung-kidney-heart" entregan tópicos de temas conocidos (pulmón, riñón, corazón) y no conocidos de esta clase. En las tablas 12 y 13 mostradas, se puede ver como se logran ver tópicos importantes por parte del set de datos entregado. Además de esto, los términos de palabras claves representan muy bien los tópicos, considerando los resultados de ambos diccionarios implementados.

Debido a la gran cantidad de palabras o términos claves existentes en los diccionarios, se ve como en esta modelación hay tópicos y términos más influyentes que otros.

Capítulo 5. Conclusiones

5.1. Conclusión

En el presente trabajo se ha propuesto un método de modelación de tópicos, con el fin de estudiar una base de datos con documentos científicos, para la extracción de información biológica a partir de grandes listas de términos resultantes del análisis de bases de datos de BioMed Central, en particular se escogen temas relacionales o clases. Ambas elecciones se realizan en base a la alta cantidad de información disponible que existe hoy en día en el ámbito científico.

En la búsqueda de información se decide trabajar solo con los abstract de la base de datos porque contiene un formato estándar de información y además porque se encuentra en idioma inglés, lo cual facilita el trabajo, ya que las bibliotecas utilizadas están diseñadas de tal modo que reconocen el idioma. Se hubiese podido trabajar agregando otros campos de la base de datos, como, por ejemplo, los keyword, pero no todos los artículos tenían keyword presentes, y si estaban muchos de ellos contenían abreviaciones o estaban en otros idiomas. Esta memoria de titulo por razones de tiempo, no contempla en sus objetivos el pre procesamiento para expandir abreviaciones o traducir dichos campos.

La metodología propuesta se basa en la extracción de tópicos a través del modelo LDA, que tiene como resultado un conjunto de términos o identificadores claves que son capaces de extraer información biológica relevante dentro de las clases modeladas. El modelo LDA no modifica la definición de las palabras claves para representar al documento por el tipo de diccionario utilizado. Las palabras claves varían según el grado de probabilidad de pertenecer a cada tópico, encontrando en cada tópico, términos que son característicos de cada uno de ellos, lo cual permite separabilidad de los mismos.

Al utilizar el modelo LDA se puede observar que la determinación de palabras claves dentro de un texto es fundamental para el buen desempeño del modelo LDA, donde el pre procesamiento del texto juega un rol fundamental en la segmentación, normalización y eliminación del ruido.

Sobre la modelación de las clases descritas en la sección 4.1.2, se tiene que para la primera clase seleccionada “lung”, existe separabilidad entre los tópicos modelados tomando como tema principal el pulmón. Sin embargo, existen diferencias en la modelación utilizando los diccionarios

descritos en la sección 4.1.3. Con el diccionario NER se puede encontrar información más completa del tema modelado, porque cada palabra clave reconocida considera sustantivos y adjetivos como entidad reconocida. Por otro lado, el diccionario Topia contiene menos palabras claves que el diccionario NER, considerando como palabras claves términos que no son relevantes dentro del tema, como por ejemplo; "cd".

La evaluación del modelo LDA, para conocer la capacidad del modelo de dar resultados correctos (TP) e incorrectos (PF) de los documentos de la clase. Debido a lo anterior, la medida de evaluación utilizada fue la precisión. Para el caso de la clase "lung" se obtuvo una precisión de 72 % usando el diccionario Topia y un 55% usando el diccionario NER. Esto indica que los documentos acertados (TP) por cada tópico son siempre mayores a los erróneos (FP).

En la segunda modelación de extracción de tópicos que involucra la clase "kidney-heart" se obtienen tópicos que hacen mención a temas relacionados al riñón y corazón. Por tanto, haciendo uso del diccionario Topia se obtienen ocho de diez tópicos modelados, entregan información de dicha clase. Sin embargo, al igual que la modelación de la clase "lung" este diccionario contiene términos que no son considerados como palabras claves, por ejemplo, "p", "ci", "il". En el caso de los resultados de la modelación de la clase "kidney-heart" con el diccionario NER los tópicos que hacen mención a los temas modelados son seis, y estos son fáciles de identificar porque al igual que sucede en la modelación con topia cada tópico contiene términos como kidney, heart, aki, ventilation, etc. La evaluación de esta clase entrega como precisión un 75.8 % usando el diccionario Topia y un 57% usando el diccionario NER. Estos valores corresponden al porcentaje de casos correctamente clasificados.

Finalmente, en la última modelación de extracción de tópicos de la clase "lung-kidney-heart", los resultados obtenidos son diversos, ya que más de la mitad de los documentos de la clase fueron escogidos al azar. La modelación de la clase con el diccionario Topia entrega 6 tópicos que hacen mención a riñón, corazón y pulmón, los tópicos restantes entregan términos como gen, genotipo, ADN. Luego, en la modelación con el diccionario NER se obtienen tópicos que entregan información de la clase, pero al igual que el que con el diccionario anterior se descubren nuevos términos que hacen mención al sistema nervioso. La evaluación del modelo se hace a través de la precisión, la cual para el diccionario de Topia es de un 40% y un 33% usando el diccionario NER. Estos porcentajes de precisión se calculan en base a solo los temas conocidos de la clase como son el pulmón, riñón y corazón.

A pesar de que el diccionario topia contiene menos palabras claves y algunos son muy cortos como por ejemplo “cd”, resulta ser mejor al utilizar el modelo LDA para modelar tópicos, ya que se obtienen menor cantidad de FP.

El uso de entidades como diccionario de entrada parece tener más palabras relacionadas a los tópicos, pero esto no es suficiente para disminuir el número de falsos positivos obtenidos.

Finalmente, el método LDA parece ser útil para modelar y descubrir tópicos en un corpus, pero es muy sensible al diccionario de entrada utilizado. Los resultados podrían mejorarse al trabajar un poco más en la estrategia de seleccionar el diccionario de entrada de tal forma que sus palabras estén estrechamente relacionadas con las clases y su extensión sea acotada.

5.2. Trabajo futuro.

Como trabajo futuro se propone trabajar con técnicas de pre procesamiento de selección de palabras claves para mejorar el diccionario de entrada del modelo LDA. Del mismo modo hacer un filtrado de las clases “lung”, “kidney-heart” y “lung-kidney-heart” hacerlo considerando todas las variaciones de las palabras incluyendo sinónimos de las palabras para tener mayor cobertura en los documentos obtenidos.

También se propone el agregar técnicas de expansión de abreviaciones y traducción para procesar y agregar keywords al trabajo. Estas mismas pueden ser utilizadas también como goldstandard, sí que la cantidad de documentos del corpus que las contienen es adecuada para la aplicación del LDA.

Finalmente, se propone el etiquetar una cantidad de documentos para generar un goldstandard que permita evaluar adecuadamente el método obteniendo no solo precisión, sino que también recall (o cobertura).

Bibliografía

[1] Ananiadou S., Kell D., Tsujii J. (2006). Text mining and its potential applications in systems biology. TRENDS in Biotechnology. 24 (12): 1-9

[2] Uramoto N., Matsuzawa H., Nagano T., Murakami A., Takeuchi H., Takeda K. (2004). A text-mining system for knowledge discovery from biomedical documents. IBM systems journal, 43(3): 1-18

[3] Erhardt R., Schneider R., & Blaschke C., (2006). Status of text-mining techniques applied to biomedical text. Drug Discovery Today. 11 (7/8): 1-11

[4] Moreno M., & López V., Uso de Técnicas no Supervisadas en la Construcción de Modelos de Clasificación en Ingeniería del Software: 143-153. Tendencias de la Minería de Datos en España. 84-688-8442-1

[5] Blei David M. (2012). Probabilistic Topic Models. Communications of the ACM. 55 (4):1-8

[6] Gordon M., Dumais S., (1998). Using Latent Semantic Indexing for Literature Based Discovery. Journal of the American Society for Information Science. 49(8):674–685.

[7] Dumais S., Furnas G., & Landauer T., (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. 41(6):391-407.

[8] Sista S., Schwartz R., Leek T., & Makhoul J. (2002). An Algorithm for Unsupervised Topic Discovery from Broadcast News Stories. Morgan Kaufmann Publishers Inc. San Francisco. pp 110-114

[9] Rodríguez F. & Bautista S., (2006). Modelos ocultos de Markov para el análisis de patrones espaciales. Ecosistemas 15 (3): 68-75

[10] Blei M., Ng A., Jordan M., (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (2003) 993-1022

[11] Ruiz S., Campos Y., (2010). Clasificación de malformaciones craneales causadas por craneosinostosis primaria utilizando kernels no lineales. Revista Mexicana de Ingeniería Biomédica. 31(1): 15 – 29.

[12] Bisgin H., Liu Z., Fang H., Xu X., & Tong W. (2011). Mining FDA drug labels using an unsupervised learning technique – topic modeling. BMC Bioinformatics, 12(Suppl 10):S11

[13] Hu W., (2012). Unsupervised Learning of Two Bible Books: Proverbs and Psalms. Sociology Mind. 2 (3): 325-334

[14] Rodríguez S., (2012), Estudio de técnicas no supervisadas para descubrir tópicos en videos deportivos. Máster en sistemas inteligentes SIE043. Universidad Jaume I

[15] Dueñas R., & Velásquez J,. (2013). Una aplicación de Web Opinion Mining para la extracción de tendencias y tópicos de relevancia a partir de las opiniones consignadas en blogs y sitios de noticias. Revista Ingeniería de Sistemas, 27.

[16] Seiter J., Amft O., Rossi M., & Tröster G., (2014). Discovery of activity composites using topic models: An analysis of unsupervised methods. 15: 215–227

[17] Ghahramani Z., (2004). Unsupervised Learning. Gatsby Computational Neuroscience Unit University College London, UK

[18] Rodríguez F., (2013), Nuevas fuentes de información para entrenamiento de etiquetadores gramaticales. Tesis presentada para optar al título de Licenciado en Ciencias de la Computación. Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales Departamento de Computación. Buenos Aires.

[19] Lara Neves M., (2013), Minería de Texto aplicada a Bioinformática Funcional. Memoria para Optar al Grado de Doctor. Universidad Complutense De Madrid Facultad De Informática. Departamento de Arquitectura de Computadores y Automática. Madrid.

[20] Kim J., Zhang Z., Park J., See-Kiong Ng., (2006), BioContrasts: extracting and exploiting protein–protein contrastive relations from biomedical literature. Bioinformatics 22 (5): 597–605

[21] Porter M F., (1980), An algorithm for suffix stripping. pp 130-137.

[22] Lista de paquetes, no nativos, de Python instalados:
Accedido en octubre, 2014.TermExtract: https://pypi.python.org/pypi/topia.termextract/1.1.0
Accedido en octubre, 2014. NLTK: http://www.nltk.org/
Accedido en octubre, 2014. Gensim: http://radimrehurek.com/gensim/

[23] Troyano J., Díaz V., Enríquez F., Barroso J., & Carrillo V., (2003), Identificación de Entidades con Nombre basada en Modelos de Markov y Árboles de Decisión. Procesamiento del lenguaje natural, ISSN 1135-5948 (31):235-242

[24] Gómez R., (2003), La evaluación en recuperación de la información [en línea]. "Hipertext.net", núm. 1. Consultado en Diciembre, 2014. <http://www.hipertext.net>

[25] Avila J. L. (2013). Modelos de Aprendizaje Basados en Programación Genética para Clasificación Multi-Etiqueta. Memoria optar al grado de Doctor Ingeniero en Informática. Universidad de Córdoba. Departamento de Informática y Análisis Numérico. Córdoba.

ANEXO. Conjunto de Etiquetas Penn Treebank [extraído de [18]]

Etiqueta	Descripción	Ejemplo
NNPS	Proper noun, plural	*Vikings*
PDT	Predeterminer both	*the boys*
POS	Possessive ending	*friend's*
PRP	Personal pronoun	*I, he, it*
PRP$	Possessive pronoun	*my, his*
RB	Adverb	*however, usually, naturally, here, good*
RBR	Adverb, comparative	*better*
RBS	Adverb, superlative	*best*
RP	Particle	*give up*
SYM	Symbol	*+, %,&*
TO	To	*to go, to him*
UH	Interjection	*uhhuhhuhh*
VB	Verb, base form	*take*
VBD	Verb, past tense	*took*
VBG	Verb, gerund/present participle	*taking*
VBN	Verb, past participle	*taken*
VBP	Verb, sing. present, non-3d	*take*
VBZ	Verb, 3rd person sing. present	*takes*
WDT	Wh-determiner	*which*
WP	Wh-pronoun	*who, what*
WP$	Possessive wh-pronoun	*whose*
WRB	Wh-abverb	*where, when*
$	Dollar sign	*$*
#	Pound sign	*#*
"	Left quote	*(' or ")*
"	Right quote	*(' or ")*
(	Left parenthesis	*([, (, {, i)*
)	Right parenthesis	*( ],), }, ¿)*
,	Comma	*,*
.	Sentence-final punc	*(. ! ?)*

Etiqueta	**Descripción**	**Ejemplo**
:	Mid-sentence punc	*(: ; ... -)*
CC	Coordinating conjunction	*and*
CD	Cardinal number	*1, third*
DT	Determiner	*the*
EX	Existential	*there is*
FW	Foreign word	*d'hoevre*
IN	Preposition/subordinating conjunction	*in, of, like*
JJ	Adjective	*green*
JJR	Adjective, comparative	*greener*
JJS	Adjective, superlative	*greenest*
LS	List marker	*1)*
MD	Modal	*could, will*
NN	Noun, singular or mass	*table*
NNS	Noun plural	*tables*
NNP	Proper noun, singular	*John*