PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA – FACULTAD DE LETRAS - BIBLIOTECAS UC

Metodología para la construcción automática de un corpus de dominio específico

Fabiola Berta Araya Araya

Proyecto de tesis para optar al Grado de Magister en Procesamiento y Gestión de la Información

Profesor Supervisor:

Dr. César Aguilar

Santiago de Chile, Enero 2018

PONTIFICIA UNIVERSIDAD CATOLICA DE CHILE

ESCUELA DE INGENIERIA – FACULTAD DE LETRAS - BIBLIOTECAS UC

Metodología para la construcción automática de un corpus de dominio específico

Fabiola Berta Araya Araya

Proyecto de tesis presentado a la Comisión integrada por:

Profesor Supervisor:

Dr. César Aguilar

Profesor Integrante:

Dra. Olga Acosta

Director del Programa:

Sr. Mauricio Arriagada Benítez

Para completar las exigencias del Grado de Magister en Procesamiento y Gestión de la Información

Santiago de Chile, Enero, 2018

DEDICATORIA

A mi mamá por su incondicional apoyo.

AGRADECIMIENTOS

Deseo expresar mi agradecimiento a Fondecyt por la beca otorgada ya que fue una valiosa ayuda, tanto en el aporte económico como también en el valor otorgado a mi tesis.

Un agradecimiento especial a mis Profesores Guías Dra. Olga Acosta López y Dr. César Aguilar por su apoyo incondicional en el procesos de inicio, desarrollo y finalización de esta tesis, por la amistad que se ha gestado, paciencia, consejos, y por la orientación dada en seguir avanzando y profundizando en estos nuevos conocimientos adquiridos.

Quiero también dar las gracias al exdirector del Programa Don Jorge Gana Leay por su dedicación, carisma y motivación en la entrega de contenidos clase a clase.

Un agradecimiento al actual Director Sr. Mauricio Arriagada Benítez por consolidar y otorgar la relevancia que tiene este Magister en una sociedad sobrepasada de información.

Por último, y no menor quiero agradecer a la asistente del MPGI la Sra. Carla Manríquez por su excelente gestión en todo el quehacer administrativo del Magister.

Quiero agradecer también, a mis partners Carri, Ignacia y Rodrigo por lo que compartimos, reímos, por las largas horas de estudio de los sábados, por la incondicional ayuda entre nosotros y la bella amistad que aún tenemos.

INDICE DE FIGURAS, TABLAS E IMÁGENES

Figura 1. Estructura del Corpus Brown

Figura 2. Representación de documentos mediante una matriz palabras-documentos

Figura 3. Perspectiva de modelo generativo y estadística inferencial en el modelado de tópicos.

Figura 4. Distribución de palabras de cuatro tópicos extraídos del corpus TASA

Figura 5. Tres tópicos relacionados con la palabra PLAY

Figura 6. Metodología para generar de forma automática un corpus de dominio específico.

TABLAS

Tabla 1. Etiquetado PoS

Tabla 2. Texto plano

Tabla 3. Texto vertical

Tabla 4. Matriz de confusión corpus de prueba

IMÁGENES

Imagen 1. Definición de parámetros

Word Cloud

Resultados con el conjunto de prueba distribución de palabras

Imagen 2. Definición de parámetros

Word Cloud

Resultados con el conjunto de prueba distribución de palabras

RESUMEN

Los corpus son un recurso hoy en día inigualable para cualquier estudio lingüístico en general y en lingüística computacional. Ahora bien, con la incorporación de los computadores cada vez con mayor capacidad de almacenamiento y procesamiento, el acceso a los datos es rápido y fiable, así como su manipulación, extracción e identificación de información relevante de enormes cantidades de textos.

La contribución concreta de este trabajo fue generar una metodología para la construcción automática de un corpus de dominio específico mediante el entrenamiento de un modelo de tópicos que fuera capaz de discriminar los documentos del dominio previamente definido de otro que no lo fuera. Para ello fue relevante la utilización de herramientas como MALLET, el módulo LDA Python y librerías como NLTK, GENSIM, entre otras mencionadas en el cuerpo del trabajo. La implementación de un programa computacional con una interfaz de fácil manipulación permitió comprobar el funcionamiento del corpus de entrenamiento y de prueba. Además, con estos resultados se evaluó el clasificador a través de una matriz de confusión.

ABSTRACT

The corpus is an unparalleled resource today for any linguistic study in general and in computational linguistics. However, with the incorporation of computers with increasing storage and processing capacity, access to data is fast and reliable, as well as its manipulation, extraction and identification of relevant information from enormous amounts of texts.

The concrete contribution of this work was to generate a methodology for the automatic construction of a specific domain corpus by training a topic model that was able to discriminate the documents of the previously defined domain from another that was not. For this, the use of tools such as MALLET, the LDA Python module and libraries such as NLTK, GENSIM, among others mentioned in the body of work was relevant.

The implementation of a computer program with an easy-to-use interface allowed checking the operation of the training and test corpus. In addition, with these results, the classifier was evaluated through a confusion matrix.

INDICE GENERAL

DEDICATORIA

AGRADECIMIENTOS

INDICE DE FIGURAS, TABLAS E IMÁGENES

RESUMEN

ABSTRACT

1 INTRODUCCION

2 ESTADO DEL ARTE

2.1. ¿QUÉ ES UN CORPUS?

2.1.1 Tipos de corpus

2.1.2 Características de los corpus

2.1.3 Diseño de un corpus

2.2 GENERACION DE CORPUS

2.2.1 Primera Generación de corpus

2.2.1.1 Corpus Brown

2.2.1.2 Corpus Lancaster-Oslo Bergen (LOB)

2.2.1.3 Otras propuestas

2.2.1.4 Corpus London-Lund (LLC)

2.2.2 Segunda generación de Mega-Corpus

2.2.2.1 Proyecto COBUILD

2.2.2.2 Longman Corpus Network

2.2.2.3 British National Corpus (BNC)

2.2.2.4 Corpus Internacional de Ingles

2.2.3 Corpus Especializados

2.2.3.1 Corpus Históricos

2.2.3.2 Corpus para Propósitos Específicos

2.2.3.3 Corpus Internacional/Multilingües

2.3 ¿QUE ES UN COPUS GOLDEN ESTANDAR?

3 MODELADO DE TOPICOS

3.1 Modelos Generativos

3.2 Modelado de tópicos probabilísticos

3.3 Asignación de Dirichlet Latente

4 METODOLOGIA

4.1 Pre-procesamiento de los textos

4.2 Etiquetado de partes de la oración (etiquetado PoS)

4.3 Eliminación de Ruidos en los Textos

4.3.1 Lista de palabras funcionales (stopwords)

4.3.2 Filtrado de nombres y adjetivos no relevantes

4.3.3 Filtrado de palabras comunes y no polisémicas

4.4 Herramientas Computacionales

4.4.1 SKETCH ENGINE

4.4.1.1 Corpus en texto plano vs corpus Vertical

4.4.2 PYTHON

4.4.3 NLTK

4.4.4 GEMSIM

4.4.5 MALLET

4.4.6 Modulo LDA Python

5 RESULTADOS

5.1 Colección de documentos

5.2 Interfaz principal del programa

5.2.1 Escenario 1

5.2.1.1 Interpretación de los resultados escenario 1

5.2.2 Escenario 2

5.2.2.1 Interpretación de los resultados escenario 2

6 CONCLUSIONES Y TRABAJO FUTURO

7 BIBLIOGRAFIA

1.- INTRODUCCION

Actualmente, existe una gran cantidad de recursos computacionales que facilitan la identificación y extracción de información relevante de enormes cantidades de textos. Tiempo atrás pensar en la convivencia entre las herramientas computacionales y la lingüística era casi imposible. Con todo, en la actualidad dada la simbiosis de estas dos disciplinas existen grandes ventajas.

Ahora bien, para poder aplicar herramientas computacionales es indispensable disponer de un material digital debidamente organizado, a este recurso se le denomina c*orpus.* Para la lingüística contemporánea, el diseño y uso de corpus lingüísticos se ha convertido en un tema fundamental, partiendo por indicar que el *corpus lingüístico* es un conjunto amplio y estructurado de ejemplos reales de uso de la lengua, que estos pueden ser textos, o muestras orales, que por lo general son transcritas. Dichos corpus se caracterizan por ser un conjunto de textos relativamente grandes y de acuerdo a su estructura, variedad y complejidad deben reflejar una lengua, y en cuanto a su uso deben ser representativos de la realidad.

La disponibilidad de grandes corpus en diferentes lenguas, así como la capacidad de recolección y validación de nuevos recursos de este tipo, ha beneficiado a muchos investigadores de diferentes áreas, ya que tienen la oportunidad de realizar análisis con datos reales y exhaustivos de sus áreas de interés, de esta forma logran la máxima cercanía con la realidad. Actualmente, las abrumadoras cantidades de información textual demandan la aplicación de herramientas computacionales que faciliten las actividades de recopilación y organización de los textos en formatos electrónicos, así como su procesamiento con fines específicos.

Uno de los problemas al que nos enfrentamos en el campo de la extracción de información es cuando deseamos comparar diferentes enfoques, por ejemplo, para la extracción terminológica, y no existe disponibilidad de corpus relacionados con áreas de conocimiento específico. La escasez de recursos enfocados a este tipo de tareas es evidente para el español, no para lenguas como el inglés, donde existe una gran cantidad de recursos que se ponen a disposición de los investigadores interesados en

desarrollar algoritmos para mejorar el desempeño de los sistemas actuales. La recopilación de nuevos documentos es una tarea que se puede agilizar con las herramientas actuales (*crawlers)*, plataformas de recolección y análisis de corpus, etc. Sin embargo, la validación de estos documentos para conformar un corpus de dominio específico es una tarea que normalmente se deja a expertos humanos. Lograr la participación de tales expertos para la recolección y validación de documentos es difícil, esto todavía se complica más si les solicitamos hacer estas dos tareas para enormes cantidades de documentos.

Dado lo anterior, el objetivo de este trabajo es proponer una metodología para la construcción automática de *corpus* de dominio específico mediante el entrenamiento de un modelo de tópicos que discierna entre documentos relevantes y no relevantes a un dominio previamente definido. Por ejemplo, si el objetivo es construir un corpus sobre temas de oftalmología, se deben seleccionar documentos representativos de dicha área, no necesariamente grandes cantidades de estos, además de otras áreas de la medicina, como la cardiología, ginecología, etc., que permitan generar un modelo que caracterice el vocabulario relevante para cada una de las áreas. Indudablemente, dicho modelo también será de utilidad como clasificador de documentos de diversas áreas, en este caso concreto, del área de la medicina.

Esta investigación se enmarca en el proyecto extracción de contextos definitorios en el área de la biomedicina, cuyo objetivo es el análisis, reconocimiento y extracción de fragmentos textuales específicos de textos en el área de biomedicina.

2.- ESTADO DEL ARTE

El siguiente apartado presenta una recopilación de lo que se ha realizado hasta ahora sobre el tema de la construcción de corpus.

2.1. ¿Qué es un CORPUS?

Definir el concepto de corpus, tal y como se usa hoy en día en el ámbito de la lingüística, no es una tarea sencilla. Básicamente, se puede señalar una definición de corpus como: "cualquier colección que contenga más de un texto". Otras definiciones de corpus propuestas en los últimos años, según, Richard Xiao (Xiao, 2008) : "Un corpus puede ser definido como una colección de textos auténticos leído automáticamente por máquinas (incluyendo traducción de textos orales), esto es, una muestra representativa de una lengua natural especifica o de una variedad de lenguas."

En términos simples, se puede considerar que un corpus es una colección de texto en formato digital, aunque autores como Geoffrey Leech, (Pérez Hernández, 2002) recalca en la habilidad que poseen los computadores para buscar, recuperar, ordenar y hacer cálculos sobre cantidades masivas de texto ha dado la oportunidad de comprender y de explicar el contenido de esos corpora de formas que no eran imaginables en el periodo que él denomina "*pre-computacional*".

Torruela y Llisterri, (Torruela, 1999) indica que la "característica más importante de los corpus es que están compuestos por datos reales y, por lo tanto sus resultados son empíricos y la función principal de un corpus es establecer la relación entre la teoría y los datos, el corpus debe mostrar a pequeña escala como funciona una lengua natural." Quizá la definición más estandarizada es la que ofrece el grupo de trabajo dedicado a los corpora textuales de EAGLES (Expert Advisory Group on Language Engineering Standards):

*"*Corpus es una colección de partes del lenguaje que son seleccionadas y ordenadas de acuerdo a un criterio lingüístico explícito con el fin de ser usado como ejemplo de un lenguaje.

Según Atkins, (Atkins, 1991) un corpus es un conjunto de textos reales, legibles por un computador y que son un modelo de la realidad y por tanto deberían mostrar los rasgos más característicos y destacados de esta.

Para la construcción de un corpus se deben considerar varias etapas:

- Planificación, *fase* que involucra dos áreas, una el *diseño lingüístico* y la otra es el *proyecto de costos y administración*. El diseño lingüístico está orientado a definir el tipo de corpus que se construirá, la medida del corpus, se refiere a que tan amplio será, esto dependerá de su propósito, cobertura idiomática, periodo temporal, durante que años, épocas, etc. La planificación de costos y la administración involucra las especificaciones y el diseño, selección de las fuentes, obtención de los permisos, captura de datos, etiquetado y procesamiento del corpus.
- A*utorizació*n, en ocasiones este punto puede ser un obstáculo porque se requiere de los permisos de copyright para los textos que serán digitalizados, apoyarse en la legislación vigente aún si los datos están siendo guardados en un computador o si esos son usados sólo por investigadores universitarios.
- R*epresentatividad,* referida a que un corpus está diseñado para representar un lenguaje en particular o una variedad de lenguajes.
- *Captura de datos*, fase que consume tiempo y costos al momento de determinar la cantidad de texto con los cuales se trabajará los documentos impresos debe ser convertido a un formato electrónico.
- *Manejo de los textos,* la existencia de un corpus amplio no asegura la demanda de datos lingüísticos, por ello es necesario trabajar con herramientas computacionales para el procesamiento del corpus que serán esenciales.
- *Etiquetado y anotaciones del corpus*, el etiquetado corresponde a un sistema de estándar de códigos insertos en los documentos que están almacenados electrónicamente y que proveen de información acerca del texto en sí mismo. Cada información lingüística debe ser primero codificada dentro del corpus, este es el proceso de anotación de un corpus.

- *Retroalimentación del usuario y desarrollo del corpus,* la construcción del corpus debe ser representativa, se evaluarán sus debilidades y fortalezas, la retroalimentación será para mejorar el corpus. Es importante que el usuario del corpus haga mención de sus resultados, comentarios y advertencias al equipo de trabajo que está desarrollando el corpus.

2.1.1 Tipos de corpus

Atkins (Atkins, 1991), junto con Torruella y Llisterri (Torruela, 1999), han propuesto varios principios de clasificación de los distintos corpus en función de una serie de criterios, aunque no está del todo clara ya que en la práctica se difiere bastante de la clasificación.

La clasificación de los corpus se centra en:

- La forma de presentación de la lengua
- El número de lenguas a que pertenecen los textos
- El tamaño, cantidad y distribución de textos que componen el corpus
- Los límites del corpus
- El grado de especificidad de los textos y variedad lingüística
- El periodo temporal que abarca los textos
- El tratamiento aplicado al corpus

Estos criterios están determinados por el o los objetivos del corpus. Según Villayandre (Villayandre Llamazares, 2008) existe una tipología de corpus, que son:

1.- Según la modalidad de la lengua, se pueden distinguir tres tipos:

- *Corpus orales*: Son aquellos que recogen muestras de la lengua hablada, que pueden ser, transcripciones ortográficas de grabaciones y grabaciones.
- *Corpus escritos*: Son aquellos que están conformados por textos que muestran la lengua escrita.

- *Corpus mixtos*: Son aquellos que combinan ambas modalidades, aunque hay una clara tendencia a favorecer la lengua escrita.

2.- De acuerdo al número de lenguas, los corpus se clasifican en monolingües, bilingües, y multilingües:

- Los *corpus monolingües* son aquellos conformados por textos de una lengua y que es capaz de dar cuenta de dicha lengua o de su variedad lingüística.
- Los *corpus bilingües son aquellos conformados por textos de dos lenguas.*
- *Los corpus multilingües* son aquellos conformados por textos de más de dos lenguas.

3.- Cantidad, tamaño y distribución de los tipos de textos

- *Corpus grandes*: estos corpus son de grandes proporciones y por lo general no responden a equilibrios ni representatividad. Actualmente, la tendencia es a corpus grandes ya que con la ayuda de recursos computacionales son manejables, y estos si garantizan la representatividad de los datos.
- *Corpus representativos*: recogen la misma proporción de diferentes géneros de textos (por ejemplo, el corpus Brown).
- *Corpus piramidales*: son aquellos que están conformados por textos distribuidos en diferentes niveles, por ejemplo un nivel consta de pocas variedades temáticas pero con mucho textos, un segundo nivel, textos más variados temáticamente, pero menos cantidad de cada uno.
- *Corpus Léxicos*: son aquellos que recogen fragmentos de textos muy pequeños. Era lo habitual en los primeros corpus, debido a las limitaciones de tamaño que los medios técnicos de la época imponían. Hoy en día han vuelto a cobrar importancia debido a lo cuidado de su diseño.

4.- Limites del corpus

- *Corpus abiertos*: son corpus dinámicos que se mantienen en constante crecimiento, esto es a través de una constante introducción de nuevas cantidades de textos.
- *Corpus cerrados*: son aquellos que cuentan con un finito número de textos, esto es establecido previo a la recopilación del corpus, una vez alcanzado el número el corpus se da por finalizado, no se incorporan nuevos textos posteriormente.

5.- Especificidad de los textos, los corpus pueden ser generales o especializados.

- *Corpus generales*: corresponde a los que reflejan la lengua o a la variedad lingüística de la forma más equilibrada posible, vale decir, cuanto más tipos de textos, modalidades, género, materia, es mucho mejor. Además, deben ser lo más grandes posible, ya que de esta forma reflejan las variedades relevantes de la lengua y su vocabulario.
- *Corpus Especializados*: son aquellos que recopilan textos específicos que puedan aportar en la descripción de un tipo de lengua o dominio de conocimiento en particular.

6.- Periodo temporal que abarcan los textos,

- *Corpus periódicos o cronológicos*: son aquellos conformados por textos de un periodo específico o de una época concreta con el objeto de estudiar la lengua producida durante ese periodo en particular.
- *Corpus diacrónicos o históricos*: son aquellos corpus que incluyen textos de épocas temporales sucesivas, cuya finalidad es estudiar la evolución de la lengua en un periodo largo, pero determinado.
- *Corpus sincrónicos*: su finalidad permite el estudio de una o más variedades lingüísticas en el momento presente, sin prestar atención a su evolución excepto en lo que se refiere a los cambios rápidos que ocurren actualmente. Es el caso del *Corpus of Contempary American English*, de más de trescientos ochenta y

cinco millones de palabras procedentes de textos de diferentes fuentes de los años 1990 a 2008.

7.- Tratamiento aplicado al corpus

- *Corpus simples*, estos son los que no presentan anotaciones ni codificaciones, un corpus así no ofrece ninguna posibilidad de estudio lingüístico.
- *Corpus verticales*: son el resultado de disponer en forma de columnas las palabras de un texto ordenadas alfabéticamente o de frecuencias.
- *Corpus anotados o codificados*: son corpus conformados por textos a los que se les han añadido de forma manual o automática, determinada información, como etiquetas, anotaciones o codificaciones.

2.1.2 Características de los corpus

- Deben estar formados por un conjunto de datos lingüísticos naturales.
- El contenido del corpus debe ser cuidadosamente escogido según criterios establecidos y objetivos del estudio.
- El corpus debe ser representativo de una lengua o variedad de esta.
- El corpus debe tener como finalidad ser objeto de estudio lingüístico.

Aunado a las características mencionadas anteriormente los corpus debe presentar ciertos requisitos para que cumplan su función como tal. Autores como Villayandre (Villayandre Llamazares, 2008) , así como Torruela y Llisterri (Torruela, 1999) indican los siguientes:

- **Formato electrónico:** un corpus, para ser una herramienta útil al lingüista, debe estar digitalizado, es decir, los textos de que consta tienen que estar en formato electrónico (corpus automatizado). El hecho que para los primeros corpus no se pudiera disponer de un computador motivo la crítica de las seudotécnicas: el procesamiento de los datos debía efectuarse de forma manual, con los errores y problemas que eso ocasionaba. Sin embargo, el empleo del computador permite automatizar tareas tales como:

  1.-Búsqueda de información: un corpus informatizado permite localizar de forma rápida secuencia de palabras o los siguientes en decimas de segundos.

  2.- Recuperación de información: un corpus informatizado permite obtener todos los casos de una palabra, secuencia de palabras, etc. Registrados en un corpus, normalmente con su contexto inmediato anterior y posterior, lo que se conoce como concordancia.

  3.- Computo de la frecuencia de aparición de una palabra, secuencia de palabras, etc.

  4.- Clasificación de los datos contenidos en el corpus según diferentes criterios: orden alfabético, frecuencia de aparición, autor, procedencia geográficas, tema, medio de publicación, etc.

- **Autenticidad de los datos:** los textos deben ser una muestra real de la lengua que será objeto de estudio.

- **Criterios de selección:** los textos que son parte del corpus deben ser elegidos de acuerdo a un criterio lingüístico, no perdiendo de vista el objetivo de este.

- **Representatividad:** este es un requisito importante ya que el corpus debe responder a la lengua bajo estudio. La selección de los textos, basada en criterios adecuados, debe responder a parámetros estadísticos que garanticen que los textos “representan” la variedad de lengua en estudio, vale decir, una muestra representativa.

- **Tamaño:** por lo general, los corpus constan de un tamaño finito, que se suele medir en millones de palabras o de formas y que se fija antes de empezar la recolección de los textos (por ejemplo un millón de palabras). Una vez alcanzado ese número, se da por terminada la recopilación del corpus, que no es más que el primer paso de todo el proceso. Sin embargo, también existen corpus abierto, como el del Proyecto COBUILD. https://collins.co.uk/page/The+Collins+Corpus, dirigido por J. Sinclair en la Universidad de Birmingham, de especial interés para la lexicografía, o el CREA de la Real Academia Española, http://corpus.rae.es/creanet.html. En el pasado se pensaba que el tamaño era muy importante: mientras mayor fuera el corpus, más posibilidades tenia de reflejar el funcionamiento real de la lengua en todas sus variedades, pero en la actualidad se priman los criterios de diseño, es decir, el tamaño solo es importante en la medida en que así lo exija la finalidad del corpus.

2.1.3 Diseño de un corpus

Una vez delimitados los distintos tipos de corpus y su aplicación, se debe considerar el diseño del corpus teniendo presente o los siguientes aspectos estipulados por Atkins (Atkins, 1991) y Torruela y Llisterri (Torruela, 1999):

- Finalidad: es un aspecto que se debe definir cuándo se empieza el diseño de este, ¿Cuál es la finalidad concreta del corpus?
- Límites: una vez especificada la finalidad del corpus, se debe establecer claramente los límites temporales, geográficos y/o lingüísticos que el corpus considerará.
- Tipo de corpus: una vez que se han definido los dos aspectos anteriores, se debe considerar los siguientes parámetros:

  1) El porcentaje y la distribución de los diferentes tipos de textos que lo componen.

  2) La especificidad de los textos.

3) La cantidad de textos que se tome de cada documento para formar las muestras.

4) Codificación y anotación.

5) La documentación que lo acompañe.

- Proporciones de los diferentes grupos temáticos del corpus.
- Población y muestra: como la finalidad del corpus es describir una lengua es muy importante considerar que la muestra deberá ser representativa de la población.
- Número y longitud de los textos de la muestra: La selección de la muestras de los textos se puede realizar de tres formas:

  1) Al azar

  2) Dividendo los textos en tres partes de extensión parecida y extrayendo de cada una de ellas las muestras en número y proporciones aproximadamente parecidas.

- Captura de los textos y etiquetado
- Procesamiento del corpus: el corpus por sí solo no es suficiente para facilitar datos sobre el comportamiento del lenguaje. Para aprovechar toda la información que contienen es necesario trabajar con herramientas computacionales. Actualmente se trabaja con programas de alta complejidad destinados a la lingüística de corpus.
- Crecimiento del corpus y feedback: para tener un corpus equilibrado, primero se debe procurar un corpus representativo, después de utilizarlo, hay que analizar los resultados y detectar los puntos débiles respecto a su representatividad. A la vista de estos análisis se debe ir reajustando las proporciones del corpus, lo ideal es trabajar con expertos en estadísticas que aporten métodos para el equilibrio del corpus y colaborar con los usuarios finales, que son los que detectarán este tipo de problemas.

- Hardware y software: este punto es muy importante al momento del diseño del corpus, la infraestructura informática que se va a necesitar para poder desarrollar y explotar el corpus.
- Aspectos legales: este punto es aún difícil de resolver, el derecho de autor. No está aún definida la normativa a que está sujeta las reproducciones ni los documentos de internet.
- Presupuesto y etapas: una vez definidos los aspectos anteriormente mencionados solo se debe establecer las diferentes etapas en que se va a realizar el proyecto y como se va a llevar a cabo su mantenimiento.

2.2 Generación de corpus

2.2.1 Primera Generación de corpus

2.2.1.1 Corpus Brown

Kennedy (Kennedy, 1998), indica que el corpus más importante pre-electrónico generado específicamente para descripciones gramaticales fue el *Survey Corpus***,** creado en el *Survey of English Usage* (SEU). Marcó la transición entre el período temprano de los corpus no-automatizados y el desarrollo de modernos corpus lingüísticos. Cuando Randolph Quirk fundó el SEU en 1959, su principal objetivo fue reunir 200 ejemplos, cada uno con alrededor de 5.000 palabras representativas del habla y de la escritura de la lengua inglesa, para formar un corpus de un millón de palabras las cuales podrían ser usadas como base para la descripción gramatical. Dos años más tarde, en 1961, se gesta la compilación del primer corpus legible por un computador para una investigación lingüística a cargo de Nelson Francis y Henry Kuçera, denominado *Brown Corpus***.**

(http://clu.uni.no/icame/manuals/BROWN/INDEX.HTML), el cual se ha convertido en una referencia para muchos corpus posteriores, ya que fue el primer corpus computacional compilado para una investigación. Cabe señalar también que tras la creación de este corpus se dio una gran por parte de aquellos lingüistas que] estaban afiliados al paradigma generativista dominante en Estados Unidos, el cual era liderado por Noam Chomsky. Este último, junto con otros, consideraba que el uso de los corpus, junto con la aplicación de modelos estadísticos a los estudios lingüísticos no aportaba resultados relevantes.

La estructura del corpus Brown consiste en 500 ejemplos cada uno con alrededor de 2.000 palabras escritas en inglés. El resultado del corpus contiene aproximadamente 1.014.300 palabras.

	Broad text category	Text category letter and description ("genre")		Number of texts			
				Brown	Frown	LOB	FLOB
Informative	Press	A	Press: Reportage	44	same as Brown		
		B	Press: Editorial	27	"	"	"
		C	Press: Reviews	17	"	"	"
	General Prose	D	Religion	17	"	"	"
		E	Skills, Trades and Hobbies	36		38	
		F	Popular Lore	48		44	
		G	Belles Lettres, Biographies, Essays	75		77	
	Learned writing	H	Miscellaneous: Government documents, industrial reports etc.	30	same as Brown		
		J	Science	80	"	"	"
Imaginative	Fiction	K	General Fiction	29	"	"	"
		L	Mystery and Detective Fiction	24	"	"	"
		M	Science fiction	6	"	"	"
		N	Adventure and Western	29	"	"	"
		P	Romance and Love story	29	"	"	"
		R	Humour	9	"	"	"

Figura 1. Estructura del Corpus Brown, de acuerdo con Nelson Francis y Henry Kucera. Fuente: (Kennedy, 1998)

La muestra fue seleccionada con la condición que fuera razonablemente representativa de la lengua inglesa americana. El corpus no fue planificado para ser representativo de supuestos modelos de calidades estilísticas, pero ha servido, hasta hoy como un

estándar de comparación para estudios y análisis de la lengua inglesa. Establecer las categoría y sus subdivisiones, fue una importante unión con el *Survey of Englsih Usage Corpus*, que había ya comenzado en Londres. Las categorías para el corpus Brown y el número de la muestra de cada una de estas fue ampliamente discutida en la Conferencia dictada en la Universidad de Brown en 1963, donde se describieron los métodos de muestras utilizados, así como los detalles bibliográficos, permiso de *copyright*, convención de códigos como abreviaturas, fórmulas, palabras extranjeras, signos de exclamación y puntuación.

La cuidadosa planificación de la estructura del corpus Brown, con la selección de las categorías de textos para representar una amplia gama de aspectos estilísticos de la escritura del inglés americano, así como una colección estándar para una investigación basada en corpus, fueron aspectos relevantes para catalogarlo como un corpus modelo.

Otra característica importante es el acceso libre que tiene este corpus, ya que brinda una base sólida para otras investigaciones. Por más de 30 años el corpus Brown ha estado disponible para todo el mundo, sin que los investigadores tengan la obligación de pagar por derecho de autor o cualquier otro costo que involucre la compilación de corpus.

2.2.1.2 Corpus Lancaster-Oslo Bergen (LOB)

Entre 1970 y 1978 un corpus de escritura en ingles británico fue compilado en la Universidad de Lancaster y la Universidad de Oslo, con la colaboración del Centro Computacional para las Humanidades de Bergen, Noruega, el cual, dan origen al *Corpus Lancaster-Oslo/Bergen* (LOB), intentó ser la contraparte del Corpus Brown. Este corpus contiene 500 textos con cerca de 2.000 palabras cada uno. Las categorías de ambos corpus fueron las mismas, así como también, la estratificación de la muestra. Hubo algunas pequeñas diferencias en el número de textos en la misma categoría, la muestra del corpus LOB es más agrupada dentro de las categorías que el corpus Brown.

El Corpus LOB contiene textos publicados en 1961, el mismo año que fueron incluidos en el corpus Brown, aunque el primero fue compilado una década después**.** Esto fue una ventaja para sus creadores**,** ya que fueron capaces de desarrollar sistemas de cómputo más avanzados para su consulta, así como codificar los ítems con etiquetas al inicio de la oración y en abreviaturas. El corpus LOB llego a estar disponible en cintas magnéticas para almacenaje computacional, y luego en cartuchos, disquete, microfichas, CD-ROM, con versiones en MS-DOS, y plataformas en Unix y Macintosh.

2.2.1.3 Otras propuestas

Se generaron tres corpus basados en el corpus Brown, los cuales son similares en estructura y tamaño: *Kolhapur Corpus of Indian English* conformado por material escrito en inglés publicados en India en 1978, mientras que las muestras vienen de los mismos 15 géneros usados por el corpus Brown y LOB. Igualmente**,** el *Wellington Corpus of Written New Zealand English***,** junto con el *Australian Corpus of English (ACE)* (también conocido como *Macquarie Corpus of Written Australian English)* tomaron como referencia el año 1986 para la selección del material escrito, 25 años después del material con el que estaban conformados el corpus Brown y LOB. Se incorporó material escrito de cuatro años más (1986-1990) ya que comparativamente la cantidad de publicaciones de New Zealand en relación a Estados Unidos y Gran Bretaña estaba por debajo del promedio. A pesar de esto, un estudio ha demostrado que ambos corpus pueden ser usados como base para el estudio de ciertas tendencias léxicas y sintácticas de las diferentes variedades del inglés.

Esta primera generación de corpus nacidos del corpus Brown fueron logros importantes, ya que se generaron una gran variedad de estudios**.** Con todo**,** ellos presentaron algunas desventajas**,** en particular**,** su tamaño -aproximadamente un millón de palabras- que de alguna manera era arbitrario y restrictivo**.** Si bien esta cantidad fue un criterio muy amplio a principios de los 60**,** un corpus debería ser lo suficientemente amplio para proveer de un número sustancial de ejemplos, así como también características lingüísticas particulares.

2.2.1.4 The London-Lund Corpus (LLC)

El *London-Lund Corpus* (LLC), compilado por la University College London en 1959, es parte de la primera generación moderna de corpus pero no hubo intención de que este fuera computarizado. Conformado por la mitad de documentos orales los que fueron transcritos y la otra mitad de documentos escritos, ambos archivos fueron analizados.

En 1975, el *Survey of Spoken English* (SSE) fue instalado en la Universidad Lund de Suecia por Jan Svartvik, inicialmente para hacer disponible en formato electrónico la parte hablada del SEU corpus. Originalmente fueron 87 textos transcritos con un total de 435,000 palabras que fueron luego complementadas por 13 textos más con el fin de completar el Corpus Lund (LLC) de 500.000 palabras. Este corpus electrónico de la lengua inglesa fue ampliamente usado hasta la mitad de los años 90.

2.2.2 Segunda generación de Mega-Corpus

Como se ha mencionado anteriormente, la mayoría de los corpus de la primera generación aptos para ser leídos por maquinas fueron modelados por el corpus Brown y SEU en tamaño y/o en la forma de representar una variedad particular de un lenguaje. Hasta 1980, un millón de palabras –el tamaño del corpus SEU y Brown- fue visto como una medida estándar para los corpus. Sin embargo, en la década de los 60 y 70, estos corpus se volvieron pequeños para el análisis semántico y léxico. Afortunadamente, el desarrollo de la tecnología para la captura de textos y almacenaje llegó en un buen momento e impulsó la creación de corpus más grandes. Así, los corpus de los años 90 incrementaron su tamaño significativamente, algunos llegaron hasta los 100 millones de palabras e incluso más.

2.2.2.1 Proyecto COBUILD

El primer gran corpus, legible por un computador, basado en un proyecto del corpus lexicográfico *American Heritage Project* de los años 70, fue también el gran proyecto de mega-corpus, creado por la Universidad de Birmingham en 1980 y financiado por la editorial Collins.

Los logros más importantes del proyecto COBUILD han sido la creación y el análisis de un corpus electrónico de textos contemporáneos, el *Collins Corpus*. Posteriormente, tal corpus fue la base para la elaboración del diccionario monolingüe para el alumno**,** el cual fue] publicado en 1987.

2.2.2.2 Longman Corpus Network

Longman Corpus Network es una base de datos comercial que está constituida por tres grandes corpus**:** a) el corpus del lenguaje inglés Longman/Lancaster; b) el corpus oral Longman**;** y c) el corpus de principiantes Longman. Cada uno tiene diferentes características y propósitos, aunque fueron compilados juntos como una base confiable para la descripción del inglés, especialmente para la redacción de diccionarios para hablantes no nativos. Todos los diccionarios Longman se compilan utilizando el Longman Corpus Network, una enorme base de datos de 330 millones de palabras de una amplia gama de fuentes reales, como libros, periódicos y revistas. Toda la información de los diccionarios, incluyendo ejemplos, se basa en este corpus.

2.2.2.3 British National Corpus (BNC)

El Corpus Nacional Británico (BNC), es un corpus de textos de 100 millones de palabras de muestras del inglés escrito y hablado**,** proveniente de una amplia gama de fuentes. El corpus cubre el inglés británico de finales del siglo XX**,** incluyendo una gran variedad de géneros, con la intención de que sea una muestra representativa del inglés británico de esa época.

La creación del BNC involucró a tres editores**:** Oxford University Press, como colaborador principal, Longman y W. & R Chambers, así como dos universidades**:** la Universidad de Oxford y la Universidad de Lancaster**,** así como la Biblioteca Británica. Este proyecto se inició en 1991 bajo la dirección del consorcio de BNC, y el proyecto finalizó en 1994.

El BNC se diferenció con respecto a corpus existentes en ese momento debido a que sus datos eran accesibles no solo a la investigación académica**,** sino también a usos comerciales y educativos. El corpus se enfocaba únicamente en el inglés británico y no se extendía para cubrir otras variantes. Esto se debió a que una buena parte significativa del financiamiento era otorgado por el gobierno británico, que estaba interesado en apoyar la documentación de su propia variedad lingüística.

2.2.2.4 Corpus Internacional de Inglés (ICE)

El más ambicioso proyecto de corpus para el estudio comparativo del inglés a nivel mundial fue el *International Corpus in English* (ICE). Básicamente el ICE es un conjunto de corpus que presenta distintas variantes del inglés**:** se incluyen más de 20 países donde este idioma es la primera o segunda lengua oficial.

El proyecto comenzó en 1990 con el objetivo de recopilar material para el estudio comparativo del inglés en todo el mundo. Veintitrés equipos de investigaciones del mundo prepararon corpus electrónicos de su propia variedad nacional o regional. Para asegurar la compatibilidad entre los corpus componentes, cada equipo sigue un diseño de corpus común así como un esquema común para la anotación gramatical.

Cada corpus contiene un millón de palabras en 500 textos de 2000 palabras**,** producidos después de 1989, siguiendo la metodología de muestreo utilizada por el corpus Brown. El ICE contiene un 60% del inglés oral o hablado transcrito. El investigador principal del proyecto, Sydney Greenbaum, insistió en la primacía de la palabra hablada, contando con la colaboración de Randolph Quirk y Jan Svartvik.

2.2.3 Corpus Especializados

Estos corpus recogen textos que pueden aportar datos para la descripción de un tipo particular de lengua, por ejemplo el *Corpus Técnico do Galego* (CTG) del Seminario de Lingüística Informática de la Universidad de Vigo, que contiene textos de tipo jurídicos-administrativos, informáticos y telecomunicaciones, ecología y ciencias ambientales, economía, sociología y medicina. Igualmente, cabe mencionar *Corpus Técnico Especializado,* desarrollado por el Instituto Universitario de Lingüística Aplicada de la Universidad Pompeu Fabra, que consta de textos en catalán, castellano, inglés, francés y alemán. Las temáticas de tales documentos giran en torno a la economía, derecho, medio ambiente, medicina e informática**.** Este corpus fue construido con el fin de estudiar cómo funciona la lengua en cada una de esas áreas**,** así como extraer información útil para detectar neologismos, elaborar diccionarios y tesauros, estudiar la variación lingüística**,** etc.

2.2.3.1 Corpus históricos

El uso de colecciones de textos para el estudio de la lengua no es un área nueva. La lingüística histórica vio con prontitud el potencial y la utilidad de los corpus históricos computarizados, tal es el caso del corpus diacrónico con textos en ingles de diferentes periodos compilados por la Universidad de Helsinki. *The Helsinki Corpus of English Texts,* que contienen textos desde el periodo antiguo, medio y moderno, con un total de 1,5 millones de palabras.

Otro corpus histórico, es el reciente *Lampeter Corpus of Early Modern English Tracts,* esta colección consiste en folletos y panfletos publicados entre 1640 y 1740 de seis dominios diferentes. El corpus Lampeter puede ser visto como un ejemplo de un corpus que cubre más de un área especializada.

2.2.3.2 Corpus para propósitos específicos

Existe una amplia y gran variedad de corpus altamente especializados que son creados para propósitos específicos. Muchos de estos son usados para trabajar en sistemas de lenguaje hablado. Ejemplos como estos son, el *Air Traffic Control Corpus, ATC0*, creado para ser usado en el área de reconocidas expresiones de dominio para el control de tráfico aéreo

Muchos de los corpus especializados, pertenecen al *Centre for Spoken Language Understanding*, CSLU, en Oregón. Estos no son de uso restringido dentro de una materia o área, son especializados por sus contenidos.

2.2.3.3 Corpus Internacional/multilingües

Existe una amplia variedad de corpus en Ingles. Hasta ahora mucho de los corpus son ciertamente del lenguaje inglés, por varias razones. Hay sin embargo, un número creciente de corpus disponibles en otras lenguas. Algunos de ellos son corpus monolingües, colecciones de textos en una lengua. Como por ejemplo *Oslo Corpus of Bosnia text* y el *Contemporary Portuguese Corpus.*

Un número de corpus multilingües también existen. Muchos de estos son corpus paralelos, corpus con el mismo texto en diferentes lenguas. Estos corpus son con frecuencia usados en el área de traducción.

2.3 ¿Qué es un CORPUS GOLDEN ESTANDAR?

El término *Gold-Standard* fue originalmente acuñado en economía, principalmente en los sistemas monetarios, donde el valor de la unidad monetaria está basado en una cantidad física de oro. Si bien ya no es usado hoy en día, aún se sigue considerando como un sistema extremadamente estable para cuestiones financieras. El significado del término ha sido transformado para denotar procedimientos científicos o colecciones que son aceptadas como estándares (Wissler, 2014). Básicamente un *Corpus Gold*

Estándar (o CGS) es el corpus que esta validado por expertos, o bien por revisores que determinan que el corpus permite realizar un trabajo con la certeza que será el mejor de su clase.

La construcción de un CGS es un proceso que consume mucho tiempo y bastante dedicación, el cual generalmente es realizado por expertos. El tamaño, la calidad y la mayoría de las tareas específicas del CGS influyen directamente en el desarrollo del algoritmo de procesamiento de lenguaje natural basado en aprendizaje automático.

Por lo tanto, los costos de la creación de un corpus, el que sea realmente confiable para entrenar y evaluar con aprendizaje automático son altos. Recientes investigaciones en esta área intentan determinar el tamaño de un corpus, evaluando la influencia de anotadores no-expertos en la calidad del CGS**,** o sustituirlo generando automáticamente un corpus estándar plata (*Silver Standard Corpus*, SSC) (Wissler, 2014). Un SSC describe un nivel respecto a la calidad de anotación de corpus que se encuentra entre un CGS creado manualmente**,** y la salida no revisada de un procesamiento automático (Eckart, 2016). Basado en artículo

https://www.linguistics.rub.de/konvens16/pub/12_konvensproc.pdf).

Los siguientes pasos proveen un vistazo general a los procedimientos para determinar el número de documentos necesarios para un CGS, basado en lo que ha mencionado Juckett (Juckett, 2012):

- Pre-selección de un corpus de trabajo con aquellos documentos del área de estudio.
- Selección y creación de un corpus comparativo que contenga palabras comúnmente usadas.
- Extracción de token, vale decir, convertir el corpus de estudio y el corpus comparativo en tokens. Un token es considerado como cualquier secuencia adyacente de caracteres alfanuméricos comenzando con una letra y ocurrencia entre espacios, entre corchetes, signos de interrogación y puntos.
- Restar del conjunto de tokens del corpus de comparación con el corpus de

estudio creando un conjunto aparte.

- Calcular la frecuencia de ocurrencia de los tokens del conjunto que se generó en el paso anterior.
- Calcular las probabilidades de los tokens único en función de los diferentes tamaños del CGS.
- Integrar (suma de los pesos) la probabilidad dentro de un valor para cada tamaño del corpus.
- Seleccionar el tamaño del corpus dependiendo de una probabilidad aceptable.

3.- Modelado de tópicos

El modelado de tópicos o temas es una tarea que consiste en identificar los conceptos implícitos en una colección de documentos, así como determinar los tópicos que los documentos abordan principalmente. Este tipo de modelado tiene muchas aplicaciones. Por ejemplo, se puede usar en recuperación de información (IR, por sus siglas en inglés) para determinar grupos de documentos similares a una consulta específica, para identificar individuos influyentes en las plataformas sociales, por mencionar algunas aplicaciones (Jónsson, [2015?]). Los algoritmos que se han usado tradicionalmente en el modelado de tópicos incluyen Análisis Probabilístico de Semántica Latente (pLSA) y Asignación de Dirichlet Latente (LDA).

De acuerdo con (Steyvers & Griffiths, n.d.) el enfoque LSA hace tres supuestos fundamentales: i) la información semántica se puede derivar de una matriz de co-ocurrencia palabra-documento (ver figura 2); ii) la reducción de dimensionalidad es una parte esencial de esta derivación; y iii) las palabras y documentos se pueden representar como puntos en un espacio euclidiano. En el caso particular de este trabajo nos enfocamos en el modelado de tópicos basado en la técnica LDA, que de acuerdo con los autores, es consistente sólo con los dos primeros supuestos de LSA debido a que describe una clase de modelos estadísticos donde las propiedades semánticas de las palabras y los documentos se expresan en términos de tópicos probabilísticos.

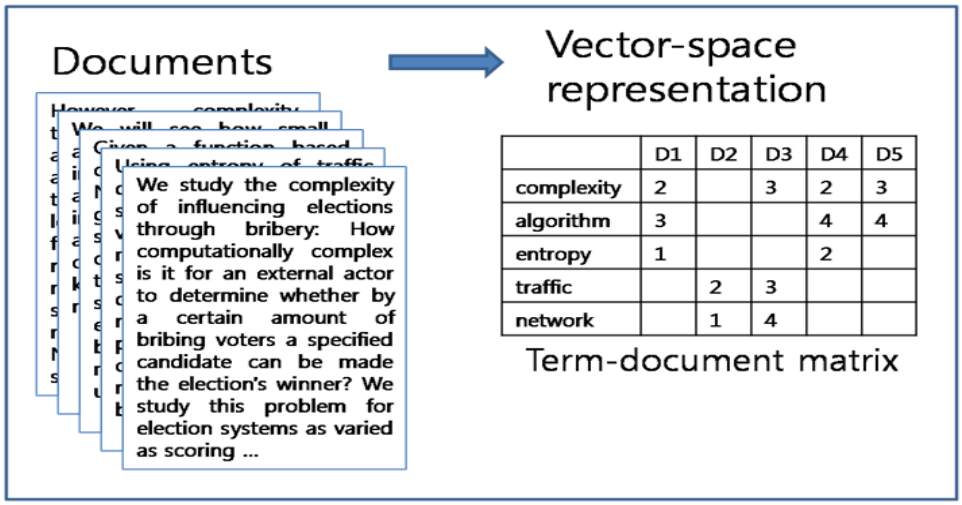

Figura 2. Representación de documentos mediante una matriz palabra-documento.

Los modelos de tópicos se basan en la idea que los documentos son una mezcla de tópicos. En este contexto, un *tópico* es una distribución de probabilidad sobre un vocabulario fijo. Así, un modelo de este tipo es un *modelo generativo* para documentos porque especifica un procedimiento probabilístico mediante el cual se pueden generar los documentos (imagen izquierda de la figura 2). En este sentido, entendemos por *documento* una secuencia de n palabras denotadas por $w_i = (w_1, w_2, \ldots, w_n)$ donde w_i es la i-ésima palabra en la secuencia. Para elaborar un nuevo documento, seleccionamos una distribución de tópicos. Posteriormente, para cada palabra de un documento particular, seleccionamos un tópico al azar de acuerdo con esta distribución y, finalmente, se extrae una palabra.

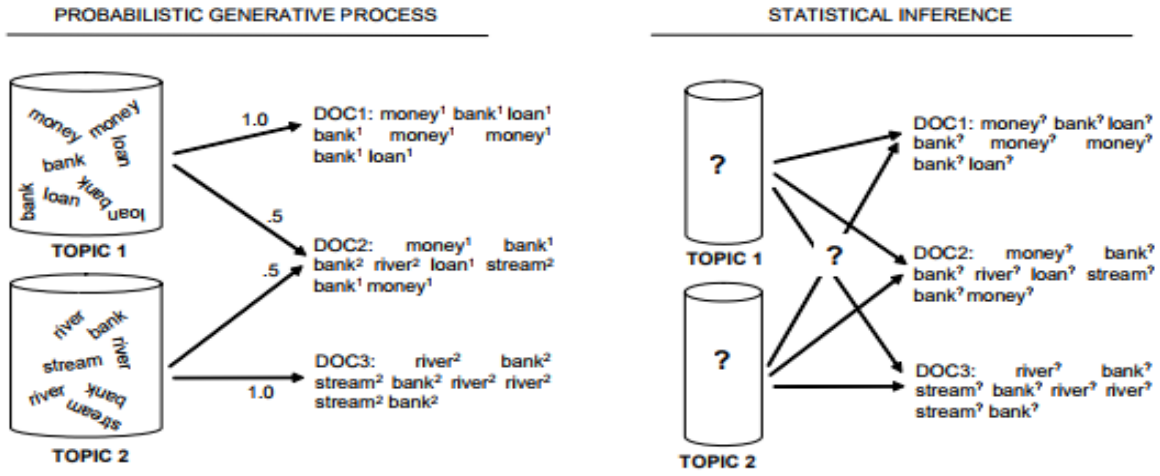

Figura 3. Perspectivas de modelo generativo y estadística inferencial en el modelado de tópicos.

Dado que partimos generalmente de documentos que ya han sido elaborados por sus correspondientes autores, requerimos de técnicas que nos ayuden a inferir el conjunto de tópicos responsable de generar la colección de documentos analizada, para ello se pueden usar técnicas estadísticas con la finalidad de invertir el proceso (imagen derecha de la figura 2). La figura 4 muestra cuatro tópicos de ejemplo que se derivaron del corpus TASA, de una colección aproximada de 37,000 textos de materiales educativos recolectados por la Touchstone Applied Science Associates (Steyvers & Griffiths, n.d.).

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figura 4. Distribución de palabras de cuatro tópicos extraídos del corpus TASA.

De este modo, cada tópico queda conformado por un conjunto de palabras temáticamente relacionadas. Por ejemplo, en el caso del tópico 247 de la figura 4, se trata de un tópico relacionado con el uso de drogas. Dado lo anterior, es posible generar cada documento a partir de la combinación de palabras de los distintos tópicos. Debido a que pueden existir palabras polisémicas, es decir, con varios significados, no existe noción de exclusividad de tópico, por lo tanto, una palabra puede pertenecer a más de un tópico a la vez (por ejemplo, el caso de la palabra PLAY en la figura 5) y su significado se puede derivar a partir del contexto donde se encuentra inserta. Finalmente, los tópicos no presuponen algún orden en el que las palabras deban aparecer en un documento.

Topic 77

word	prob.
MUSIC	.090
DANCE	.034
SONG	.033
PLAY	.030
SING	.026
SINGING	.026
BAND	.026
PLAYED	.023
SANG	.022
SONGS	.021
DANCING	.020
PIANO	.017
PLAYING	.016
RHYTHM	.015
ALBERT	.013
MUSICAL	.013

Topic 82

word	prob.
LITERATURE	.031
POEM	.028
POETRY	.027
POET	.020
PLAYS	.019
POEMS	.019
PLAY	.015
LITERARY	.013
WRITERS	.013
DRAMA	.012
WROTE	.012
POETS	.011
WRITER	.011
SHAKESPEARE	.010
WRITTEN	.009
STAGE	.009

Topic 166

word	prob.
PLAY	.136
BALL	.129
GAME	.065
PLAYING	.042
HIT	.032
PLAYED	.031
BASEBALL	.027
GAMES	.025
BAT	.019
RUN	.019
THROW	.016
BALLS	.015
TENNIS	.011
HOME	.010
CATCH	.010
FIELD	.010

Figura 5. Tres tópicos relacionados con la palabra PLAY.

La representación mediante tópicos probabilísticos tiene ventajas sobre representaciones puramente espaciales. Cada tópico es individualmente interpretable porque proporciona una distribución de probabilidad sobre palabras que son temáticamente relacionadas (ver figura 5), lo que al ojo humano resulta más transparente. Esto contrasta con los ejes arbitrarios de una representación espacial.

3.1 Modelos generativos

Un modelo generativo para documentos consiste de reglas de muestreo probabilístico simples que describen cómo se pueden generar las palabras de los documentos basado en variables latentes (es decir, no observadas directamente). El objetivo de ajustar un modelo generativo es encontrar el mejor conjunto de variables latentes que expliquen los datos observados, asumiendo que el modelo realmente generó los datos.

El modelo generativo descrito anteriormente no hace supuesto alguno respecto al orden de las palabras en los documentos. La única información relevante al modelo es la frecuencia de ocurrencia de las palabras en el documento. A esto se le denomina *supuesto de bolsa de palabras (*en inglés*, bag-of-words assumption),* y es común en muchos modelos estadísticos del lenguaje, incluyendo LSA.

3.2 Modelado de tópicos probabilístico

En la literatura sobre este tipo de trabajos se da cuenta del uso de una variedad de modelos de tópicos probabilísticos para analizar el contenido de documentos y el significado de las palabras. Todos estos modelos usan la misma idea fundamental: un documento es una mezcla de tópicos. Sin embargo, hacen supuestos estadísticos ligeramente diferentes.

A continuación, se presenta una descripción formal de la notación generalmente usada en estos modelos probalísticos (extraída de (Steyvers & Griffiths, n.d.):

P (z) ← corresponde a la distribución de tópicos z en un documento particular.

P (w | z) ← corresponde a la distribución de probabilidad de las palabras w dado el tópico z.

P (z_i = j) ← corresponde a la probabilidad de que el j-ésimo tópico fuese muestreado para la i-ésima palabra.

P (w_i | z_i = j) ← corresponde a la probabilidad de la palabra w_i dado el tópico j-ésimo. Así, el modelo específica la siguiente distribución de palabras dentro de un documento:

$$P(w_i) = \sum_{j=1}^{T} P(w_i \mid z_i = j) P(z_i = j)$$

Donde T se refiere al número de tópicos. Con el ánimo de simplificar la notación, (Steyvers & Griffiths, n.d.) remplazan los términos de la expresión anterior por: $\emptyset^{(i)} = P(w \mid z = j)$ para referir a la distribución multinomial de palabras dado el tópico j y $\theta^{(d)} = P(z)$ para referir a la distribución multinomial de tópicos para el documento d. Además, asumimos que la colección consiste de D documentos y cada documento d consiste de n_d palabras. Así, N es el número total de palabras generada por la colección de documentos (es decir, $N = \sum n_d$). Los parámetros ϕ y θ indican qué palabras son importantes para qué tópico y qué tópicos son importantes para un documento particular, respectivamente.

3.3 Asignación de Dirichlet Latente

La técnica LDA para el modelado de tópicos la propuso D. Blei (Blei, 2012) al introducir una Dirichlet previa sobre θ (que corresponde a una extensión de la técnica ya planteada por Hofmann (1999; 2001), (Hofmann, T., 1999)). Como una previa conjugada para la multinomial, la distribución Dirichlet es una elección conveniente, lo que simplifica el problema de inferencia estadística. La densidad de probabilidad de una distribución Dirichlet T-dimensional sobre la distribución multinomial p= (p_1,…,p_T) se define como:

$$\mathrm{Dir}(\alpha_1,\ldots,\alpha_T) = \frac{\Gamma\left(\sum_j \alpha_j\right)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1}$$

Los parámetros de esta distribución se especifican por $\alpha_1 \ldots \alpha_T$. Cada hiperparámetro α_j se puede interpretar como una frecuencia previa del número de veces que el tópico j se muestrea en un documento, antes de observar cualquier palabra real de ese documento. Es conveniente usar una distribución Dirichlet simétrica con un hiperparámetroα, tal que $\alpha_1 = \alpha_2 = \ldots = \alpha_T = \alpha$. Al considerar una Dirichlet previa sobre la distribución de tópicosθ, el resultado es una distribución de tópicos *suavizada,* con la cantidad de suavizada determinada por el parámetroα.

Un ejemplo simple de resultados generados por la aplicación de LDA

Se tiene el siguiente conjunto de oraciones:

1.- Me gusta comer brócolis y plátanos

2.- Yo odio los plátanos y las espinacas al desayuno

3.- Las chinchillas y los gatos son simpáticos

4.- Mi hermana adopto un gato ayer

5.- Este hermoso hámster come ruidosamente un brócoli

La aplicación de LDA produce la distribución de palabras siguiente:

Tópico A: 30% brócoli, 15% de plátanos, 10% desayuno, 10% de masticar

Tópico B: 20% chichillas, 20% de gatos, 20% hermoso, 15% hámster

Interpretación: El tópico A está relacionado con alimentos y el tópico B con animales.

Ahora, la distribución de tópicos dentro de documentos es la siguiente:

Oraciones 1 y 2: 100% del tópico A
Oraciones 3 y 4: 100% del tópico B
Oración 5: 60% del tópico A, 40% del tópico B

4.- METODOLOGIA

La construcción de un corpus de dominio específico para fines de extracción terminológica y conceptual es una tarea que consume tiempo y esfuerzo, donde generalmente se debe involucrar a expertos en los temas de interés para recolectar la información relevante y representativa del área en la que estamos interesados. La participación de estos expertos se vuelve aún más difícil si le agregamos el requerimiento de obtener grandes cantidades de información. Así, el énfasis de este trabajo es aportar una metodología que reduzca el trabajo del experto a considerar sólo un conjunto representativo de documentos, no necesariamente enorme si consideramos las tendencias de tamaño de corpus actuales, pero sí representativo de las áreas que tengan conceptualizaciones en común (por ejemplo, en las áreas de la medicina, paciente-enfermedad-tratamiento, son conceptos comunes con el mismo significado) y generar con ello un modelo que permita diferenciar los documentos relevantes al dominio de los que no lo son.

Dado lo anterior, en esta sección describimos la metodología derivada de esta investigación para generar de forma automática un corpus de dominio específico. Dicha metodología contempla la implementación de un modelo de tópicos donde se conoce *a priori* el número de tópicos y lo que se desea es derivar el vocabulario que caracteriza a cada uno de los tópicos.

4.1 Pre-procesamiento de los textos

La producción científica en español que se considera puede contener información gráfica o en forma de tablas, así como el apartado de referencias, que no son relevantes para los propósitos del corpus. Así, se debe eliminar esta información de los documentos.

4.2 Etiquetado de partes de la oración (etiquetado PoS)

La tarea de etiquetado gramatical o de partes de la oración es el proceso de asignar una categoría gramatical o parte del discurso a cada token en un corpus. Por ejemplo, una salida del etiquetador proporciona comúnmente la forma de la palabra tal y como ocurre en el discurso real, la etiqueta PoS y el lema con la siguiente estructura:

Palabra real	Etiqueta PoS	Lema
Síntomas	NC	Síntoma

Tabla 1. Etiquetado PoS

Las etiquetas usadas por TreeTagger para el caso del español se basan en el conjunto definido en Penn Treebank[3]. La metodología que proponemos en este trabajo requiere etiquetado gramatical porque consideramos sólo nombres y adjetivos para la fase de modelado de tópicos. La consideración de sólo nombres y adjetivos se asume porque la extracción terminológica se enfoca generalmente en estas dos categorías para la conformación de términos, es decir, etiquetas lingüísticas que tengan potencial de hacer alusión a conceptos relevantes del dominio.

4.3 Eliminación de *ruido* en los textos

4.3.1 Lista de palabras funcionales (stopwords)

Las palabras de clase cerrada (artículos, conjunciones, etc.) tienen una función gramatical en los textos, para los fines de nuestro trabajo no aportan significado relevante debido a que el objetivo es caracterizar las áreas de la medicina que

[3] www.ims.uni-stuttgart.de/ftp/pub/corpora/spanish-tagset.txt

consideramos en términos de un vocabulario específico. Así, proponemos el uso de la lista stopword considerada en NLTK para el español. Dicha lista puede ser invocada a partir de los comandos:

```
from nltk.corpus import stopwords
stop = stopwords. words ('spanish')
```

Si bien el filtrado de otras categorías diferentes a nombres y adjetivos se realiza con el etiquetado gramatical, sabemos que la precisión de los etiquetadores no es del 100%, por lo que puede darse el caso que categorías de este tipo sean etiquetadas como nombres y adjetivos erróneamente, por ello es de importancia considerar esta lista como una segunda etapa de reducción de ruido.

4.3.2 Filtrado de nombres y adjetivos no relevantes

Acosta, Aguilar, Infante (2015) proponen un método basado en la comparación entre corpus (corpus de referencia, de aproximadamente 5 millones de tokens vs. corpus de medicina extraído de MedLinePlus en español) para identificar las palabras más relevantes del dominio analizado, así como asignarles una ponderación que refleje su relevancia o su potencial de ser un término del dominio (en inglés, *termhood*). Aunado a lo anterior, implementan un conjunto de heurísticas lingüísticas para obtener adjetivos calificativos que generalmente no participan en la construcción de términos (por ejemplo*, frecuente, importante, bajo*, etc.). Así, dado que trabajamos con textos de sub-áreas de la medicina, en este trabajo consideramos la lista de nombres con ponderación muy baja para el dominio y los adjetivos obtenidos con las heurísticas lingüísticas. Dicha lista de adjetivos y nombres se incorpora a la stopword de NLTK con el objetivo de filtrar estos elementos del corpus ya lematizado y filtrado.

4.3.3 Filtrado de palabras comunes y no polisémicas

La modelación de tópicos se enfoca en determinar la distribución de palabras que caracteriza a cada tópico. Una vez generada esta caracterización, podemos derivar la distribución de tópicos correspondiente a cada documento. En este proceso de ajustar un modelo que haga emerger los diferentes tópicos implícitos en los textos es posible encontrar palabras que sean polisémicas, es decir, que tengan diferentes significados de acuerdo al tópico donde se encuentren. Un ejemplo de esta situación es la palabra *lengua*. *Lengua* en textos de lingüística tiene un significado diferente a la misma palabra encontrada en textos de medicina. Así, en este trabajo proponemos el filtrado de palabras que sean vocabulario en común entre las sub-áreas consideradas y que realmente tienen el mismo significado. Ejemplos de este tipo de palabras pueden ser: *paciente, diagnóstico, tratamiento, enfermedad*, etc.

4.4 Herramientas computacionales

A continuación se describe, a grandes rasgos, las herramientas que se proponen para cada fase de la aplicación de la metodología.

4.4.1 SKETCH ENGINE

Sketch Engine es una herramienta desarrollada por Lexical Computing Limited, empresa fundada en 2003 por el lexicógrafo e investigador Adam Kilgarriff (1960-2015). Se trata de un software comercial, escrito en C++, Python, JavaScript, jQuery, bajo el sistema operativo Linux, Mac OS X.

Sketch Engine realiza análisis textual en línea, básicamente toma como entrada un corpus en cualquier idioma, con un cierto nivel de anotación lingüística para su posterior análisis.

Sketch Engine permite trabajar con los siguientes tipos de corpus ((Universidad Complutense de Madrid.Facultad de Filologia, 2011):

• Corpus que vienen integrados en la herramienta.

• Corpus disponibles en instituciones o en grupos de investigación que hay que “subir” previamente a la herramienta.

• Corpus construidos a partir de textos seleccionados y recopilados manualmente

• Corpus construidos automáticamente con la herramienta WeBootCat, la cual permite la búsqueda y recopilación automática de documentos disponibles en la Web.

4.4.1.1 Corpus en texto plano vs. corpus vertical

El *corpus en texto plano*, es un archivo de texto sin etiquetas, es decir, está conformado por archivos formados exclusivamente por textos, sólo caracteres, sin ningún formato, no proveen de información relevante en relación al documento como tipo de letras, tamaño, formas etc. Ver tabla 2.

```
Síndrome de Down
Es un trastorno genético en el cual una persona tiene
47 cromosomas en lugar de los 46 usuales.

Causas
En la mayoría de los casos, el síndrome de Down ocurre cuando hay una
copia extra del cromosoma 21. Esta forma de síndrome de Down se
denomina trisomía 21. El cromosoma extra causa problemas con la forma
en la que se desarrollan el cuerpo y el cerebro.
```

Tabla 2. Texto plano

En cambio, el *corpus texto vertical* está conformado por la forma de la palabra real, la etiqueta gramatical que le corresponde y el lema (la forma canónica de la palabra, tal como aparecería en un diccionario). Este formato es el mejor para preservar la mayor cantidad de información posible.

Forma de la palabra	**Etiqueta**	**Lema**
La	DA	el
casa	NC	casa
roja	AQ	rojo

Tabla 3. Texto vertical

La plataforma Sketch Engine se utilizará para construir y analizar el corpus en términos de la construcción de listas de palabras asociadas con una palabra específica (opción tesauro), así como explorar este tipo de asociaciones visualmente (wordcloud de palabras asociadas), explorar concordancias utilizando el lenguaje de consulta de la herramienta (CLQ), obtención de estadísticas del corpus a medida que se va incorporando información (número de tokens, distribución de frecuencias de palabras, etc.)

4.4.2 Python

Python es un lenguaje de programación de alto nivel correspondiente al paradigma orientado a objetos, que se puede usar para realizar todo tipo de tareas en múltiples plataformas.

Las características de Python son las siguientes:

- La cantidad de librerías que contiene, tipos de datos y funciones incorporadas en el propio lenguaje que ayudan a realizar muchas tareas habituales sin necesidad de tener que programarlas desde cero.
- La sencillez y velocidad con la que se crean los programas. Un programa en Python puede tener de 3 a 5 líneas de código menos que su equivalente en Java o C.
- La cantidad de plataformas en las que se puede desarrollar, como Unix, Windows, Mac.

Con este lenguaje se pretende desarrollar el script para implementar la metodología propuesta. Es decir, el algoritmo contemplará desde la fase de entrada de documentos en texto plano, que ya han pasado por la fase de eliminación de información no relevante (gráficos, tablas, etc.), hasta la construcción del modelo de tópicos y la fase de prueba para medir desempeño.

4.4.3 NLTK

La herramienta de Procesamiento de Lenguaje Natural o más comúnmente NLTK es un conjunto de bibliotecas y programas para el procesamiento del lenguaje natural. NLTK fue creado en 2001 como parte de un curso de lingüística computacional del Departamento de Ciencias de la computación e informática de la Universidad de Pennsylvania. Desde entonces, se ha ido desarrollando y expandiendo.

Específicamente NLTK es una colección de herramientas que permiten hacer varias tareas: desde procesar uno o varios documentos para fines sencillos, como contar palabras, agrupar tokens, observar distribuciones de frecuencia, entre otras. También se puede realizar análisis sintáctico, desambiguación de frases y tareas de extracción de información.

Además, NLTK está destinado a apoyar la investigación y la enseñanza en PLN o áreas muy relacionadas, que incluyen: la lingüística empírica, las ciencias cognitivas, la inteligencia artificial, la recuperación de información, y el aprendizaje máquina.

NLTK se usará en las primeras fases del procesamiento textual para segmentar en tokens los textos, así como realizar el filtrado de la lista stopword para el español incorporada en el módulo.

4.4.4 GENSIM

Es una librería diseñada para extraer automáticamente la semántica de los documentos de la forma más eficiente y con la menor complicación posible. Básicamente, la librería está diseñada para procesar textos sin estructura. Los algoritmos en gensim, como análisis de semántica latente, LDA y proyecciones aleatorias, descubren la estructura semántica de los documentos al examinar los patrones estadísticos de co-ocurrencia de las palabras dentro de un corpus de documentos. Estos algoritmos no son supervisados, lo que significa que no hay intervención humana, sólo se requiere un corpus de documentos sin estructura. Una vez que se encuentran estos patrones estadísticos, cualquier documento de texto plano se puede expresar de manera sucinta en la nueva representación semántica y se puede explorar su similitud con otros documentos.

Gensim, se usará para generar el diccionario de términos para la modelación de tópicos, así como su vectorización considerando la frecuencia de ocurrencia.

4.4.5 MALLET

MALLET ((McCallum, 2002) es una herramienta computacional basada en Java para el procesamiento estadístico del lenguaje natural, clasificación de documentos, agrupamiento o clustering, modelado de tópicos, extracción de información, y otras aplicaciones a los textos con aprendizaje automático.

MALLET incluye complejas herramientas para la clasificación de documentos: eficientes rutinas para transformar un texto a un conjunto de "rasgos", una amplia variedad de algoritmos como Naive Bayes, Entropía máxima, árboles de decisión, y de específicos códigos para evaluar el desempeño del clasificador utilizando varias métricas de uso común.

Además de clasificación, MALLET incluye herramientas para el etiquetado secuencial en aplicaciones tales como la extracción de entidades nombradas en un texto. El conjunto de herramientas de MALLET como el modelado de tópicos contiene eficientes implementaciones, basadas en muestreo Gibb usando LDA, Pachinko Allocation y LDA jerárquico.

Muchos de los algoritmos que utiliza MALLET dependen de una optimización numérica, por tal motivo, MALLET incluye una implementación eficiente de memoria limitada de BFGS, entre otros métodos de optimización.

Además de las complejas herramientas de aprendizaje automático, MALLET incluye rutinas capaces de transformar los textos en representaciones numéricas que luego pueden ser procesadas de manera eficiente.

Mallet está considerado en el grupo de herramientas que se utilizará para realizar el modelado de tópicos como una exploración de la técnica, aunque Mallet está programado en Java, se pueden encontrar implementaciones en Python.

4.4.6 Módulo LDA Python

El módulo LDA de Python implementa la Latent Dirichlet Allocation (LDA) mediante muestreo Gibbs. El módulo es rápido y ha sido probado en Linux, OS X y Windows. Se requiere contar con una versión Python 2.7 o 3.3+, además de requerir de los paquetes Numpy y pbr. Una de las ventajas de este módulo, indiscutiblemente es que puede ser importado directamente en un Script Python para desarrollar el script completo.

Lda tiene partes esenciales escritas en C vía Cython. Si se requiere trabajar con un corpus muy grande se recomienda usar implementaciones más sofisticadas, por ejemplo, la implementada en hca y Mallet. Hca está escrito completamente en C y MALLET en Java. A diferencia de lda, hca puede usar más de un procesador a la vez. Tanto Mallet como hca implementan la técnica de modelación de tópicos de forma robusta comparada con el LDA estándar[4].

[4] https://pypi.python.org/pypi/lda

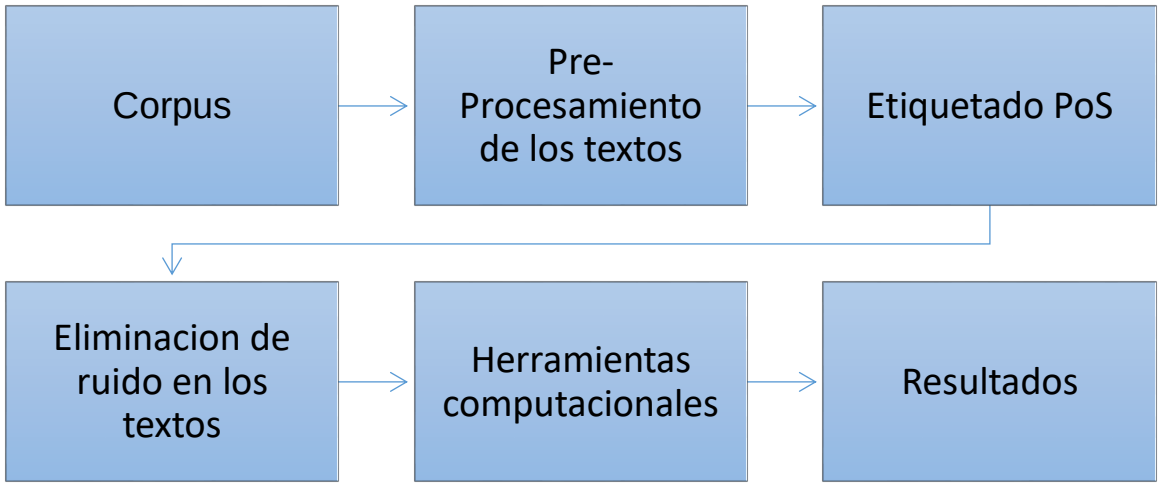

Figura 7 Metodología para generar de forma automática un corpus de dominio específico.

5.- Resultados

5.1 Colección de documentos

La fuente base del corpus es SciELO. SciELO es un proyecto de biblioteca electrónica, iniciativa de la Fundación para el apoyo a la investigación científica del estado de Sao Paulo, Brasil y del Centro Latinoamericano y del Caribe de información científica de la salud (BIREME). Dicho repositorio permite la publicación de ediciones completas de las revistas científicas mediante una plataforma de software que posibilita el acceso a través de distintos mecanismos, incluyendo listas de títulos, materias, índice de autores y un motor de búsqueda. El proyecto SciELO tiene como objetivo el desarrollo de una metodología común para la preparación, almacenamiento, diseminación y evaluación de la literatura científica en formato electrónico.

Desde el sitio www.scielo.org se eligieron las colecciones en español de revistas específicas del área médica, concretamente de las áreas de oftalmología, ginecología, y cardiología. El corpus está conformado de artículos en español y consiste de una colección de documentos del dominio de la medicina, específicamente de enfermedades, cirugías, diagnósticos, tratamientos, alteraciones y temas relacionados, para las tres áreas consideradas. Este corpus está conformado por 246 artículos, lo que arroja un total de 1.026.954 tokens, 871.110 palabras y 33.000 oraciones.

5.2 Interfaz principal del programa

Dada la escasez de aplicaciones con una interfaz gráfica amigable se decidió programar una aplicación con el lenguaje de programación Python, utilizando como base para el diseño y programación de la interfaz gráfica el módulo tkinter. Actualmente, lo más común es encontrar librerías disponibles para interactuar vía comandos, o bien invocarlas en aplicaciones específicas, caso de Mallet (Java) o lda (Python). En este apartado se muestra la interfaz para definir los parámetros relevantes para entrenar un modelo con la técnica LDA, así como los diferentes resultados obtenidos del entrenamiento y prueba del modelo.

El procesamiento se hace con 246 artículos en total. Donde 222 (aproximadamente 90%) se usan como conjunto de entrenamiento y 24 (10% restante) de prueba. Es importante mencionar que se incluyeron 10 documentos más en el conjunto de prueba que corresponden a las áreas de traumatología y pediatría con el propósito de explorar el desempeño del modelo con documentos de otras áreas no usadas en el entrenamiento.

El análisis de resultados se realizó considerando dos escenarios diferentes. El primer escenario sólo contempla un filtrado de palabras no relevantes básico (stopwords) basado en la lista disponible para el español en NLTK, mencionada ya en el capítulo anterior. Por otro lado, el segundo escenario considera etiquetado gramatical (en este caso se usa el etiquetador TreeTagger), además del filtrado de palabras no relevantes de NLTK más las extraídas de forma automática vía el enfoque de contraste de corpus mencionado también en el capítulo anterior. Cabe mencionar que a medida que se realizaban las exploraciones de resultados, variando número de tópicos y número de palabras más relevantes por tópico, palabras con significado común (ejemplo, enfermedad-paciente-diagnóstico, etc.) fueron incorporadas a la stopword para eliminar *ruido* y dejar una distribución de palabras por tópico más depurada.

Finalmente, el desempeño del modelo se evalúa mediante matrices de confusión como herramienta básica para analizar el clasificador.

5.2.1 Escenario 1: Sólo se realiza el filtrado de la lista de palabras no relevantes (Stopwords) disponible en NLTK. Los parámetros definidos corresponden a 3 tópicos, 20 palabras por tópico y 400 iteraciones.

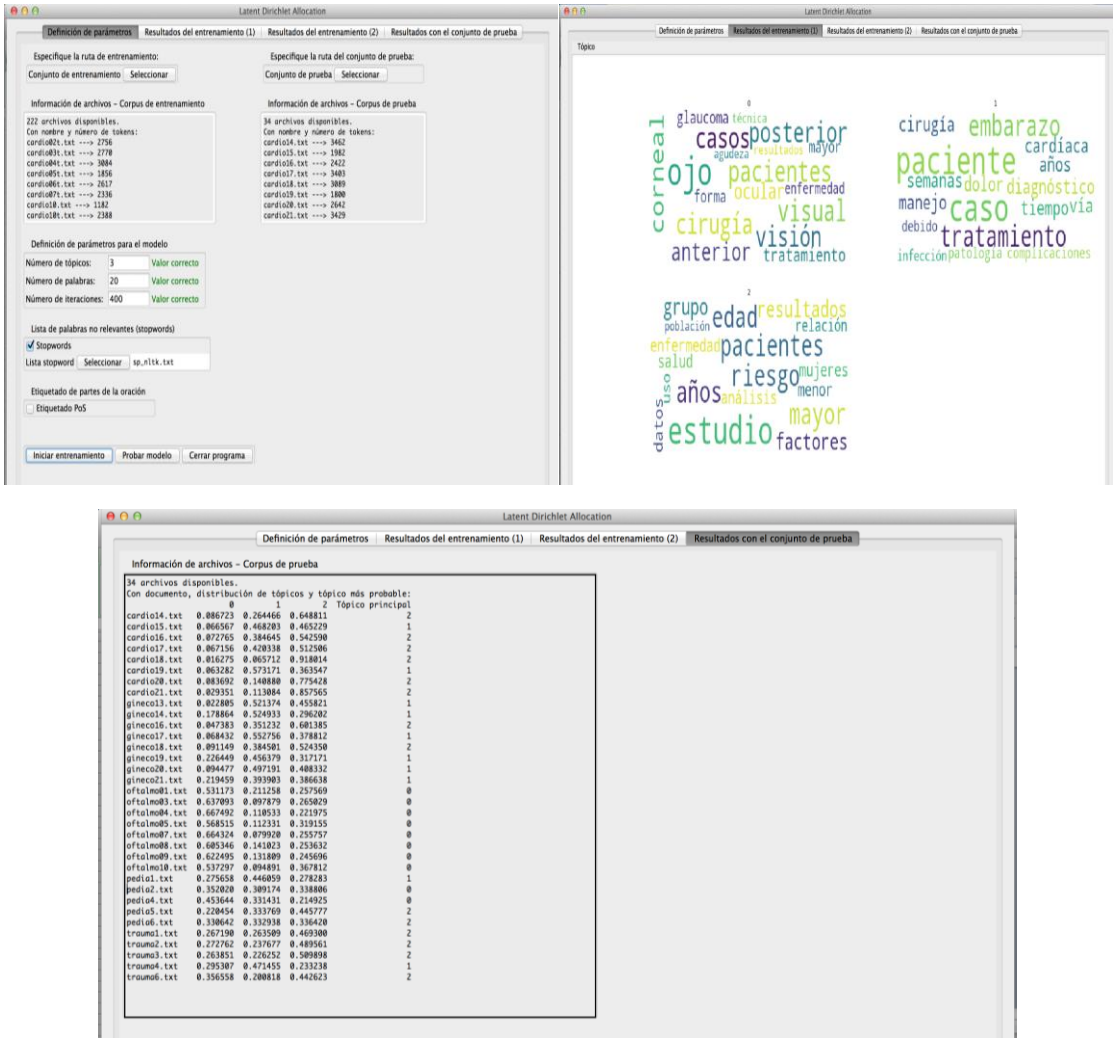

Imagen 1. Definición de parámetros; Word cloud; Resultados con el conjunto de prueba distribución de tópicos.

5.2.1.1 Interpretación de resultados del escenario 1: De las imágenes anteriores se deriva que sólo el tópico 0 se encuentra más cercano a una caracterización del área de la oftalmología, seguido del tópico 1, que podría ser una incipiente descripción del área de la ginecología (embarazo, infección, etc.). El tópico que definitivamente se encuentra ausente es el correspondiente al área de cardiología.

En este escenario de análisis observamos que el área que tiene una precisión perfecta es justamente oftalmología, siendo todos los casos de prueba clasificados correctamente en este tópico. No sucede lo mismo con el tópico incipiente de ginecología, donde se clasifica sólo el 75% como perteneciente a esta clase. Finalmente, los documentos de cardiología se clasifican en las clases 1 (25%) y 2 (75%), donde podemos observar que el tópico 2 sólo contiene términos genéricos que podrían aplicar a cualquier área de la medicina y donde la mayoría no necesariamente denota un concepto del dominio en cuestión (es el caso de *años, uso, datos, factores,* etc.). Omitimos la matriz de confusión debido a que sólo un tópico podría realmente considerarse caracterizado (tópico con encabezado 0).

5.2.2 Escenario 2: Se realiza el filtrado de la lista de palabras no relevantes (Stopwords) disponible en NLTK, más los adjetivos y nombres extraídos automáticamente vía el enfoque de comparación entre corpus mencionado en el capítulo anterior. Además, en las diferentes exploraciones realizadas al conjunto de documentos se determinó un grupo de nombres y adjetivos comunes a las 3 áreas, que no presentan polisemia, por tanto se incluyeron en la lista de palabras no relevantes. Por otro lado, se aplica etiquetado gramatical para considerar sólo nombres y adjetivos. Los parámetros definidos corresponden a 3 tópicos, 20 palabras por tópico y 400 iteraciones.

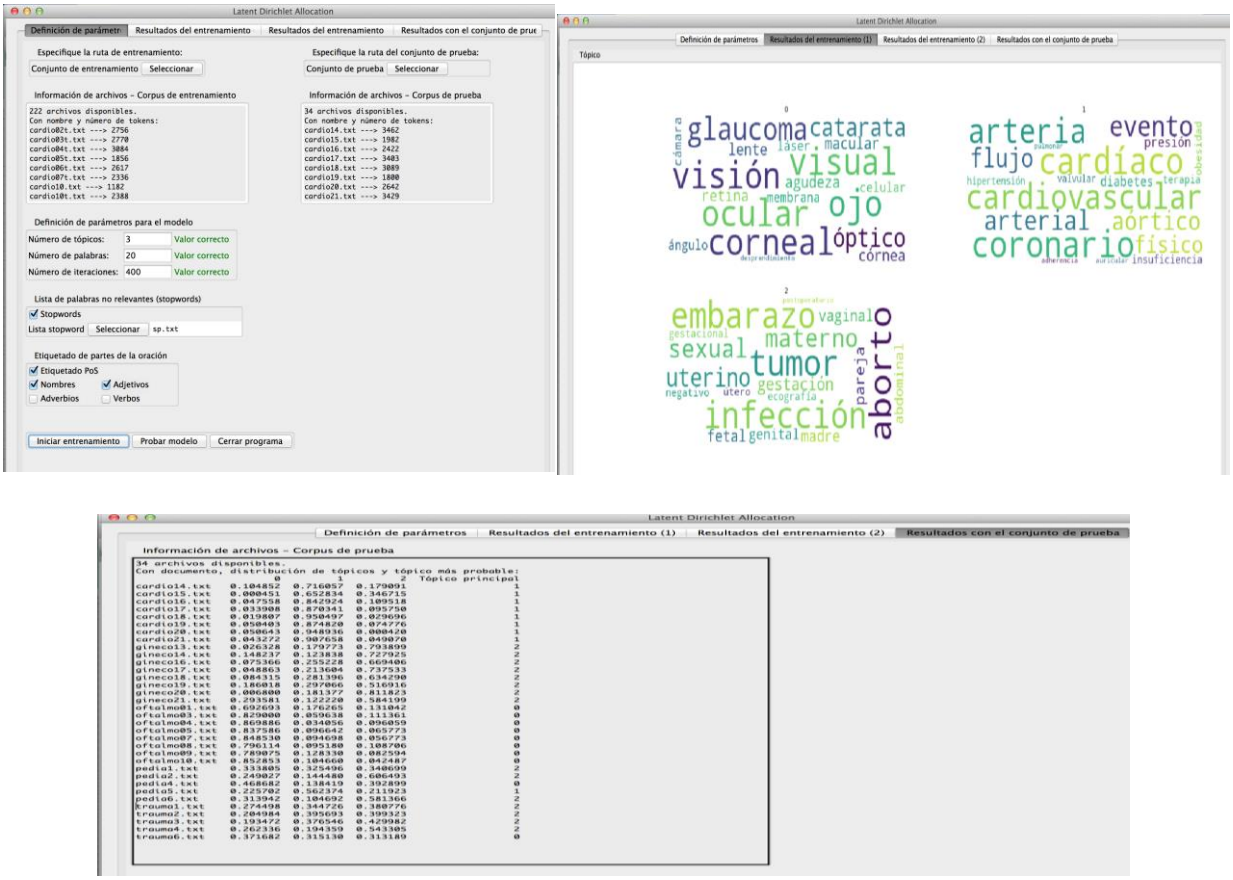

Imagen 2. Definición de parámetros; Word cloud; Resultados con el conjunto de prueba distribución de palabras.

5.2.2.1 Interpretación de resultados escenario 2: A partir de los resultados anteriores es posible observar la mejora obtenida con la aplicación del etiquetado gramatical, la lista de palabras no relevantes y de aquellos términos que no necesariamente tienen un significado diferente (polisemia) en cada sub-área. Algunos términos que podríamos considerar genéricos aún aparecen, por ejemplo, en el tópico 1, que caracteriza al área de cardiología, donde tenemos el caso de palabras como *terapia, flujo y físico*. De la matriz de confusión puede verse que la precisión[5] (la precisión o valor predictivo positivo definido como la proporción de verdaderos positivos contra todos los resultados positivos (tanto verdaderos positivos, como falsos positivos) es perfecta (100%) para el conjunto de prueba, es decir, el modelo muestra

[5] Fórmula para obtener precisión $\frac{VP}{VP+FP}$

una excelente capacidad de generalización, misma que puede deducirse a partir de observar las distribuciones de palabras para cada tópico. Ver imagen 2.

Es importante mencionar que algunos artículos de un área determinada, por ejemplo, cardiología, reportan experimentos donde podrían mencionar pacientes mujeres con algún cuadro específico ginecológico y ello sesgar la asignación del tópico principal, es decir, el tópico principal podría ser cardiología, sin embargo, se etiqueta como de ginecología. Desde nuestro punto de vista, lo anterior más que afectar el modelado puede enriquecerlo, ya que observando la distribución de tópicos para cada documento (que la aplicación desarrollada también despliega) podría ser de gran ayuda para un usuario final que busca ese tipo de casos en cardiología.

Escenario 2: Matriz de confusión – Corpus de prueba

		Clase estimada			Total
		Cardiología	Ginecología	Oftalmología	
Clase real	Cardiología	8	0	0	8
	Ginecología	0	8	0	8
	Oftalmología	0	0	8	8
Total		8	8	8	24

Tabla 4. Matriz de confusión corpus de prueba.

6. - Conclusiones y Trabajos a Futuro

- Los corpus son un recurso hoy en día inigualable para cualquier estudio lingüístico en general y en lingüística computacional. Ahora bien, con la incorporación de los computadores cada vez con mayor capacidad de almacenamiento y procesamiento, el acceso a los datos es rápido y fiable, así como su manipulación, extracción e identificación de información relevante de enormes cantidades de textos.

- La contribución concreta de este trabajo fue generar una metodología para la construcción automática de un corpus de dominio específico mediante el entrenamiento de un modelo de tópicos que fuera capaz de discriminar los documentos del dominio previamente definido de otro que no lo fuera. Para ello fue relevante la utilización de herramientas como MALLET, el módulo LDA Python y librerías como NLTK, GENSIM, entre otras mencionadas en el cuerpo del trabajo.

- La implementación de un programa computacional con una interfaz de fácil manipulación permitió comprobar el funcionamiento del corpus de entrenamiento y de prueba. Además, con estos resultados se evaluó el clasificador a través de una matriz de confusión.

- Este trabajo de investigación no estuvo exento de dificultades, primero la falta de expertos que se involucren en los temas de interés para la recolección de información relevante y representativa del área específica que se está trabajando. Por ello la construcción de un corpus de dominio específico para cualquier persona interesada en realizar una investigación es una tarea que consume tiempo y esfuerzo.

- La metodología planteada en este trabajo sólo consideró 3 sub-áreas de la medicina, por cuestiones de tiempo básicamente. Lo ideal, y esto queda como trabajo a futuro, es contemplar conjuntos representativos de la mayor cantidad posible de las áreas de la medicina y con ello entrenar un modelo que logre discernir documentos del área (s) de interés del resto automáticamente.

7. - Bibliografía

Aguilar, C. A., 2008. *Analisis linguistico de definiciones en contextos definitorios,* Mexico: UNAM.

Aguilar, C. A., 2009. Reconocimiento de definiciones asociadas a frases predicativas en contextos definitorios. *Procesamiento del Lenguaje Natural,* Volumen 43, pp. 151-158.

Atkins, S. C. J. O. N., 1991. *Corpus design criteria.* [En línea]
Available at: http://www.natcorp.ox.ac.uk

Baeza Yates, R. ,. R. N. B. e., 2011. *Modern information retrieval: the concepts and technology behind search.* Harlow, England: Pearson.

Bennett, G. R., 2010. *An Introduction to corpus linguistics,* Michigan: s.n.

Bird, S., 2009. *Natural Language Processing with Python.* 1st ed. Sebastopol, CA: O'Reilly.

Blei, D., 2012. Probabilistics topic models. *Communications of the ACM,* 55(4), pp. 77-84.

Bramer, W. M. [., 2016. Comparing the coverage, recall and precision of searches for 120 systematics reviews in Embase, MEDLINE, and Google Scholar.. *Systematic Review,* Volumen 5:39, pp. 2-7.

Chang, J., s.f. *Reading tea leaves: how human interpret topic models,* Princeton: Princeton University.

Chen, E., 2011. *Introductopn to Latent Dirichlet Allocation,* s.l.: s.n.

Chuang, J., 2012. Termite: visualization Techniques for Assessing Textual Topic Models. *AVI,* pp. 21-25.

Clark, S., 2013. *Topic modelling and latent dirichlet allocation,* s.l.: Lent.

Deleger, L., 2010. *Building Gold Standard Corpora for Medical Natural Language Processing Task,* Cincinnati: Division of Medical Informatics, Cincinnati Children's Hospital Medical Center.

EAGLES, s.f. *Introduccion a las etiquetas EAGLES.* [En línea]
Available at: http://nlp.lsi.upc.edu/freeling-old/doc/tagsets/tagset-es.html
[Último acceso: 10 noviembre 2016].

Eckart, K. G. M., 2016. *Creating silver standard annotations for a corpus of non-standard data.* Stuttgart, s.n.

Graham, S., Weingart, S. & milligan, I., 2012. *The Programming Historian.* [En línea]
Available at: http://programminghistorian.org/lessons/topic-modeling-and-mallet
[Último acceso: 4 November 2016].

Gries, S. T., 2009. What is corpus linguistics?. *Language and Linguistics Compass,* Volumen 3, pp. 1-17.

Griffiths, T. L., 2007. Topics in semantic representation. *Psychological Review,* 114(2), pp. 211-244.

Hanani, U., 2001. Information filtering: overview of issue, research and system. *User Modeling and User-Adapted Interaction,* pp. 203-259.

Hofmann, T., 1999. *Probabilistic latent semantic analysis.* s.l., s.n.

Hohmann, A., 1996. *Term extraction with standard tools for corpus exploration: experience from German.* Stuttgart, International Congress on Terminology and Knowledge Engineering.

Infante Bradley, T. E., 2015. *Diseno y evaluacion de un metodo de extraccion automatica terminos para el area de medicina,* Santiago: Pontificia Universidad catolica de Chile.

Jónsson, E. S. J., [2015?]. *An Evaluation of Topic Modelling Techniques for Twitter,* Toronto: University of Toronto.

Juckett, D., 2012. A method for determining the number of documents needed for a gold standard corpus. *Journal of Biomedical Informatics,* pp. 460-470.

Kennedy, G., 1998. *An introduction to corpus linguistics.* London, England: Logman.

Kilgarriff, A., 2003. *The Sketch Engine,* Brighton: University of Brighton.

Kilgarriff, A., 2013. *The Sketch Engine: ten years on,* Brighton: Lexical Computing.

Kors, J. A., 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association Advance Access,* pp. 1-11.

Lluis, P., s.f. *Analizadores Multilingues en FreeLing,* s.l.: Universitat Politecnica de Catalunya.

McCallum, A. K., 2002. *MALLET: Machine Learning for Language Toolkit.* [En línea]
Available at: http:// mallet.cs.umass.edu

McEnery, T. & Wilson, A., 2001. *Corpus Linguistics: an introduction.* 2nd. ed. Edinburgh: Edinburgh University Press.

online, S. L. & BIREME, s.f. *SciELO.* [En línea]
Available at: http://www.scielo.org
[Último acceso: 10 Junio 2017].

Oronoz, M., 2015. On the creation of a clinical gold standard corpus in spanish: Mining adverse drug reactions. *Journal of Biomedical Informatics,* Volumen 56, pp. 318-332.

Padro, L., 2010. *Analizadores multilingues en Freeling,* Barcelona: Universitat Politecnica de Catalunya.

Pérez Hernández, C., 2002. Explotación de los córpora textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento. *Estudios de linguistica del Español,* p. 2.3.

Perez Hernandez, C., 2010. *Linguistica computacional y linguistica de corpus. Potencialidades para la investigacion textual,* Malaga: Universidad de Malaga.

Raschka, S., 2015. Giving computers the ability to learn from data. En: *Python Machine Learning.* Birmingham: Packt Publishing, pp. 1-16.

Rea Rizzo, C., 2010. Getting on with corpus compilation: from theory to practice. *ESP World,* 9(1), pp. 1-23.

Rojo, G., 2008. *Linguistica de corpus y linguistica del espanol.* Motevideo, XV Congreso de la ALFAL.

Steyvers, m. & Griffiths, T., s.f. *Probabilistic topic models,* California: University of California.

Torruela, J. L. J., 1999. Diseño de corpus textuales y orales. En: *Filología e informática. Nuevas tecnologías en los estudios filológicos..* Barcelona: Milenio, pp. 45-77.

Underwood, T., 2012. *Topic modeling made just simple enough.* [En línea]
Available at: http://tedunderwood.com

Universidad Complutense de Madrid.Facultad de Filologia, 2011. *Taller sobre herramientas de analsis textual: la herramienta Sketch Engine,* Madrid: s.n.

Villayandre Llamazares, M., 2008. Lingusitica con corpus (I). *E.H. Filologia,* Volumen 30, pp. 329-349.

Wiebe, J. & Bruce, R., 1998. *Development and Use of a Gold-Standard Data Set for subjectivity Classification,* Las Cruces: New Mexico State University.

Wissler, L., 2014. *The Gold standard in corpus annotation.* [En línea]
Available at: http://ieee-student-conference.de

Xiao, R., 2008. Corpus creation. En: *Handbook of natural language processing.* Boca Raton, FL: CRC Press, pp. 147-165.