PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
FACULTAD DE INGENIERÍA
ESCUELA DE INGENIERÍA INFORMÁTICA

APLICACIÓN Y EVALUACIÓN LDA PARA ASIGNACIÓN DE TÓPICOS EN DATOS DE TWITTER

BRUNO CHANDÍA SEPÚLVEDA

Profesor Guía: **Rodrigo Alfaro Arancibia**

Carrera: **Ingeniería Civil en Informática.**

Marzo, 2016

Índice

Resumen

El presente informe presenta el análisis del algoritmo para clasificación de tópicos latentes -Latent Dirichlet Allocation- aplicado al problema de modelado de tópicos del microblog Twitter. Planteándose así los problemas y las nuevas oportunidades que presenta el Latent Dirichlet Allocation para modelar bases de datos de Twitter y analizando los resultados obtenidos por este algoritmo frente al desafío del formato de noticias rápidas y del problema del lenguaje natural de dicho microblog. Así, en este documento se presentan el estado del arte de las técnicas para modelo de tópicos, el marco teórico y marco conceptual relevante de los modelos de tópicos, los fundamentos del modelo de tópico usado, la descripción de la implementación del algoritmo y el análisis de los resultados obtenidos.

Palabras Clave: Twitter, estadística y probabilidad, minería de textos, Latent Dirichlet Allocation, Modelo de Bayes, Modelo de Tópicos, Minería de Datos, modelado de documentos.

Lista de Figuras

Lista de Tablas

1 Introducción

El concepto de modelo de tópicos o temas (más conocido como *topic model* en inglés) es una de las técnicas probabilísticas e informáticas que ha ido apareciendo con más fuerza en las últimas décadas, anunciando las diversas aplicaciones útiles que puede tener y los diversos beneficios en el tratamiento de textos en la web para inferir, concluir, analizar y comparar datos e información de distinta índole. Es en este mismo sentido en el que el presente estudio se embarca en el análisis del algoritmo Latent Dirichlet Allocation que entregan una clasificación de temas para documentos elaborados a partir de publicaciones de usuarios de Twitter.

La masiva popularidad que Twitter tuvo en sus inicios ha decaído. Sin embargo en la actualidad, Twitter sigue siendo una de las redes sociales más utilizadas y por lo tanto una de las redes sociales que presenta un campo fértil para trabajar con clasificación de textos y modelo de tópicos. Se trata de una red social en la que se escriben mensajes de no más de 140 caracteres los cuáles son mostrados en la página-Twitter de cada seguidor del escritor. La influencia que ha tenido Twitter en los medios de comunicación ha sido realmente importante, es más muchos en la web proponen que, ya hace tiempo, Twitter ha revolucionado la forma de hacer periodismo ya que se presenta como un "periódico de noticias cortas" que muchas veces anuncia hechos mucho antes que los medios tradicionales de prensa los publiquen.

Sin embargo, a pesar de la riqueza de Twitter como fuente de información de tendencias, opiniones, política y cultura entre otros, todavía queda mucho campo por investigar para crear herramientas informáticas que faciliten la extracción de información desde Twitter, debido principalmente al problema que presenta el trabajo con lenguaje natural y muchas veces muy informal. En este documento se presentan una de las técnicas actuales más importantes dentro del campo de modelos de tópicos: Latent Dirichlet Allocation (LDA, "*asignación latente dirichlet"*).

Primero se presenta la descripción del problema y los objetivos generales y específicos de este trabajo. Se explica la relevancia del estudio y trabajo que se está realizando (como Twitter, por ejemplo, ha entrado a competir con los medios periodísticos tradicionales) y las dificultades que aún se tienen en el trabajo con la minería de datos para esta plataforma. Y, por último se muestran estadísticas que confirman y complementan las observaciones dadas.

Como segunda parte, de acuerdo a la metodología del trabajo especificada se define el plan de trabajo para este estudio en el que se distinguen las fases de comprensión y descripción del problema, descripción y análisis del algoritmo LDA, la metodología que se utilizará para implementar este algoritmo y la metodología utilizada para evaluar dicho algoritmo.

Luego se describe el estado del arte de los modelos de tópicos enfocándose en los conceptos de Twitter y de los modelos de tópicos más relevantes para poder entender las técnicas utilizadas y los distintos beneficios o desventajas que pueden presentar estos en comparación con el LDA.

El marco teórico define los términos que serán la base para poder entender el posterior análisis de los algoritmos seleccionados y las suposiciones relevantes previas al análisis de documentos con datos de Twitter a través de LDA para luego poder entender de mejor manera la evaluación realizada a los resultados de dicho algoritmo. Así también se presentará el modelo de análisis en el que se explican los distintos escenarios en los que se evaluará los resultados y las características de los datos utilizados para la evaluación de este.

Finalmente, luego de aplicar dos pequeños software en código Java y Python para el pre-procesamiento más un software en Python para el procesamiento con LDA, se realiza la evaluación de los resultados obtenidos sobre la base de datos Twitter.

1.1 Descripción del Problema

Hace ya varios años Twitter se ha convertido en un medio que entrega noticias rápidas, "transmitiendo" información de opinión pública, de entretención e incluso de marketing. Es más, no es difícil encontrar en la web, artículos que hablan sobre el periodismo alternativo que ofrece Twitter y las distintas posibilidades que esto significa para los medios de comunicación. Esto debido a que en los últimos años Twitter se ha hecho una cada vez más útil plataforma popular para que los usuarios de internet se comuniquen e informen entre sí.

1.1.2 El fenómeno Twitter

Varios estudios han examinado a Twitter desde distintas perspectivas, incluyendo las características topológicas de Twitter como red social o medio de prensa [1], los tweets como sensores sociales de los eventos en tiempo real [2] o el pronóstico de taquilla de nuevas películas [3] entre otros. Aún así, a pesar de las distintas aplicaciones que puedan existir para el uso de esta fuente de información, el manejo de las grandes cantidades de datos en Twitter aún es limitado debido a las particularidades de este mismo: mensajes cortos existentes en grandes cantidades, con lenguaje informal.

En una publicación sobre modelos de tópicos, Wayne Xin Z. et al, han trabajo comparando el contenido de Twitter con los contenidos del periódico New Yok Times usando un modelo de tópicos sin supervisión en el que se usa la técnica LDA y una modificación de esta misma para encontrar tópicos en el New York Time y en Twitter respectivamente. Luego de realizar clasificación de tópicos mediante LDA se utilizan diversas técnicas para evaluar y comparar los tópicos de Twitter y de New York Times teniendo en cuenta una taxonomía en la que se clasifican tweets según tipos de tópicos y categorías de tópicos. Finalmente el modelo planteado estudia la relación entre los tweets y retweets realizados junto con las proporciones de tipo de tópicos y categorías de tópicos [4].

1.1.3 Algoritmo a trabajar: LDA

El algoritmo con el que se trabajará para analizar datos de Twitter emplea el modelo de bolsa de palabras (bag of words). La propuesta hecha por Davi Blei et al en [5] conduce a distribuciones mezcladas (*mixture models*) para unidades estructurales más grandes asumiendo así que cada documento es una mezcla de un pequeño número de categorías y la aparición de cada palabra en un documento se debe a una de las categorías a las que el documento pertenece.

Una de los modelos antecesores del LDA es el algoritmo pLSA -probabilistic Latent Semantic Allocation- propuesto por Hoffman [6], el cual se basa en una mezcla de descomposición, derivada de un modelo de clases latentes, pero con unos fundamentos estadísticos sólidos. Así el LDA toma estos conceptos desde el pLSA para luego transformarse en un modelo de tópicos que permite que conjuntos de observaciones puedan ser explicadas por grupos de observaciones latentes o subyacentes que describen por qué algunas partes de los datos son similares, pero a la vez mejora el modelo pLSA, ya que Blei considera que el trabajo de Hoffman es incompleto porque no da un modelo probabilístico a nivel de documento y esto genera una serie de problemas como la falta de claridad a la hora de asignar una probabilidad a los documentos.

Se plantea el análisis de documentos formados por datos de Twitter para tres distintos escenarios (ver secciones 1.2 y 6.1) a partir de los beneficios y limitaciones de este algoritmo y del hecho de que muchos de los trabajos realizados con estos algoritmos se basan en análisis de documentos que no poseen la estructura de Twitter (donde no se tiene las estructura tweets formando una sola oración que no necesariamente se relaciona con el siguiente tweet realizado por un usuario).

1.1.4 El presente desafío LDA: Análisis de datos de Twitter

El uso del algoritmo nombrado se ha ya realizado en varias investigaciones. En estas se han presentado problemáticas similares a las de este trabajo:

- De manera subjetiva, el LDA no se comporta bien con documentos muy pequeños o muy grandes. O dicho de otra forma, este algoritmo no es objetivamente certero.
- ¿Cómo crear un documento con tweets para una correcta recuperación de información?
- Como elegir la creación del “documento” como el conjunto de muchos tweets (ver 6.1)

De esta manera se pueden ver los desafíos que presenta el actual trabajo de investigación, en el que existe una plataforma social del tipo microblog con muchas oportunidades para extraer información con distintos motivos, pero que a la vez presenta grandes desafíos por la estructura que posee esta misma.

1.2 Metodología del Trabajo

La realización de este trabajo consta de las siguientes fases: comprensión y descripción del problema, marco teórico del problema propuesto, descripción y aplicación del algoritmo y la evaluación de los resultados arrojados por el algoritmo.

En la fase de comprensión y descripción del problema se describe la investigación llevada a cabo, dando énfasis a las características del microblog Twitter y de los elementos especiales a tener en cuenta al momento de plantear un modelo de clasificación, análisis y comparación de los resultados de los algoritmos.

En la fase de descripción e implementación de los algoritmos se comenta y estudia los escenarios para comparación de resultados obtenidos por el algoritmo donde se explica la técnica LDA utilizada y los resultados obtenidos.

Se analizará cada *corpus* de acuerdo a estos tres escenarios. Seguida a esta etapa se implementarán las técnicas y algoritmos necesarios para así finalmente evaluar los resultados obtenidos.

1.2.1 Primer escenario: corpus por mes

Se define de la siguiente manera:

- Corpus: formado por todos los tweets que pertenecen a un mismo mes.
- Documentos: el *corpus* se divide en diez documentos, donde cada documento estará formado por todos los tweets que pertenecen a las fechas resultantes de dividir el mes correspondiente en diez fechas equidistante

1.2.2 Segundo escenario: corpus por bimestre

Se define de la siguiente manera:

- Corpus: formado por todos los tweets que pertenecen a un mismo bimestre.
- Documentos: el *corpus* se divide en diez documentos, donde cada documento estará formado por todos los tweets que pertenecen a las fechas resultantes de dividir el bimestre correspondiente en diez fechas equidistantes

1.2.3 Tercer escenario: corpus por totalidad de datos

Se define de la siguiente manera:

- Corpus: formado por todos los tweets de la base de datos.
- Documentos: el *corpus* se divide en diez documentos, donde cada documento estará formado por todos los tweets que pertenecen a las fechas resultantes de

dividir el período que abarca a todos los tweets de la base de datos en diez fechas equidistantes

1.3 Factibilidad de la Investigación e Implementación

Dado que existen ya bastantes herramientas de licencias no pagadas para ejecutar algoritmos de modelos de tópicos, no está contemplado realizar un informe de costos económicos en este trabajo.

A pesar de que existe una gran documentación de estas técnicas y herramientas, se debe tener en cuenta la curva de aprendizaje de todos los conceptos que involucran a cada técnica cómo por ejemplo las palabras vacías en el pre-procesamiento de datos, la matriz de bolsa de trabajo utilizada en el algoritmo, ciertos fundamentos estadísticos bayesianos y los conceptos que explican el porqué el LDA funciona intuitivamente.

Por último hay que agregar que los fundamentos de todas las áreas de estudio que se manejarán en este proyecto forman parte de la malla curricular del autor de este informe y estudio, por lo que el tiempo de investigación que se presentará en la sección 3 (Plan de Trabajo) está dentro de los tiempos suficientes para poder realizar dicha investigación.

2 Definición de Objetivos

2.1 Objetivo General

Evaluar el comportamiento y los resultados del algoritmo LDA para clasificación de tópicos en conjunto de datos de la plataforma microblog Twitter.

2.2 Objetivos Específicos

Investigar y comprender el estado del arte de las técnicas utilizadas para la clasificación de tópicos.

Establecer los fundamentos teóricos y conceptuales de las técnicas utilizadas para la clasificación de tópicos con LDA.

Establecer los escenarios en los que se implementará y analizará el algoritmo de clasificación de tópicos LDA.

Realizar una evaluación del comportamiento del algoritmo LDA para los distintos escenarios planteados.

3 Plan de Trabajo

Como se definió en la metodología, las etapas que comprenden el plan de trabajo están dadas por las siguientes fases: comprensión y descripción del problema, conocimiento del estado del arte de la tecnología relacionada al algoritmo LDA, comprensión de los conceptos teóricos del algoritmo, implementación de los algoritmos utilizados y evaluación de los resultados obtenidos.

Tareas	Septiembre	Octubre	Noviembre	Diciembre
Comprensión y descripción del problema 1. Estudio red social Twitter. Características topológicas y conceptos relevantes 2. Estudio del arte sobre publicaciones y trabajos relacionados: clasificación de textos, análisis de contenidos, minería de textos, entre otros.				
Comprensión Fundamentos Teóricos 3. Estudio LDA 4. Propuesta de casos y escenarios de prueba				
Comprensión fundamentos teóricos y pre-procesamiento de base de datos a trabajar 5. Estudio de técnicas similares 6. Estudio de los fundamentos teóricos del algoritmo 7. Pre-procesamiento de datos a implementar				
Implementación y evaluación del algoritmo según los tres escenarios planteados				

Tabla 1. Plan de trabajo

4 Estado del Arte

Si bien existe un gran número de investigaciones que abarcan modelos de tópicos, técnicas de Recuperación de Información entre otras más que abordan problemas prácticos como por ejemplo a través de minería de datos, el problema de analizar, procesar y clasificar grandes cantidades de texto en un corpus de datos provenientes de Twitter aún presenta investigaciones tempranas y con varias problemáticas a solucionar. En cuanto a corpus adquiridos desde Twitter referentes a noticias y todo tipo de discusión que tenga que ver con eventos, sucesos y hechos que puedan catalogarse como noticias cortas presenta varias oportunidades de análisis y minería de datos. Un corpus creado desde tweets de un número determinado de usuarios de Twitter presenta severas particularidades que deben ser manejadas con cuidado al momento de realizar técnicas para el manejo de datos y minería de textos.

4.1 Twitter: Datos sobre su relevancia

Según un artículo de la web Schools.com para el 2012 cerca de la mitad de las personas del continente norteamericano obtienen información sobre las noticias locales desde un dispositivo móvil, el 46% de ellos se informan de las noticias a través de internet al menos 3 veces a la semana. De hecho, las fuentes de noticias en internet han superado los ingresos de publicidad que reciben las fuentes de noticias en diarios impresos.

Algunas preguntas relevantes en torno al fenómeno Twitter son: ¿Hay cada vez más gente que está utilizando las redes sociales como sus principales fuentes de información de noticias rápidas? ¿Puede la gente confiar en que las noticias que las redes sociales entregan son precisas y objetivas? En la actualidad los medios de comunicación sociales reflejan una fuerte presencia como fuentes de noticias online.

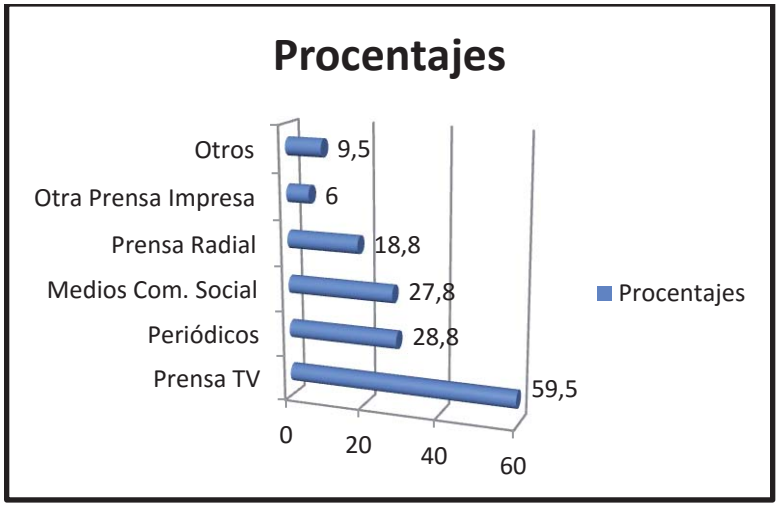

Figura 1. Los medios de comunicación online

Desde 2009, el tráfico a sitios de noticias de plataformas de medios sociales se ha incrementado un 57% [7]. Además, 9% de los adultos que reciben las noticias en un dispositivo móvil digital usa Facebook o Twitter para informarse.

Si se revisan los sucesos mundiales importantes de los últimos años se pueden encontrar varios que han sido infomados primero por medios de comunicación social online:

- El levantamiento en Egipto
- El anuncio de la Boda Real
- Campaña presidencial Newt Gingrich en Estados Unidos
- Manifestantes asesinados en Bahrein
- Hillary Clinton decide no quedarse con un segundo puesto
- Muere Whitney Houston
- Osama Bin Laden: allanamiento y muerte

Como ejemplo, un dato interesante es que la primera persona que informó a través de Twitter sobre el allanamiento y muerte de Osama fue su vecino, el cual sin saberlo "twitteó" sobre uno de los hecho históricos más importantes de esta década.

En la Tabla pueden observarse algunos puntos a favor y en contra respecto al hecho de reportar noticias a través de medios de comunicación online.

A favor	En contra
Inmediatez	Poco espacio para escribir una noticia
Todos pueden ser reporteros	Inexactitud
Las noticias se expanden rápidamente	La integridad periodística se puede ver comprometida
Las cámaras de los dispositivos móviles van donde no pueden llegar las cámaras de TV	Reporteros inexpertos ponen sus vidas en riesgo

Tabla 2. Tabla de pros y contras de reportear vía Twitter

Otro dato importante es que el 49,1% de las noticias rápidas que personas han escuchado noticias a través de medios sociales de comunicación online provenían de una fuente falsa [7].

4.2 Investigaciones afines

Actualmente existen varios estudios relacionados con la aplicación de modelo de tópicos a corpus con grande cantidad de datos de Twitter. A continuación se nombran algunos.

4.2.1 Modelo de Tópico en Textos de Área de la Salud

Otro estudio realizados el 2011 presenta un modelo de tópicos que usa el modelo probabilístico LDA para descubrir aspectos de las enfermedades como un nuevo tipo de modelo de tópicos para Twitter que asocia tópicos que tienen que ver con síntomas, tratamientos y palabras generales relacionadas con las enfermedades. En este se forma una colección de 1,6 millones de tweets en numerosos temas de discusiones en Twitter relacionadas con la salud. En comparación con los modelos de tópicos conocidos a la fecha el modelo propuesto aísla los tópicos referidos a enfermedades más coherentes y más conocidos, como la gripe, infecciones y la obesidad entre otros. En los análisis y comparaciones finales realizados en dicha investigación los resultados obtenidos por el modelo propuesto coincidieron con los resultados obtenidos al realizar una búsqueda a través de Google Flu Trends [8].

Los investigadores de salud pública dedican recursos considerables para realizar vigilancia de la población en cuanto a la salud, que requiere encuentros clínicos con profesionales de la salud. El modelo presentado por Michael J. Paul en el 2011, entrega una alternativa de bajo costo para el seguimiento de las tendencias de salud pública a través de Twitter.

Del mismo modo, los tweets mencionan temas de salud relacionados con temas como "Tengo fiebre de 102,5 - tengo gripe - tengo dolor en los ojos - me duele la garganta de tanto tomar Tylenol". Estos Tweets indican que los usuarios tienen una enfermedad (gripe), los síntomas asociados (fiebre, etc) y los tratamientos (Tylenol).

En la Tabla 3 se puede ver alguno de los resultados obtenidos por este modelo.

Enfermedades	**24.4%**	**8.5%**
Salud	11.3%	16.6%
No Relacionado	52.6%	58.1%
No en Inglés	10.5%	16.7%
Ambiguo	1.1%	0.1%

Tabla 3. Porcentaje de tópicos encontrados en documentos de salud.

4.2.2 Modelo de Tópicos para Monitoreo de Sismos

Varios estudios más han considerado el uso de Twitter para el seguimiento de los diversas tendencias, incluyendo el seguimiento de monitoreo de sismos.

Cuando un terremoto ocurre, la gente hace muchos comentarios en Twitter (tweets) relacionados al terremoto, lo que permite la detección de terremotos hechos de manera rápida,

simplemente mediante la observación de dichos tweets. En la publicación "Earthquake shakes Twitter users: real-time event detection by social sensors" (2010) se describe una investigación de la interacción en tiempo real de eventos tales como terremotos en Twitter, y se propone un algoritmo para monitorear los tweets y detectar un tópico relativo a dicho evento. Para detectar un evento de destino, se debe idear un clasificador de los tweets sobre la base de características de palabras clave en un tweet, el número de palabras, y su contexto. En este estudio se considera que cada usuario de Twitter puede tratarse como un sensor para así aplicar búsqueda mediante técnicas de filtrado: *filtrado Kalman y el filtrado de partículas*, que son dos técnicas ampliamente utilizadas para la estimación de la ubicación en todas partes.

Para este modelo se construyó un sistema de notificación de terremoto en Japón ya que con los numerosos terremotos y el gran número de usuarios de Twitter en todo el país, se puede detectar terremotos a través de de monitoreo de tweets con alta probabilidad (96% de terremotos de Agencia Meteorológica de Japón JMA – se detectan eventos con tres o más grados en la escala de intensidad sísmica). El sistema propuesto en esta investigación detecta terremotos con prontitud y envía e-mails a los usuarios registrados y las notificaciones se entregan mucho más rápido que los anuncios que se emiten por la JMA.

4.2.3 Modelo de Tópicos para Seguimiento de Eventos

Otros estudios más han considerado el uso de Twitter para el seguimiento de las diversas tendencias, incluyendo el seguimiento de las noticias, opiniones y en general de eventos relevantes. Ciertos estudios como el llamado "Joint Topic Modeling for Aligning Events and their Twitter Feedback" demuestran como durante los eventos norteamericanos, tales como el Superbowl, los debates presidenciales y las primaria, entre otros, Twitter se ha convertido en la plataforma de facto para las multitudes para compartir perspectivas y comentarios acerca de ellos.

En este estudio se tuvieron que resolver dos problemas fundamentales de la investigación que han estado recibiendo cada vez más atención en los últimos años. Uno de ellos es para extraer los temas tratados por el evento y los tweets de la otra es para segmentar el evento. En este estudio, estos problemas fueron tratados por separado y estudiados en el aislamiento. Para superar dichos problemas se supuso que estos problemas eran de hechos, interdependientes y que debían ser abordados conjuntamente.

Para lo anterior se desarrolló un modelo bayesiano que realiza el modelado de los temas y el caso de la segmentación en un marco unificado. Finalmente este estudio evaluó el modelo propuesto, tanto cuantitativa como cualitativamente en dos conjuntos de datos de tweets a gran escala asociadas con dos eventos desde diferentes ámbitos para mostrar que mejora significativa se observó respecto a los modelos de referencia.

4.3 Modelo de Tópicos: ¿Cómo se llegó al modelo LDA?

Un modelo científico se define como una representación abstracta, conceptual, gráfica, física, matemática, de fenómenos, sistemas o procesos a fin de analizar, describir o explicar esos fenómenos o procesos. Un modelo permite determinar un resultado final a partir de unos

datos de entrada y es importante tener en cuenta que se considera que la creación de un modelo es una parte esencial de toda actividad científica.

Los modelos de tópicos son modelos matemáticos estocásticos debido a la existencia de incertidumbre al momento de formular respuestas o salidas de dichos modelos, es decir, esto implica que los resultados o salidas son probabilidades.

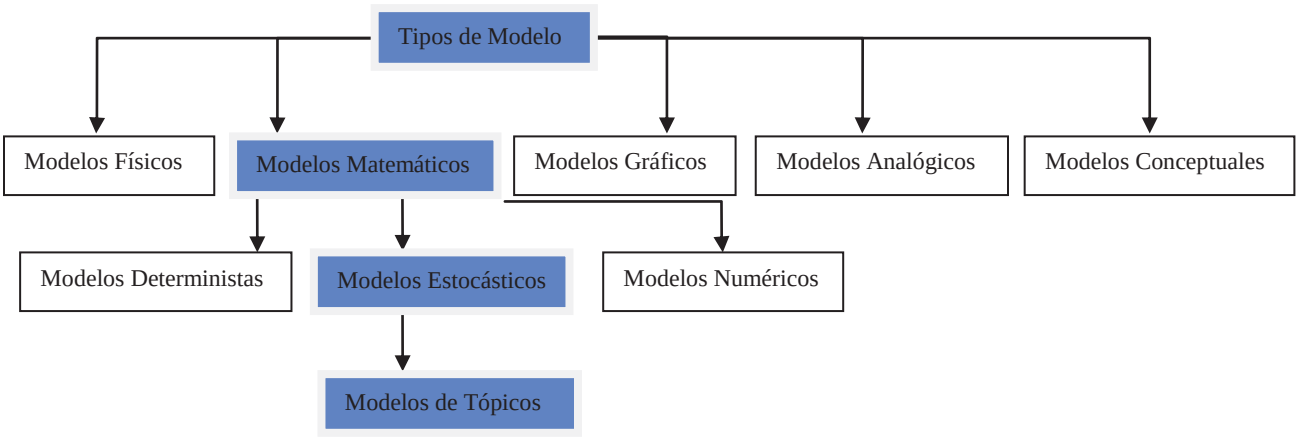

Figura 2. Mapa conceptual tipos de modelos

Davi M. Blai define el concepto y la función de los modelos de tópicos de la siguiente manera:

"Los modelos de tópicos son un conjunto de algoritmos que proporcionan una solución estadística al problema de la gestión de grandes archivos de documentos. Con los recientes avances científicos en apoyo de los componentes de la máquina de aprendizaje flexibles sin supervisión para el modelado, algoritmos escalables para la inferencia posterior y con el mayor acceso a modelos de conjuntos de datos masivos (dataset), los modelos de tópicos prometen ser un componente importante para la síntesis y la comprensión de nuestros crecientes archivos digitalizados de la información." [9]

La idea de los modelos de tópicos y los conceptos involucrados en estos irán quedando más claros con la explicación de los siguientes puntos.

4.3.1 Modelo Antecesor LDA I: Latent Semantic Analysis

El Análisis Semántico Latente (LSA, por sus siglas en inglés) es un método de indexación y recuperación de datos que utiliza una técnica matemática llamada Descomposición de Valor Singular (SVD) para identificar patrones en las relaciones entre los términos y conceptos contenidos en una colección estructurada de texto. El LSA se basa en el principio de que las palabras que se utilizan en los mismos contextos tienden a tener significados similares. Una característica clave de LSA es su capacidad para extraer el contenido conceptual de un cuerpo de texto mediante el establecimiento de asociaciones entre los términos que aparecen en contextos similares [10].

La clave de la innovación de LSA fue el uso de SVD para descomponer la matriz original de términos-contextos. Si se tiene que las filas son, una para cada término y que las columnas son, una para cada documento, la aplicación de la SVD nos daría como resultado las matrices graficadas en la siguiente imagen:

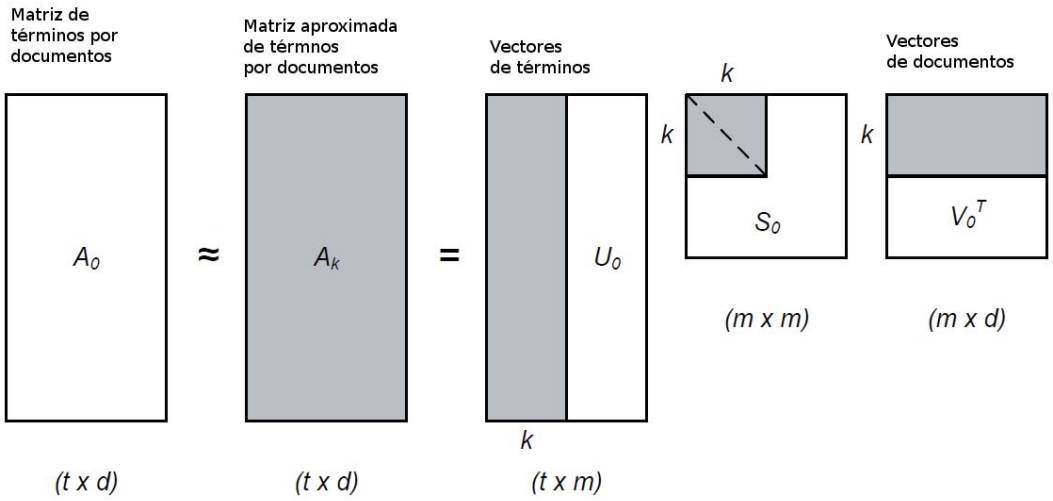

Figura 3. SVD de la matriz de términos-contextos LSA.

El estudio de esta técnica no es el propósito de este trabajo pero es pertinente presentarla debido a que es uno de los modelos en el cual, se basa LDA para trabajar en clasificación de temas latentes.

4.3.2 Modelo Antecesor a LDA II: Probabilistic Latent Semantic Analysis

En comparación con LSA, el PLSA se basa en una mezcla de descomposición, derivada de un modelo de clases latentes, pero con unos fundamentos estadísticos sólidos. Principalmente está basado en un modelo estadístico llamado modelado de aspecto (aspect model), que es un modelo de latencia variable para datos concurrentes, que asocia variables de clases latentes $Z_k, k \in \{1,2,\dots K\}$, con cada *observación*, donde K es el número de clases latentes. El número de clases latentes, es similar al número de dimensiones para el LSA, un parámetro que tiene que ser seleccionado previamente. La *observación* consiste en la aparición de una palabra $w_j, j \in \{1,2,\dots,N\}$, en un particular documento $d_k, k \in \{1,2,\dots,M\}$ donde N es el número de palabras y M el número de documentos y K el número de clases latentes a ser halladas.

Estas clases latentes pueden ser entendidas como los temas, categorías o tópicos que componen el texto. Las distribuciones de probabilidad que asocian las variables latentes con las palabras y documentos describen cuánto están relacionadas estas con los temas. El modelo generativo para la *observación* se muestra gráficamente en la **Fig. 4** y se define a continuación: 1) Obtener un documento d_i en el que la aparición de una palabra será observada con probabilidad $P(d_i)$; 2) cuando el documento d_i es conocido, seleccionar el tema de la palabra con probabilidad $P(z_k|d_i)$. Esta distribución de probabilidad es también

una medida del grado en que el documento es relevante para cada tema; y 3) cuando el tema es conocido, seleccionar la palabra w_j; cuya aparición es observada con probabilidad $P(w_j|d_i)$.

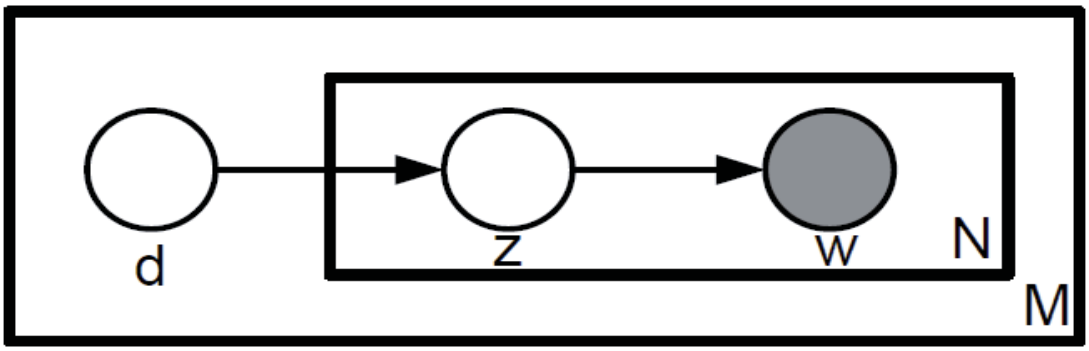

Figura 4. Modelo de grafo para PLSA

En vez de utilizar el modelo convencional del algoritmo de Maximización de la Expectativa (*Expectation Maximization*), Hoffman propuso un algoritmo de EM Templado para calcular las distribuciones de probabilidad a partir del los datos de entrenamiento. El algoritmo alterna entre estos dos pasos: un paso de *expectativa* (*E*) donde las probabilidades posteriores son calculadas en base a las actuales estimaciones de los parámetros; y un paso de *maximización* (*M*) en el que los parámetros son actualizados en función de la minimización de los criterios y la dependencia de las probabilidades posteriores calculadas en el paso *E*.

4.3.3 Latent Dirichlet Allocation

Esta es una de las técnicas que viene adquiriendo fuerza en el campo de los modelos de tópicos. Esta técnica está compuesta por conceptos de Modelos Bayesianos y se basa en el proceso probabilístico genérico que permite inferir tópicos de un documento en base a una distribución a posteriori obtenida por el LDA.

Los modelos bayesianos pueden servir para indicar cómo debemos modificar nuestras probabilidades subjetivas cuando recibimos información adicional de un experimento. La estadística bayesiana está demostrando su utilidad en ciertas estimaciones basadas en el conocimiento subjetivo a priori y el hecho de permitir revisar esas estimaciones en función de la evidencia empírica. Esto último es lo que está abriendo nuevas formas de hacer conocimiento aplicado a la teoría de la información. Una aplicación de esto son los clasificadores bayesianos que son frecuentemente usados en implementaciones de filtros de correo basura o spam, que se adaptan con el uso.

Sea $\{A_1, A_3, \dots, A_i, \dots, A_n\}$ un conjunto de sucesos mutuamente excluyentes, tales que la probabilidad de cada uno de ellos es distinta de 0. Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i) * P(A_i)}{P(B)} \qquad (4.1)$$

Donde:

- $P(A_i)$ son las probabilidades a priori.
- $P(B|A_i)$ es la probabilidad de B en la hipótesis A_i (conocimiento adicional sobre las probabilidades a priori de A_i)
- $P(B|A_i)$ son las probabilidades a posteriori.

La técnica LDA presenta un modelo para clasificar corpus mediante un proceso genérico probabilístico basado en el Teorema de Bayes. Es decir, esta técnica parte de una suposición subjetiva a priori $P(A_i)$, para luego agregar la probabilidad de un evento B conocido dado la probabilidad de esta suposición a priori $P(B|A_i)$.

Una de las cosas que hace Blai *et al* en su propuesta de LDA es enfatizar que en el anterior modelo PLSA, Hoffman comete el error de no tomar en cuenta el peso y la importancia de los documentos. Esto, ya que no le asigna una distribución de probabilidad a los documentos. El PLSA por una parte descubrirá estructuras latentes de tópicos en el corpus, pero por otra al obviar que un documento representa al menos un tópico del conjunto de datos pierde de vista el peso que tienen los documentos de un corpus para asignar el tópico al que pertenece una palabra del documento. Así, al agregar importancia a un documento el LDA tiene resultados más equilibrados al generar una tensión entre el tópico al que pertenece un término del vocabulario versus el tópico al que pertenecen las palabras de cada documento.

5 Antecedentes

En este capítulo se presentarán y definirán los conceptos más relevantes que se utilizarán en este trabajo y específicamente los conceptos más importantes que describen las características principales de Twitter y del modelo y algoritmo LDA.

5.1 Características Principales de Twitter

Twitter es la principal plataforma mundial que usa el sistema de Microblogging. Los Microblogs o NanoBlogs son servicios que permiten a los usuarios enviar y publicar mensajes breves (generalmente de no más de 140 caracteres). Twitter es sin duda el Microblog más famoso y popular existente sin embargo, también se usan otros tipos de Microblogging entre los usuarios como Tumblr que permite compartir textos cortos, videos e imágenes; o Identi.ca el cual permite compartir textos de no más de 140 caracteres, igual que Twitter y que permite relacionarse con las cuentas de Twitter. Así, una de las cosas que se han impuesto más fuertemente entre la jerga de los usuarios de internet son los términos que se usan habitualmente dentro de Twitter y que se explicarán de manera breve, a continuación.

5.1.1 Los Seguidores o Followers

Son todos los usuarios que han decidido ser parte de la red de un usuario y además son quienes son informados por cada vez que dicho usuario escriba un nuevo “tweet” (esto es un mensaje).

5.1.2 Los Hashtags

Son una forma muy particular de asignar un tema al tópico y se representan por el signo # antes de la palabra que describe el tema. Por ejemplo #terremoto, #copaAmerica, #ToleranciaCero.

5.1.3 Retweet

Se hace un *retweet* sobre un mensaje al compartir el mensaje de cierto usuario. El mensaje que se ha *retweeteado* podrá ser leído por todos los seguidores del usuario que lo ha *retweeteado*.

5.1.4 Menciones

En un mensaje se hace un mención a un usuario cuando se escribe su nombre de usuario en el mensaje con el propósito de mencionar a este usuario. Un ejemplo de esto sería: “@diputadoLarrain es mi político favorito”.

5.2 Clasificación de Textos

La capacidad de poder encontrar información pertinente y correcta a partir de técnicas de clasificación de textos es utilizada en distintas áreas de las ciencias de la computación. Una de las formas de clasificación más usadas es verificar la coincidencia de patrones dentro de un documento o incluso dentro de un sitio en internet.

El conocimiento es hoy la principal fuente de competitividad y ventaja. El éxito o fracaso de una empresa, por ejemplo, puede depender en la capacidad de encontrar información relevante en el momento adecuado. La explosión del uso de la web 2.0 hace ya bastantes años y el uso creciente de tecnologías de internet como un canal básico de comunicación multiplica las fuentes de información y disponibilidad de estas mismas. Sin embargo mientras más aumentan las fuentes de información, también aumentan a la vez la complejidad del orden y estructura de esta información y debido a que el valor principal no está en los datos en sí sino en la capacidad de gestionar adecuadamente algún procedimiento que permita obtener el conocimiento fundamental para los objetivos de una organización, la clasificación de textos es una de las áreas relevantes en los últimos avances de la informática.

5.2.1 Definición de Clasificación de Textos

Según [11] la clasificación de texto es el acto de tomar un conjunto de documentos de texto etiquetados, luego realizar un aprendizaje de la correlación existente entre el contenido de los documentos y sus etiquetas correspondientes para, finalmente, realizar una predicción de las etiquetas de un conjunto de documentos de prueba sin etiquetar, intentando conseguir el mejor resultado posible.

La clasificación de texto se ha estudiado ampliamente y es una de las especialidades más antiguas en las ciencias de la computación para recuperación de información: Los enfoques estadísticos clásicos de técnicas de cooperación se basan en los modelos de máquinas de aprendizaje tales como los modelos generativos, máquinas de Naive Bayes o en modelos discriminantes, tales como máquinas de soporte vectorial.

Una definición, complementaria a la anterior, es la dada en [12], donde se explica que la clasificación de texto es la tarea de asignar un valor booleano a cada $< d_j, c_i > \epsilon D \times C$, donde D es el dominio de los documentos y $C = \{c_1, \dots, c_{|c|}\}$ es el conjunto de categorías predefinidas. Un valor T asignado a $< d_j, c_i >$ indica que el documento d_j pertenece a la categoría c_i y un valor F indica que d_j no pertenece a la categoría c_i.

Así, el conocimiento exógeno no está disponible y no se puede usar, por lo tanto, la clasificación debe llevarse a cabo en base a conocimiento endógeno del documento (no existe metadata que diga si el tipo de documento, la fecha de publicación, fuente de publicación, entre otros).

5.2.2 Preparación de los Datos: Palabras Vacías

Un texto posee muchas palabras no relevantes llamadas palabras vacías o *stop words*. Estas son palabras sin significado como los artículos, los pronombres o las preposiciones y necesitan ser procesadas de acuerdo a algún criterio para ser eliminadas de la bolsa de palabras antes o después del procesamiento de los datos. La definición del conjunto de palabras que se consideran como palabras vacías están dadas por decisión y criterios humanos que varía según el enfoque que se le da a cada modelo específico. Un criterio de elección sencillo puede ser eliminar todos los artículos, adverbios y preposiciones más algunos pares de palabras que se consideren irrelevantes como por ejemplo, "el que", "tener qué". Algunas organizaciones han dejado disponible conjuntos de palabras en la web que pueden ser descargadas y utilizadas para la eliminación de estas en un documento.

5.2.3 Text Corpus

Los usos tradicionales para este término han sido mayormente para investigaciones lingüísticas sobre textos impresos como artículos de periódicos o libros. Sin embargo, con el creciente uso de recursos de información en la web se ha incrementado también el uso de *corpus* para entrenamiento de datos en tareas de procesamiento de lenguaje natural (NLP por sus siglas en inglés, Natural Language Processing). Esta investigación usará esta misma técnica para realizar y utilizar el algoritmo LDA.

5.2.4 Bag of Words (BoW)

Comúnmente se utiliza una matriz BoW para representar a cada documento como un vector. Este vector está compuesto a su vez por los términos de un diccionario. Así, se puede definir el *corpus* D_i como un conjunto de documentos de la siguiente manera:

$$D_i = (d_1, d_2, \dots, d_i, \dots, d_m) \tag{5.1}$$

Cada documento d_i puede definirse con una matriz BoW la que a su vez se puede definir de diferentes maneras a conveniencia:

- Una matriz BoW binaria para indicar que una palabra del vocabulario aparece o no en el documento
- Una matriz BoW con las frecuencias de cada término
- Una matriz BoW con las frecuencias de cada término normalizadas usando la técnica tf-idf

Para este trabajo se usará la matriz BoW que representa las frecuencias de cada término del vocabulario en cada documento. Así se tiene:

$$d_i = (t_1, t_2, \dots, t_j, \dots, t_n) \tag{5.2}$$

Donde t_j representa la frecuencia de cada término del documento d_i.

5.3 Aplicaciones de Modelos de Tópicos

A medida que aparece más información en la web, se hace más difícil acceder a lo que estamos buscando. Por esto es que se necesitan mejores herramientas para poder organizar, buscar y entender estas grandes cantidades de información. Los modelos de tópicos proveen métodos para buscar, organizar y resumir largos archivos electrónicos [13]. Con los modelos de tópicos, básicamente se busca:

- Descubrir patrones ocultos de temas que dominan una colección de datos
- Hacer anotaciones (clasificación) de los documentos de acuerdo a estos temas
- Utilizar estas anotaciones para organizar, resumir y buscar textos

Un modelo de tópicos aceptable podría descubrir los tópicos en un *corpus.* Por ejemplo, en una revista científica inglesa se debería poder descubrir los temas más relevantes inmersos y ocultos en tal documento y mediante un modelo de tópicos sería posible encontrar las siguientes distribuciones de palabras relativas a un tema específico.

"Genetics"	**"Evolution"**	**"Disease"**	**"Computers"**
human	evolution	disease	Compute
genome	Evolutionary	host	Models
Dna	Species	bacteria	Information
genetic	organisms	diseases	Data
genes	life	resistance	Computers
Sequence	Origin	bacterial	System
gene	Biology	new	Network
molecular	Groups	Strains	Systems
sequencing	phylogenetic	control	Model
map	living	infectious	Parallel

Tabla 4. Tópicos hallados con técnicas y modelos de tópicos

Otro uso que se le podría dar a un modelo de tópicos es descubrir la evolución de los temas en una misma fuente de datos a través del tiempo. Véase el ejemplo de la evolución de los tópicos a través del tiempo en la figuras 5 y 6.

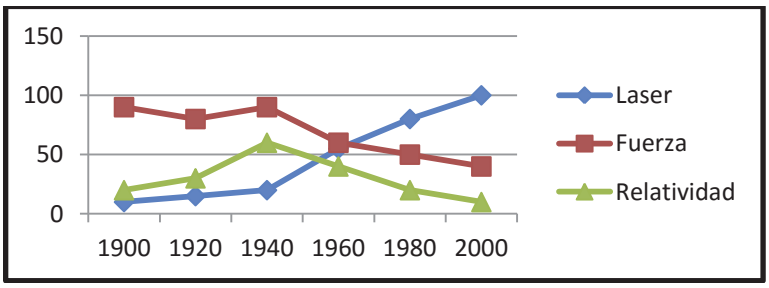

Figura 5. Evolución de los temas en la física teórica a través del tiempo

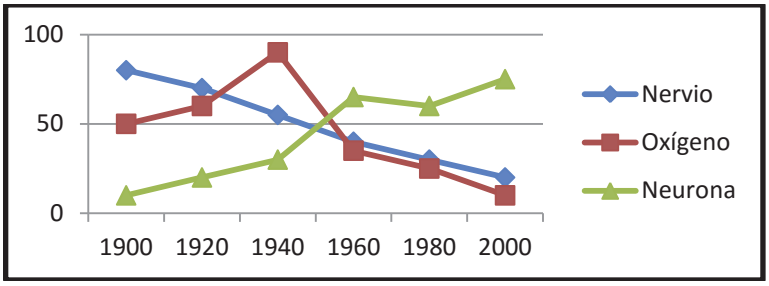

Figura 6. Evolución de los temas en la neurociencia a través del tiempo

Finalmente, como se ha comentado mientras tenemos textos más grandes se hace mucho más difícil la tarea de estudiar y navegar a través de estos textos. Por esto varios investigadores de máquinas de aprendizaje han desarrollado modelos de tópicos probabilísticos que tienen como finalidad clasificar grandes archivos con grandes cantidades de datos. Los algoritmos de modelos de tópicos son métodos estadísticos que analizan las palabras de textos originales para descubrir los temas que "corren" a través de ellos, la forma en que están conectados unos con otros y como van cambiando a través del tiempo. Así los modelos de tópicos nos permiten organizar y resumir archivos a una escala que sería imposible clasificar humanamente [14].

5.4 El Modelo de Tópicos LDA

LDA es un modelo generativo que permite que conjuntos de observaciones puedan ser explicados por grupos inadvertidos que describen por qué algunas partes de los datos son similares. Por ejemplo, si las observaciones son palabras en documentos, cada documento es una mezcla de categorías y la aparición de cada palabra en un documento se debe a una de las categorías a las que el documento pertenece. Entonces, la intuición detrás de la técnica LDA es que cada documento siempre exhibe múltiples temas en su cuerpo.

En la figura 7 se muestra un documento donde se pueden separar y distinguir las palabras que tienen que ver con análisis como "computer" y "prediction", las palabras que tienen que ver con biología evolutiva como "life" y "organism" y las palabras que hablan sobre genética como "sequenced" y "genes". Si alguien se tomara más tiempo y resaltara todas las palabras que tienen que ver con análisis, biología evolutiva y genética, se daría cuenta que todo el documento posee estos tres temas mezclados pero en diferentes proporciones.

La técnica LDA es un modelo estadístico para colección de documentos que intenta capturar esta intuición. En esta se define un tema para pasar a ser una distribución sobre un diccionario de palabras fijado. Por ejemplo en el artículo analizado en la Figura 7 el tema "genética" tiene un vocabulario de palabras que poseen una alta probabilidad de pertenecer al tema "genética".

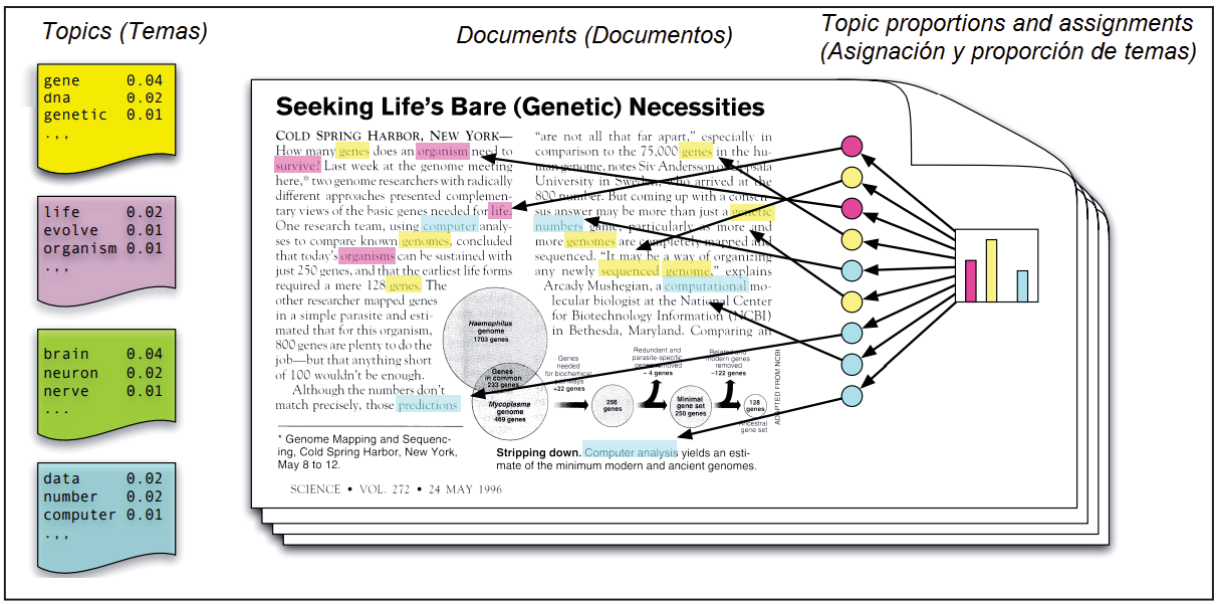

Figura 7. La intuición detrás de la técnica LDA

El proceso generativo LDA sobre el cual se realiza *data learning* es el siguiente:

1. Se elige aleatoriamente una distribución sobre los tópicos.
2. Por cada palabra en el documento
    a. Se elige aleatoriamente un tópico desde la distribución de tópicos establecidos en el paso 1.
    b. Se elige aleatoriamente una palabra de la correspondiente distribución en el diccionario.

Se debe tener siempre en cuenta que los algoritmos no tienen información externa sobre los temas encontrados y los artículos no están etiquetados con los temas o palabras claves (es un proceso de aprendizaje no-supervisado). La distribución de temas encontrados surge mediante el cálculo de la estructura de temas ocultos que probablemente generan la colección de documentos observados.

La utilidad de los modelos de tópicos se deriva de la propiedad que infiere la estructura oculta que se asemeja a la estructura temática de la colección. Esta estructura oculta interpretable clasifica cada documento en la colección mediante una minuciosa tarea y cada clasificación puede ser utilizada para ayudar a otras tareas como la recuperación de información, búsqueda, y la exploración de corpus. De esta forma, el modelo proporciona una solución algorítmica a la gestión, organización y anotación de grandes archivos de textos.

5.4.1 Estructura de Tópicos Ocultos y Tópicos Observados

El modelo de tópicos LDA asume que en un corpus existe una "estructura oculta de tópicos" y una Estructura de "variables observadas". Lo que hace LDA es que atrapa dicha intuición en un "modelo de variables ocultas". En un modelo de variables ocultas se practica y

postula una estructura oculta en los datos observados en el corpus correspondiente. Luego de obtener dicha estructura, se aprende de esta misma usando "probabilidades a posteriori".

En el LDA se tiene que los datos observados son las palabras observadas en cada documento y las variables ocultas representan los tópicos en sí mismos y además representan como cada documento exhibe las variables ocultas (tópicos ocultos). Lo anterior es explicado para poder entender el siguiente proceso, que es el proceso generativo realizado por LDA para conseguir probabilidades a posteriori dado ciertos sucesos a priori.

Se puede revisar este proceso observando primero el siguiente modelo gráfico. En este se define su notación de la siguiente manera:

- K: es el número específico de tópicos
- D: es el número específico de documentos
- N: es la cantidad de palabras en el documento $d \in D$
- $\vec{\alpha}$: es el vector positivo de parámetros α de dimensión K
- $\overrightarrow{\beta_{(1:K)}}$: representa los tópicos K de la estructura de tópicos ocultos
- $\overrightarrow{\theta_{(1:D)}}$: representa la proporción de tópicos por documentos $d \in D$
- $\overrightarrow{W_{(d,n)}}$: representa las variables observadas y no oculta dada la asignación del documento $d \in D$
- $\overrightarrow{Z_{(1:D,1:N)}}$: representa la asignación del tópicos oculto $\beta_{(k\prime)}$ a la palabra $\overrightarrow{W_{(d,n)}}$

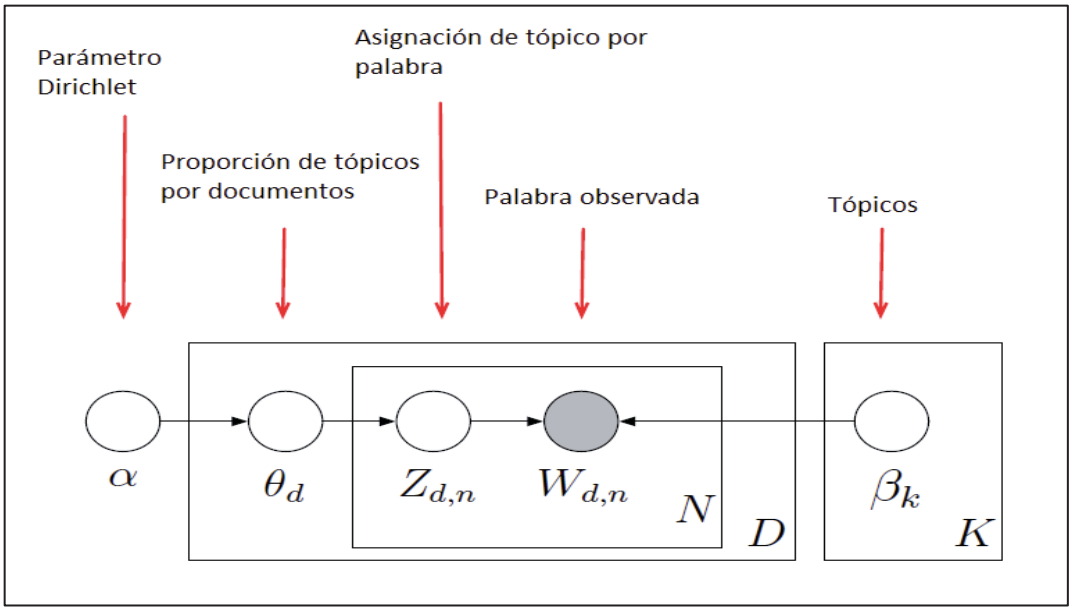

Figura 8. Modelo grafo del LDA.

Así, como resultado, el LDA devuelve una distribución de "probabilidades a posteriori" que servirán como una estructura de proporción de tópicos para poder realizar tareas de Recuperación de Información y Minería de Textos entre otros, mediante el siguiente proceso:

Desde una colección de documento, se infiere

- $Z_{d,n}$: Asignación de un tópico k para cada palabra $W_{d,n}$
- θ_d : Proporción de tópicos por cada documento d
- β_k : Distribución de tópicos por corpus

5.5 LDA con técnica de muestreo Gibbs Sampling

Aunque LDA es un modelo relativamente simple, la inferencia exacta es generalmente intratable. Una solución para esto, es usar algoritmos de inferencia aproximada tales como el algoritmo Gibbs Sampling. Este es un algoritmo para obtener una secuencia de observaciones que son a su vez aproximaciones de una distribución de probabilidad multivariable. Así dicha secuencia es usada para aproximar la distribución conjunta.

Lo que hace este algoritmo en el LDA es reflejar la distribución "a posteriori" obtenida en el muestreo *i-ésimo* y reflejar el efecto de este cambio en el corpus en el siguiente *(i+1)-ésimo* muestreo. Así en cada nuevo cálculo para saber el siguiente resultado *a posteriori* carga con una herencia del estado anterior.

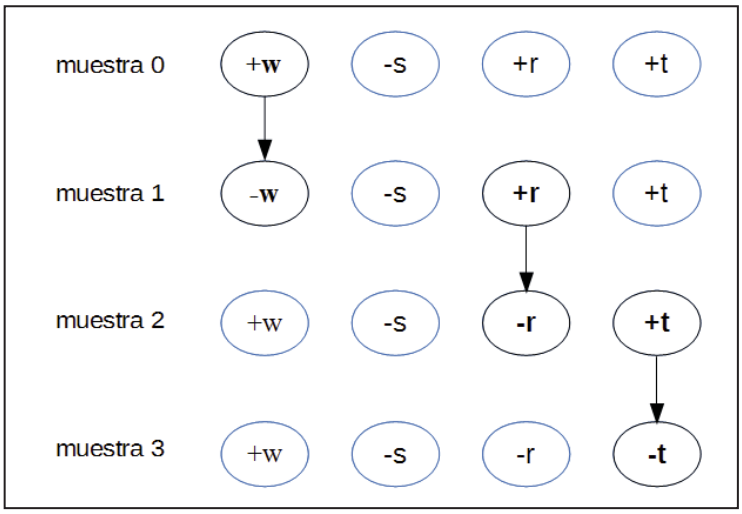

Figura 9. Ejemplo gráfico proceso Gibbs Sampling

- Cada muestra es dependiente de la muestra anterior
- Se maximizará la condición actual en la siguiente muestra

5.6 Implementación

Para la implementación de la etapa de pre-procesamiento se utilizó Python y Java, ambos lenguajes de programación multiplataforma y de código abierto. En el caso de Python se utilizó la librería NLTK (http://www.nltk.org) un conjunto de librerías para el análisis de textos y procesamiento de lenguaje natural.

El código que desarrolla el LDA a utilizar es una implementación que utiliza la técnica de muestreo “Gibbs Sampling”, y ha sido desarrollado por Nakatani Shuyo bajo la licencia de código abierto MIT (https://github.com/shuyo/iir/blob/master/lda/lda.py).

5.7 Ejemplo Básico LDA + Gibbs Sampling

Si se generaliza el procedimiento que realiza este algoritmo, se podría explicar con el siguiente ejemplo:

1. Sea $T = (t_1, t_2, t_3, t_4)$ el vocabulario de términos hallados en el corpus.
2. Sea $K = (k_1, k_2, k_3, k_4)$ los tópicos de todo el corpus.
3. Sea la matriz inicial de términos por tópicos para *d-ésimo* documento, la siguiente:

Términos x Tópicos	k_1	k_2	k_3	k_4
t_1	2	2	2	4
t_2	4	1	1	5
t_3	1	1	1	1
t_4	1	1	1	1

4. Muestreo para el tópico k_1:

Términos x Tópicos	$\boldsymbol{k_1}$	k_2	k_3	k_4
t_1	**1**	2	2	5
t_2	**3**	1	1	6
t_3	**1**	1	1	1
t_4	**1**	0	2	1

5. Muestreo para el tópico k_2:

Términos x Tópicos	k_1	$\boldsymbol{k_2}$	k_3	k_4
t_1	1	**1**	2	6
t_2	3	**0**	1	7
t_3	2	**0**	1	1
t_4	1	**0**	2	1

6. Muestreo para el tópico k_3:

Términos x Tópicos	k_1	k_2	$\boldsymbol{k_3}$	k_4
t_1	1	1	**1**	7
t_2	3	0	**0**	8
t_3	2	1	**0**	1
t_4	2	0	**1**	1

7. Muestreo para el tópico k_4:

Términos x Tópicos	k_1	k_2	k_3	$\boldsymbol{k_4}$
t_1	1	1	1	**7**
t_2	3	0	0	**8**
t_3	3	1	0	**0**
t_4	3	0	1	**0**

8. Finalmente

Términos x Tópicos	k_1	k_2	k_3	$\boldsymbol{k_4}$
t_1	0.1	0.1	0.1	**0.7**
t_2	0.27	0	0	**0.72**
t_3	0.75	0.25	0	**0**
t_4	0.75	0	0.25	**0**

Este proceso se repite por cada documento, “n” número de iteraciones correspondientes.

6 Experimentos y Resultados

Los experimentos se hicieron utilizando una base de datos facilitada por Analitic.cl para el análisis de este algoritmo, empresa dedicada a la recolección, clasificación y análisis de información de redes sociales.

Los datos constan de 65377 tweets de usuarios chilenos, en los que se mencionan a las tiendas comerciales Ripley y Falabella. Estos datos han sido recolectados por Analitic.cl entre las fechas: 06 de noviembre del 2013 y 28 de febrero del 2014.

Total Tweets en Base de Datos	65377
Total Palabras en Base de Datos	1062216
Palabras por Tweets Aprox	16

Tabla 5. Totales tweets y palabras a usar

Los tweets, en su mayoría, son de la siguiente forma:

Fecha Publicación	Autor	Cuerpo
06-11-2013 13:36	@pau_mb	Gracias a todos los que han comprado, va poquito, pero se agradece. Tienen todavía un año para ayudarnos CÓDIGO NOVIOS FALABE
06-11-2013 13:38	@ElDrZombie	RT @pau_mb: Gracias a todos los que han comprado, va poquito, pero se agradece. Tienen todavía un año para ayudarnos CÓDIGO NOVIOS FALABELL?
06-11-2013 13:43	@neomagnus	Ganas de renovar tu guardaropa o cambiar tu tele? Recuerda el codigo novios de Falabella 685349-08 cuando pagues y nos ayudas a sumar puntos
06-11-2013 13:45	@ElDrZombie	RT @neomagnus: Ganas de renovar tu guardaropa o cambiar tu tele? Recuerda el codigo novios de Falabella 685349-08 cuando pagues y nos ayuda?
06-11-2013 13:46	@gutierrezchile	RT @Dan_Escudero: SE NECESITA PROMOTOR (A) PARA FUJIFILM EN FALABELLA #VALDIVIACL (HORARIO MALL) Contactarse por interno! Sólo VERDADEROS I?
06-11-2013 14:00	@Spirallingg	Tener contrato indefinido y un sueldo mayor al que piden y no puedo sacar la maldita tarjeta Falabella! Me estoy perdiendo muchos puntos :(
06-11-2013 14:02	@PaolaAraneda	Queridos amigos de @RipleyChile el maquillador de NARS Alto Las Condes , Atiende increíble, felicítenlo de mi parte.

Tabla 6. Ejemplo tweets base de datos Ripley/Falabella

La cantidad de palabras en cada corpus y en cada documento varía dependiendo de cada escenario y se presentan en los siguientes puntos junto con el análisis de resultados en cada escenario.

6.1 Creación de Escenarios

Como se describió en la sección 1.2 acerca de la metodología del trabajo, para el proceso y análisis de resultados de los datos Twitter se ha creado tres "escenarios" para revisar el comportamiento del algoritmo LDA y la nueva información entregada por este mismo.

La actual sección se dedica a mostrar los escenarios y los resultados del pre-procesamiento en cada escenario.

6.1.1 Primer escenario: corpus por mes

- Corpus: formado por todos los tweets que pertenecen a un mismo mes.
- Documentos: el *corpus* se divide en diez documentos, donde cada documento estará formado por todos los tweets que pertenecen a las fechas resultantes de dividir el mes correspondiente en diez fechas equidistantes
- Datos después del pre-procesamiento:

	Corpus Nov 2013	Corpus Dic 2013	Corpus Ene 2014	Corpus Feb 2014
Total Palabras	105187	165390	160600	51539
Total Vocabulario	1950	2240	1892	1248

Tabla 7. Totales palabras y vocabulario a usar primer escenario

6.1.2 Segundo escenario: corpus por bimestre

- Corpus: formado por todos los tweets que pertenecen a un mismo bimestre.
- Documentos: el *corpus* se divide en diez documentos, donde cada documento estará formado por todos los tweets que pertenecen a las fechas resultantes de dividir el bimestre correspondiente en diez fechas equidistantes
- Datos después del pre-procesamiento:

	Nov-Dic 2013	Ene-Feb 2014
Palabras Corpus	267316	218365
Total Vocabulario	3615	2766

Tabla 8. Totales palabras y vocabulario a usar segundo escenario

6.1.3 Tercer escenario: corpus por totalidad de datos

- Corpus: formado por todos los tweets de la base de datos.
- Documentos: el *corpus* se divide en diez documentos, donde cada documento estará formado por todos los tweets que pertenecen a las fechas resultantes de

dividir el período que abarca a todos los tweets de la base de datos en diez fechas equidistantes

- Datos después del pre-procesamiento:

	Nov-Dic2013 Ene-Feb2014
Palabras Corpus	375261
Total Vocabulario	10217

Tabla 9. Totales palabras y vocabulario a usar tercer escenario

6.2 Análisis LDA Primer Escenario (Corpus por Mes)

Según los escenarios anteriores, a continuación se muestra la lista de los tópicos hallados. Además, como se verá en los gráficos, no todos los tópicos estaban en cada mes. El siguiente es el listado de los tópicos latentes hallados en los tweets correspondientes a cuatro meses (noviembre 2013, diciembre 2013, enero 2014 y febrero 2014):

"ciber monday; concierto; huelga trabajadores; navidad; tiendas cencosud; ripley one direction (1D) fans; liquidación-falabella gift card; candado puente virtual; novela avenida brasil; toselli puma evopower; cambio gabinete bachelet; macarena pizarro"

El nombre de cada tópico ha sido puesto de forma manual. Para un explicación de cómo se forma un tópico y del diccionario de palabras encontrados es importante revisar el Anexo A. Así se irá indicando el anexo que se debe revisar en cada resultado.

6.2.1 Resultado LDA Tweets Noviembre 2013

En este mes el LDA no encontró los siguientes tópicos latentes: *ripley one direction (1D) fans, liquidación-falabella gift card, candado puente virtual, novela avenida brasil, toselli puma evopower, cambio gabinete bachelet, macarena pizarro* (ver Anexo A.1)

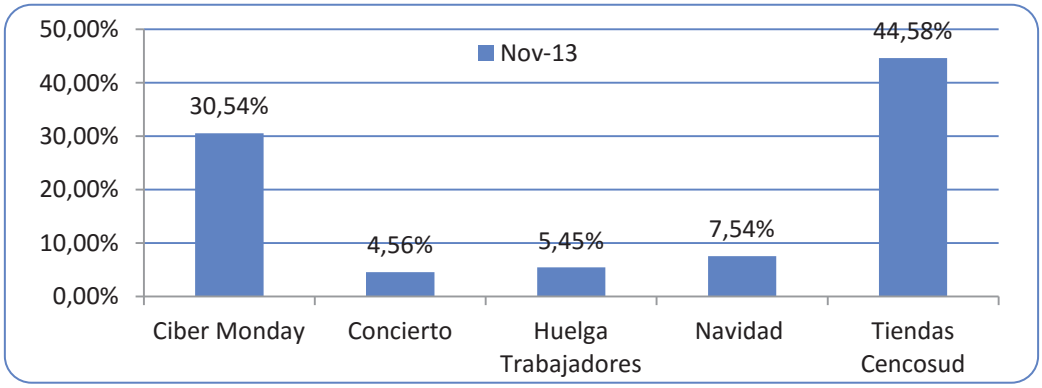

Figura 10. Gráfico tópicos latentes noviembre 2013

6.2.2 Resultado LDA Tweets Diciembre 2013

En este mes el LDA no encontró los siguientes tópicos latentes: *ciber monday, concierto, tiendas cencosud, liquidación-falabella gift card, candado puente virtual, novela avenida brasil, toselli puma evopower, cambio gabinete bachelet, macarena pizarro* (ver Anexo A.2)

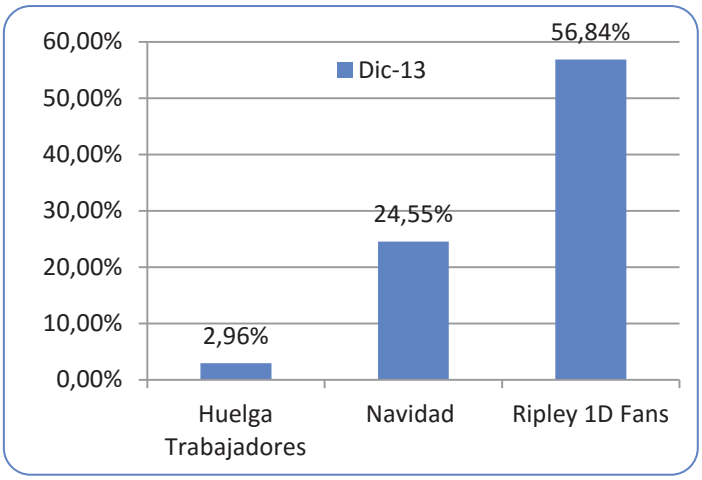

Figura 11. Gráfico tópicos latentes diciembre 2013

6.2.3 Resultado LDA Tweets Enero 2014

En este mes el LDA no encontró los siguientes tópicos latentes: *ciber monday, concierto, huelga trabajadores, navidad, ripley one direction (1D) fans, candado puente virtual, novela avenida brasil, toselli puma evopower, cambio gabinete bachelet, macarena pizarro* (ver Anexo A.3)

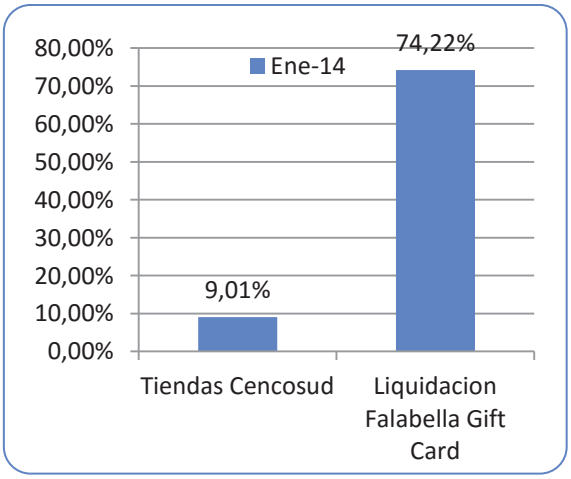

Figura 12. Gráfico tópicos latentes enero 2014

6.2.4 Resultado LDA Tweets Febrero 2014

En este mes el LDA no encontró los siguientes tópicos latentes: *ciber monday, concierto, huelga trabajadores, navidad, ripley one direction (1D) fans, liquidación-falabella gift card, candado puente virtual, novela avenida Brasil* (ver Anexo A.4)

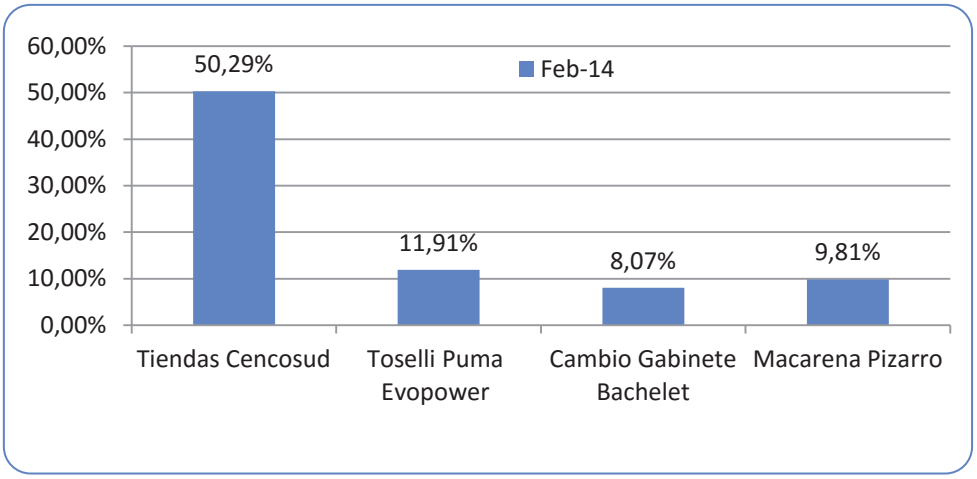

Figura 13. Gráfico tópicos latentes febrero 2014

6.2.5 Análisis Total

En el siguiente gráfico muestra el resultado de todos los meses procesados por el LDA en este escenario.

Se puede observar que, proporcionalmente, el tópico que tiene más peso en todo este primer escenario es el tema que tiene que ver con "liquidación falabella gift card" con un 74,22% de peso en cuanto a lo que se estuvo hablando en el mes de enero, el segundo con más peso es el tópico "ripley 1d fans" con un 56,84% de peso en cuanto a lo que se estuvo hablando en diciembre del 2013 y el tópico "tiendas cencosud" con un 50,29% de peso en cuanto a lo que se estuvo hablando el mes de febrero del 2014.

Además el tópico que más se repitió durante todos los meses de este escenario fue el tópico "tiendas cencosud" con un 44,58% para el mes de noviembre del 2013 y un 9,01% para enero del 2014 y un 50,29% para febrero del 2014. Los tópicos "navidad" y "huelga trabajadores" se repitieron también con menor porcentaje en noviembre y diciembre del 2013.

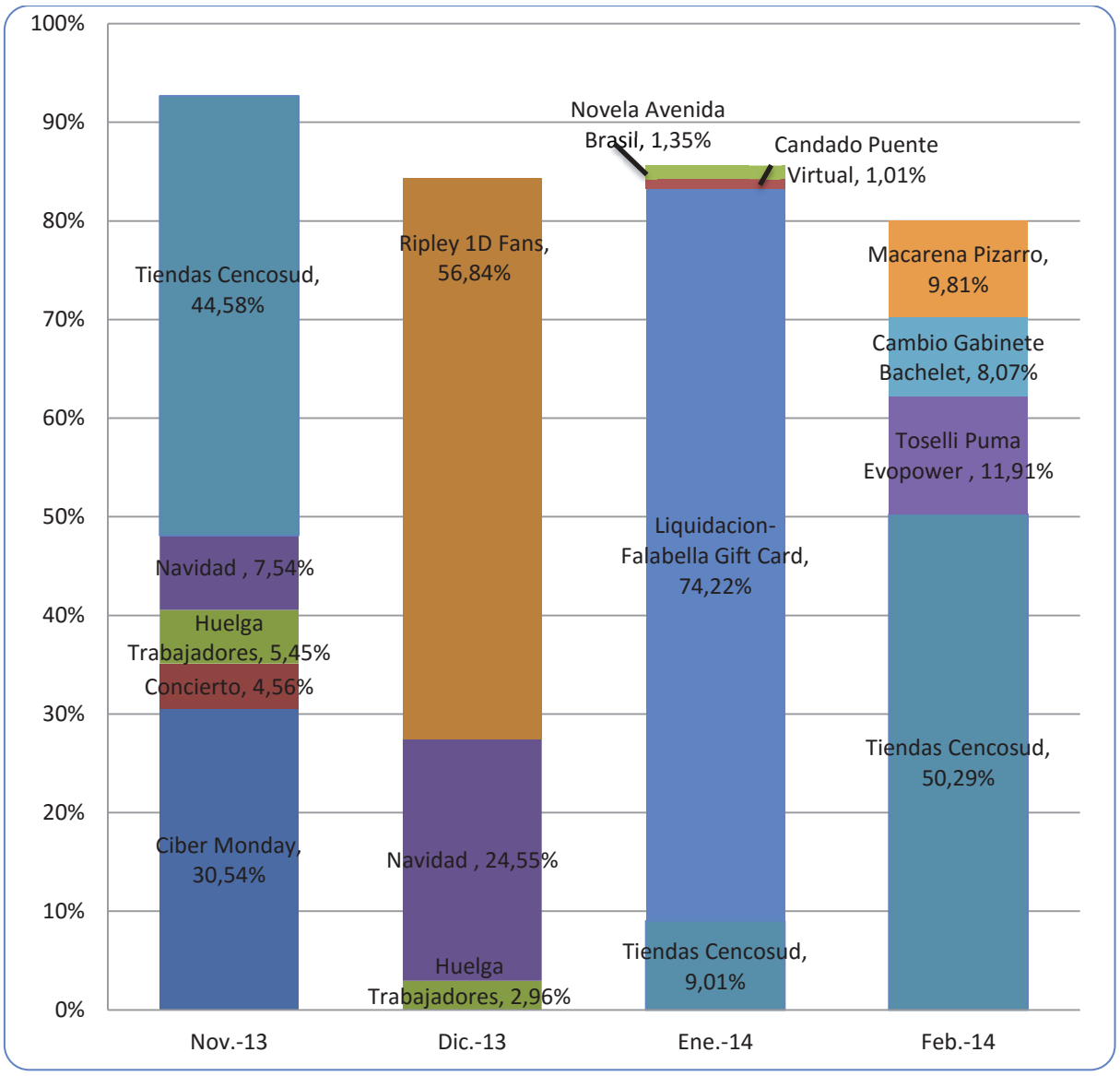

Figura 14. Gráfico tópicos hallados en primer escenario

6.3 Análisis LDA Segundo Escenario (Corpus por Bimestre)

Al igual que en el análisis del 1er escenario (corpus por mes), en el procesamiento de los datos de cada bimestre, los tópicos latentes que se encontraron no siempre se hallan en cada corpus -en este caso, en cada bimestre- procesado por el LDA. En el siguiente listado se muestran los tópicos latentes hallados en los tweets correspondientes a los dos bimestres (noviembre-diciembre del 2013 y enero-febrero del 2014):

"ciber monday; concierto; huelga trabajadores; navidad; tiendas cencosud; ripley one direction (1D) fans; liquidación falabella gift card; candado puente virtual, novela avenida brasil, toselli puma evopower, cambio gabinete bachelet, macarena pizarro"

6.3.1 Resultado LDA Tweets Noviembre-Diciembre 2013

En este bimestre el LDA no encontró los siguientes tópicos latentes: *tiendas cencosud, liquidación-falabella gift card, candado puente virtual, novela avenida brasil, toselli puma evopower, cambio gabinete bachelet, macarena pizarro* (ver Anexo A.5)

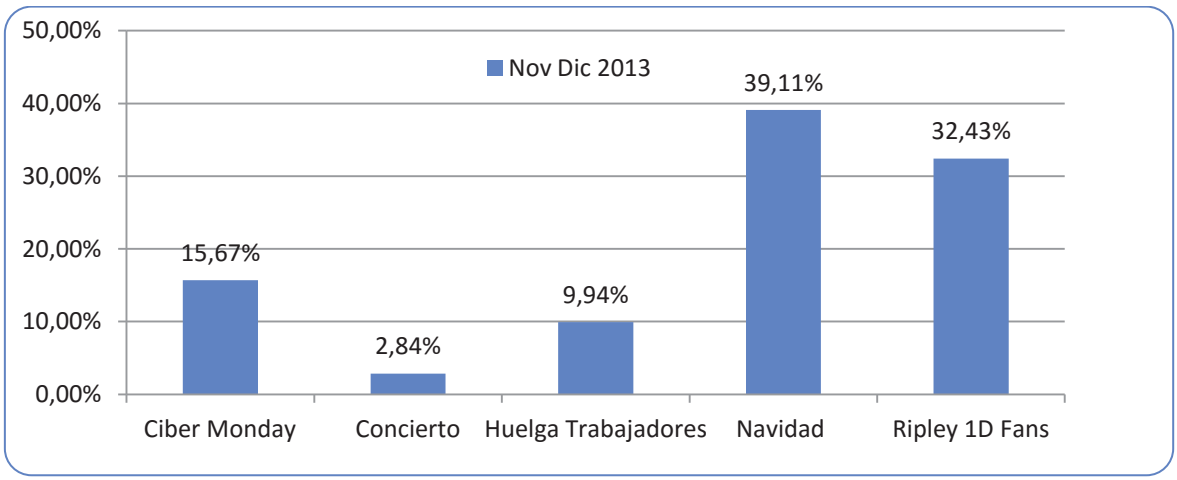

Figura 15. Gráfico tópicos latentes noviembre diciembre 2013

6.3.2 Resultado LDA Tweets Enero-Febrero 2014

En este bimestre el LDA no encontró los siguientes tópicos latentes: *ciber monday, concierto, huelga trabajadores, navidad, ripley one direction (1D) fans* (ver Anexo A.6)

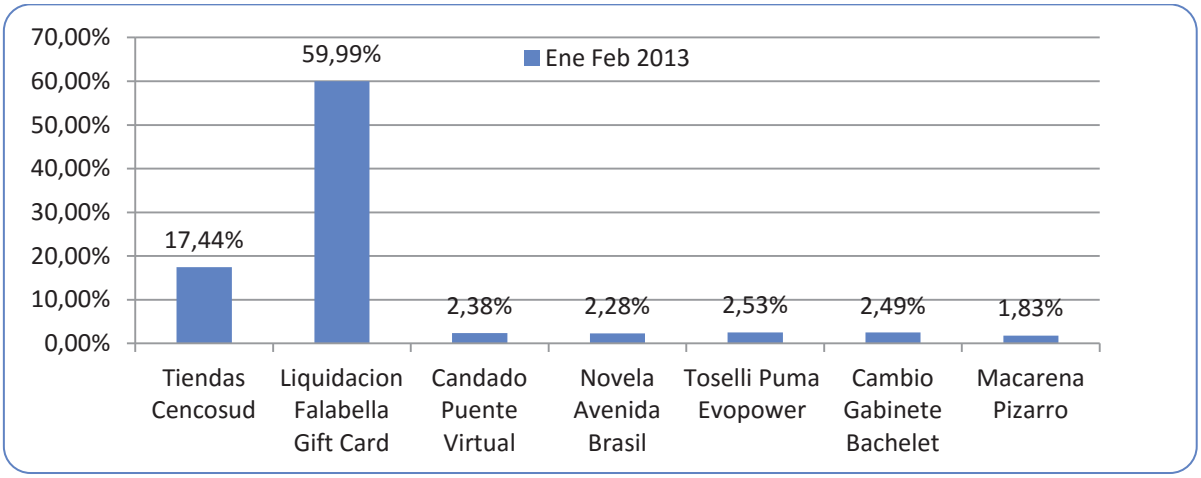

Figura 16. Gráfico tópicos latentes enero febrero 2014

6.3.3 Análisis Total

Se puede observar que, proporcionalmente el tópico que tiene más peso en todo este segundo escenario es el tema que tiene que ver con "liquidación falabella gift card" con un 59,99 % de peso en cuanto a lo que se estuvo hablando en el bimestre de enero-febrero del 2014, el segundo con más peso es el tópico "ripley 1d fans" con un 32,43% de peso y el tercer tópico "navidad" con un 39,11% de peso en cuanto a lo que se estuvo hablando en el bimestre noviembre-diciembre del 2013. Esto se mantiene muy parecido a lo arrojado por el primer escenario.

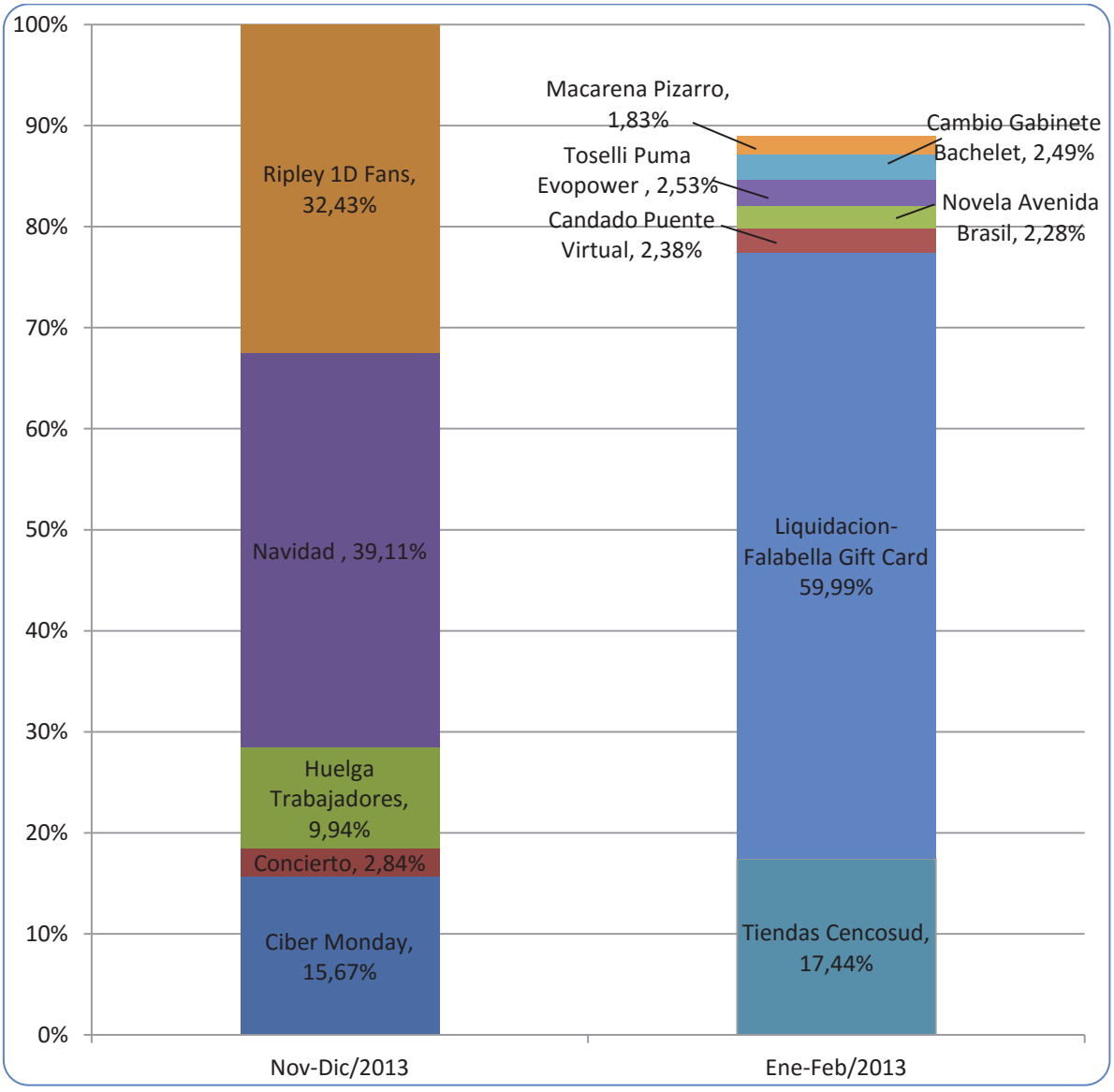

Figura 17. Gráfico tópicos hallados en segundo escenario

Un dato muy importante a mencionar es que el tópico *tiendas cencosud* solo apareció en el segundo bimestre (enero-febrero 2014) revisado, aún cuando en el primer escenario, este

tópico sí aparecía en el mes de noviembre 2013, mes correspondiente en este segundo escenario, al primer bimestre. Esto se debe en gran parte a que el LDA agrega importancia al peso que tiene un tópico en todo el corpus, y si bien en el primer escenario el tópico "*tiendas cencosud*" tiene un peso en el mes de noviembre 2013, en este segundo escenario el corpus no tiene solo al mes de noviembre sino que también al mes de diciembre (del 2013) y esto hace que este tópico se "diluya" y pierda importancia ya que una de las cosas que hace LDA es asignar una palabra a un tópico según la relevancia que el tópico del corpus más relevante.

6.4 Análisis LDA Tercer Escenario (Todos los Meses)

Finalmente se realiza un análisis en el que un solo corpus es la base de dato que contiene todos los tweets de los meses noviembre 2013, diciembre 2013, enero 2014 y febrero 2014. Los tópicos hallados en este corpus fueron los siguientes:

"novela avenida Brasil; ripley onedirection fans; macarena pizarro; liquidación falabella gift card; tiendas cencosud; navidad; candado puente virtual" (ver Anexo A.7)

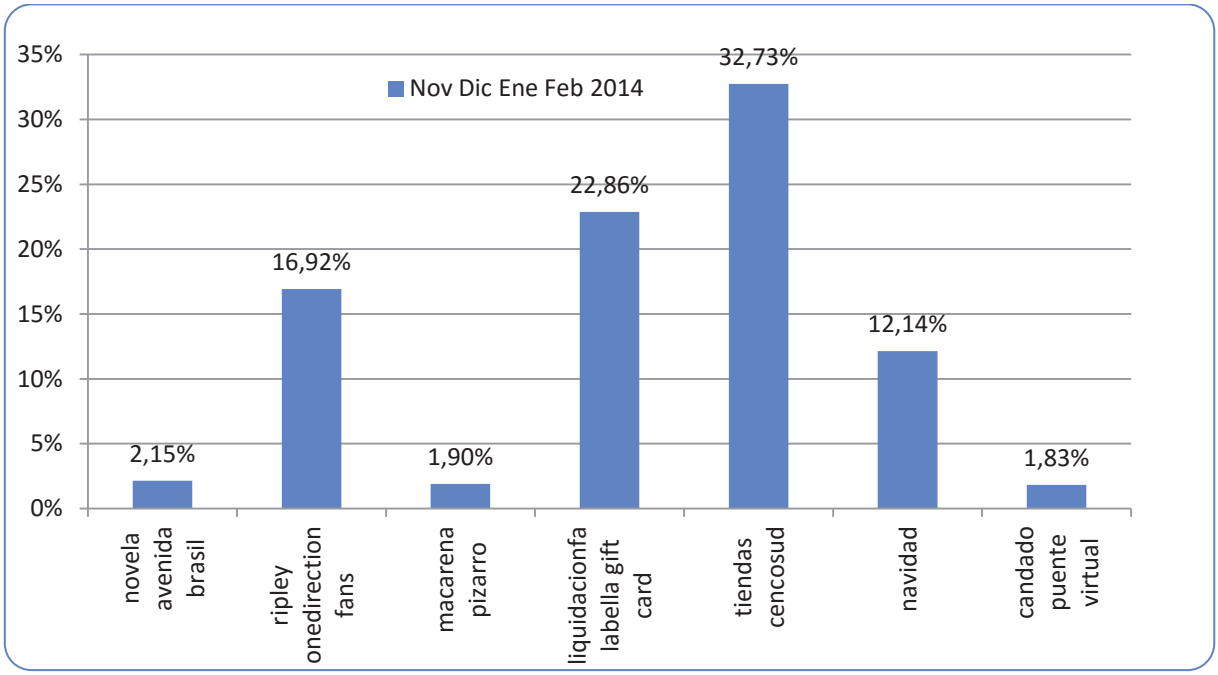

Figura 18. Gráfico tópicos hallados tercer escenario

Se puede observar que mientras que en el primer y segundo escenario se encontraron 12 tópicos, en este último escenario solo quedaron siete de dichos tópicos. Los tópicos que han quedado son los que han tenido más relevancia en los cuatro meses revisados y, a pesar de que algunos tópicos que no aparecen acá tenían probablemente más peso porcentual en un mes, los que aparecen en este corpus son los que tienen más presencia en los cuatro meses. Así se puede concluir que para un correcto uso del algoritmo no basta con solo revisar la totalidad de

la base de datos, sino que con este experimento se puede apreciar como el LDA descubre tópicos relevantes en ciertos períodos que no se descubrirían tan claramente revisando toda la base de datos o, siendo más estrictos, como se pude ver en este experimento; hay tópicos que no aparecerán al procesar toda una base de datos, pero que si aparecerán al revisar solo una parte de esta.

7 Fundamentos del Funcionamiento

En esta sección se revisará el porqué el modelo LDA funciona y cómo lo hace. Si el proceso generativo del LDA no recibe más información de entrada que el texto propiamente tal, el valor alfa de Dirichlet que sirve como factor de suavidad y el número de tópicos que desea encontrar; ¿cómo logra identificar grupos de palabras que tienen que ver con un mismo tópico relativamente bien?

Si bien el LDA es un modelo y por lo tanto, es una explicación lo aproximada a la realidad (entrega resultados aproximados), se sigue haciendo necesaria la explicación más acertada al porqué y cómo esta técnica logra formar estos grupos de palabras se describen a continuación.

7.1 El Principio de Caja de Dirichlet

Este establece que si n elementos se distribuyen en m lugares distintos y si n > m, luego existirá al menos un lugar m con más de un elemento, o dicho de otra forma, existirá al menos un grupo de elementos del mismo tipo. Así por ejemplo si se tienen trece personas y estas se separan por los meses en que nacieron; al menos dos personas estarán en el mismo grupo.

Este principio básico e intuitivo explica porqué si tenemos un texto con cinco palabras y elegimos que el algoritmo entregue cuatro tópicos, éste devolverá al menos un grupo de palabra con dos palabras pertenecientes a un mismo tipo. Así, la probabilidad de que grupos de palabras que pertenecen a un mismo tópico se sitúen juntas es maximizada cuando se dividen todas las palabras del corpus entre el número de tópicos.

7.2 Coocurrencia de Palabras en un Texto

Otro punto importante que revela el LDA, es el sencillo hecho de que las palabras que pertenecen a un mismo tópico *coocurren* y no así las palabras que tienen que ver con distintos tópicos. Así, si podemos observar que las palabras fútbol, balón, ciencias y medicina hay una probabilidad mayor de que fútbol y balón o ciencias y medicina sean palabras *coocurrentes* ya que tienen que ver con deporte o ciencias respectivamente; pero por el contrario, hay una probabilidad más baja de que fútbol y medicina *coocurran* o de que balón y ciencias *coocurran*.

7.3 Tópicos por Palabras v/s Tópicos por Documentos

Este es tal vez la mejor ventaja de este modelo. El LDA, en el proceso generativo, asume intuitivamente que, una palabra pertenece a un tópico, y que un documento pertenece al menos a un tópico.

Como se ha explicado en la sección 5.4 durante el proceso de asignación de tópicos por palabras, se realiza una asignación de una palabra a un tópico y la vez se asume que esta

palabra pertenece a un documento. En este proceso lo que el algoritmo va a decidir es, a qué tópico pertenece cada palabra tomando en cuenta los siguientes factores:

- Una palabra pertenece a un tópico, por lo que, en estricto rigor, si en el corpus se tiene un diccionario de mil palabras; potencialmente, existirán mil tópicos
- Un documento pertenece a un tópico, por lo que, en estricto rigor, las palabras de *x* documento, pertenecen al tópico de dicho documento.

Analizando los dos puntos anteriormente mencionados, si por ejemplo tenemos un corpus de diez documentos, y en todo el corpus (en todos los documentos), se tiene un vocabulario de mil palabras, según el razonamiento anterior, se tiene un corpus con mil posibles tópicos ya que hay un vocabulario de mil palabras y a la vez se tiene que en el corpus hay diez tópicos ya que hay diez documentos y en estricto rigor cada documento habla de un tópico.

En el siguiente modelo de grafo se puede ver esto. En este grafo $\boldsymbol{\pi}$ es el tópico asignado a un documento ***d.***

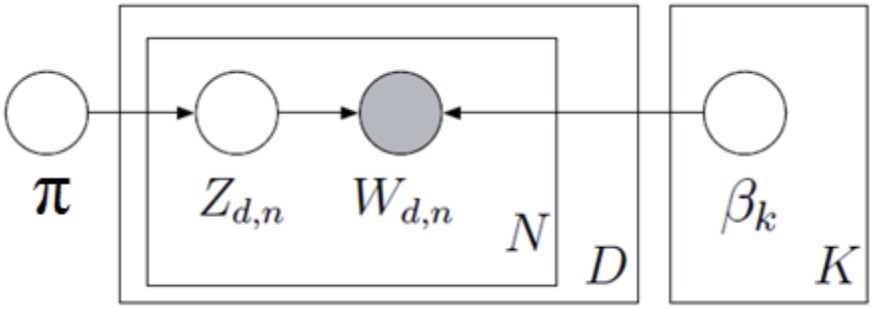

Figura 19. Modelo de grafo para explicar asignación de tópicos por documento

7.4 Proporción de tópicos por Documentos

Si bien el punto 7.3 explica la intuición de que cada documento habla de un tópico, esta intuición estricta no siempre será verdad y es por esto que finalmente el LDA agrega una variable α *a priori* para flexibilizar esto agregando la intuición de que un documento puede estar hablando de más de un tema.

El siguiente modelo de grafo muestra esto. El LDA agrega la proporción en la que varios tópicos están en un documento y esta proporción (θ_d) es fijada con el parámetro *Dirichlet* α. Este factor va a establecer *a priori* en cómo va a influir la proporción de tópicos θ_d de cada documento en la asignación de tópicos de cada palabra.

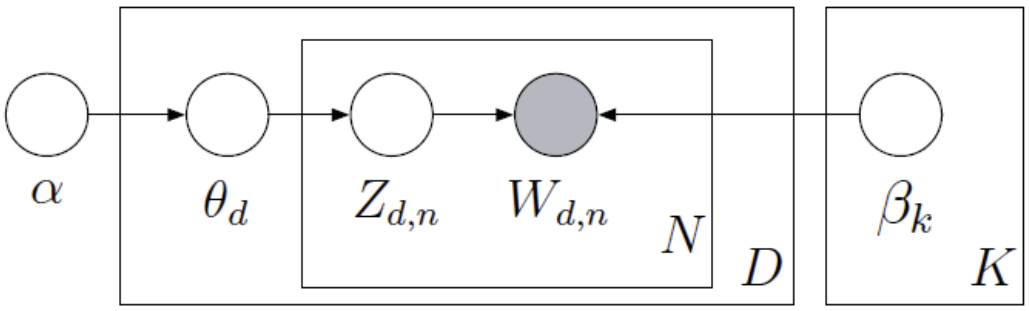

Figura 20. Modelo de grafo para explicar proporción de tópicos por documento

Para ver el proceso en que estas asignaciones se van realizando revisar la sección 5.4.1.

8 Conclusiones

En este informe se ha presentado el problema de estudio y los objetivos principales de este trabajo. Además se han presentado estadísticas sobre el aumento del uso de las redes sociales online, particularmente Twitter, en la que se pudo ver estudios y recopilaciones de datos, más estadísticas que demuestran el aumento del uso de Twitter como un medio periodístico.

El estado del arte de los modelos de tópicos que se presenta introduce a la técnica LDA que permitirá entender mejor el modelo de tópicos y los fundamentos teóricos involucrados. Es por esto que se ha introducido a este tema presentando las características generales más relevantes para esta técnica y por lo que se han mencionado sus modelos antecesores (LSA y pLSA).

Las técnicas de extracción de información no son tan nuevas y no se realizan solamente para clasificar archivos. También es común verlas en otras áreas de la informática como las búsquedas de contenidos, la clasificación de imágenes, el estudio de web semánticas o la minería de textos. Por esto mismo el campo de investigación y de aplicación es amplio, aunque, aún así sigue habiendo mucho por intentar hacer en esta área ya que el crecimiento de los datos y las nuevas formas en que estos se presentan en internet no se detiene y por esto se puede decir que es muy probable que este campo de investigación no pare de tener cada vez más adeptos.

Se ha definido también los antecedentes fundamentales de Twitter: el significado de los hashtags y las menciones, entre otros, ayudando así a la comprensión de la base de datos con la que se trabajará. Además, estos conceptos fueron utilizados en su mayoría para el pre-procesamiento de los datos, buscando en todo el texto las palabras que eran hashtags o menciones y quitándolas de la base de datos a trabajar. Luego de esto también se definieron conceptos básicos acerca de la clasificación de textos: el concepto de palabras vacías, *text corpus* y *bag of words*. Estos últimos conceptos sobre clasificación de textos han sido tratados para poder entender mejor el pre-procesamientos de los datos, el procesamiento y los resultados de cada experimento.

Para los experimentos realizados se definieron tres escenarios que han ayudado a ver el comportamiento del LDA, al procesar solo partes de la base de datos y al procesar toda la base de datos. Si bien estos escenarios parecen ser elecciones muy obvias, gracias a estos mismos escenarios se puede ver mejor la intuición de este modelo de tópico y como “esconde o muestra” un tópico dependiendo del tamaño de la base de datos.

Finalmente de este estudio se desprende que el LDA es una eficaz herramienta estocástica para descubrir grupos latentes de palabras que describen cuáles son los tópicos de un texto. Además se ha podido ver como este algoritmo, puede entregar mejor información si el corpus a trabajar es dividido convenientemente en distintos corpus para revisar más a fondo cada parte del texto y comparar los resultados de cada experimento. Teniendo estos resultados, las aplicaciones que pueden entregar los resultados del LDA son varios: recuperación de

información en búsqueda inteligente sobre datos, reconocimiento de patrones en imágenes, búsqueda de textos en la web, entre otros.

8.1 Trabajos Futuros

Luego de que el modelo LDA fue publicado y dado a conocer, mayormente por David Blei, han surgido muchas modificaciones y propuestas de mejoras de este mismo. Solo por nombrar algunas de estas, el mismo autor de LDA usado en este trabajo, ha publicado un software de código abierto que utiliza *Labeled LDA*.

Un trabajo interesante sería procesar bases de datos compuestas por tweets, como es el caso de esta investigación y comparar entre las diferentes modificaciones del LDA que han ido apareciendo en diferentes publicaciones académicas.

Otro trabajo interesante a realizar, enfocándose en esta investigación, sería procesar una base de datos que se componga de más meses y poder ver que tópicos evolucionan o van reapareciendo a lo largo del tiempo.

Referencias

[1] Kwak, H. Lee, C., Park, H., moon S. *What is Twitter, a social network or a news media?* Proceedings of the 19th WWW (2010)

[2] Sakaki, T., Okazaki, M. *Earthquake shakes Twitter users: real-time event detection by social sensors.* Proceedings of the 19th WWW (2010)

[3] Asur, S., Huberman, B.A. *Predicting the future with social media.* WI-AT (2010)

[4] Wayne Xin Z., Jing Jiang, Jianshu Weng, Jing He y Es-Peng Lim: *Comparing Twitter Traditional Media Using Topic Models.* In Proceedings of the 33rd European Conference on Information Retrieval (2011)

[5] Blei, D.M., A.Y. Ng, and M.I. Jordan, *Latent dirichlet allocation.* J. Mach. Learn. p. 993-1022. (2003)

[6] Hofmann, T., *Probabilistic latent semantic indexing*, in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM: Berkeley, California, United States. p. 50-57. (1999)

[7] Kristin Marino, *Social Media: The new news source.* Disponible vía web en http://www.schools.com/visuals/social-media-news.html. Revisada por última vez el 10 de noviembre del 2015.

[8] Michael J. Paul, Mark Dredze. A Model for Mining Public Health Topics from Twitter, Johns Hopkins University, (2011)

[9] David M. Blai, *Probabilistic topic models: Surveying a suite of algorithms that offer a solution to managing large document archives.* Princeton University (2012)

[10] Deerwester, S., et al, *Improving Information Retrieval with Latent Semantic Indexing*. Proceedings of the 51st Annual Meeting of the American Society for Information Science 25, pp. 36–40 (1988)

[11] Gabriel Dulac-Arnold, Ludovic Denoyer, Patrick Gallinari. *Text Classification: A Sequential Reading Approach.* Lectures Notes in Computer Science Journal. (2012)

[12] Fabrizio Sebastiani. *Machine Learning in Automated Text Categorization* (2002)

[13] David M. Blei. *Introduction to Probabilistic Topic Models.* Princeton University (2011)

[14] David M. Blei, Charla Magistral: *Topic Models.* Disponible vía web en www.videolectures.net/mlss09uk_blei_tm. Videolectures, Department of Computer Science, Princeton University. Revisada por última vez el 10 de noviembre del 2015.

Anexos

A: Tabla diccionario de tópicos hallados

Cada columna de cada tabla representa el diccionario de un tópico y poseen el siguiente formato:

<< **nombre_topico$_i$**>>
<<**numero_topico$_i$**>> (<<**total_palabras_topico$_i$**>>)
<< w_i>>: <<p_i >> (<<$cont_i$>>)
…
…
…
<< w_n>>: <<p_n >> (<<$cont_n$>>)

- $nombre\_topico_i$: nombre del tópico (escogido arbitrariamente) que identifica el tópico descrito por el grupo de palabras
- $numero\_topico_i$: número del tópico i ($i \in N; 0 < i < 20$)
- $total\_palabras\_topico_i$: total de palabras asignadas a este tópico
- w_i: i-ésima palabra de la lista de tópicos
- p_i: porcentaje de pertenencia del w_i en el tópico nombre_topico$_i$; ($p \in R; 0 < p < 1$)
- i: $i \in N; 0 < i < 20$
- n: $n = 10$

Se buscan $\boldsymbol{n}$ tópicos –representados en cada cada columna- por cada corpus y de estos $\boldsymbol{n}$ solo se describen los que tienen importancia. Los números de tópicos que no aparecen en los resultados, es porque tienen un número muy bajo de palabras encontradas. Para darle "relevancia" a un diccionario de tópico en esta investigación el **total_palabras_topico$_i$** debe ser mayor a 1000.

Por ejemplo, la siguiente columna muestra que el **total_palabras_topico$_i$** asignado para las palabras asignadas al tópico 19 es nulo, por lo tanto no se le da importancia a dicho diccionario de tópico en este trabajo y no aparece entre los tópicos hallados.

topic: 19 (0 words)
los: 0.000098 (0)
breve: 0.000098 (0)
buscarla: 0.000098 (0)
indignado: 0.000098 (0)
nonos: 0.000098 (0)
…
importan: 0.000098 (0)

A.1 Diccionario Tópicos Encontrados: 11/2013

"ciber monday"	*"concierto"*	*"huelga trabajadores"*
topic: 2 (1591 words)	**topic: 6 (1310 words)**	**-- topic: 7 (1405 words)**
antofagasta: 0.016173 (41)	ticketek: 0.014223 (32)	trabajadores: 0.015336 (36)
raja: 0.013055 (33)	laferte: 0.013348 (30)	huelga: 0.012815 (30)
abogadaflaite: 0.010717 (27)	santiago: 0.012910 (29)	novios: 0.011975 (28)
producto: 0.010327 (26)	concierto: 0.012473 (28)	corriente: 0.011134 (26)
ganar: 0.008769 (22)	vespucio: 0.012035 (27)	usen: 0.010714 (25)
monday: 0.008769 (22)	iphone: 0.010722 (24)	robaron: 0.010714 (25)
descubre: 0.008379 (21)	sala: 0.008972 (20)	devolver: 0.010714 (25)
amigas: 0.007989 (20)	recarga: 0.008534 (19)	millones: 0.010714 (25)
deseo: 0.007599 (19)	preventa: 0.008096 (18)	plata: 0.009874 (23)
cyber: 0.007210 (18)	ropa: 0.008096 (18)	cuenta: 0.008613 (20)

"navidad"	"ciber monday"	"tiendas cencosud"
-- topic: 8 (3719 words)	**topic: 9 (13481 words)**	**topic: 12 (4619 words)**
ripleychile: 0.030784 (144)	falabella: 0.156717 (2265)	cencosud: 0.041026 (229)
comercial: 0.022262 (104)	cybermonday: 0.071493 (1033)	ripleychile: 0.040669 (227)
suben: 0.020132 (94)	sala: 0.040226 (581)	luce: 0.022435 (125)
pascua: 0.016510 (77)	comprar: 0.018850 (272)	venta: 0.013854 (77)
navidad: 0.014167 (66)	ofertas: 0.017605 (254)	lanzamiento: 0.011709 (65)
heller: 0.012463 (58)	cybermondaycl: 0.015668 (226)	locale: 0.010636 (59)
feliz: 0.012037 (56)	cibermonday: 0.013800 (199)	utilidades: 0.009206 (51)
estadio: 0.011611 (54)	pagina: 0.013247 (191)	ripley: 0.008849 (49)
lafundacionsol: 0.011184 (52)	entrar: 0.011864 (171)	compra: 0.007776 (43)
cencosud: 0.010332 (48)	cybermondaychile: 0.009442 (136)	tiendas: 0.007776 (43)

"huelga trabajadores"	"tiendas cencosud"	"concierto"
topic: 15 (1284 words)	**topic: 16 (17378 words)**	**topic: 17 (940 words)**
demanda: 0.008189 (18)	falabella: 0.199068 (3653)	puntos: 0.008094 (15)
ganancias: 0.008189 (18)	ripley: 0.144527 (2652)	cod: 0.007572 (14)
trabajadores: 0.007747 (17)	ripleychile: 0.052662 (966)	beindex: 0.006527 (12)
millones: 0.007304 (16)	chile: 0.015284 (280)	amigas: 0.006527 (12)
video: 0.006861 (15)	tiendas: 0.013268 (243)	fiesta: 0.006005 (11)
ayudar: 0.005976 (13)	comprar: 0.011306 (207)	ciudad: 0.005483 (10)
roja: 0.005976 (13)	cmr: 0.009944 (182)	yankee: 0.004961 (9)
contra: 0.005976 (13)	tarjeta: 0.009889 (181)	dar: 0.004961 (9)
respetan: 0.005976 (13)	comercial: 0.009290 (170)	hoyts: 0.004439 (8)
valdivia: 0.005533 (12)	compra: 0.009236 (169)	concierto: 0.004439 (8)

A.2 Diccionario Tópicos Encontrados: 12/2013

"ripley onedirection fans"	"ripley onedirection fans"	"navidad"
topic: 1 (16211 words)	**topic: 8 (35896 words)**	**topic: 9 (1553 words)**
ripleychile: 0.141683 (2455)	ripleychile: 0.130390 (4826)	feliz: 0.018144 (48)
dripley: 0.135336 (2345)	dripley: 0.124446 (4606)	pascuero: 0.009540 (25)
conocer: 0.073712 (1277)	niall: 0.060474 (2238)	viejito: 0.008043 (21)
niall: 0.038226 (662)	conocer: 0.046858 (1734)	onedirection: 0.008043 (21)
encantaria: 0.037476 (649)	harry: 0.034242 (1267)	tiendas: 0.007295 (19)
harry: 0.030091 (521)	louis: 0.018762 (694)	polar: 0.006921 (18)
gustaria: 0.018666 (323)	poder: 0.018438 (682)	abcdin: 0.006921 (18)
ripley: 0.015896 (275)	ojos: 0.016763 (620)	almacenes: 0.006547 (17)
encanta: 0.015608 (270)	persona: 0.015547 (575)	regalo: 0.005799 (15)
cumplen: 0.012723 (220)	risa: 0.014656 (542)	wallstickers: 0.005425 (14)

"huelga trabajadores"	"judicial"	"navidad"
topic: 14 (2714 words)	**topic: 18 (1855 words)**	**topic: 19 (20957 words)**
boleta: 0.027517 (105)	fallida: 0.029748 (88)	falabella: 0.206029 (4548)
ripleychile: 0.022561 (86)	ejecutivos: 0.015294 (45)	ripley: 0.113263 (2500)
cmr: 0.014997 (57)	corte: 0.012941 (38)	ripleychile: 0.043099 (951)
huelga: 0.013432 (51)	quedan: 0.006555 (19)	navidad: 0.021719 (479)
mefascinaripley: 0.011607 (44)	caso: 0.005210 (15)	comprar: 0.011392 (251)
marca: 0.011346 (43)	entel: 0.004874 (14)	comercial: 0.011120 (245)
adornos: 0.011346 (43)	bci: 0.004874 (14)	tienda: 0.008810 (194)
posible: 0.011346 (43)	cmrchile: 0.004538 (13)	feliz: 0.007632 (168)
sernac: 0.010563 (40)	diario: 0.004538 (13)	compra: 0.007587 (167)
unid: 0.010563 (40)	banco: 0.004538 (13)	tiendas: 0.007315 (161)

A.3 Diccionario Tópicos Encontrados: 1/2014

"tiendas cencosud"	"liquidacionfalabella gift card"	"liquidacionfalabella gift card"
topic: 4 (8109 words)	**topic: 5 (48767 words)**	**topic: 6 (12590 words)**
falabella: 0.170348 (1542)	card: 0.146350 (7275)	ruleta: 0.121639 (1646)
ripley: 0.110381 (999)	gift: 0.143252 (7121)	card: 0.108341 (1466)
ripleychile: 0.060906 (551)	ruleta: 0.138344 (6877)	participando: 0.100288 (1357)
chile: 0.010657 (96)	participando: 0.131766 (6550)	liquidacionfalabella: 0.083149 (1125)
comprar: 0.010105 (91)	liquidacionfalabella: 0.090208 (4484)	gift: 0.076204 (1031)
compra: 0.008338 (75)	liquidaalabella: 0.023404 (1163)	falabella: 0.023013 (311)
debora: 0.007565 (68)	falabella: 0.020065 (997)	ripley: 0.020575 (278)
cencosud: 0.006902 (62)	liquidacionfala: 0.016082 (799)	tiendas: 0.020427 (276)
banco: 0.006792 (61)	ganar: 0.015660 (778)	belsport: 0.020132 (272)
tienda: 0.006571 (59)	ripleychile: 0.012703 (631)	sparta: 0.020132 (272)

"candado puente virtual"	"perrito ripley"	"liquidacionfalabella gift card"
topic: 9 (911 words)	**topic: 11 (1639 words)**	**topic: 13 (5415 words)**
qdmryocgow: 0.012655 (23)	perrito: 0.036170 (93)	liquidacionfalabella: 0.068936 (438)
virtual: 0.009424 (17)	comercial: 0.024178 (62)	falabella: 0.045983 (292)
puente: 0.008885 (16)	vive: 0.022631 (58)	ganar: 0.039066 (248)
participando: 0.008347 (15)	dripley: 0.016441 (42)	card: 0.037494 (238)
ganar: 0.008347 (15)	ayudar: 0.012959 (33)	ripley: 0.032149 (204)
viajes: 0.007808 (14)	felices: 0.009865 (25)	gift: 0.024603 (156)
posible: 0.007270 (13)	encanta: 0.009091 (23)	queremos: 0.020830 (132)
candado: 0.006731 (12)	colegio: 0.008704 (22)	ripleychile: 0.019730 (125)
gopro: 0.006193 (11)	escolares: 0.008704 (22)	hastleta: 0.017057 (108)
comparte: 0.006193 (11)	cola: 0.008317 (21)	tweet: 0.015171 (96)

"novela avenida_brasil"
topic: 19 (1216 words)
benicio: 0.017808 (38)
avenidabrasil: 0.016883 (36)
hijo: 0.014107 (30)
primer: 0.013645 (29)
esperan: 0.011795 (25)
murilo: 0.009482 (20)
padre: 0.008557 (18)
fallo: 0.008557 (18)
perrito: 0.008557 (18)
debora: 0.007169 (15)

A.4 Diccionario Tópicos Encontrados: 2/2014

"toselli puma evopower"	"cambio gabinete bachelet"	"tiendas cencosud"
topic: 1 (2638 words)	**topic: 2 (1788 words)**	**-- topic: 11 (11142 words)**
ripleychile: 0.062385 (203)	primer: 0.025498 (61)	falabella: 0.225608 (2654)
toselli: 0.043685 (142)	gobierno: 0.022181 (53)	ripley: 0.122429 (1440)
parque: 0.034488 (112)	gabinete: 0.021766 (52)	ripleychile: 0.063700 (749)
puma: 0.032649 (106)	cambio: 0.021766 (52)	debora: 0.013301 (156)
arauco: 0.030809 (100)	bachelet: 0.020937 (50)	compra: 0.012281 (144)
evopower: 0.027131 (88)	ripley: 0.019693 (47)	banco: 0.011346 (133)
mefascinaspor: 0.022226 (72)	pablo: 0.014303 (34)	comprar: 0.011006 (129)
corner: 0.016094 (52)	mefascinaspor: 0.012645 (30)	chile: 0.009476 (111)
indexsummer: 0.011496 (37)	corner: 0.011401 (27)	comercial: 0.009136 (107)
vida: 0.010883 (35)	vivo: 0.011401 (27)	tarjeta: 0.008627 (101)

"macarena pizarro"
-- topic: 19 (2173 words)
verano: 0.021988 (61)
indexsummer: 0.021273 (59)
feliz: 0.020200 (56)
pizarro: 0.018413 (51)
empleados: 0.016625 (46)
whatsapp: 0.015552 (43)
vestido: 0.015195 (42)
macarena: 0.014480 (40)
cumple: 0.013050 (36)
precio: 0.012335 (34)

A.5 Diccionario Tópicos Encontrados: 11-12/2013

"navidad"	"concierto"	"ripley onedirection fans"
topic: 0 (6473 words)	**topic: 1 (3796 words)**	**topic: 4 (43392 words)**
vespucio: 0.024817 (205)	iphone: 0.011154 (62)	ripleychile: 0.159415 (7205)
plaza: 0.022523 (186)	puntoticket: 0.009905 (55)	conocer: 0.064680 (2923)
trabajadores: 0.016847 (139)	ticketek: 0.009191 (51)	niall: 0.064237 (2903)
productos: 0.016001 (132)	entradas: 0.009012 (50)	harry: 0.039613 (1790)
pascua: 0.015639 (129)	concierto: 0.007228 (40)	encantaria: 0.023618 (1067)
venta: 0.013828 (114)	hites: 0.006871 (38)	encanta: 0.018529 (837)
comercial: 0.012016 (99)	preventa: 0.006335 (35)	poder: 0.017489 (790)
suben: 0.012016 (99)	led: 0.006157 (34)	louis: 0.017489 (790)
sfvespucio: 0.010084 (83)	dondatos: 0.005621 (31)	ojos: 0.016184 (731)
feliz: 0.009842 (81)	sala: 0.005443 (30)	risa: 0.015078 (681)

"huelga trabajadores"	"ciber monday"	*"huelga trabajadores"*
topic: 7 (3548 words)	**topic: 8 (18396 words)**	**topic: 12 (5703 words)**
alianza: 0.010923 (58)	falabella: 0.079912 (1614)	trabajadores: 0.037081 (278)
valdivia: 0.010737 (57)	cybermonday: 0.050759 (1025)	plaza: 0.022036 (165)
trabajadores: 0.010737 (57)	sala: 0.029376 (593)	call: 0.021104 (158)
video: 0.010363 (55)	ripleychile: 0.013488 (272)	center: 0.020571 (154)
cmr: 0.009803 (52)	ofertas: 0.013042 (263)	avenida: 0.017908 (134)
ayudar: 0.009803 (52)	comprar: 0.012597 (254)	pedrojdavis: 0.016843 (126)
millones: 0.009803 (52)	cencosud: 0.011706 (236)	suelo: 0.016577 (124)
huelga: 0.008122 (43)	pagina: 0.010172 (205)	comen: 0.016310 (122)
mayores: 0.007189 (38)	cibermonday: 0.009825 (198)	segundo: 0.016044 (120)
ganancias: 0.006442 (34)	entrar: 0.008835 (178)	florespineda: 0.015911 (119)

"huelga trabajadores"	"navidad"	"ciber monday"
topic: 13 (4044 words)	**topic: 15 (45848 words)**	**topic: 17 (2569 words)**
boleta: 0.017004 (99)	falabella: 0.189391 (9025)	antofagasta: 0.007655 (33)
queremos: 0.010852 (63)	ripley: 0.111603 (5318)	amigas: 0.006741 (29)
huelga: 0.008801 (51)	ripleychile: 0.046742 (2227)	seguros: 0.005827 (25)
mefascinaripley: 0.007605 (44)	navidad: 0.012013 (572)	monday: 0.005598 (24)
posible: 0.007605 (44)	comprar: 0.010314 (491)	descubre: 0.005598 (24)
adornos: 0.007434 (43)	tiendas: 0.010062 (479)	ibaruguv: 0.004227 (18)
sindicato: 0.007092 (41)	comercial: 0.009097 (433)	cyber: 0.003770 (16)
teatro: 0.006921 (40)	chile: 0.008698 (414)	cod: 0.003770 (16)
beindex: 0.006750 (39)	compra: 0.008383 (399)	presto: 0.003770 (16)
marca: 0.006750 (39)	tienda: 0.007879 (375)	ganar: 0.003770 (16)

A.6 Diccionario Tópicos Encontrados: 1-2/2014

"toselli puma evopower"	"macarena pizarro"	"novela avenida_brasil"
topic: 0 (3170 words)	**-- topic: 3 (2260 words)**	**-- topic: 9 (2817 words)**
ripleychile: 0.031078 (141)	pizarro: 0.013862 (50)	perrito: 0.021310 (89)
toselli: 0.029102 (132)	feliz: 0.012215 (44)	avenidabrasil: 0.019881 (83)
puma: 0.028223 (128)	vestido: 0.011941 (43)	benicio: 0.014167 (59)
arauco: 0.027345 (124)	whatsapp: 0.011941 (43)	murilo: 0.013690 (57)
parque: 0.025807 (117)	empleados: 0.011392 (41)	vive: 0.011310 (47)
corner: 0.016143 (73)	macarena: 0.011392 (41)	esperan: 0.008929 (37)
evopower: 0.011970 (54)	cumple: 0.010843 (39)	total: 0.008929 (37)
brasil: 0.008017 (36)	indexsummer: 0.010019 (36)	recuerdo: 0.007262 (30)
vemos: 0.007797 (35)	verano: 0.008647 (31)	deborah: 0.007262 (30)
amor: 0.006699 (30)	similar: 0.008372 (30)	primer: 0.007024 (29)

"tiendas cencosud"	"liquidacionfalabella gift card"	"candado puente virtual"
-- topic: 10 (21592 words)	**-- topic: 13 (74266 words)**	**-- topic: 14 (2952 words)**
falabella: 0.184309 (4234)	liquidacionfalabella: 0.155435 (11758)	ripleymeregalaunagopro: 0.015340 (66)
ripley: 0.104222 (2394)	card: 0.136611 (10334)	puente: 0.012111 (52)
ripleychile: 0.055343 (1271)	participando: 0.132804 (10046)	virtual: 0.010727 (46)
banco: 0.009902 (227)	gift: 0.132143 (9996)	cencosud: 0.010265 (44)
chile: 0.009815 (225)	ruleta: 0.124674 (9431)	ganar: 0.009573 (41)
compra: 0.009728 (223)	falabella: 0.022545 (1705)	sella: 0.009573 (41)
comprar: 0.009510 (218)	ripleychile: 0.016900 (1278)	gopro: 0.009112 (39)
debora: 0.009380 (215)	ganar: 0.015724 (1189)	candado: 0.009112 (39)
tarjeta: 0.006899 (158)	ripleymeregalaunagopro: 0.009537 (721)	ubica: 0.008651 (37)
falabellaayuda: 0.006638 (152)	ripley: 0.009418 (712)	dondatos: 0.008420 (36)

"cambio gabinete bachelet"
-- topic: 15 (3080 words)
comercial: 0.019382 (86)
directioners: 0.016693 (74)
gobierno: 0.011091 (49)
primer: 0.010643 (47)
asumir: 0.010419 (46)
song: 0.010195 (45)
gabinete: 0.009971 (44)
cambio: 0.009971 (44)
dripley: 0.009523 (42)
bachelet: 0.008626 (38)

A.7 Diccionario Tópicos Encontrados: 11-12/2013; 1-2/2014

"novela avenida_brasil"	"ripley onedirection fans"	"macarena pizarro"
topic: 2 (5966 words)	**topic: 3 (46964 words)**	**topic: 8 (5263 words)**
debora: 0.019808 (178)	conocer: 0.063238 (3162)	debora: 0.023289 (193)
corner: 0.015036 (135)	gustaria: 0.047041 (2352)	atrapado: 0.007643 (63)
parque: 0.013816 (124)	harry: 0.036903 (1845)	pizarro: 0.007282 (60)
arauco: 0.013261 (119)	encantaria: 0.029744 (1487)	vina: 0.007282 (60)
manana: 0.012262 (110)	encanta: 0.017407 (870)	verano: 0.006800 (56)
estara: 0.010820 (97)	poder: 0.016787 (839)	macarena: 0.006078 (50)
miercoles: 0.009488 (85)	louis: 0.016587 (829)	vestido: 0.005356 (44)
mexico: 0.006825 (61)	ojos: 0.015127 (756)	cumpleanos: 0.005356 (44)
vida: 0.005715 (51)	risa: 0.014487 (724)	cumple: 0.005356 (44)
brasil: 0.005382 (48)	sueno: 0.013727 (686)	empleados: 0.005115 (42)

"liquidacionfalabella gift card"	"navidad"	"tiendas cencosud"
topic: 9 (63440 words)	**topic: 16 (21903 words)**	**topic: 11 (90827 words)**
card: 0.194832 (12953)	sala: 0.026114 (651)	falabella: 0.177587 (16670)
gift: 0.193253 (12848)	pagina: 0.019300 (481)	ripley: 0.095848 (8997)
participando: 0.185356 (12323)	falabella: 0.018859 (470)	region: 0.011244 (1055)
ganar: 0.023546 (1565)	ofertas: 0.011003 (274)	comprar: 0.010498 (985)
falabella: 0.011153 (741)	cencosud: 0.010201 (254)	chile: 0.009817 (921)
queremos: 0.008400 (558)	comprar: 0.008317 (207)	compra: 0.008932 (838)
usando: 0.008205 (545)	navidenas: 0.007996 (199)	comercial: 0.008304 (779)
gopro: 0.008084 (537)	entrar: 0.007876 (196)	tiendas: 0.007143 (670)
foto: 0.007934 (527)	monday: 0.007515 (187)	tienda: 0.007004 (657)
hero: 0.007769 (516)	oferta: 0.006353 (158)	compre: 0.006504 (610)

"navidad"	"candado puente virtual"
topic: 14 (11803 words)	**topic: 17 (5088 words)**
vespucio: 0.015456 (229)	moreno: 0.011004 (89)
navidad: 0.014783 (219)	separadas: 0.010020 (81)
plaza: 0.013974 (207)	gaspar: 0.009406 (76)
trabajadores: 0.012089 (179)	cuerdas: 0.009283 (75)
pascua: 0.010877 (161)	virtual: 0.007070 (57)
productos: 0.009530 (141)	canciller: 0.006824 (55)
venta: 0.008115 (120)	peru: 0.006578 (53)
catalogo: 0.007374 (109)	candado: 0.006209 (50)
fusion: 0.007172 (106)	director: 0.005963 (48)
multas: 0.006903 (102)	gabriel: 0.005717 (46)