Departamento de Estadística e Investigación Operativa Aplicadas y Calidad

Detección automática de temas a partir de información recuperada de Twitter generada durante incendios forestales

Trabajo Fin de Máster

Autor:
Humberto Pino Chandia
Tutores:
Ángeles Calduch Losa
Elena del Val Noguera
Rocio Poveda Bautista

24 de noviembre de 2020

El conocimiento es la mejor inversión que se puede hacer

Abraham Lincoln

...a toda mi familia...

Resumen

El gran volumen de información generada diariamente en Twitter y la facilidad con la que se recupera por medio de su API han generado el interés de muchos trabajos en los que se recopila y analiza dicha información. Además, a esto se agrega que en situaciones tales como el huracán Sandy, según fuentes de Twitter Guskin [2012], el uso de la red social ha llegado a duplicar su uso normal. Esto ha llevado a que se realicen diversos análisis para conocer qué información se comparte durante situaciones de emergencia en Twitter.

Lo que se expone en el presente trabajo es el análisis realizado a información recuperada desde Twitter, generada en situaciones de emergencia. En este caso se han seleccionado cuatro incendios forestales ocurridos en Chile en el mes de Enero del año 2020. Utilizando el algoritmo *Latent Dirichlet Allocation* se han obtenido los principales tópicos de cada uno de los conjuntos de datos recuperados por medio de la API proporcionada por Twitter, es decir, se han obtenido los principales temas de los que hablan los usuarios de Twitter cuando suceden situaciones de emergencia.

Palabras clave: Redes Sociales, Twitter, Emergencias, Latent Dirichlet Allocation, Recuperación de Información.

Abstract

The large volume of information daily generated on Twitter and its ease provide for its retrieval through its API have generated the interest of many works related to the collection and analysis of such information. In addition to this, in situations such as Hurricane Sandy, according to Twitter sources Guskin [2012], the use of the social network has even doubled its normal use. This has led to various analyzes being carried out to find out what information is shared during emergency situations on Twitter.

In this work is exposed the analysis carried out on information retrieved from Twitter, generated in emergency situations. In this case, four forest fires occurred in Chile in January 2020 have been selected. Using the algorithm textit Latent Dirichlet Allocation, the main topics of each of the recovered data sets have been obtained through the API provided by Twitter, that is, we have obtained the main topics that Twitter users talk about in emergency situations.

Keywords: Social Networks, Twitter, Emergencies, Latent Dirichlet Allocation, Information Retrieval.

Resum

El gran volum d'informació que es genera diàriament a Twitter i la facilitat que hi ha per a la recuperació d'aquesta per mitjà de la seva API, ha generat l'interès de molts treballs relacionats amb la recollida i anàlisi d'eixa informació. A més d'això, es afegeix que en situacions com l'huracà Sandy, segons fonts de Twitter Guskin [2012], l'ús de la xarxa social ha fins i tot duplicat el seu ús normal. Això ha portat al fet que es realitzen diverses anàlisis per conèixer quina informació es comparteix durant situacions d'emergència a Twitter.

El que s'exposa en el present treball és l'anàlisi realitzada a informació recuperada des de Twitter, generada en situacions d'emergència. En aquest cas s'han seleccionat quatre incendis forestals ocorreguts a Xile en el mes de gener de l'any 2020. Utilitzant l'algoritme textit Latent Dirichlet Allocation s'han obtingut els principals tòpics de cadascú dels conjunts de dades recuperades per mitjà de l'API proporcionat per Twitter, és a dir, s'han obtingut els principals temes de què parlen els usuaris de Twitter quan succeeixen situacions d'emergència.

Paraules clau: Xarxes socials, Twitter, emergències, Latent Dirichlet Allocation, recuperació d'informació.

Índice

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN

1.1. Introducción

Las redes sociales como LinkedIn[1], Facebook[2] y Twitter[3] tienen millones de usuarios y se encuentran entre los sitios más populares de la web Lusoli et al. [2012] y el número de personas que utilizan las redes sociales en línea como una nueva forma de comunicación aumenta continuamente del Val et al. [2015], lo que ha generado un incremento en el volumen de información que se genera. Cada día, millones de usuarios comparten opiniones sobre diferentes aspectos de la vida cotidiana Pak and Paroubek [2010] y cada mensaje que un usuario escribe en estas redes y sus interacciones con otros usuarios dejan una huella digital que queda registrada del Val et al. [2015].

Hoy en día, compartir información vía redes sociales es algo normal incluso durante la aparición repentina de una situación de crisis, las personas afectadas publican información útil en Twitter que puede utilizarse para el conocimiento de la situación y otros esfuerzos humanitarios de respuesta ante desastres, si se procesa de manera oportuna y eficaz Imran et al. [2016]. En tales situaciones, la identificación de tweets que reportan información relevante y procesable es extremadamente importante para la coordinación efectiva de las operaciones de socorro posteriores al desastre Basu et al. [2018].

El presente trabajo busca recuperar y analizar información recuperada desde Twitter que se ha generado con motivo de incendios forestales ocurridos en Chile. El análisis que se ha realizado consta de dos partes, la primera se enfoca en reconocer tipos de usuarios que emiten tweets utilizando hashtags relacionados con las emergencias y en la segunda parte se han detectado de forma automática los principales temas sobre los que se habla en una situación de emergencia, para lo que se ha utilizado del algoritmo LDA. Como resultados de este trabajo se ha podido detectar cuatro tipos de usuarios que emiten tweets durante situaciones de emergencias, además de obtener temas de manera automática, los cuales han agrupado en cuatro grandes tópicos que se presentarán mas adelante.

[1] https://es.linkedin.com/
[2] https://www.facebook.com/
[3] https://twitter.com/

1.2. Motivación

Twitter permite compartir información en tiempo real dándole un ordenamiento cronológico en lo que se conoce como *timeline*. Esta característica permite a Twitter ser una red social que puede ser útil en casos en las que otras redes sociales no, como por ejemplo, situaciones de emergencia. En concreto, con motivo del *Huracán Sandy*, según Twitter, la gente envió más de 20 millones de tweets sobre la tormenta, dónde más de un 30 % de la información compartida correspondió a noticias e información Guskin [2012]. Este tipo de utilización de Twitter depende en gran parte de lo que las personas comunican durante la ocurrencia de situaciones extraordinarias, es decir, de qué hablan las personas durante la ocurrencia de una emergencia. Para esto, el contenido generado por los usuarios en sitios de microblogging, como Twitter, es usado ampliamente con la finalidad de obtener información procesable durante desastres naturales [Basu et al., 2020]. En otras palabras, se recupera de Twitter información generada con motivo de situaciones de emergencia, la cual posteriormente es analizada por medio de herramientas estadísticas y de minería de datos. Es así, que considerando la experiencia de trabajos realizados en que se ha recuperado información generada con motivo de catástrofes naturales, el presente trabajo busca realizar un análisis de la información que se genera en Twitter durante emergencias sucedidas en Chile.

1.3. Hipótesis

Como se ha mencionado anteriormente, los usuarios de Twitter generan un gran volumen de información en tiempo real, relacionada con una amplia gama de tópicos y temas, entre los que es posible encontrar información relacionada con situaciones de emergencia y catástrofes naturales. Si bien, esta información es de mucha utilidad en el momento en que se genera, ya que permite a los usuarios conocer detalles de lo que está sucediendo, dicha información también puede presentar una utilidad posterior a la situación en la cual se generó. Es decir, se puede obtener información útil si se analizan los tweets que se generan en situaciones de emergencia. Esto nos permite afirmar que utilizando herramientas estadísticas y software de programación, es posible analizar y encontrar patrones de similitud en la información recuperada desde Twitter que comparten los usuarios en situaciones de emergencia, los cuales pueden ser de utilidad tanto en el momento, como en situaciones de emergencias que pudiesen suceder en el futuro.

1.4. Objetivos

Los objetivos que se han planteado para la realización del presente trabajo se detallan a continuación:

Objetivo General

Analizar la información recuperada desde Twitter, que se origina a causa de incendios forestales ocurridos en Chile, poniendo énfasis en los tipos usuarios y en la detección de temas de los cuales se habla en una situación de emergencia.

Objetivos Específicos

Para la consecución del objetivo general, se han planteado además los siguientes objetivos específicos:

- Realizar un estado del arte relacionado a la recuperación de información de Twitter, lo cual permita construir el marco teórico del trabajo a realizar.
- Detectar los tipos de usuarios que comparten información en Twitter relacionada con incendios forestales ocurridos en Chile.
- Detectar tópicos con la información recuperada, sintetizando temas de mención recurrente entre los distintos conjuntos de datos recuperados desde Twitter.
- Determinar si existe diferencia entre los tópicos detectados a medida que transcurre el tiempo en una situación de emergencia.

En resumen, lo que el presente trabajo busca es identificar los grupos de usuarios que comparten información en una situación de emergencia y los principales tópicos de la información que dichos usuarios comparten.

1.5. Descripción de la Metodología

La metodología a utilizar en el presente trabajo fin de máster se ha basado en el **Método Científico**, con lo que se tendrán en cuenta las siguientes etapas a desarrollar para la resolución del problema planteado. [4]:

1. Búsqueda y descubrimiento del problema.
2. Documentación y definición del problema.
3. Planteamiento de respuestas probables.
4. Deducción de consecuencias de las hipótesis planteadas.
5. Diseño del procedimiento para la verificación de las hipótesis.
6. Comprobación de las hipótesis.
7. Contraste con la realidad.
8. Establecer conclusiones sobre resultados de la investigación.
9. Determinar conclusiones y generalizar resultados.

Lo primero que se realizará será poner en contexto el problema que se busca solucionar, planteando claramente la *hipótesis* y los *objetivos*, tanto general como específicos, de la investigación a realizar. Esta primera etapa será la base de la investigación y dará la pauta de lo que se llevará a cabo posteriormente, por lo que se considera de una importancia fundamental. Con la hipótesis y los objetivos definidos, se realizará la revisión del estado del arte, cuya finalidad será conocer el estado actual, en la comunidad científica, respecto de la recuperación y análisis de información de Twitter relacionada con situaciones de emergencia como son los incendios. Es por esto que se revisarán tanto revistas y artículos científicos publicados en el último tiempo. En conjunto con la construcción del estado del arte, se realizará una revisión bibliográfica que permita desarrollar el modelado estadístico que sirva de sustento matemático para el trabajo que se realizará.

La metodología a desarrollar se enfocará en las tareas necesarias para identificar a los tipos de usuarios que comparten información durante situaciones de emergencia, y qué tipo de información comparten. Por ello, será necesaria la recuperación de información desde Twitter, para posteriormente extraer los tópicos de los cuales más hablan los usuarios. Esto se explicara en detalle en los siguientes capítulos.

[4]extraído desde los apuntes de la asignatura Metodologías de la Investigación dictada por el profesor de la UPV, Doctor Bernabé Hernandis Ortuño)

1.6. Estructura del Documento

El presente trabajo fin de máster está compuesto por cinco capítulos. El contenido de cada uno de ellos se presenta a continuación:

Capítulo 1 - Introducción

El capítulo 1 lo componen la introducción y la motivación que han generado la investigación. Además de eso, en este capítulo se detallan los objetivos, tanto general como específicos, que se buscan conseguir con la investigación realizada.

Capítulo 2 - Estado del Arte

El capítulo 2 presenta el desarrollo del estado del arte, el cual permitirá conocer la situación actual en la comunidad científica de la temática que se pretende desarrollar. Para ello, se ha realizado la revisión de revistas y artículos científicos recientes.

Capítulo 3 - Modelado Estadístico

En el capítulo 3 se presenta el desarrollo del modelado estadístico, es decir, el soporte matemático sobre el que se sostiene el trabajo. En él se presenta cómo trabaja la extracción de tópicos mediante la utilización del algoritmo *Latent Dirichlet Allocation* (LDA), además de explicar cómo se obtiene el número de tópicos que se obtendrán para cada set de datos.

Capítulo 4 - Metodología

La metodología utilizada en la presente investigación es presentada en el capítulo 4. Comenzando con la elección del software a utilizar y una breve explicación respecto del API proporcionada por Twitter para recuperación de información. En dicho capitulo se detalla además la secuencia a realizar para la obtención de los tópicos para cada uno de los conjuntos de datos.

Capítulo 5 - Resultados

El capítulo 5 presenta los resultados obtenidos después de realizar la metodología detallada en el capítulo anterior. Primero se presenta el proceso de recuperación de datos desde Twitter por medio del API. Posteriormente

se presentan los resultados obtenidos al extraer los tópicos para cada uno de los conjuntos de datos.

Capítulo 6 - Conclusiones

Finalmente se presentan las conclusiones y líneas de trabajo futuras que se generan como consecuencia del trabajo de investigación.

Capítulo 2

ESTADO DEL ARTE

Twitter genera un gran volumen de información cada vez que se produce una situación de emergencia, ejemplo de esto son los más de 20 millones de tweets que se compartieron con motivo del Huracán Sandy, cantidad que duplica su uso en los días previos Guskin [2012], en los que aproximadamente un 34 % de los tweets emitidos correspondieron a noticias e información, seguido por un 24 % por fotos y vídeos relacionados con el huracán Guskin [2012], en los que podemos encontrar información compartida por organizaciones de gobierno, ONG, personas naturales, etc. Dicha información es de fácil acceso y recuperación por medio del API de Twitter, la cual ofrece un amplio acceso a los datos de Twitter que los usuarios han decidido compartir con el mundo Twitter [2020].

Esto ha llevado a una gran cantidad de científicos de datos a darle utilidad a estos grandes volúmenes de información generados en momentos de alta complejidad como emergencias, catástrofes naturales o eventos inesperados.

Lo que se propone en el presente documento es darle utilidad, por medio de técnicas estadísticas, a la información que es posible recuperar desde la red social Twitter. Considerando que la amplitud de los temas que se tratan en Twitter es de la mayor diversidad posible, es que el presente trabajo se enfocará en analizar la información que se genera con motivo de la ocurrencia de catástrofes naturales y emergencias. Para ello, lo primero será explicar de forma muy sencilla, ya que no es la finalidad última de este trabajo, en qué consiste Twitter y sus principales características. Posteriormente se mencionarán trabajos en los que se ha utilizado la extracción de información desde Twitter con motivo de la ocurrencia de situaciones de emergencia.

2.1. La Red Social Twitter

Nacida originalmente bajo el nombre de *twttr* en el año 2006, Twitter, como actualmente se conoce, es una red social creada por Jack Dorsey inicialmente con la finalidad de comunicarse con un pequeño grupo de personas por medio del servicio de mensajería corta de telefonía celular *SMS*, por sus siglas en inglés. Según datos informados oficialmente por Twitter en su *Q1 2019 Earnings Report* Twitter [2019], la red social cuenta con aproximadamente 330 millones de usuarios mensuales, de los cuales unos 145 millones usa el servicio diariamente, lo que convierte a Twitter en una de las redes sociales más populares de la actualidad.

Desde su creación en el año 2006, Twitter ha sufrido una serie de cambios hasta convertirse en lo que conocemos a día de hoy, donde su principal finalidad es permitir a sus usuarios registrados publicar breves actualizaciones de estado en tiempo real, las que pueden ser mensajes, noticias, enlaces

a otras páginas web, fotografías e incluso vídeos. Dichas actualizaciones de estado se conocen comúnmente con el nombre de *tweet* y se encuentran en lo que se define como *timeline* o *línea de tiempo*, la cual puede estar definida como de acceso público, es decir, que cualquier usuario puede acceder a ella, o privado, donde sólo usuarios autorizados podrán acceder a ver el contenido existente en ella. Estas actualizaciones o *tweets*, cubren una variada gama de temas, entre los que podemos encontrar comentarios sobre eventos recientes o en curso y temas emergentes, actividades personales, política y muchos otros [Ma et al., 2014]. En ellos, por medio de 280 caracteres (en un principio solo se permitían 140) es posible encontrar interacciones entre distintos tipos de usuarios que pueden ser personas, organizaciones u otro tipo de entidades, pudiendo o no estar conectadas entre sí.

Entre las principales características de Twitter podemos mencionar:

1. Permite a sus usuarios suscribirse a las actualizaciones de otros usuarios. Los suscriptores son conocidos como *seguidores* o *followers* y a la acción de suscribirse a las actualizaciones de alguien se le conoce como *seguir* o *following*.

2. Permite a sus usuarios darle a tweets individuales la etiqueta de *me gusta* o *like* para mostrar preferencia o afinidad por ellos.

3. Permite a sus usuarios reenviar tweets generados originalmente por otros usuarios. Esta acción se conoce con el nombre de *retweet* o *retuitear*.

4. Permite el envío de mensajes directos a un usuario determinado. Este mensaje solo puede ser visualizado por el usuario al que se le ha enviado.

5. Genera los llamados *trending topics* o simplemente *trends* que corresponde a un nombre, frase o tema que se menciona a un ritmo mayor que otros temas [Annamoradnejad and Habibi, 2019], es decir, se vuelven populares, ya sea por un esfuerzo concentrado de los usuarios o debido a la ocurrencia de un evento que incita a las personas a hablar de ello.

Qué es un Tweet

La unidad básica de Twitter se denomina *tweet* y consiste en un mensaje de 280 caracteres de longitud ubicado en una linea temporal de mensajes. Cada tweet contiene al menos tres atributos, el contenido del tweet, su autor (también conocido como usuario) y su hora de publicación [Ma et al., 2014]. Puede ser enviado directamente desde el website de Twitter, aplicaciones externas compatibles principalmente en teléfonos inteligentes y tabletas. Dentro

de las principales acciones permitidas por medio de tweets podemos encontrar:

1. Los tweets son por defecto públicos, es decir, cualquier usuario puede acceder a ellos, con la consideración que todos los usuarios pueden restringir la entrega de los tweets que emite solo a sus seguidores.

2. Los tweets permiten incluir en su contenido una mención a uno o varios usuarios de Twitter. Para ello se debe anteponer el símbolo @ (arroba) seguido de la cuenta que se desea mencionar dentro del tweet. Esto se conoce coloquialmente como *arrobar*.

3. Dentro de un tweet es posible incluir una *etiqueta* a una o varias palabras relacionadas con un tema. Para ello se debe anteponer el símbolo *numeral* # seguido de lo que se desea etiquetar. Esto se conoce como *hashtag* y permite a los usuarios un descubrimiento de contenido más rápido o rastrear eventos específicos en tiempo real [Annamoradnejad and Habibi, 2019].

Como se ha mencionado anteriormente, en la actualidad Twitter admite hasta 240 caracteres en cada tweet. Dicha cantidad permite que sólo sea posible abordar un tema o tópico. Esto, que pudiese considerarse como una limitación, permite obtener un ordenamiento en los temas que se están tratando en el instante en que se está utilizando la red social y además posibilita a sus usuarios expresar su opinión respecto de temas tan diferentes como son deporte, política, actualidad, emergencias, o lo que esté ocurriendo en el momento de emitir el tweet. Si bien la limitación de caracteres hace de Twitter una red social de expresión de opiniones limitada, lo anterior permite a Twitter ser utilizada como una herramienta en situaciones de alta complejidad como pudiesen ser situaciones de emergencia en donde ya se ha demostrado que los sitios de microblogging son una fuente de información útil durante un desastre [Basu et al., 2020]. Dicha información recuperada desde tweets puede ser de gran utilidad si se analiza de una manera adecuada, utilizándose para establecer protocolos de acción ante emergencias, establecer canales de información y/o comunicación que pudiesen aportar en situaciones de emergencias, entre otras iniciativas que pudiesen ser utilizadas. Es aquí donde la *recuperación de información* que es posible realizar desde Twitter toma un rol fundamental.

2.2. Trabajos Relacionados

Hoy en día es posible encontrar bastantes trabajos relacionados con la extracción de información desde Twitter para su posterior análisis, esto es debido principalmente al gran volumen de información disponible en Twitter y su sencilla forma de acceder a ella por medio del API proporcionada por la propia red social. Considerando que este es un tema en el cual ya se cuenta con algunos trabajos al respecto, a continuación se presenta una serie de estudios relacionados, los cuales han servido de guía para la realización de este trabajo.

En el año 2009, en el artículo presentado por [Mills et al., 2009] se plantea que las redes sociales se están convirtiendo en un medio de comunicación fundamental, mientras que en el año 2010, [Li and Rao, 2010] afirman que durante el terremoto ocurrido durante el año 2008 en China Twitter superó ampliamente a los canales de información convencional. En los años posteriores no han sido pocas las investigaciones que se han enfocado en la utilización de Twitter como herramienta para el análisis de información generada durante catástrofes naturales, por lo que este análisis apuntará principalmente a los últimos 5 años. En el trabajo realizado por [Takahashi et al., 2015] se analizaron 10.147 tweets obtenidos entre dos fechas determinadas relacionados con el tifón "Haiyan", los cuales fueron obtenidos por medio de una serie de hashtags utilizando el software cualitativo "NVivo". El análisis consistió en plantear 5 preguntas relacionadas a la utilización que los usuarios le dan a Twitter y se monitorearon sus respuestas, antes, durante y después del tifón. Posteriormente, el trabajo realizado por [Imran et al., 2016] plantea la utilización de Twitter como un canal activo de comunicación durante eventos de emergencias, donde son los usuarios quienes publican información. El estudio también plantea que si dicha información es procesada de manera efectiva y oportuna pudiese resultar útil. En el año 2017 [Basu et al., 2017] propone la importancia de diseñar y evaluar sistemas de recuperación de información utilizados durante situaciones de desastres, y busca agruparlos en 5 temas que para los autores son las principales necesidades durante desastres. Este estudio fue realizado con información obtenida desde Twitter con motivo del terremoto ocurrido en Nepal en el año 2015. Del año 2018 es posible destacar el trabajo realizado por [Goswami et al., 2018], en el cual se utilizan técnicas de minería de datos diseñadas para desarrollar una estrategia adecuada de gestión de desastres basada en los datos recopilados de los desastres. Por su parte, en el año 2019, [Pourebrahim et al., 2019] plantea entre otras cosas el aumento sustancial en el número de tweets y usuarios únicos durante la ocurrencia del huracán Sandy, en las cuales, una gran cantidad de publicaciones contenían información de primera mano sobre el huracán mostrando la in-

tensidad del evento en tiempo real. Otro punto importante que se plantea en este estudio corresponde a la importancia de las agencias gubernamentales como fuente de información y de comunicación unidireccional. Finalmente, en el año 2020, [Basu et al., 2020] presentan una actualización a lo ya realizado el año 2017, además de presentar la terminología que allí se define como *IRMiDis* (Information Retrieval from Microblogs during Disasters) o recuperación de información de microblogs durante desastres.

En el Cuadro 2.1 se presenta un resumen de los trabajos mencionados anteriormente, todos enfocados a Twitter, y su principal tópico de desarrollo.

Trabajo	**Técnica Utilizada**	**Tipo Desastre**	**Objetivo**
[Takahashi et al., 2015]	Chi-cuadrado /Regresión Logística	Tifón	Examinar el uso de Twitter durante y después del tifón
[Imran et al., 2016]	Naive Bayes/-Support Vector Machine /Random Forrest	Emergencias Varias	Machine Learning utilizando anotaciones humanas
[Basu et al., 2017]	Indri/ word2vec	Terremoto	Diseñar y evaluar sistemas de recuperación de información
[Goswami et al., 2018]	Data Mining	Emergencias Varias	Desarrollo de una estrategia de gestión de desastres utilizando data mining
[Pourebrahim et al., 2019]	Support Vector Machine / Método Louvain	Huracán	Explorar dinámicas de comunicación en Twitter
[Basu et al., 2020]	Neural Networks / Support Vector Machine / Kernel Lineal	Terremoto	Proponer metodologías de recuperación de información

Cuadro 2.1: Cuadro resumen de trabajos relacionados con la presente propuesta.

Capítulo 3

MODELADO ESTADÍSTICO

El presente trabajo busca en primer lugar recuperar la información generada en Twitter durante situaciones de emergencias, y posteriormente utilizar esta información para detectar los temas o tópicos más frecuentes de los cuales se habla en Twitter en este tipo de situaciones. Para esto, hay dos herramientas que serán claves. La primera de ellas corresponde al API proporcionada por Twitter para la recuperación de información, la cual no será explicada con propiedad, ya que no es la finalidad de este trabajo. La segunda es el algoritmo *Latent Dirichlet Allocation* (LDA), que es uno de los algoritmos utilizados para realizar trabajos de análisis de textos y obtención de tópicos. Además, como herramientas paralelas, nos apoyaremos de una serie de estadísticos para definir la cantidad adecuada de temas a obtener por medio del algoritmo LDA.

3.1. Extracción de tópicos utilizando Latent Dirichlet Allocation

Para la extracción de tópicos desde los tweets obtenidos se utilizará el algoritmo *Latent Dirichlet Allocation* o *LDA* por su siglas en inglés. Dicho algoritmo fue planteado por Blei, Ng y Jordan en su estudio del año 2003 [Blei et al., 2003]. Se encuentra dentro de los llamados *Algoritmos de Aprendizaje no Supervisado*, es decir, no existen clases predefinidas previamente ni tampoco se conoce el número de grupos en los cuales el algoritmo clasificará los datos. Se utiliza principalmente para agrupar y reducir el número de dimensiones de los datos analizados. En palabras simples, el algoritmo LDA permite que grandes grupos de palabras sean explicadas por grupos de palabras que no son observadas en primera instancia, es decir, en base a pequeños grupos de temas. Para una mejor comprensión del funcionamiento del algoritmo LDA, es preciso mencionar terminología relevante y necesaria de conocer previamente a la descripción de cada una de las técnicas.

- **Palabra**: Considerada como la unidad básica en el estudio de datos discretos, se encuentra definida como un ítem de un vocabulario indexado por $(1, ..., V)$. En notación vectorial, es posible representar cada palabra como un vector con una única componente igual a uno $(w^v = 1)$ y todas las otras componentes iguales a cero $(w^u = 0)$, considerando $v \neq u$.

- **Documento**: Se considera un *documento* como una secuencia de N palabras, denotadas como w, $w = (w1, w2, ..., w_N)$, donde w_N es la palabra n-ésima en la secuencia.

- **Dataframe**: Es una estructura de datos etiquetados con columnas de tipos potencialmente diferentes.
- **Corpus**: Es una colección de M documentos denotados como $D = w_1, w_2, ..., w_M$.
- **Tópico**: Corresponde a una palabra o frase altamente repetida en un momento determinado en una red social.

La finalidad del algoritmo LDA es que los documentos se vean representados como el resultado de mezclas aleatorias de temas latentes, en donde cada tema se caracteriza por una distribución sobre las palabras. La forma en que funciona el algoritmo LDA para cada documento d dentro del corpus es la siguiente:

1. Elige $N \sim$ Poisson (x)
2. Elige $\theta \sim$ Dir α
3. Para cada una de las N palabras w_n:
   (a) Elige un tema $Z_n \sim$ Multinomial θ
   (b) Elige una palabra w_n desde $p(w_n|Z_n, \beta)$, que es la probabilidad multinomial condicionada sobre el tema Z_n

Una variable aleatoria Dirichlet θ k-dimensional, puede tomar valores en el rango de $(k-1)$-simplex, en donde simplex es un vector-k dentro de la variable θ aleatoria que se incluye en la expresión $(k-1)$-simplex si $\theta_i \geq 0, \sum_{i=1}^{k} \theta_1 = 1$, y tiene la siguiente densidad de probabilidad en este simplex:

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} ... \theta_k^{\alpha_k - 1} \tag{3.1}$$

En la ecuación 1, el parámetro α es un vector-k con componentes $\alpha_i > 0$, y donde $\Gamma(x)$ es la función Gamma. Dados los parámetros α y β, la distribución conjunta de una mezcla de temas θ, un conjunto N de temas z y un conjunto N de palabras w, están dadas por:

$$p(\theta, z, w|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(z_n|\theta) p(w_n|z_n, \beta) \tag{3.2}$$

donde $p(z_n|\theta)$ es simplemente θ_i para un único i tal que $z_i^n = 1$. Integrando sobre θ y sumando sobre z se obtiene la distribución marginal de un

documento:

$$p(w|\alpha,\beta) = \int p(\Theta|\alpha)(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\Theta)p(w_n|z_n,\beta))d\theta \tag{3.3}$$

Finalmente, tomando el producto de la probabilidad marginal de un solo documento, se tiene la probabilidad marginal de un corpus:

$$p(|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d,\alpha)(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_{dn}|\Theta_d)p(w_{dn}|z_{dn},\beta))d\theta_d \tag{3.4}$$

3.2. Determinación de Número de tópicos

Un punto importante en el presente trabajo es la elección de la cantidad de tópicos que se obtendrán por medio del algoritmo LDA. Para ello, se utilizarán cuatro métricas presentadas en estudios en los que se aborda la obtención de una cantidad de tópicos adecuada para el algoritmo LDA. Estas cuatro métricas tienen como finalidad buscar el número óptimo de tópicos, para ello dos de estas métricas tienen finalidad maximizar y otras dos tienen como finalidad minimizar. Dichas métricas las encontramos en los siguientes trabajos:

1. Métricas de maximización
    - Accurate and effective latent concept modeling for ad hoc information retrieval [Deveaud et al., 2014].
    - Finding scientific topics [Griffiths and Steyvers, 2004].
2. Métricas de minimización
    - On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations [Arun et al., 2010].
    - A density-based method for adaptive LDA model selection [Cao et al., 2009].

A continuación, se procederá con la presentación y explicación de cada una de estas métricas.

Accurate and effective latent concept modeling for ad hoc information

La métrica presentada en el estudio desarrollado por Deveaud, San Juan y Bellot [Deveaud et al., 2014] en el año 2014 propone un método que estima automáticamente el número de conceptos latentes en función de sus distribuciones de palabras, es decir, considerando que los temas que el algoritmo LDA extrae están constituidos por las n palabras con mayor probabilidad, se define un operador $argmax[n]$ que produce los argumentos top-n que obtienen el valor de n más grande. Con esto se obtiene el set W_k de las n palabras con la probabilidad más alta $P_{TM}(w|k) = \phi_{k,w}$. Por lo tanto, el set queda representado por:

$$W_k = argmax[n]\phi_{k,w} \tag{3.5}$$

Finalmente, lo que este trabajo propone es una heurística que estima el número de conceptos latentes de una consulta maximizando la divergencia de información D entre todos los pares $(k_i; k_j)$ de los temas del LDA. El número de conceptos $\hat{K}$ estimado por este método viene dado por la siguiente fórmula:

$$\hat{K} = \underset{k}{argmax} \frac{1}{K(K-1)} \sum_{(k,k')\in T_K} D(k||k') \tag{3.6}$$

donde K corresponde al número de tópicos dados como un parámetro a LDA y T_K es el set de K tópicos modelados por LDA.

Finding Scientific Topics

El estudio presentado por Griffiths, Steyvers, Blei, y Blei en el año 2004 [Griffiths and Steyvers, 2004] busca tratar cada tema de un documento como una distribución de probabilidad sobre las palabras, lo cual permite considerar un documento como una mezcla probabilística de temas, con lo que, si se consideran T temas, es posible escribir la probabilidad de la *i-ésima* palabra en un documento como:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j), \tag{3.7}$$

donde z_i es una variable latente que indica el tema desde el cual se dibujó la palabra i y $P(w_i|z_i = j)$ es la probabilidad de la palabra w_i bajo el tema j. $P(z_i = j)$ da la probabilidad de elegir una palabra de los j temas en el documento actual, lo cual variará a través de los diferentes documentos. Intuitivamente, $P(w|z)$ indica qué palabras son importantes para un tema, mientras que $P(z)$ es la prevalencia de esos temas dentro de un documento.

On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations

El estudio presentado por Arun, Suresh, Veni Madhavan y Narasimha Murty en el año 2010 [Arun et al., 2010] nos da un enfoque respecto de LDA como un método de factorización de matriz no negativa que divide una matriz de frecuencia Documento-Palabra M en una matriz Tema-Palabra $M1$ de orden $T * W$ y una matriz de Documento-Tema $M2$ de orden $D * T$ donde D, T y W representan el número de documentos, temas y palabras respectivamente. Ambas matrices, $M1$ y $M2$ son matrices estocásticas donde la k-ésima fila en $M1$ es una distribución sobre palabras en el k^{th} tema y n^{th} fila en $M2$ es una distribución de temas en el n^{th} documento. Si no se tratara de matrices estocásticas, sino de recuentos representados, es decir, si el elemento $(i, j)^{th}$ de la matriz $M1$ indica el número de veces que la palabra j se ha asignado al tema i y si el $(i, j)^{th}$ elemento en la matriz $M2$ indica el número de veces que el tema j es asignado a la palabra en el documento i, entonces está claro que:

$$\sum_{v=1}^{W} M1(t, v) = \sum_{d=1}^{D} M2(d, t) \forall t = 1 to T \tag{3.8}$$

Esto no es más que el número de palabras asignadas a cada tema visto de dos maneras diferentes: una como suma de filas sobre palabras y otra como suma de columnas sobre documentos.

A density-based method for adaptive LDA model selection

El estudio presentado por Cao, Xia, Li, Zhang y Tang [Cao et al., 2009] en el año 2009, propone seleccionar de forma adaptativa el número más apropiado de temas en función de la densidad del tema, buscando que la similitud sea lo más grande posible en el intra-grupo, y a su vez lo más pequeña posible entre grupos. Para comprender de mejor manera dicha técnica, es necesario conocer previamente tres definiciones:

- **Densidad del tema**: Dado un tema Z y la distancia r, al calcular la distancia promedio del coseno entre Z y los otros temas, el número de temas dentro del radio de r desde Z es la densidad de Z, llamada $Densidad(Z, r)$.
- **Cardinalidad modelo**: Dado un modelo de tema M y un número entero positivo n, el número de temas cuyas densidades de tema son menores que n es la cardinalidad de M, llamada $Cardinalidad(M, n)$.

- **Muestra de referencia**: Dado un tema Z, radio r y umbral n, si la $Densidad(Z, r) \leq n$, entonces llame al vector de distribución de palabras de Z como muestra de referencia del tema Z.

En base a las 3 definiciones presentadas por los autores, se describe el método de la siguiente manera:

1. Proporcionar un K_0 arbitrario, inicializar las estadísticas suficientes mediante un método aleatorio, y utilizar el algoritmo variacional EM para estimar los parámetros del modelo, y obtener el modelo inicial LDA (α, β).

2. Respecto de la *matriz* $\hat{a}$ de distribución de temas del modelo antiguo como resultado de un clúster, se calcula secuencialmente el coseno promedio del modelo $r1 = ave\_dis(\beta)$; las densidades de todos los temas $Densidad(Z, r1)$ y la cardinalidad del modelo antiguo $C = Cardinality(LDA, 0)$.

3. A continuación, se debe re-estimar el parámetro k del modelo basado en C utilizando la siguiente formula de actualización:

$$k_{n+1} = k_n + f(r) \times (k_n - C_n) \tag{3.9}$$

   dónde $f(r)$ es la dirección cambiante de r. Si la dirección es negativa (contraria a la dirección anterior), entonces $f_{n+1}(r) = -1 * f_n(r)$ de lo contrario, $f_{n+1}(r) = f_n(r).f_0(r) = -1$. Cuando la dirección de convergencia es negativa, se clasifican los temas por las densidades y se extraen los temas anteriores de K' como muestras de referencia para inicializar las estadísticas suficientes. Cuando la dirección de convergencia es positiva, inicializamos las estadísticas suficientes mediante el método de *sembrado*.

4. Se deben repetir los pasos (2) y (3) hasta que la distancia promedio del coseno y la cardinalidad del modelo LDA converjan.

3.3. Paquete LDA Tuning

Las cuatro técnicas señaladas anteriormente se encuentran agrupadas en el paquete estadístico *LDA Tuning*, el cual puede calcular las 4 métricas de manera simultánea y no requiere de entrenar modelos LDA de manera simultánea. Este cálculo se realiza de forma paralela, por lo que en ocasiones requiere de recursos computacionales considerables para obtener un buen rendimiento.

Capítulo 4

METODOLOGÍA

La metodología presentada a continuación tiene como finalidad dar respuesta a las tres principales preguntas que se han realizado en base a los objetivos planteados en el Capítulo 1. La primera está enfocada a la detección de los tipos de usuarios que comparten en Twitter información relacionada con incendios forestales ocurridos en Chile.

- **Pregunta 1**: ¿Cuáles son los tipos de usuarios de Twitter que usan la red social durante un incendio?

La segunda, está enfocada a la detección de tópicos utilizando la información recuperada desde Twitter. Ambas preguntas están directamente relacionadas con los objetivos planteados para la investigación.

- **Pregunta 2**: ¿Con qué fines los usuarios de Twitter usaron la red social durante un incendio?

Finalmente, se buscará dar respuesta a la tercera pregunta planteada que tiene como finalidad determinar si existe diferencia entre los tópicos detectados a medida que transcurre el tiempo en una situación de emergencia.

- **Pregunta 3**: ¿Con qué fines usaron la red social los usuarios de Twitter durante y después de un incendio? ¿Existen diferencias o similitudes?

Para obtener las respuestas de las preguntas planteadas anteriormente, la metodología a seguir se esquematiza en la Figura 4.1.

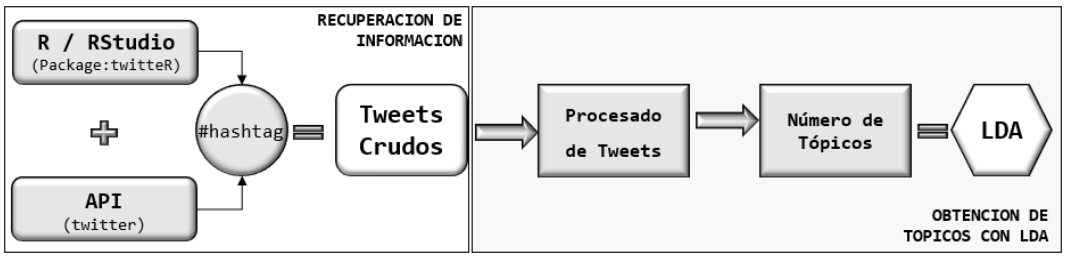


Figura 4.1: Esquema de la metodología.

La metodología a seguir se explicará en mayor detalle en los siguientes apartados.

4.1. Software a Utilizar

R R Core Team [2020] es un lenguaje de programación y entorno para gráficos y computación estadística que proporciona una amplia variedad de técnicas estadísticas entre las que podemos encontrar modelación lineal y no lineal, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupación, etc. **R** se encuentra disponible como Software Libre bajo los términos de la "Licencia Pública General" GNU de la "Free Software Foundation" en forma de código fuente.

Se ha utilizado además el software **R-Studio** RStudio Team [2015] que consiste en un entorno de desarrollo integrado (IDE) para el lenguaje de programación utilizado por R. Incluye una consola, un editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas para el trazado, el historial, la depuración y la administración del espacio de trabajo. Se encuentra disponible en código abierto y en ediciones comerciales.

Paquete twitteR

El Paquete *twitteR* Gentry [2015] es un paquete que permite el acceso a Twitter por medio de la utilización de su API, proceso que se detallará más adelante. Para más información es posible revisar la documentación existente en el website https://cran.r-project.org/web/packages/twitteR/twitteR.pdf.

Paquete tm

El Paquete *Text Mining Infrastructure en R* Feinerer and Hornik [2019], conocido comúnmente como *tm*, es un paquete utilizado en el software R para labores relacionadas a minería de texto. Entre las principales características del paquete *tm* podemos encontrar:

1. Ofrecer funcionalidad en la administración de documentos de texto.
2. Abstraer el proceso de manipulación de documentos.
3. Facilitar el uso de formatos de texto heterogéneos.

El paquete *tm* presenta un gran número de funciones que pueden ser utilizadas en tareas de minería de texto, de entre las que destacan *Corpus*, *findFreqTerms*, *TermDocumentMatrix*, por mencionar sólo algunas. El paquete *tm* está disponible gratuitamente bajo la Licencia Pública General de GNU (GPL). Para más información es posible revisar la documentación existente en el website https://cran.r-project.org/web/packages/tm/index.html.

Paquete topicmodels

Como se presenta en la descripción del paquete *topicmodels* Grün and Hornik [2011], proporciona una interfaz para el código C para modelos de asignación de Dirichlet latente (LDA) y modelos de temas correlacionados (CTM) de David M. Blei y coautores, y el código C ++ para ajustar modelos LDA utilizando muestreo de Gibbs de Xuan-Hieu Phan y co-autores. Para más información es posible revisar la documentación existente en el website https://cran.r-project.org/web/packages/topicmodels/index.html.

4.2. API de Twitter

Twitter proporciona a usuarios, ya sean particulares o empresas, acceso programático a sus datos mediante el uso de su **API** o *Application Programming Interface* (interfaces de programación de aplicaciones), esto permite a programas informáticos comunicarse entre sí, para con esto poder intercambiar información (enviar y recibir). Para la utilización de esta herramienta es necesario conectarse por medio del uso de credenciales que solicita Twitter, con las que por medio de la utilización del paquete **twitteR** del software R es posible obtener información desde la red social. En las siguientes secciones se explicará este proceso con mayor detalle.

4.3. Recuperación de información

Como se mencionaba anteriormente, para realizar la obtención de información desde Twitter, se necesita la API de Twitter, a la cual se conecta por medio de la previa obtención de una serie de credenciales que es posible obtener por medio de una cuenta en la red social. Las credenciales requeridas por Twitter son:

- Consumer key
- Consumer secret
- Acces token
- Acces secret

Una vez se cuenta con dichas credenciales, nos conectamos al API de Twitter utilizando el paquete estadístico *twitteR* del software R. Al unir ambos requerimientos será posible obtener la información de Twitter, la que

en este caso denominaremos como *tweets crudos* ya que corresponden a los tweets tal y como se muestran en Twitter y no se ha realizado ningún tratamiento de ellos. El siguiente esquema ejemplifica el proceso de obtención de datos desde Twitter.

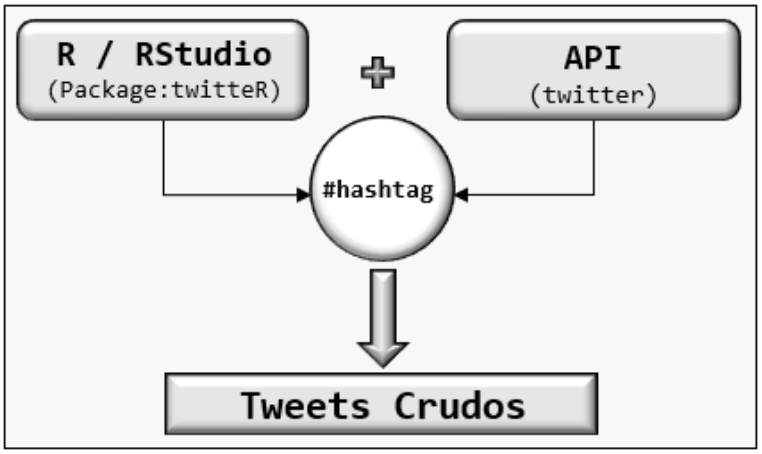


Figura 4.2: Esquema de recuperación de información.

Con lo que el código requerido por R queda estructurado de la siguiente manera:

```
#Instalar Paquete "twitteR"
install.packages(twitteR)
library(twitteR)

#Claves API del Usuario
consumer_key <- 'valor 1'
consumer_secret <- 'valor 2'

#Token de Acceso y Token de acceso secreto
acces_token <- 'valor 3'
acces_secret <- 'valor 4'

setup_twitter_oauth(consumer_key,
                    consumer_secret,
                    acces_token,
                    acces_secret)
```

Código 4.1: Código acceso API.

Una vez estando conectados al API de Twitter, del paquete *twitteR*, utilizaremos la función **searchTwitter** especificando como parametro de entrada el término o hashtag que se desea buscar en Twitter, el número de tweets requeridos y la fecha desde la cual han sido emitidos dichos tweets.

```
#Se obtienen "tweets crudos"
Chigu_tweets<- searchTwitter("topico_buscado",
                             n=5000,
                             since="aaa-mm-dd")
```

Código 4.2: Obtención de tweets.

La información que se rescata para cada uno de los tweets obtenidos con el API de Twitter corresponde a 16 campos, entre los que podemos encontrar:

Campo	**Descripción**
text	Texto del tweet
created	Fecha en que se emitió el tweet
id	Número identificador único de cada tweet
screenName	Nombre del usuario que emitió el tweet

Cuadro 4.1: Campos obtenidos desde tweets recuperados.

De los 16 campos entregados por el API de Twitter, el análisis se realizará utilizando el campo **text**, ya que en él se encuentra almacenado todo el texto con el que se trabajará en las etapas posteriores. Ya con la información obtenida desde Twitter, se debe proceder a procesarla para su utilización, para lo que será necesario realizar un trabajo de procesado de tweets el cual se detallará a continuación.

4.4. Procesado de tweets

Previo a la obtención de tópicos, es necesario hacer un trabajo de procesado de los tweets obtenidos. Esto con la finalidad de eliminar palabras, caracteres o símbolos que generen ruido en la extracción de los diferentes tópicos. Para ello, se realizará la siguiente secuencia de actividades con los tweets.

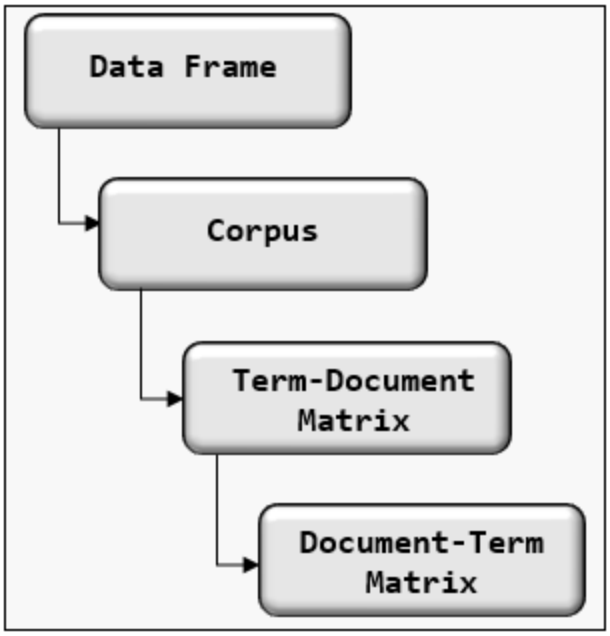


Figura 4.3: Esquema procesado de tweets.

A continuación se detallará cada una de las etapas realizadas en el trabajo de pre-procesado de los tweets obtenidos.

Creación del dataframe

Como se mencionaba anteriormente, al recuperar tweets crudos, el ordenamiento que presentan no es el adecuado para la extracción de tópicos. Es por esto que lo primero que haremos en el proceso de pre-tratamiento de datos será convertirlos en un *dataframe*. Un dataframe es una estructura de dos dimensiones que por su ordenamiento se asemeja a una matriz, pero a diferencia de las matrices, un dataframe admite datos de distinto tipo en sus filas, manteniendo el mismo tipo de datos en sus columnas, mientras que las matrices deben contener igual tipo de datos tanto en sus filas como en sus columnas. Una de las características más importantes de un dataframe es que éste se encuentra compuesto por vectores, cuyo nombre corresponde al nombre de cada una de las diferentes columnas. Para transformar nuestros datos a una estructura de dataframe utilizaremos el siguiente código:

```
1 #Creacion del Dataframe
2 DataFrame1 <-as.data.frame(tweets_crudos)
```

Código 4.3: Creación de data-frame.

Con lo anterior ya se cuenta con un ordenamiento adecuado para trabajar los datos de manera más fácil, con lo que podemos avanzar al segundo paso de nuestro pre-tratamiento de datos. Lo siguiente, será realizar la primera limpieza de texto de los datos, para lo cual utilizaremos la función **gsub**. Esta función tiene como finalidad modificar la ocurrencia de una expresión regular por otro argumento definido. Reemplazaremos retweets, referencias a otras cuentas existentes en el texto, símbolos de puntuación, números, links a paginas web y palabras de menos de tres letras, para que no afecten al análisis a realizar.

```
#Se obtiene el texto de los tweets
txt = DF1$text
txt

#remover retweets
txtclean = gsub("(RT|via)((?:\\b\\W*@\\w+)+)", "", txt)
#remover otras cuentas de Twitter
txtclean = gsub("@\\w+", "", txtclean)
#remover simbolos de puntuacion
txtclean = gsub("[[:punct:]]", "", txtclean)
#remover numeros
txtclean = gsub("[[:digit:]]", "", txtclean)
#remover links
txtclean = gsub("http\\w+", "", txtclean)
#remove words less than 3 letters
txtclean = gsub('\\b\\w{1,3}\\b','',txtclean)

txtclean
```

Código 4.4: Limpieza de texto.

Creación y ordenamiento del corpus

Un *corpus* es una colección de documentos que contiene texto (en lenguaje natural), lo que es uno de los requerimientos del algoritmo LDA. Para la construcción de dicho corpus utilizaremos el paquete de R *tm*. Una vez construido el corpus, será necesario realizar un ordenamiento que consistirá en eliminar caracteres que pueden afectar al resultado de los tópicos proporcionados por LDA.

```
#Construir un corpus
library(tm)
corpus = Corpus(VectorSource(txtclean))
```

Código 4.5: Construcción de corpus.

Lo primero que haremos será convertir todas las letras de nuestro corpus a minúsculas.

```
#Convertir todo a letras minusculas
corpus <- tm_map(corpus,
                 content_transformer(tolower))
```

Código 4.6: Convertir corpus a letras minúsculas.

A continuación se eliminarán las URL de las websites que pudiesen encontrarse en el corpus.

```
#Remover URL's
removeURL <- function(x) gsub("http[^[:space:]]*",
             "", x)
corpus <- tm_map(corpus,
              content_transformer(removeURL))
```

Código 4.7: Eliminar URL's del corpus.

Posteriormente, se procederá a remover todo otro carácter distinto a letras que pudiese haber, respetando los espacios existentes.

```
#Remover todo excepto letras y espacios
removeNumPunct <- function(x)
                  gsub("[^[:alpha:][:space:]]*",
                  "", x)
corpus <- tm_map(corpus,
                  content_transformer(removeNumPunct))
```

Código 4.8: Remover números del corpus.

El concepto de *palabra vacía* o *stopword* se le atribuye a Hans Peter Luhn, quien fuese uno de los pioneros en trabajos de recuperación de Información (IR). Una *stopword* se entiende como aquella palabra sin significado entre las que podemos encontrar artículos, preposiciones y pronombres, por mencionar

algunos. En este caso, se procederá a la eliminación de dichas palabras del corpus, ya que se considera que su aportación no es relevante para el estudio. Además de las palabras vacías, se eliminará cualquier espacio vacío extra que aún pudiese existir en el corpus.

```
#se remueven stopwords (palabras vacias)
myStopwords <- c(stopwords('spanish'), "vacia1",
                 "vacia2")
corpus <- tm_map(corpus, removeWords, myStopwords)

#se remueven espacios en blanco extras
corpus <- tm_map(corpus, stripWhitespace)
```

Código 4.9: Remover stopwords y espacios en blanco del corpus.

Finalmente, se procederá a remover los caracteres especiales que pudiesen existir en el documento. Es importante mencionar que se busca eliminar todas aquellas letras que tengan algún carácter tal como tilde, diéresis, virgulilla, etc. por mencionar algunos, ya que el algoritmo LDA hace distinción si una palabra está escrita con o sin tilde, es decir, LDA reconocerá como dos tópicos distintos *revolucion* y *revolución*. Para esto, se deberá crear un listado de caracteres a reemplazar, que estará compuesto por todas aquellas letras que pudiesen ser escritas utilizando un carácter especial, y lo reemplazamos por la letra sin dicho carácter.

```
#Remover caracteres especiales
unwanted_array = list(listado_caracteres_a_reemplazar)

chartr(paste(names(unwanted_array), collapse=''),
       paste(unwanted_array, collapse=''),
       corpus)
```

Código 4.10: Remover caracteres especiales del corpus.

Una vez realizado esto, se contará con un corpus ordenado y sin caracteres que puedan afectar el análisis a realizar.

Creación de la Term-Document Matrix

Una vez que tenemos los datos agrupados en un corpus ordenado y limpiado de términos que no aportan al análisis que se desea realizar, se procede

a construir la Matriz *Term-Document Matrix*, con la cual se le dará un ordenamiento de *documentos* como filas y *palabras* como columnas. Para esto se utilizará el paquete *tm* de R.

```
#Term-Document Matrix
tdm <- TermDocumentMatrix(corpus,
                          control=list(bounds = list
                              (global = c(50,Inf))))

dim(tdm)
```

Código 4.11: Creación de la Term-Document Matrix.

Creación de la Document-Term Matrix

Al igual que la TDM, la *Document-Term Matrix* se utilizará para darle ordenamiento a nuestro corpus. En este caso, la *Document-Term Matrix* no es más que la matriz transpuesta de la *Term-Document Matrix*.

```
#Document-Term Matrix
dtm <- as.DocumentTermMatrix(tdm,
                             control=list(bounds = list
                                 (global = c(50,Inf))))
dim(dtm)
```

Código 4.12: Creación de la Document-Term Matrix.

4.5. Elección de número de tópicos

Como se mencionó anteriormente, para la elección de número de tópicos se utilizarán las métricas desarrolladas por Arun Arun et al. [2010], Cao-Juan Cao et al. [2009], Deveaud Deveaud et al. [2014] y Griffiths Griffiths and Steyvers [2004], todas ellas agrupadas en el paquete *ldatuning* de R. Es importante mencionar que es posible seleccionar la métrica que se estime conveniente, pudiendo incluso seleccionar las cuatro métricas de una sola vez. Esto dependerá de la capacidad del equipo con que nos encontremos trabajando, ya que la principal limitación para esto es el tiempo que demorara en calcular el numero óptimo de tópicos.

```
#obtener el numero de topicos
library(ldatuning)
optimal.topics <- FindTopicsNumber(dtm.new,
                                   topics = 2:15,
                                   metrics =
                                   c("Arun2010",
                                   "CaoJuan2009",
                                   "Deveaud2014",
                                   "Griffiths2004"))

FindTopicsNumber_plot(optimal.topics)
```

Código 4.13: Número de tópicos.

Una vez obtenidas las métricas que buscamos, R nos permitirá representar los resultados por medio de gráficos, lo cual permitirá apreciar con mayor facilidad el numero de tópicos que se deben seleccionar.

4.6. Algoritmo Latent Dirichlet Allocation

Después de obtener el número de tópicos más adecuado utilizando las métricas mencionadas en la sección anterior, se procederá a la obtención del número de tópicos utilizando el algoritmo LDA.

```
#obtencion de topicos
library(topicmodels)
lda <- LDA(dtm.new, k = n) #encontrar "n" topicos
term <- terms(lda, 7) # primeros 7 terminos de cada
   topico
(term <- apply(term, MARGIN = 2, paste, collapse = ", "
   ))
```

Código 4.14: Algoritmo LDA en R.

Con esto, LDA nos entregará un cuadro con cada uno de los n tópicos que se han solicitado, además de presentarse los siete términos más importantes de cada tópico.

Capítulo 5

RESULTADOS

En el presente capítulo se busca exponer los resultados obtenidos después de la realización de cada una de las etapas de la metodología planteada en el capítulo anterior. Se ha recuperado información desde Twitter por medio de su API, con la cual se ha obtenido un número de tópicos en cuatro situaciones de emergencia relacionadas a Incendios ocurridos en Chile. Para todo esto se ha utilizado el algoritmo LDA como principal herramienta, además de paquetes pertenecientes al software estadístico R.

5.1. Datos recuperados desde Twitter

Primero se ha procedido a obtener los datos con los cuales se trabajará. Esto se ha realizado mediante la API que proporciona Twitter para la recuperación de información desde su plataforma. Para ello ha sido necesario contar con una cuenta de Twitter y obtener las credenciales solicitadas para ingresar a su API. Luego de contar con las credenciales, se deben introducir en el software R, donde por medio del paquete estadístico *twitteR* se procederá a la recuperación de los datos necesarios para la realización del estudio. Como se mencionaba anteriormente, estos datos corresponden a *tweets* recuperados que han sido generados con motivo de situaciones de emergencias ocurridas en Chile en el mes de Enero del año 2020. Si bien en un principio la primera opción fue realizar la recuperación de información mediante la geolocalización de tweets utilizando las coordenadas de la ubicación de la emergencia, esta opción fue descartada luego de realizar las primeras pruebas, debido a la baja cantidad de tweets que el API de Twitter era capaz de recuperar. Por ello, se realizó la recuperación de la información mediante la utilización de los *hashtags* correspondientes al nombre del lugar físico, es decir, la localidad que se ha visto afectada por dichos incendios. Es importante mencionar que la recuperación de los tweets se ha realizado en una fecha posterior a la ocurrencia de las emergencias. Este dato es de mucha importancia, ya que en los cuatro casos de recuperación de información se le ha solicitado al API de Twitter que recupere 5.000 tweets, lo cual ha sido realizado sólo en uno de los cuatro requerimientos, ya que en los otros tres casos el API ha sido capaz de recuperar una cantidad inferior. En el Cuadro 5.1 se presenta un resumen de los datos obtenidos con cada uno de los hashtags utilizados para su recuperación.

Hashtag	Inicio	Fin	Cantidad
Chiguayante	27/01/2020	30/01/2020	5.000
Hualqui	25/01/2020	30/01/2020	2.935
Nonguen	25/01/2020	30/01/2020	4.592
Santa Juana	26/01/2020	30/01/2020	1.656

Cuadro 5.1: Resumen de la información recuperada desde Twitter.

5.2. Procesado de tweets

Se ha realizado el procesado de tweets siguiendo la metodología presentada en el capítulo anterior. Es importante mencionar que este proceso es necesario y similar en todos los dataset que se han analizado.

5.3. Dataset - Incendio Chiguayante

El primer dataset con que se trabajó fue el correspondiente al incendio ocurrido en *Chiguayante*. La información recuperada corresponde a 5.000 tweets obtenidos entre los días 27-01-2020 y 30-01-2020 con motivo del incendio forestal que ahí sucedió. Lo primero que se realizó fue el trabajo relacionado a procesado de texto y obtención de la *Document-term Matrix*. Después, se ha obtenido el siguiente cuadro resumen:

```
1 <<DocumentTermMatrix (documents: 5000, terms: 165)>>
2 Non-/sparse entries: 27758/797242
3 Sparsity           : 97%
4 Maximal term length: 19
5 Weighting          : term frequency (tf)
```

Código 5.1: Cuadro resumen DTM Chiguayante.

El Código 5.1 corresponde a un resumen obtenido junto a la *Document-term Matrix*, en el cual se entrega información que busca resumir lo que se está analizando. Lo primero que se presenta es el número de documentos, que en este caso corresponde a 5.000 (considera cada tweet como un documento), con un total de 165 términos diferentes. Se tiene una tabla con 797.242 celdas cuyo valor podría ser 0 y 27.758 celdas tienen un valor mayor que 0, lo que supone un 97 % de las filas tienen como valor el 0. *Maximal term length* indica que la palabra más larga tiene 19 caracteres de longitud. Finalmente, *term*

frquency indica que la forma de considerar los términos es por medio de su frecuencia.

Lo primero que se ha obtenido al realizar el análisis es la información de quiénes son los usuarios que han emitido tweets. Para ello se ha generado un cuadro en el cual se definen los tipos de usuarios más comunes.

Categoría de usuario	**Descripción**
Organismos de gobierno	Cuentas asociadas al gobierno
Organizaciones no gubernamentales	Cuentas asociadas a organismos no gubernamentales
Medios de comunicación	Cuentas asociadas a medios de comunicación y periodismo
Usuarios individuales	Usuarios individuales de Twitter

Cuadro 5.2: Resumen tipos de usuarios de Twitter.

Respecto de los usuarios más frecuentes que han emitido tweets utilizando el tópico *#Chiguayante* se ha obtenido el gráfico de barras que se presenta a continuación:

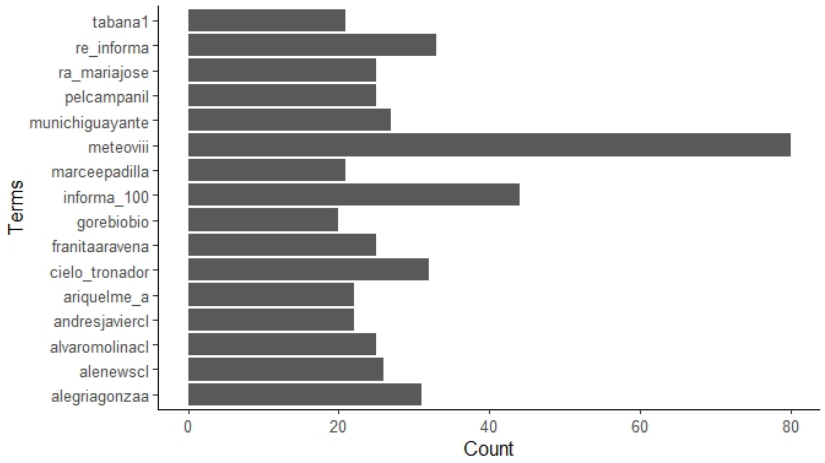


Figura 5.1: Usuarios más frecuentes - Incendio Chiguayante.

En el gráfico presentado en la Figura 5.1, primero podemos identificar a cuentas de organismos de gobierno como *@gorebiobio*, que corresponde a la cuenta oficial del Gobierno Regional del Bío-Bío (región es el equivalente geográfico a una comunidad autónoma). También encontramos tweets emitidos por la cuenta *@munichiguayante* que pertenece a la municipalidad

(ayuntamiento) de la ciudad de Chiguayante. Por otra parte, es posible encontrar cuentas asociadas a medios de comunicación tales como *@meteoviii*, cuya finalidad es entregar información meteorológica y *@Informa100*, cuya finalidad es entregar reportes de emergencias, ambas enfocadas en la Región del Bío-Bío. Las demás cuentas corresponden a usuarios individuales de Twitter.

La segunda pregunta a responder tiene relación con los fines con que los usuarios de Twitter usaron la red social durante las situaciones de emergencia en estudio. Para ello, lo primero que se ha realizado es definir cuatro finalidades en las que clasificar los resultados obtenidos. Esta clasificación se presenta en el Cuadro 5.3 a continuación:

Categoría de uso	**Descripción**
Qué sucede	Este uso incluye todos aquellos tópicos relacionados con lo que está sucediendo
Dónde sucede	Este uso incluye todos aquellos tópicos relacionados con identificar el lugar del suceso
Cuándo sucede	Este uso incluye todos aquellos tópicos que tienen como finalidad informar de la temporalidad del suceso
Otros	Se incluyen tópicos misceláneos

Cuadro 5.3: Resumen de tipos de usos de Twitter.

Con los tipos de usos ya definidos, se han obtenido dos gráficas cuya finalidad es generar una noción de lo que contiene la información que se ha recuperado. La primera de ellas corresponde a un *wordcloud*, presentado en la Figura 5.2, la cual si bien es una herramienta de carácter altamente visual y muy subjetiva en su interpretación, permite realizar una primera aproximación de hacia dónde podrían estar orientados los resultados.

Figura 5.2: Representación visual términos mas frecuentes - Incendio Chiguayante.

Del *wordcloud* presentado en la Figura 5.2 es posible apreciar que las tres palabras que más se repiten son *incendio, lonco* y *forestal*. Esto se ha corroborado con el gráfico presentada en la Figura 5.3 en el que vemos que las palabras que presentan una mayor frecuencia son *incendio, concepción, forestal, lonco* y *sector*.

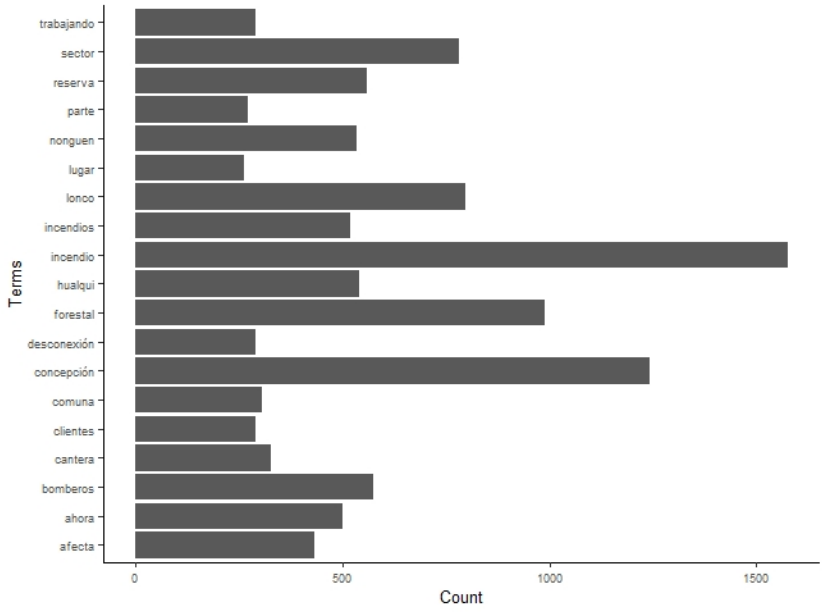
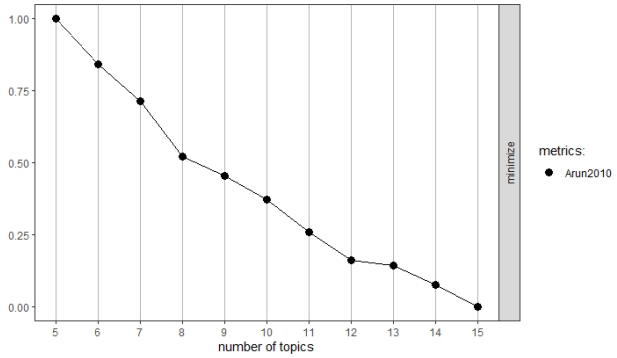
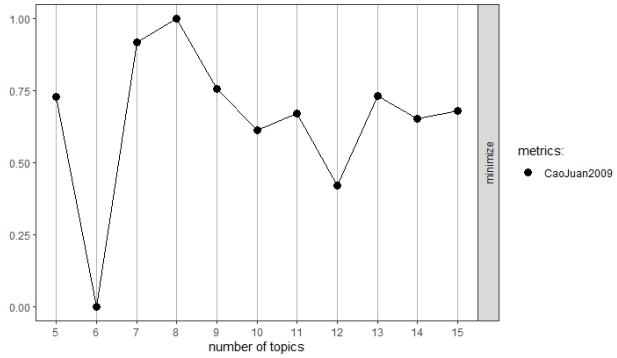


Figura 5.3: Gráfico de barras términos más frecuentes - Incendio Chiguayante.

Posteriormente, utilizando la *document-term matrix* se han obtenido los valores de las métricas de minimización, Figura 5.4, y los valores de las métricas de maximización, Figura 5.5, cuyos gráficos se presentan a continuación.

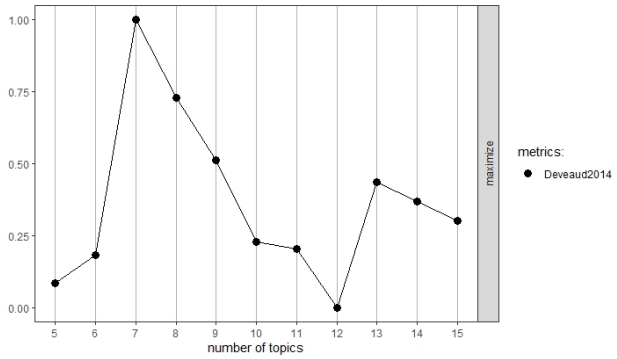


(a) Arun 2010

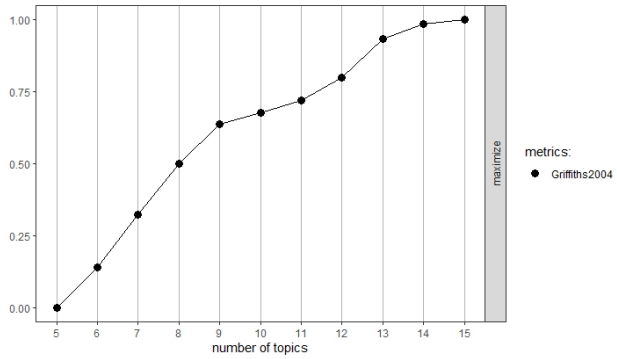


(b) CaoJuan 2009

Figura 5.4: Métricas de minimización - Incendio Chiguayante.



(a) Deveaud 2014



(b) Griffiths 2004

Figura 5.5: Métricas de maximización - Incendio Chiguayante.

Luego de obtenidas las cuatro métricas planteadas anteriormente, seleccionamos el valor más adecuado para cada una de ellas, que para fines de este estudio, se considerará aquel valor que presente un cambio en la tendencia.

Métrica	**Tipo**	**k**
Arun	Minimización	7
CaoJuan	Minimización	6
Deveaud	Maximización	7
Griffiths	Maximización	9

Cuadro 5.4: Métricas obtenidas con LDA Tuning - Incendio Chiguayante.

Como se mencionaba anteriormente, el criterio de elección del valor a considerar en cada métrica ha tomado aquel punto en que exista un cambio en la evolución de la inercia, es decir, donde se produzca un quiebre en la tendencia. Después de obtener estas métricas, se han obtenido cuatro valores para k presentados en el cuadro 5.4, en donde las métricas de *Arun* y *Deveaud* tienen un valor de k igual a 7, *CaoJuan* igual a 6 y finalmente *Griffiths* igual a 9. En este caso, el valor de k seleccionado será 7, considerando que es el que más se repite. Por otra parte, por cada uno de los tópicos se han seleccionado las cinco palabras más representativas, con lo que los tópicos obtenidos por el algoritmo LDA correspondientes al hashtag *#Chiguayante* quedan conformados según se muestra a continuación en la Figura 5.6.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	incendio	nacional	incendios	hualqui	nonguen	ahora	concepción
2	forestal	incendioforestal	forestales	concepción	sector	fuego	cerros
3	lonco	medio	comuna	desconexión	incendio	corte	momentos
4	bomberos	muestra	actualización	clientes	reserva	días	zona
5	sector	adicta	combate	parte	lonco	noche	lucen

Figura 5.6: Tópicos obtenidos - Incendio Chiguayante.

La Figura 5.6 muestra cada uno de los 7 tópicos y las 5 palabras que lo componen. Además, la Figura A.1 en anexo, muestra el valor del coeficiente *beta*, que presenta la importancia o peso de cada una de las palabras dentro del tópico al que pertenece. Con lo anterior es posible observar que en todos los tópicos obtenidos, a excepción del tópico 4 y el tópico 7, se menciona la ocurrencia de *incendios*. En todos los tópicos, con excepción del tópico 2 y el tópico 6, se hace mención a ubicaciones, las cuales si bien difieren, se refieren a zonas cercanas y relacionadas con la ocurrencia de los incendios acaecidos. Por último, los tópicos 3, 6 y 7 hacen mención a momentos en lo que la emergencia está sucediendo, utilizando palabras como *ahora, noche, días, momentos, etc.*

Resumiendo lo anterior de manera porcentual, encontramos que los tópicos detectados por el algoritmo LDA están compuestos en un 37 % por palabras relacionadas a *Dónde* está ocurriendo la emergencia, seguida por *Qué* está ocurriendo con un 34 %. Finalmente encontramos las palabras relacionadas a *Cuándo* y *Otros* ambas con un 14 % de los usos. El Cuadro 5.5 presenta estos valores totales, y el Cuadro A.1 del anexo presenta los valores obtenidos para cada tópico.

Uso	Porcentaje
Qué sucede	34 %
Dónde sucede	37 %
Cuándo sucede	14 %
Otros	14 %

Cuadro 5.5: Cuadro uso porcentual - Incendio Chiguayante.

Finalmente, se ha buscado determinar si existe diferencia entre los tópicos detectados a medida que transcurre el tiempo en una situación de emergencia. Para ello, consideraremos dos criterios de selección de tweets, *tweets emitidos durante la emergencia*, que considerará tweets emitidos el primer día de emergencia, en este caso el día 27/01/2020, y los *tweets emitidos después la emergencia*, que para este caso considerará los tweets emitidos entre los días 29/01/2020 al 30/01/2020. En otras situaciones sería importante realizar el análisis considerando además los tweets emitidos *antes*, pero teniendo en cuenta que se trata de situaciones de emergencia, la información a la que las personas se refieren antes de la ocurrencia no resulta de relevancia alguna para realizar análisis de este tipo. Por lo anterior, se han obtenido las figuras con tópicos según su emisión en el tiempo que se adjuntan a continuación:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	bomberos	incendio	concepción	sector	incendios	fuego	incendio
2	corte	forestal	hualqui	reserva	forestales	concepción	lonco
3	incendio	ahora	desconexión	lonco	concepción	días	casa
4	ahora	comuna	clientes	nonguen	horas	hace	lado
5	lonco	concepcion	parte	incendio	carabineros	noche	llego

Figura 5.7: Tópicos obtenidos el día 27/01/2020 - Incendio Chiguayante.

Al mirar con detenimiento los tópicos obtenidos por el algoritmo LDA en la Figura 5.7, es posible encontrar información bastante interesante. Lo primero es que se encuentran las palabras *incendio* y *ahora* tanto en el tópico 1 como en el tópico 2, lo cual se puede traducir concluyendo la temporalidad de la emergencia que se esta ocurriendo, es decir, *el incendio está ocurriendo ahora*. Respecto de la ubicación de la emergencia, el tópico 1 contiene la palabra *lonco*, la cual corresponde a una una ubicación especifica (Lonco es un sector que se encuentra en el límite de las comunas de Concepción y Chiguayante), mientras que en los tópicos 2 y 3 se menciona *concepción*, que corresponde a una ubicación más general. Otra cosa que es importante

mencionar es que *bomberos* aparece mencionado en el tópico 1 y *carabineros* aparece mencionado en el tópico 5, siendo las únicas dos instituciones que se mencionan en los tópicos obtenidos, no apareciendo ninguna organización perteneciente al gobierno (ministerio, onemi, subsecretaria, etc). La Figura A.2 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.7.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	incendio	gust	incendios	villuco	sector	muestra	comunas
2	forestal	humidity	comuna	incendio	nonguen	adicta	carabineros
3	santa	temperature	forestales	sector	reserva	atacama	primera
4	concepción	pressure	combate	hora	incendioforestal	catástrofes	zona
5	afecta	wind	buenas	actualización	nacional	fijado	aérea

Figura 5.8: Tópicos obtenidos entre los días 28/01/2020 - 30/01/2020 - Incendio Chiguayante.

Ahora nos enfocamos en la información mostrada en la Figura 5.8 en la cual se presentan los tópicos extraídos en el intervalo de tiempo comprendido entre los días 28 al 30 de Enero del año 2020.

Es posible apreciar que, si bien los tópicos presentados continúan entregando información similar a las figuras anteriores referente a la ubicación y a lo que está ocurriendo, las palabras que hacen mención a temporalidad ya no denotan ocurrencia en presente, a diferencia de lo que podíamos ver en la figura 5.7. En relación a las palabras relacionadas a instituciones gubernamentales, solo encontramos la palabra *carabineros* en el tópico número 7, lo cual denota una muy baja cantidad de menciones. Esto fortalece lo que se planteó anteriormente referente a la baja cantidad de menciones a instituciones gubernamentales. Por ultimo, es posible encontrar dos tópicos que no hacen mención a la emergencia, tópico 2 y tópico 6, lo que lleva a concluir que la importancia que se ha dado en este periodo de tiempo a la emergencia ha disminuido en comparación al día en que se desencadenó. La Figura A.3 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.8.

Finalmente, si se realiza la comparación porcentual entre ambos grupos de tópicos es posible apreciar que existe una disminución de las menciones en los tópicos asociados a *Qué sucede*, *Dónde sucede* y *Cuándo sucede*, mientras que los tópicos asociados a *Otros* aumentan de manera considerable. Es posible apreciar estos resultados en el Cuadro 5.6. Los Cuadros A.2 y A.3 de los anexos presentan los resultados para cada conjunto de tópicos.

Uso	Durante	Después	Variación
Qué sucede	46 %	37 %	-9 %
Dónde sucede	37 %	34 %	-3 %
Cuándo sucede	17 %	6 %	-11 %
Otros	0 %	23 %	23 %

Cuadro 5.6: Cuadro de uso porcentual Durante/Después - Incendio Chiguayante.

5.4. Dataset - Incendio Hualqui

El siguiente dataset con el que se trabajará será el correspondiente a *Hualqui*. Se ha realizado todo el procesado de texto necesario hasta tener la *Document-term Matrix*, la cual entrega el siguiente cuadro resumen.

```
<<DocumentTermMatrix (documents: 2935, terms: 99)>>
Non-/sparse entries: 15147/275418
Sparsity           : 95%
Maximal term length: 19
Weighting          : term frequency (tf)
```

Código 5.2: Cuadro resumen DTM Hualqui

El cuadro resumen 5.2 entrega la siguiente información, existen 2935 documentos (cada tweet es un documento), con un total de 99 términos diferentes. Se tiene una tabla con 275.418 celdas cuyo valor podría ser 0 y 15.147 celdas tienen un valor mayor que 0, lo que supone un 95 % de las filas tienen como valor el 0. *Maximal term length* indica que la palabra más larga tiene 19 caracteres de longitud. Finalmente, *term frquency* indica que la forma de considerar los términos es por medio de su frecuencia.

En una primera parte del análisis nos enfocaremos en los usuarios más frecuentes que han emitido tweets utilizando el tópico *#Hualqui*. Para ello se ha obtenido el gráfico de barras que se presenta a continuación:

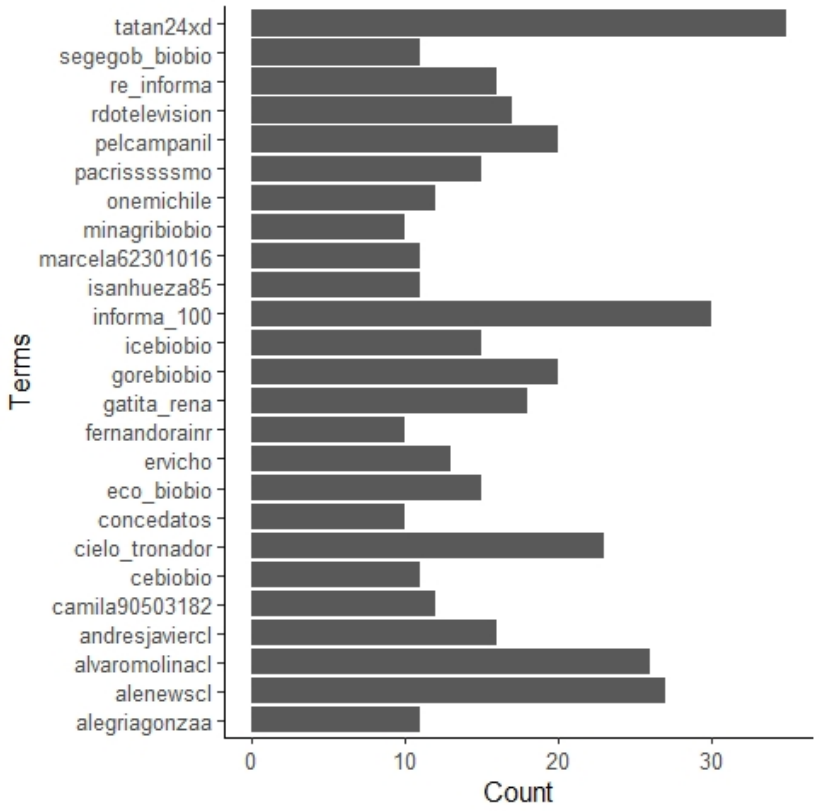


Figura 5.9: Usuarios mas frecuentes - Incendio Hualqui.

Del gráfico presentado en la Figura 5.9 podemos identificar cuentas de organismos de gobierno como son *@segregob_biobio*, *@onemichile*, *@minagribiobio* y *@gorebiobio*. También encontramos tweets emitidos por cuentas informativas como *@re_informa*, *@informa_100*, *@eco_biobio*, *@concedatos*, *@cebiobio* y *@alenewscl*. Además se han detectado tweets emitidos por organizaciones no gubernamentales como lo es *@icebiobio*. Los restantes usuarios corresponden a usuarios individuales.

Respecto de la segunda pregunta planteada, se ha obtenido el *wordcloud* presentado en la Figura 5.10, en el cual se identifican palabras como *incendio, forestales, alerta*, etc. como aquellos términos que más se presentan dentro del dataset.

Figura 5.10: Representación visual términos mas frecuentes - Incendio Hualqui.

Posteriormente, la Figura 5.11, presenta la gráfica en la que se muestran los términos más frecuentes, que como indica el wordcloud son *incendios, concepción, alerta* y *forestales.*

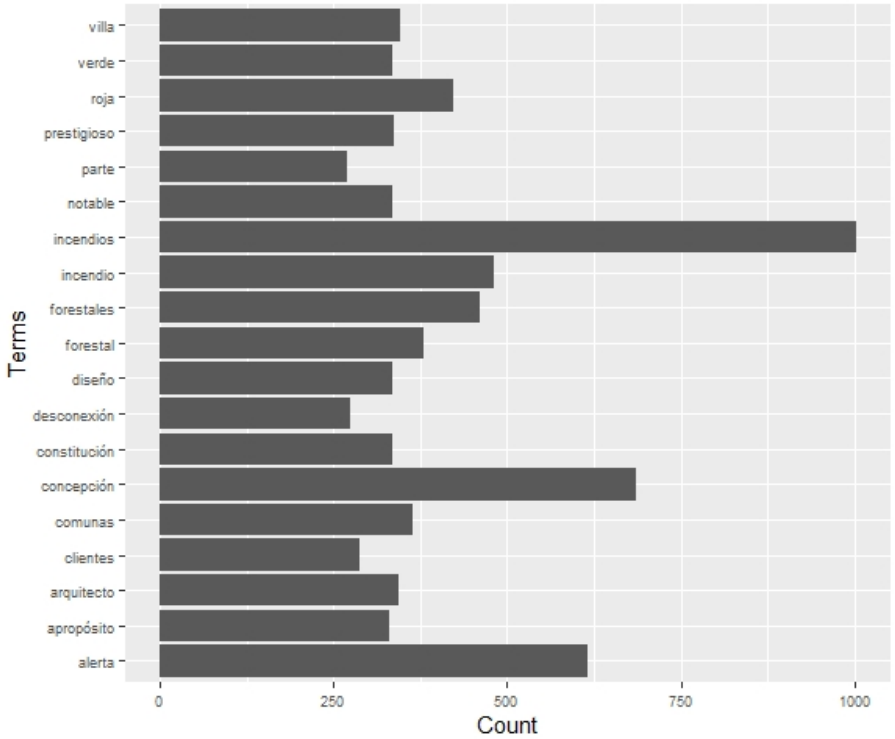
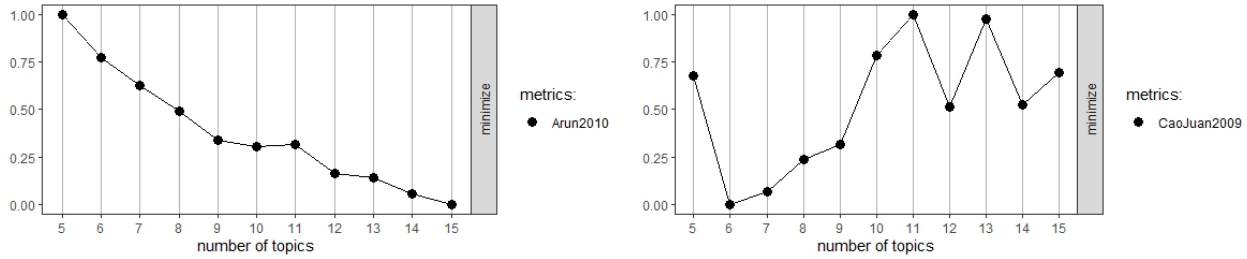


Figura 5.11: Gráfico de barras términos más frecuentes - Incendio Hualqui.

Una vez realizadas las gráficas anteriores, se han obtenido, por medio del paquete *LDA tuning* los valores de los estadísticos que nos ayudarán a definir

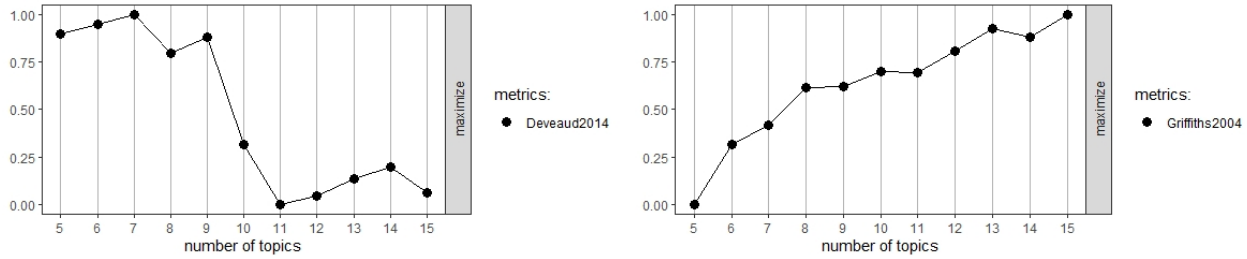
la cantidad de tópicos que se obtendrán por medio del algoritmo LDA. En la Figura 5.12 se presentan las gráficas correspondientes a los estadísticos de minimización.



(a) Arun 2010 (b) CaoJuan 2009

Figura 5.12: Métricas de minimización - Incendio Hualqui.

En la Figura 5.13 se presentan las gráficas correspondientes a los estadísticos de maximización.



(a) Deveaud 2014 (b) Griffiths 2004

Figura 5.13: Métricas de maximización - Incendio Hualqui.

En el cuadro 5.7 se muestran los valores seleccionados para cada una de las métricas anteriormente presentadas.

Métrica	**Tipo**	**k**
Arun	Minimización	5
CaoJuan	Minimización	5
Deveaud	Maximización	7
Griffiths	Maximización	7

Cuadro 5.7: Métricas obtenidas con LDA Tuning - Incendio Hualqui.

Después de obtener las métricas de *Arun, CaoJuan, Deveaud* y *Griffiths*, el valor seleccionado de k en este caso ha sido 6, y al igual que con el dataset anterior, se seleccionaron las cinco palabras más relevantes de cada uno de los 6 tópicos. Finalmente, los tópicos obtenidos para los datos correspondientes al hashtag *Hualqui* han quedado de la siguiente forma:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	ahora	alerta	villa	incendios	concepción	incendio
2	región	roja	arquitecto	forestales	clientes	forestal
3	intendente	comunas	prestigioso	curanilahue	desconexión	hectáreas
4	regional	incendios	diseño	activos	parte	bomberos
5	comuna	amarilla	notable	santajuana	informamos	sector

Figura 5.14: Tópicos obtenidos - Incendio Hualqui.

La figura 5.14 muestra cada uno de los 6 tópicos y las 5 palabras que lo componen. Además, la figura A.4 muestra el valor del coeficiente *beta* para cada una de las palabras que componen cada tópico.

Si nos enfocamos en la información entregada por los tópicos obtenidos, podemos ver que el tópico 1 no es tan claro al incluir palabras como *región, comuna* e *intendente*, lo cual no nos proporciona información muy relevante, pero por otra parte incluye la palabra *ahora*, que como mencionamos anteriormente, implica que lo que está sucediendo es en tiempo presente. Ahora, si nos enfocamos en el tópico número 2, es posible encontrar información que presenta un grado mayor de claridad, ya que encontramos palabras como *incendios*, y *alerta*, además de las palabras *amarilla* y *roja*, lo cual denota la ocurrencia de un incendio, cuya magnitud ha variado, ya que se mencionan tanto la *alerta amarilla* (nivel medio de complicación) como la *alerta roja* (nivel mayor de complicación). El tópico 3 no hace mención a la emergencia, a diferencia del tópico 4 que habla de *incendios forestales activos* además de mencionar dos ubicaciones como son *curanilahue* y *Santa Juana*. Finalmente, podemos ver que los tópicos 5 y 6 están relacionados con la emergencia y podemos encontrar palabras como *incendios, forestal, concepción* y *sector*. Es importante indicar que en el tópico 1 se menciona la palabra *intendente* y sólo en el tópico 6 se menciona la palabra *bomberos*, ya que ambas son las únicas menciones que reciben organismos gubernamentales, en el caso de intendente y una institución encargada del control de incendios como lo es bomberos.

Si se plantea lo anterior de manera porcentual, los tópicos detectados por el algoritmo LDA están compuestos en un 53 % por palabras relacionadas a

Qué está ocurriendo, seguida por un 27 % de palabras relacionadas a *Dónde* está ocurriendo la emergencia y solamente un 3 % de palabras relacionadas a *Cuándo* sucede. Las palabras relacionadas a *Otros* se llevan un 17 % de las menciones. El Cuadro 5.8 presenta estos valores totales, y el Cuadro A.4 del anexo presenta los valores obtenidos para cada tópico.

Uso	**Porcentaje**
Qué sucede	53 %
Dónde sucede	27 %
Cuándo sucede	3 %
Otros	17 %

Cuadro 5.8: Cuadro uso porcentual - Incendio Hualqui.

Al igual que con el dataset anterior, el análisis se ha realizado sin considerar el factor tiempo, por lo que el siguiente análisis se ha realzado considerando dos dataset, el primero está formado por los tweets emitidos durante la emergencia, que considerará tweets emitidos el día 25/01/2020. El segundo está compuesto por los tweets emitidos después de la emergencia, entre los días 26/01/2020 al 30/01/2020. En las Figuras 5.15 y 5.16 se muestran los temas detectados teniendo en cuenta los dos intervalos temporales mencionados.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	alerta	roja	incendio	incendios	alerta	forestal
2	roja	alerta	forestales	comunas	comunas	incendio
3	declara	amarilla	hectáreas	roja	incendios	onofre
4	incendios	forestales	sector	incendio	forestal	concepción
5	nuest	provincial	forestal	nuest	roja	sector

Figura 5.15: Tópicos obtenidos el día 25/01/2020 - Incendio Hualqui.

Si se consideran los tópicos presentados en la Figura 5.15, correspondientes sólo a los tweets emitidos el día 25 de enero, podemos ver que por ejemplo el tópico 1 se refiere a la declaración de alerta roja por incendios, lo cual es bastante similar al tópico número 2 y número 3 en los cuales se hace mención a alerta amarilla y roja en el primero, e incendio en sector forestal en el segundo. Por último, los tópicos número 4, 5 y 6, hacen referencia a lo mismo utilizando palabras como incendios forestales y alerta roja, además

de mencionar comunas, sector y concepción, con lo que se refiere a ubicación de la emergencia. Otro punto importante a comentar es que los tópicos entregados no hacen mención a bomberos o carabineros ni tampoco aluden a organizaciones gubernamentales. La Figura A.5 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.15.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	incendio	villa	región	concepción	curanilahue	alerta
2	forestal	arquitecto	ahora	clientes	incendios	incendios
3	bomberos	prestigioso	incendiosforestales	desconexión	activos	forestales
4	concepción	diseño	corte	parte	forestales	roja
5	hectáreas	notable	actualización	informamos	comunas	concepción

Figura 5.16: Tópicos obtenidos entre los días 26/01/2020 - 30/01/2020 - Incendio Hualqui.

En la Figura 5.16 se presentan los tópicos extraídos en el intervalo de tiempo comprendido entre los días 26 y 30 de enero del año 2020. Lo primero que es posible apreciar, es que a diferencia de lo que se presentaba en la Figura 5.15, alerta roja sólo se menciona en el tópico número 6, lo que se podría interpretar como una baja en la intensidad de la emergencia. Referente al tópico número 1, junto a incendio forestal y concepción se menciona bomberos, lo cual en los tópicos obtenidos en la Figura 5.15 no sucedía. Los tópicos número 3, número 4 y número 5 hacen mención a los incendios forestales y menciona ubicaciones como curanilahue y concepción. Finalmente, el tópico número 2 no presenta relación con la emergencia, lo cual podría significar una disminución en los tweets que se emiten relacionados a la emergencia.

Por otra parte, podemos afirmar que, si bien la información entregada por las Figuras 5.14, 5.15 y 5.16 es similar, podemos ver que al realizar la separación utilizando intervalos de tiempo, los tópicos que aparecen en el primer intervalo de tiempo están relacionados a la temporalidad presente del suceso. En este caso en particular, podemos mencionar que el término alerta roja se mencionaba en 4 de los 6 tópicos extraídos. Referente a la mención a la emergencia (incendio) en el primer conjunto de tópicos presentados en la Figura 5.15 se menciona en todos ellos, a diferencia de los tópicos de la Figura 5.16 en donde sólo se menciona en 4 de los 6 tópicos, es decir, ha disminuido la cantidad de tweets que hacen referencia a los incendios forestales. Por último, y al igual que el dataset de Chiguayante, no existen tópicos relacionados a instituciones gubernamentales, y sólo encontramos mencionada la palabra *bomberos* en el tópico número 6, tanto en la Figura 5.14 como de

la Figura 5.16. La Figura A.6 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.16.

Finalmente, en la comparación porcentual entre ambos grupos de temas detectados es posible apreciar que existe una disminución de las menciones en los tópicos asociados a *Qué sucede* aproximadamente de un 17 %. Los tópicos referentes a *Dónde sucede* se mantienen, mientras que los tópicos correspondientes a *Cuándo sucede* presentan un aumento de un 7 %. Los tópicos asociados a *Otros* temas aumentan en un 10 %. Estos resultados se presentan en el Cuadro 5.9, mientras que los Cuadros A.5 y A.6 de los anexos muestran los resultados para cada conjunto de tópicos.

Uso	Durante	Después	Variación
Qué sucede	70 %	53 %	-17 %
Dónde sucede	20 %	20 %	0 %
Cuando sucede	0 %	7 %	7 %
Otros	10 %	20 %	10 %

Cuadro 5.9: Cuadro de uso porcentual Durante/Después - Incendio Hualqui.

5.5. Dataset - Incendio Nonguen

El dataset con que se trabajara a continuación corresponde al asociado a *Nonguen*. Lo primero que haremos es todo el procesado de texto necesario hasta tener la *Document-term Matrix*, la cual muestra el siguiente cuadro resumen.

```
1 <<DocumentTermMatrix (documents: 4592, terms: 165)>>
2 Non-/sparse entries: 27855/729825
3 Sparsity           : 96%
4 Maximal term length: 19
5 Weighting          : term frequency (tf)
```

Código 5.3: Cuadro resumen DTM Nonguen.

El Código 5.3 entrega la siguiente información: existen 4.592 documentos, con un total de 165 términos diferentes. Se tiene una tabla con 729.825 celdas cuyo valor podría ser 0 y 27.855 celdas tienen un valor mayor que 0, lo que supone un 96 % de las filas tienen como valor el 0. *Maximal term length* indica que la palabra más larga tiene 19 caracteres de longitud. Finalmente, *term*

frquency indica que la forma de considerar los términos es por medio de su frecuencia.

Al igual que los dataset anteriores, lo primero será enfocar el análisis en los usuarios más frecuentes que han emitido tweets utilizando el tópico *#Nonguen*. Para ello se ha obtenido el gráfico de barras que se presenta a continuación:

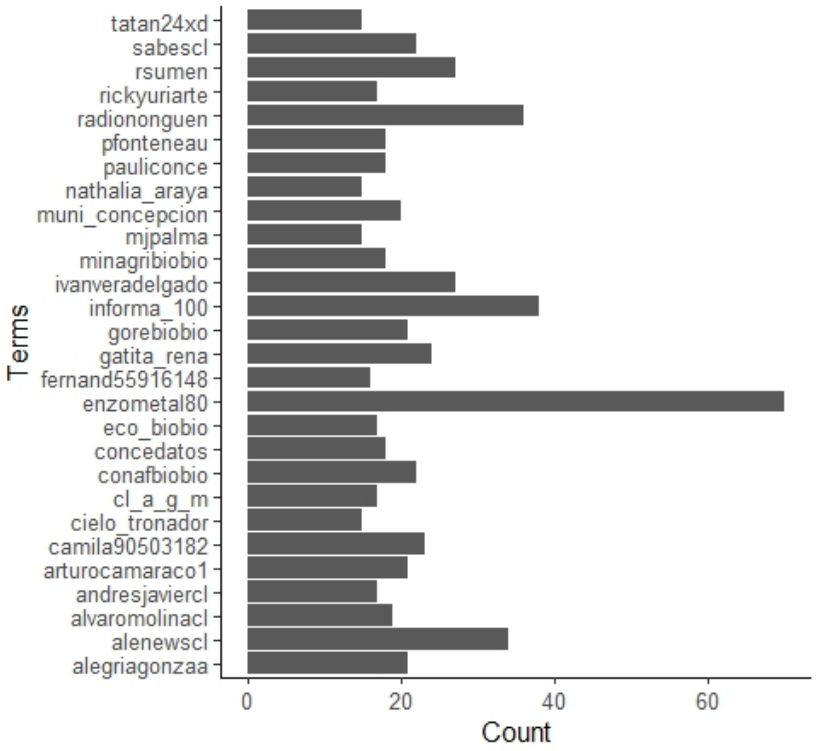


Figura 5.17: Usuarios mas frecuentes - Incendio Nonguen.

Del gráfico presentado en la Figura 5.17 podemos identificar cuentas de organismos de gobierno como son *@muni_concepcion*, *@minagribiobio*, *@gorebiobio* y *@conafbiobio*. También encontramos tweets emitidos por cuentas informativas como *@sabescl*, *@informa_100*, *@eco_biobio*, *@concedatos*, *@radiononguen* y *@alenewscl*. Los restantes usuarios corresponden a usuarios individuales.

Después se ha obtenido un *wordcloud*, el cual se presenta en la Figura 5.18.

Figura 5.18: Representación visual términos más frecuentes - Incendio Nonguen.

Es posible apreciar que las palabras que resaltan mayormente en el Wordcloud son *reserva, incendio, forestal* y *concepción*. Además del wordcloud, presentaremos un gráfico de frecuencia de palabras para corroborar lo que se ha visto por medio de la nube de palabras.

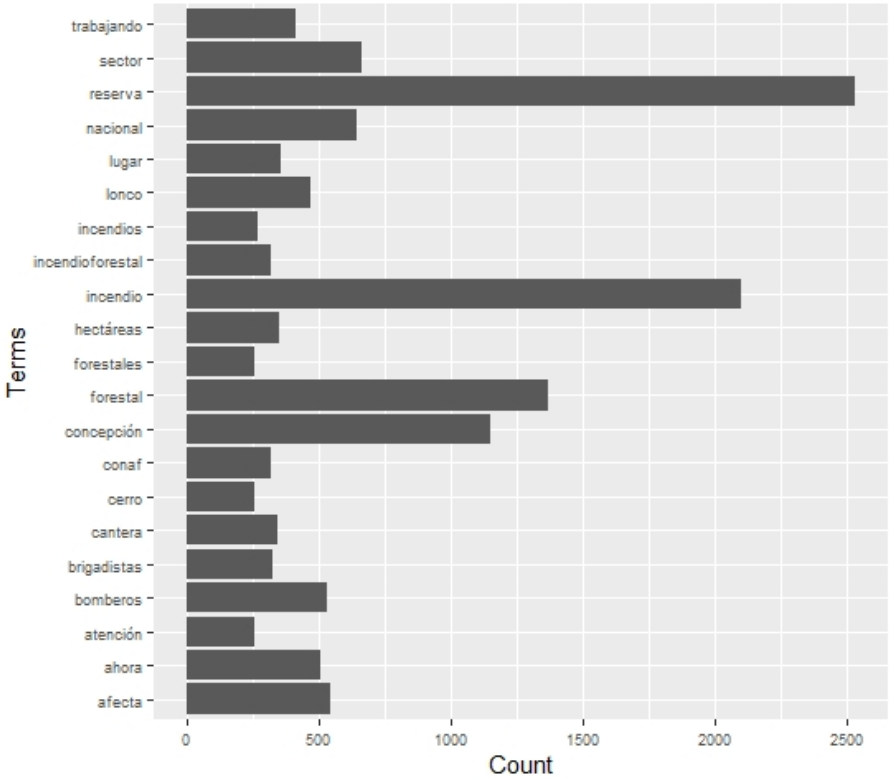
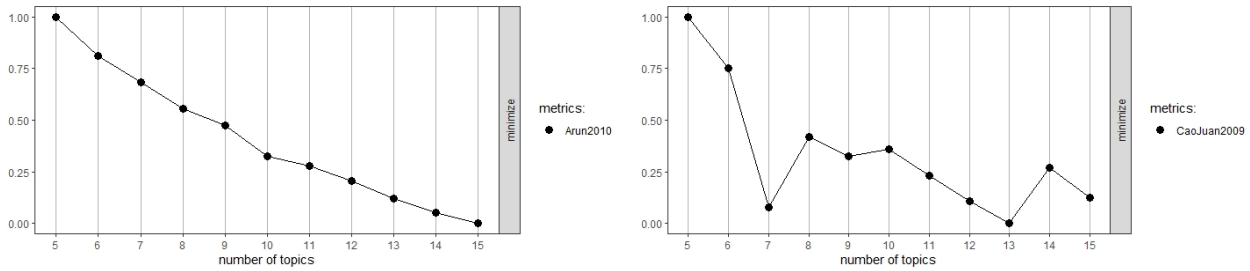


Figura 5.19: Gráfico de barras términos mas frecuentes - Incendio Nonguen.

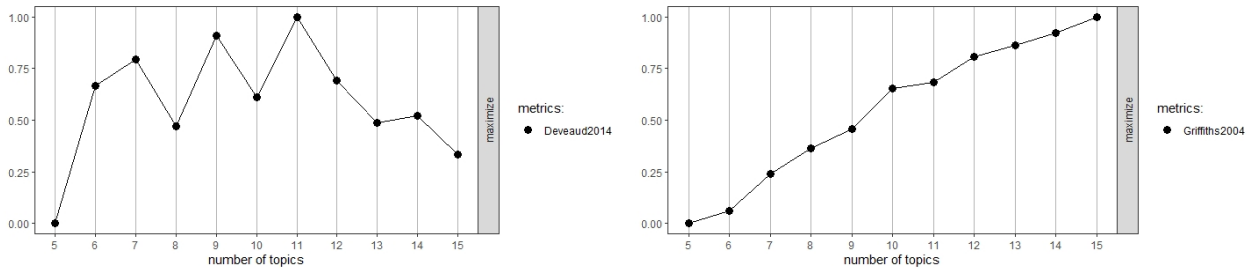
Coincidentemente, las palabras que se han mencionado en el wordcloud como las que se aprecian con mayor facilidad, son aquellas que se presentan en la Figura 5.19, *reserva, incendio, forestal* y *concepción*.

Por medio de la utilización de las 4 métricas anteriormente mencionadas, se han obtenido el número de tópicos para el algoritmo LDA. Las Figuras obtenidos se adjuntan a continuación. La Figura 5.20 presenta las métricas de minimización y la Figura 5.21 presenta las métricas de maximización.



(a) Arun 2010 (b) CaoJuan 2009

Figura 5.20: Métricas de minimización - Incendio Nonguen.



(a) Deveaud 2014 (b) Griffiths 2004

Figura 5.21: Métricas de maximización - Incendio Nonguen.

Los valores según cada una de las métricas se presentan en el cuadro 5.10 a continuación:

Métrica	Tipo	k
Arun	Minimización	6
CaoJuan	Minimización	5
Deveaud	Maximización	7
Griffiths	Maximización	7

Cuadro 5.10: Métricas obtenidas con LDA Tuning - Incendio Nonguen.

En este caso, el valor de k seleccionado será 7, con lo que los tópicos obtenidos para los datos correspondientes al hashtag *Nonguen* son:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	incendio	reserva	forestal	incendio	incendio	afecta	reserva
2	reserva	nacional	reserva	sector	forestal	brigadistas	concepción
3	sector	además	bomberos	reserva	concepción	medio	incendio
4	forestales	mapa	concepción	lonco	ahora	nacional	hectáreas
5	sábado	ilustrar	sector	trabajando	reserva	junto	forestal

Figura 5.22: Tópicos obtenidos - Incendio Nonguen.

La Figura 5.22 muestra cada uno de los 7 tópicos y las 5 palabras que lo componen. Además, la Figura A.7, en los anexos, muestra el valor del coeficiente *beta* para cada una de las palabras que componen cada tópico.

La información mostrada por los tópicos obtenidos con el algoritmo LDA se describe a continuación. Referente al tópico 1, es posible apreciar que se menciona *incendio, forestal, sector, reserva, Sábado*, es decir, se hace mención a la emergencia adjuntando además de la ubicación el día de ocurrencia. La información mostrada por los tópicos 3, 4, 5, 6 y 7 es similar y podemos ver repetidas palabras tales como *incendio, forestal, afecta* para hacer mención a la emergencia, *reserva nacional, concepción*, para hacer mención a la ubicación y palabras como *sábado, ahora* para hacer mención a cuándo está sucediendo dicha emergencia. Es importante mencionar que los organismos encargados de la contención de incendios tan sólo se ven mencionados en el tópico 3 *bomberos*, y en el tópico 6 *brigadistas*, lo cual es consecuente con los anteriores dataset, en que las menciones a los organismos encargados a trabajar en el control de la emergencia y organismos del gobierno relacionados es muy baja.

Si se presenta lo anterior de manera porcentual, los tópicos detectados por el algoritmo LDA están compuestos en un 37 % por palabras relacionadas a *Qué* está ocurriendo, seguida por un 34 % de palabras relacionadas a *Dónde*

está ocurriendo la emergencia y un 11 % de palabras relacionadas a *Cuándo* sucede. Las palabras relacionadas a *Otros* se llevan un 17 % de las menciones. El Cuadro 5.11 presenta estos valores, y el Cuadro A.7 del anexo presenta los valores obtenidos para cada tópico.

Uso	**Porcentaje**
Qué sucede	37 %
Dónde sucede	34 %
Cuándo sucede	11 %
Otros	17 %

Cuadro 5.11: Cuadro uso porcentual - Incendio Nonguen.

El análisis que permita responder a la tercera pregunta planteada será similar al realizado con los anteriores dataset de Chiguayante y Hualqui, es decir, se considerara la fecha de emisión de los tweets para dividir los datos en dos conjuntos. El primero será considerando los tweets emitidos durante la emergencia, que considerará tweets emitidos el primer día de emergencia, el día 25/01/2020. El segundo será considerando los tweets emitidos después la emergencia, entre los días 26/01/2020 al 30/01/2020. Después de realizado esto, los resultados obtenidos se muestran en la Figura 5.23.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	nacional	reserva	reserva	reserva	incendio	concepción	acceso
2	reserva	ahora	incendio	atención	forestal	reserva	sábado
3	además	bosque	bomberos	animales	cerro	hacia	preventiva
4	contexto	nativo	forestal	junto	hectáreas	incendio	nacional
5	mapa	incendio	lugar	caso	manquimávida	incendioforestal	cerrada

Figura 5.23: Tópicos obtenidos el 25/01/2020 - Incendio Nonguen.

Si se consideran los tópicos presentados en la Figura 5.23, correspondientes solo a los tweets emitidos el día 25 de enero, es posible apreciar que el tópico 1, si bien hace mención a una ubicación, *reserva nacional*, no hace mención a la ocurrencia de una emergencia, a diferencia de los tópicos 2 y 3 que hacen clara alusión a lo que sucede en ese momento, *incendio, reserva, bosque, nativo, ahora*. Por su parte, el tópico número 3 está compuesto por *incendio, lugar, reserva, forestal*, y es el único tópico en el cual se hace mención a *bomberos*, única mención a un organismo de encargado de trabajar en el control de la emergencia. Los tópicos número 4 y 7 no presentan información relevante para análisis. Por último, los tópicos número 5 y 6, si bien no

entregan información de manera tan clara, contienen palabras como *incendio forestal*, además de mencionar ubicaciones como *concepción, manquimavida, reserva* y *cerro*. Como comentario final, es posible afirmar que los tópicos obtenidos desde este dataset, si bien han entregado información, ésta no ha sido tan clara como se quisiera. Esto es debido a la existencia de tópicos que no contienen palabras relacionadas con la emergencia. La Figura A.8 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.23.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	incendio	trabajando	reserva	forestal	concepción	concepción	reserva
2	reserva	reserva	forestal	sector	forestal	lonco	incendio
3	sector	sector	nacional	lonco	afecta	incendio	forestal
4	trabajando	incendio	incendio	bomberos	incendio	forestal	brigadistas
5	nacional	seguimos	concepción	intendente	control	bomberos	monte

Figura 5.24: Tópicos obtenidos entre los días 26/01/2020 - 30/01/2020 - Incendio Nonguen.

Si nos enfocamos en la información mostrada por la Figura 5.24 en la cual se presentan los tópicos extraídos en el intervalo de tiempo comprendido entre los días 26 al 30 de enero del año 2020, lo primero que es posible apreciar es que el tópico número 1 hace referencia a un *incendio sector reserva nacional* y además menciona la palabras *trabajando*, con lo que es posible deducir que ya existe personal trabajando en el control de la emergencia. Por su parte el tópico número 2 hace referencia a un *incendio sector reserva* también mencionando que *seguimos trabajando*. El tópico número 3 hace referencia a la misma emergencia pero agrega además datos referentes a la ubicación de la emergencia, como lo es *concepción*. Referente a los demás tópicos podemos mencionar que por medio de palabras como *incendio* y *forestal* todos ellos hacen mención a la emergencia y con palabras como *sector, lonco, concepción* y *reserva* se hace referencia a la ubicación de la emergencia. A diferencia de los dataset anteriores, este set es el que presenta más menciones a organismos relacionados al control de emergencias como *bomberos* y *brigadistas*, además de mencionar por vez primera autoridades de gobierno como es *intendente*.

Puede apreciarse que la información entregada por cada una de las figuras en que se presentan los tópicos detectados por el algoritmo LDA es similar. Al realizar la partición de datos utilizando la fecha de emisión como criterio, los dataset tienen algunas diferencias. Por ejemplo, si nos enfocamos al tópico número 1, en la Figura 5.22 muestra información muy clara de lo que está sucediendo, además de dónde y cuándo está sucediendo, a diferencia del tópico número 1 entregado por la Figura 5.23 en que sólo se hace mención a

la ubicación, sin mencionar qué sucede y entregando además palabras que no tienen relación con la emergencia. Por último, el tópico número 1 mostrado en la Figura 5.24 es similar al mostrado en la Figura 5.22, ya que presenta información más clara y precisa de lo que está sucediendo, dónde está sucediendo, además de añadir la palabra *trabajando*, lo cual da a entender que por ese momento ya había personal de alguna institución trabajando para controlar la emergencia. En este dataset, los tópicos entregados en la Figura 5.23, que se han emitido en el primer día de emergencia, no están tan enfocados en la emergencia y presentan una cantidad de palabras que no tienen relación con la emergencia tales como *además, contexto, mapa, caso*, etc. por mencionar sólo algunas. Por otra parte, los tópicos presentados en la Figura 5.24 están enfocados en su totalidad a la emergencia, entregando información clara de lo que está sucediendo y dónde está sucediendo. Importante mencionar que este es el primer dataset en que se presenta una mención a un organismo gubernamental. La Figura A.9 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.24.

Por último, se ha realizado la comparación porcentual entre ambos grupos de tópicos, es posible apreciar que existe un aumento de las menciones en los tópicos asociados a *Qué sucede* y a *Dónde sucede*. En cambio los tópicos asociados a *Cuándo sucede* y *Otros* presentan un incremento. El Cuadro 5.12 presenta los valores obtenidos, mientras que los Cuadros A.8 y A.9 del anexo presentan los resultados para cada conjunto de tópicos.

Uso	**Durante**	**Después**	**Variación**
Qué sucede	51 %	54 %	3 %
Dónde sucede	31 %	46 %	14 %
Cuándo sucede	6 %	0 %	-6 %
Otros	11 %	0 %	-11 %

Cuadro 5.12: Cuadro de uso porcentual Durante/Después - Incendio Nonguen.

5.6. Dataset - Incendio Santa Juana

El último dataset a analizar será el que contiene los datos del incendio de *Santa Juana*. Para ello, lo primero será realizar todo el procesado de texto necesario hasta obtener la *Document-term Matrix*, la cual genera el siguiente cuadro resumen.

```
<<DocumentTermMatrix (documents: 1656, terms: 61)>>
Non-/sparse entries: 7634/93382
Sparsity           : 92%
Maximal term length: 19
Weighting          : term frequency (tf)
```

Código 5.4: Cuadro resumen DTM Santa Juana.

El Código 5.4 presentado muestra la siguiente información, existen 1.656 documentos, con un total de 61 términos diferentes. Se tiene una tabla con 93.382 celdas cuyo valor podría ser 0 y 7.634 celdas tienen un valor mayor que 0, lo que supone un 92 % de las filas tienen como valor el 0. *Maximal term length* indica que la palabra más larga tiene 19 caracteres de longitud. Finalmente, *term frquency* indica que la forma de considerar los términos es por medio de su frecuencia.

Al igual que los dataset anteriores, lo primero será enfocar el análisis en los usuarios más frecuentes que han emitido tweets utilizando el tópico *#SantaJuana*. Para ello se ha obtenido el gráfico de barras que se presenta a continuación:

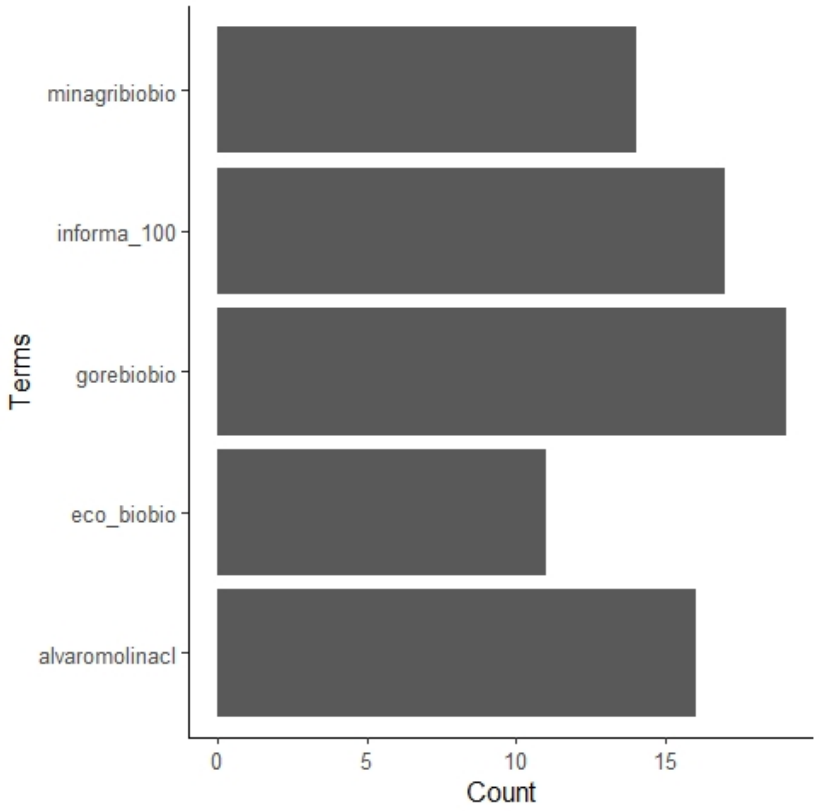


Figura 5.25: Usuarios más frecuentes - Incendio Santa Juana.

En la Figura 5.25 podemos identificar cuentas de organismos de gobierno como son *@minagribiobio* y *@gorebiobio*. También encontramos tweets emitidos por cuentas informativas como *@informa_100* y *@eco_biobio*. Los restantes usuarios corresponden a usuarios individuales.

La Figura 5.26 muestra el *wordcloud* obtenido, con el cual tendremos una primera aproximación a lo que serán nuestros resultados.

Figura 5.26: Representación visual términos mas frecuentes - Incendio Santa Juana.

A continuación, la Figura 5.27 muestra un gráfico de barras en el que se presentan los términos más utilizados.

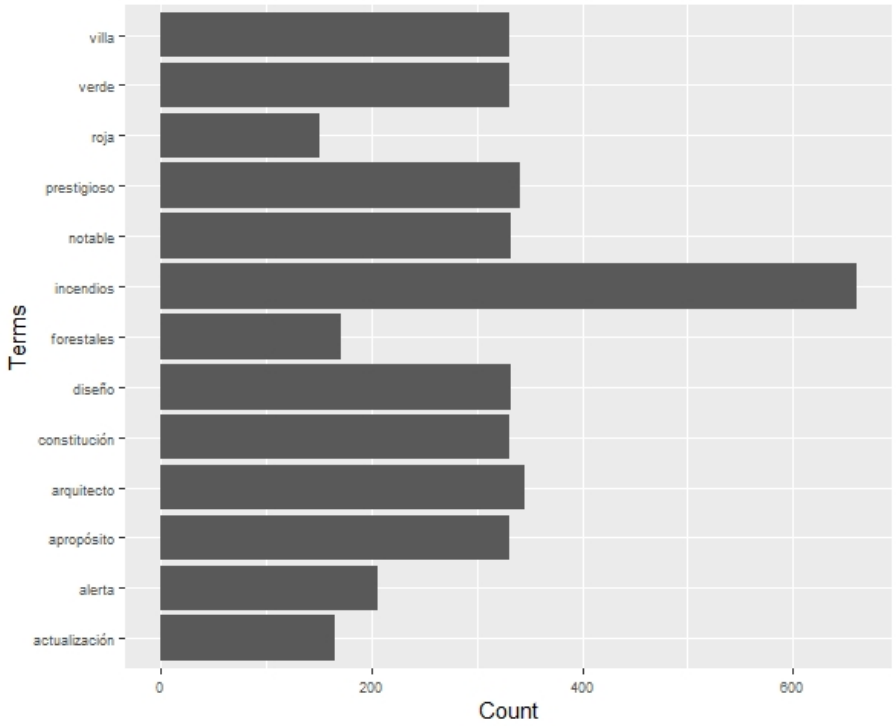
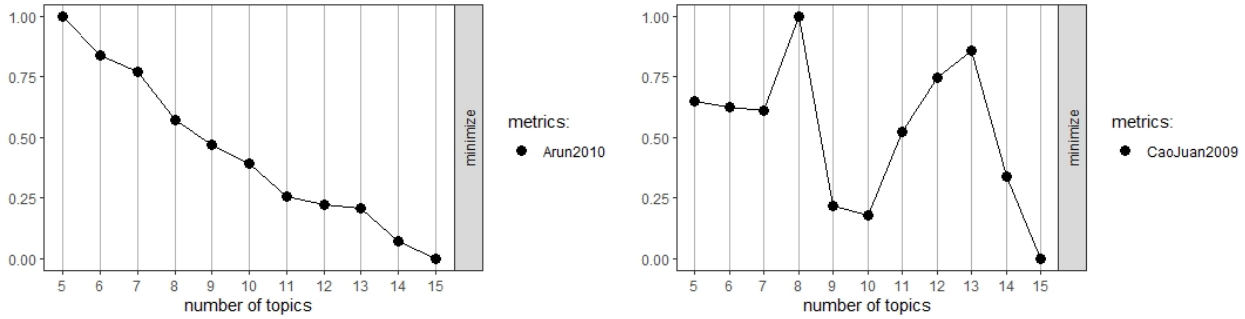


Figura 5.27: Gráfico de barras términos mas frecuentes - Incendio Santa Juana.

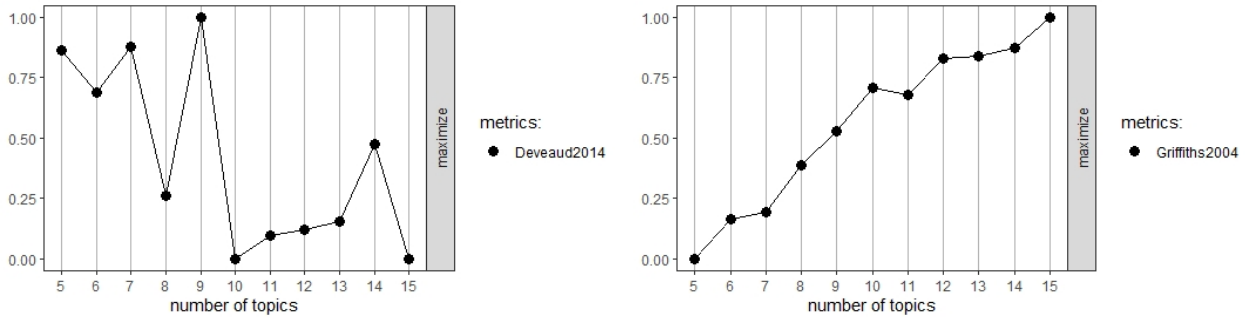
El número de tópicos a obtener con el algoritmo LDA se ha conseguido

por medio de las siguientes métricas de minimización, ver Figura 5.28 métricas *Arun y CaoJuan*, y maximización, ver Figura 5.29 métricas *Deveaud y Griffiths*.



(a) Arun 2010 (b) CaoJuan 2009

Figura 5.28: Métricas de minimización - Incendio Santa Juana.



(a) Deveaud 2014 (b) Griffiths 2004

Figura 5.29: Métricas de maximización - Incendio Santa Juana.

Después de obtenidas las cuatro métricas que se presentan en el cuadro 5.13, se procede a seleccionar el valor de k más adecuado.

Métrica	**Tipo**	**k**
Arun	Minimización	7
CaoJuan	Minimización	8
Deveaud	Maximización	7
Griffiths	Maximización	7

Cuadro 5.13: Métricas obtenidas con LDA Tuning - Incendio Santa Juana.

En este caso, el valor de k seleccionado será 7, con lo que los tópicos obtenidos por el algoritmo LDA para los datos correspondientes al hashtag *Nonguen* son:

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	incendio	incendios	arquitecto	forestal	actualización	región	alerta
2	hectáreas	intendente	prestigioso	armada	aeronaves	incendiosforestales	roja
3	sector	arquitecto	diseño	domingo	horas	activos	incendios
4	colico	prestigioso	notable	talcahuano	primeras	ahora	forestales
5	viviendas	villa	apropósito	jornada	hacia	intendente	concepción

Figura 5.30: Tópicos obtenidos - Incendio Santa Juana.

La Figura 5.30 muestra cada uno de los 7 tópicos y las 5 palabras que lo componen. La Figura A.10 en el anexo presenta el valor del coeficiente *beta* para cada una de las palabras que componen cada tópico.

Para el dataset en análisis, el algoritmo LDA presenta la siguiente información según cada uno de los tópicos. El tópico número 1 menciona palabras como *incendio, hectáreas, sector* y *viviendas*. En tanto en los tópicos número 2 y 4, encontramos palabras como *incendios y forestal*, pero encontramos también palabras como *prestigioso, arquitecto y jornada* que no tienen relación con la emergencia. Si nos detenemos ahora en el tópico número 3, la información entregada no presenta relación con la emergencia, ya que encontramos palabras como *prestigioso, arquitecto, diseño, notable*. Por último, los tópicos números 5, 6 y 7 presentan relación con la emergencia. El primero de ellos se refiere a *actualización primeras horas aeronaves hacia*, mientras el tópico número 6 presenta palabras como *incendios forestales activos ahora* además de palabras como *región e intendente*. Por último el tópico número 7, presenta información clara *alerta roja incendios forestales concepción*. Referente a menciones a organismos encargados del control de la emergencia, así como de entidades gubernamentales, sólo se mencionan estos últimos por medio de la palabra *intendente* sin entregar mayor aporte de información.

Respecto de su composición porcentual, en este caso los tópicos detectados están compuestos en un 46 % por palabras relacionadas a *Qué* está ocurriendo, seguida por sólo un 14 % de palabras relacionadas a *Dónde* está ocurriendo la emergencia y un 17 % de palabras relacionadas a *Cuándo* sucede. Las palabras relacionadas a *Otros* se llevan un 23 % de las menciones. El Cuadro 5.14 presenta estos valores, y el Cuadro A.10 del anexo presenta los valores obtenidos para cada tópico.

Uso	Porcentaje
Qué sucede	46 %
Dónde sucede	14 %
Cuándo sucede	17 %
Otros	23 %

Cuadro 5.14: Cuadro uso porcentual - Incendio Santa Juana.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	incendios	incendios	incendios	apropósito	verde	forestal	incendios
2	prestigioso	constitución	apropósito	prestigioso	diseño	constitución	diseño
3	verde	villa	arquitecto	notable	notable	arquitecto	villa
4	alerta	diseño	villa	arquitecto	prestigioso	prestigioso	verde
5	apropósito	apropósito	notable	forestales	constitución	incendios	constitución

Figura 5.31: Tópicos obtenidos el 26/01/2020 - Incendio Santa Juana.

Ahora, si nos enfocamos en los tópicos presentados en la Figura 5.31, correspondientes sólo a los tweets emitidos el día 26 de enero, de manera coincidente con lo presentado en la Figura 5.30, sólo es posible encontrar palabras dentro de los tópicos que se encuentran en relación con la emergencia en cuestión. Es así como podemos hallar la palabra *incendios* en los tópicos números 1, 2, 3 y 7, *forestales* en los tópicos número 4 y 6. Las demás palabras entregadas por cada tópico no están relacionadas con la emergencia y no presentan sentido interpretativo alguno. La Figura A.11 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.31. Lo siguiente será analizar el dataset que se ha creado a partir de los tweets emitidos entre los días 27 a 30 de enero, que se presentan a continuación.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
1	incendiosforestales	incendios	ruta	horas	arquitecto	alerta	aeronaves
2	incendio	forestales	fundación	actualización	prestigioso	concepción	primeras
3	región	despacho	monagas	aeronaves	diseño	roja	hacia
4	intendente	roja	ramírez	primeras	notable	forestales	brigadas
5	actualización	actualización	recorre	despacho	apropósito	incendios	actualización

Figura 5.32: Tópicos obtenidos entre los días 27/01/2020 - 30/01/2020 - Incendio Santa Juana.

Por otra parte, si nos enfocamos en la información entregada por la Figura 5.32, lo primero que es posible apreciar es que a diferencia de lo que se pre-

sentaba en la Figura 5.31, los tópicos número 1 y 2 están relacionados con la emergencia. El primero de ellos menciona *actualización incendios forestales región* e *intendente*, y el tópico número 2 menciona *actualización incendios forestales* además de incluir la palabra *roja*, que podría estar relacionada con una posible alerta. Por otra parte, el tópico número 4 hace mención a *primeras aeronaves actualización*, mientras que el tópico número 6 es aún más claro mencionando *alerta roja incendios forestales concepción*, mientras que el tópico número 7 menciona palabras como *primeras aeronaves, brigadas y actualización*. Los tópicos números 3 y 4 no presentan relación con la emergencia.

Esto permite afirmar que, si bien la información entregada por las tres figuras es similar, podemos ver que al realizar la separación utilizando intervalos de tiempo, los tópicos que aparecen en el primer intervalo de tiempo están relacionados con la temporalidad presente del suceso. En este caso en particular, podemos mencionar que el término *alerta roja* se mencionaba en 4 de los 6 tópicos extraídos. Referente a la mención a la emergencia (incendio), en el primer conjunto de tópicos presentados en la Figura 5.15 se menciona en todos ellos, a diferencia de los tópicos de la Figura 5.16 en donde sólo se menciona en 4 de los 6 tópicos, es decir, ha disminuido la cantidad de tweets que hacen referencia a los incendios forestales. Por otra parte, y al igual que el dataset de Chiguayante, no existen tópicos relacionados con instituciones gubernamentales, y solo encontramos mencionada la palabra *bomberos* en el tópico número 6, tanto de la Figura 5.14 como de la Figura 5.16. La Figura A.12 en el anexo, muestra el valor del coeficiente *beta* correspondientes a la Figura 5.32.

Finalmente, y al igual que en los datasets anteriores, se ha realizado la comparación porcentual entre ambos grupos de tópicos. En el Cuadro 5.15 es posible apreciar que existe un aumento de un 3 % y un 14 % respectivamente para las menciones en los tópicos asociados a *Qué sucede* y a *Dónde sucede*. Respecto de los tópicos asociados a *Cuándo sucede* y *Otros* existe una disminución de un 6 % y un 11 % en las menciones realizadas. Los Cuadros A.8 y A.9 del anexo presentan los resultados para cada conjunto de tópicos.

Uso	Durante	Después	Variación
Qué sucede	51 %	54 %	3 %
Dónde sucede	31 %	46 %	14 %
Cuándo sucede	6 %	0 %	-6 %
Otros	11 %	0 %	-11 %

Cuadro 5.15: Cuadro de uso porcentual Durante/Después - Incendio Santa Juana.

5.7. Discusión

Después de realizado el presente trabajo fin de máster se ha podido comprobar lo siguiente:

- Como se vio en asignaturas tales como *Modelos de Regresión Lineal*, *Minería de Datos* o *Análisis, Monitorización y Diagnóstico de Procesos Multi-variantes*, el proceso de pre-tratamiento de datos necesario de realizar previamente al análisis de la información representa gran parte de la carga de trabajo. En el caso de datos obtenidos desde Twitter no es diferente, ya que tareas como la eliminación de puntuación, convertir letras mayúsculas a minúsculas o remover direcciones URL, por mencionar sólo algunas, son actividades imprescindibles para la posterior construcción de las matrices *term-document matrix* y *document-term matrix* que corresponden al elemento de entrada para la realización del algoritmo LDA.

- Respecto de la elección de la cantidad de tópicos, se ha podido comprobar que, si bien pudiese resultar una tarea bastante subjetiva, existen herramientas como el *wordcloud* y el *gráfico de frecuencias*, que son útiles para tener una primera aproximación referente a la cantidad de tópicos que se deberán obtener con el algoritmo LDA. Posteriormente, para una definición más objetiva, es posible utilizar paquetes estadísticos tales como *LDA Tuning*, el cual presenta complejas métricas de optimización, minimización y maximización, que permiten definir la cantidad de tópicos que se le requerirán al algoritmo LDA. Finalmente, es importante mencionar que incluso los algoritmos más sofisticados requieren que el analista de datos cuente con nociones de conocimiento relacionadas al tema que se encuentra en estudio.

- Como se mencionaba en el capítulo 2, el algoritmo LDA es uno de las

técnicas de aprendizaje no supervisado más utilizada para realizar labores de análisis de textos. Esto es debido a que con él es posible agrupar y reducir el número de dimensiones de los datos en estudio. En el caso del presente trabajo, se ha utilizado para reducir grandes volúmenes de texto almacenado como tweets, en un cierto número de tópicos compuestos por sólo cinco palabras, los cuales permiten generarse una idea de lo que se está expresando en Twitter.

- Un punto importante que se ha podido constatar con el análisis realizado, es que en los tópicos obtenidos de ninguno de los cuatro set de datos, y bajo ninguna consideración temporal, se aprecian de manera recurrente instituciones de Chile relacionadas al control de emergencias, como pudiese ser la Oficina Nacional del Ministerio del Interior (ONEMI). Tampoco aparece mencionada la Corporación Nacional Forestal (CONAF), considerando que en este estudio los datos analizados corresponden a incendios de tipo forestal. Por otra parte, palabras como *bomberos*, *carabineros* y *brigadistas* sí aparecen mencionadas en los cuatro dataset, que finalmente son las personas encargadas de controlar y mitigar lo que sucede.

Capítulo 6

CONCLUSIONES

Luego de realizado el presente trabajo fin de máster se ha podido comprobar que es posible darle utilidad a la información que se recupera desde redes sociales, específicamente en este caso se ha utilizado Twitter principalmente por el gran volumen de información que se genera y por la facilidad de acceso y recuperación en comparación a otras redes sociales.

Las primeras etapas de la metodología necesaria para realizar la detección automática de tópicos en información recuperada desde la red social Twitter han requerido la construcción del estado del arte y del modelado estadístico que dieran el marco teórico necesario para la realización del presente trabajo fin de máster. A continuación se ha realizado la recuperación de datos desde Twitter, a los que posteriormente ha sido necesario pre-tratar y limpiar para eliminar contenido que no entrega información relevante para el objetivo de este trabajo.

Respecto de los resultados obtenidos, lo primero ha sido la detección de los tipos de usuarios que comparten información, y ha sido posible identificar cuatro grandes grupos mayoritarios de usuarios: organismos de gobierno, organizaciones no gubernamentales, medios de comunicación y finalmente usuarios individuales de Twitter. En relación a los tópicos obtenidos de manera automática por medio de la utilización del algoritmo LDA, se han logrado identificar cuatro grandes temas a los que los usuarios de Twitter se refieren con motivo de la ocurrencia de incendios, *qué sucede, dónde sucede, cuándo sucede y otros*. Finalmente, se buscó determinar si existe diferencia entre los tópicos detectados a medida que transcurre el tiempo en una situación de emergencia. Para ello, se consideraron dos criterios de selección de tweets, *tweets emitidos durante la emergencia*, que considerará tweets emitidos el primer día de emergencia y los *tweets emitidos después la emergencia*.

El desarrollo del presente trabajo presentó dos grande dificultades que hicieron modificar la primera idea de desarrollo del presente trabajo, lo que además generó modificar la metodología a desarrollar. La primera de ellas fue la obtención de tweets geoposicionados, ya que al recuperar los tweets nos enfrentamos a que la cantidad de tweets que se recuperaban no era la suficiente para realizar un buen análisis. El segundo problema con que nos enfrentamos fue que los tweets fueron recuperados tres semanas después de ocurridos los incendios por lo que en los cuatro casos le solicitamos a Twitter 5.000 tweets y sólo entregó esta cantidad en uno de los cuatro casos. Como posible trabajo futuro se podría realizar el análisis de tweets emitidos con motivo de otros tipos de catástrofes, terremotos, tsunamis, erupciones volcánicas, etc. Por otra parte, si se contara con tweets etiquetados correspondientes a alguna situación de emergencia similar, sería una idea interesante utilizarlo para construir un clasificador que permita por ejemplo diferenciar tweets sobre recursos disponibles y recursos requeridos.

Apéndice A

ANEXOS

Anexos Chiguayante

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	60 %	20 %	60 %	40 %	20 %	40 %	0 %
Dónde	40 %	0 %	20 %	60 %	80 %	0 %	60 %
Cuando	0 %	0 %	20 %	0 %	0 %	60 %	20 %
Otros	0 %	80 %	0 %	0 %	0 %	0 %	20 %

Cuadro A.1: Cuadro de uso porcentual total - Incendio Chiguayante.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	60 %	40 %	40 %	20 %	60 %	20 %	80 %
Dónde	20 %	40 %	60 %	80 %	20 %	20 %	20 %
Cuando	20 %	20 %	0 %	0 %	20 %	60 %	0 %
Otros	0 %	0 %	0 %	0 %	0 %	0 %	0 %

Cuadro A.2: Cuadro de uso porcentual durante - Incendio Chiguayante.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	60 %	0 %	80 %	20 %	20 %	20 %	60 %
Dónde	40 %	0 %	20 %	40 %	80 %	20 %	40 %
Cuando	0 %	0 %	0 %	40 %	0 %	0 %	0 %
Otros	0 %	100 %	0 %	0 %	0 %	60 %	0 %

Cuadro A.3: Cuadro de uso porcentual después - Incendio Chiguayante.

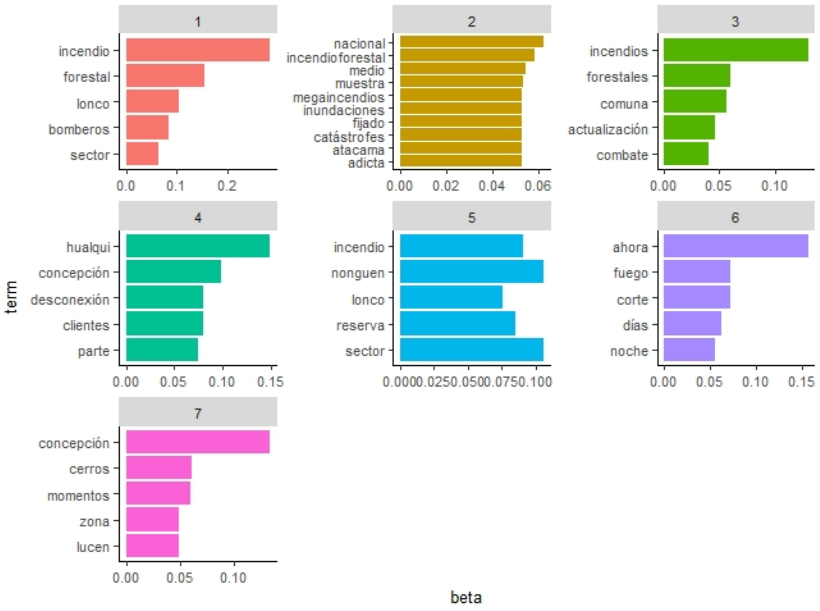


Figura A.1: Coeficiente β tópicos - Incendio Chiguayante.

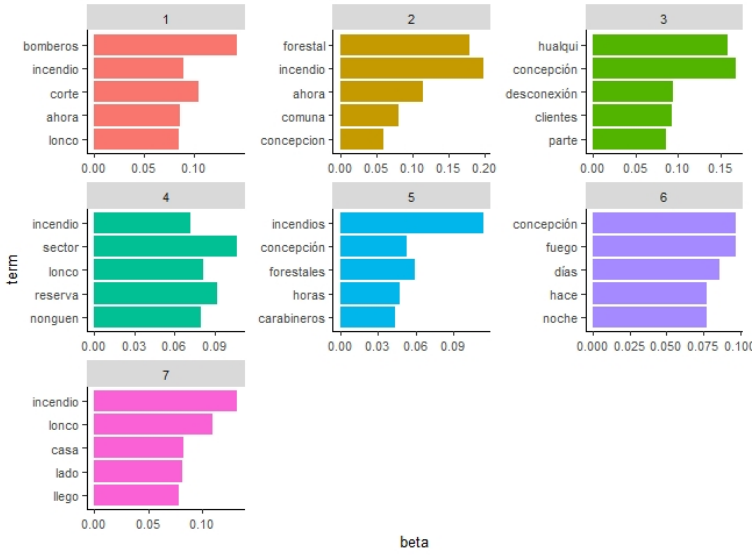


Figura A.2: Coeficiente β tópicos obtenidos el 27/01/2020 - Incendio Chiguayante.

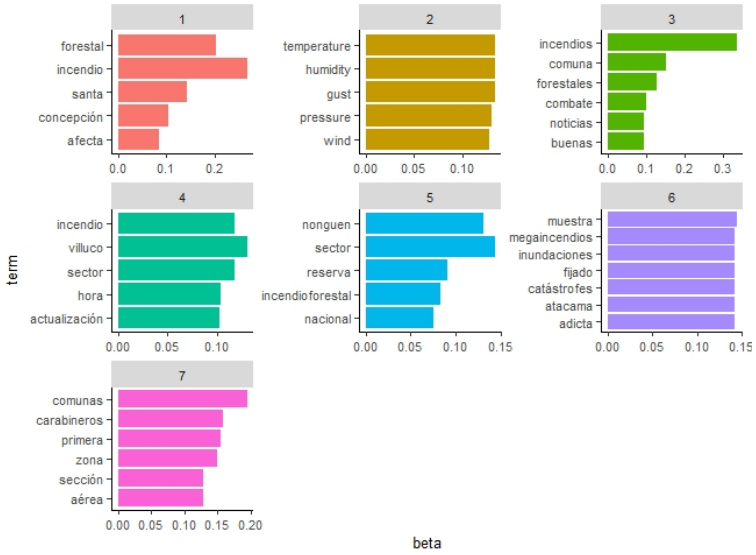


Figura A.3: Coeficiente β tópicos obtenidos entre los días 28-30/01/2020 - Incendio Chiguayante.

Anexos Hualqui

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
Qué	40 %	80 %	0 %	60 %	60 %	80 %
Dónde	40 %	20 %	20 %	40 %	20 %	20 %
Cuando	20 %	0 %	0 %	0 %	0 %	0 %
Otros	0 %	0 %	80 %	0 %	20 %	0 %

Cuadro A.4: Cuadro de uso porcentual total - Incendio Hualqui.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
Qué	80 %	80 %	80 %	60 %	80 %	40 %
Dónde	0 %	20 %	20 %	20 %	20 %	40 %
Cuando	0 %	0 %	0 %	0 %	0 %	0 %
Otros	20 %	0 %	0 %	20 %	0 %	20 %

Cuadro A.5: Cuadro de uso porcentual durante - Incendio Hualqui.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**
Qué	80 %	0 %	40 %	60 %	60 %	80 %
Dónde	20 %	0 %	20 %	20 %	40 %	20 %
Cuando	0 %	0 %	40 %	0 %	0 %	0 %
Otros	0 %	100 %	0 %	20 %	0 %	0 %

Cuadro A.6: Cuadro de uso porcentual después - Incendio Hualqui.

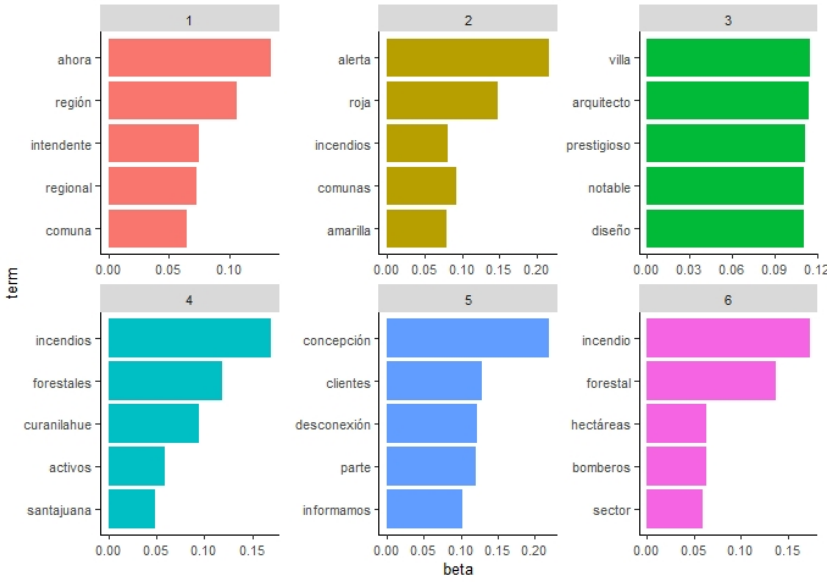


Figura A.4: Coeficiente β tópicos - Incendio Hualqui.

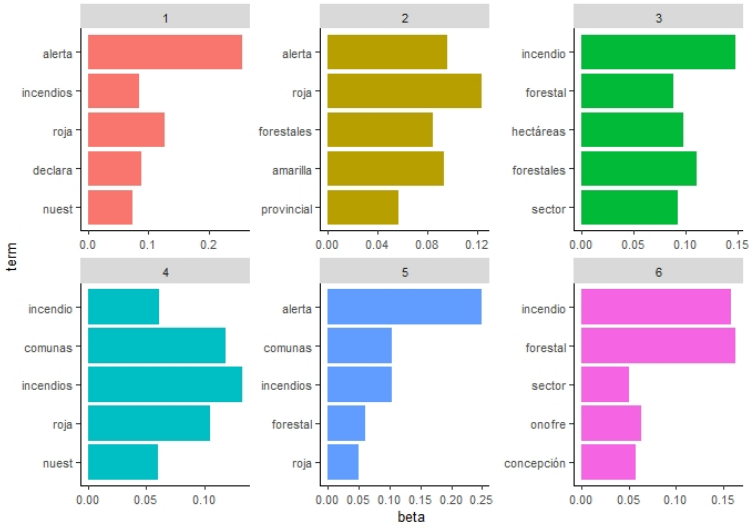


Figura A.5: Coeficiente β tópicos obtenidos el 25/01/2020 - Incendio Hualqui.

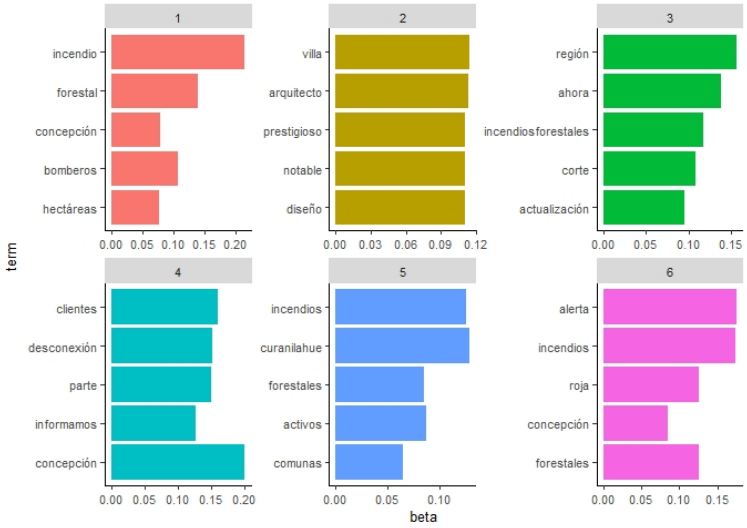


Figura A.6: Coeficiente β tópicos obtenidos entre los días 26-30/01/2020 - Incendio Hualqui.

Anexos Nonguen

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	40 %	0 %	40 %	40 %	40 %	40 %	60 %
Dónde	40 %	40 %	60 %	60 %	40 %	0 %	0 %
Cuando	20 %	0 %	0 %	0 %	20 %	0 %	40 %
Otros	0 %	60 %	0 %	0 %	0 %	60 %	0 %

Cuadro A.7: Cuadro de uso porcentual total - Incendio Nonguen.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	0 %	60 %	60 %	60 %	60 %	60 %	80 %
Dónde	40 %	20 %	40 %	20 %	40 %	40 %	0 %
Cuando	0 %	20 %	0 %	0 %	0 %	0 %	20 %
Otros	60 %	0 %	0 %	20 %	0 %	0 %	0 %

Cuadro A.8: Cuadro de uso porcentual durante - Incendio Nonguen.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	40 %	60 %	20 %	60 %	80 %	60 %	60 %
Dónde	60 %	40 %	80 %	40 %	20 %	40 %	40 %
Cuando	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Otros	0 %	0 %	0 %	0 %	0 %	0 %	0 %

Cuadro A.9: Cuadro de uso porcentual después - Incendio Nonguen.

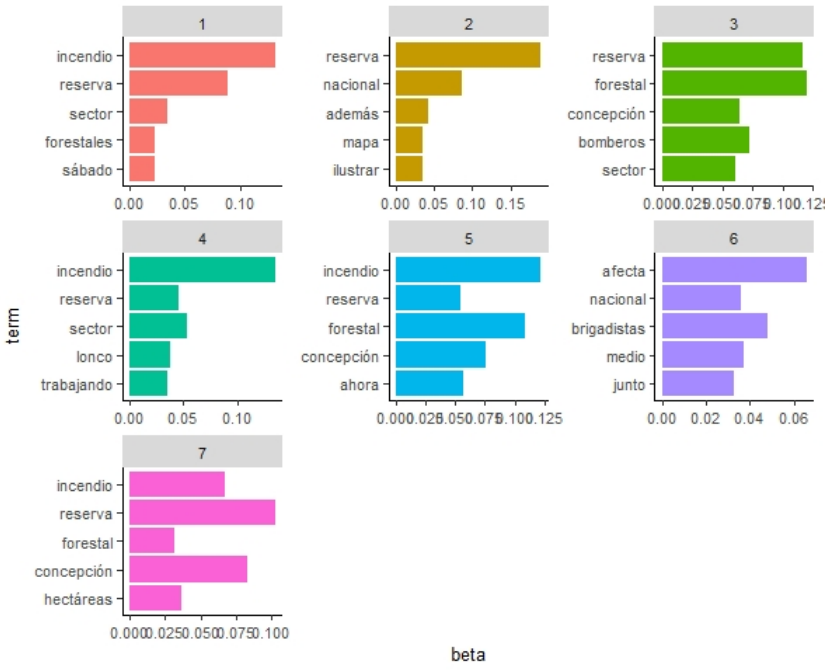


Figura A.7: Coeficiente β tópicos - Incendio Nonguen.

Imagen 5

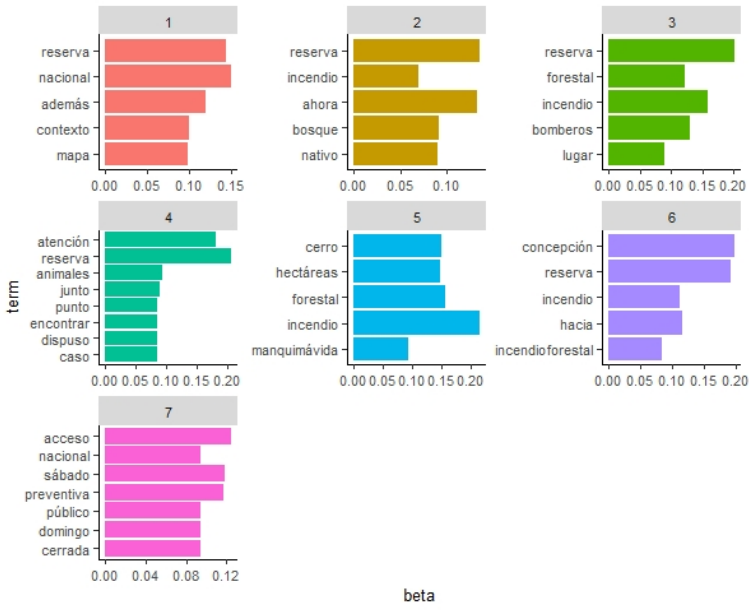


Figura A.8: Coeficiente β tópicos obtenidos el 25/01/2020 - Incendio Nonguen.

Imagen 6

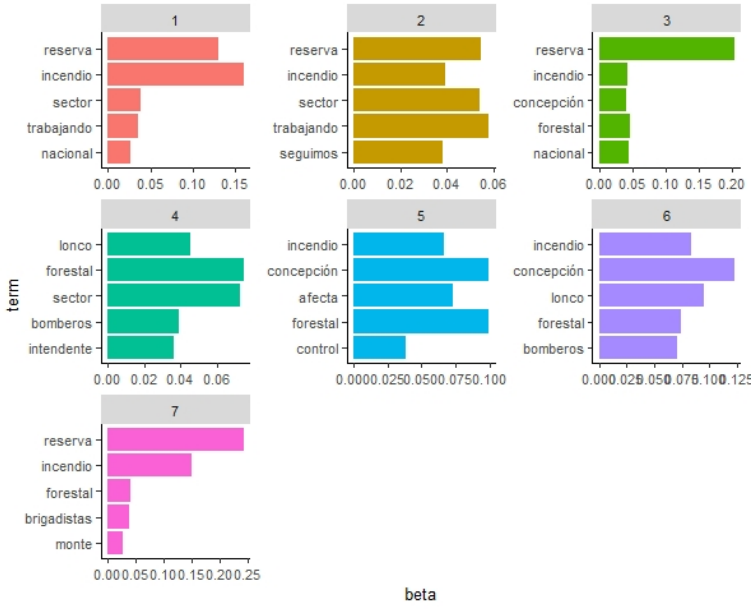


Figura A.9: Coeficiente β tópicos obtenidos entre los días 26-30/01/2020 - Incendio Nonguen.

Anexos Santa Juana

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	60 %	40 %	0 %	40 %	40 %	60 %	80 %
Dónde	40 %	0 %	0 %	20 %	0 %	20 %	20 %
Cuando	0 %	0 %	0 %	40 %	60 %	20 %	0 %
Otros	0 %	60 %	100 %	0 %	0 %	0 %	0 %

Cuadro A.10: Cuadro de uso porcentual total - Incendio Santa Juana.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	40 %	20 %	20 %	20 %	20 %	20 %	20 %
Dónde	0 %	40 %	20 %	0 %	20 %	20 %	40 %
Cuando	0 %	0 %	0 %	0 %	0 %	0 %	0 %
Otros	60 %	40 %	60 %	80 %	60 %	60 %	40 %

Cuadro A.11: Cuadro de uso porcentual durante - Incendio Santa Juana.

Uso	**T1**	**T2**	**T3**	**T4**	**T5**	**T6**	**T7**
Qué	60 %	60 %	0 %	0 %	0 %	80 %	80 %
Dónde	20 %	0 %	0 %	40 %	0 %	20 %	0 %
Cuando	20 %	20 %	0 %	40 %	0 %	0 %	20 %
Otros	0 %	20 %	100 %	20 %	100 %	0 %	0 %

Cuadro A.12: Cuadro de uso porcentual después - Incendio Santa Juana.

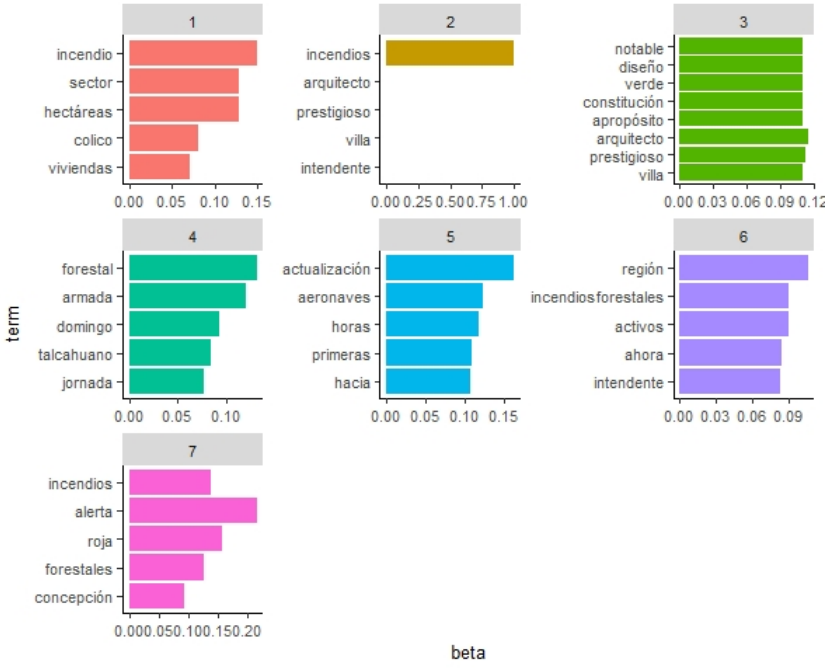


Figura A.10: Coeficiente β tópicos - Incendio Nonguen

Imagen 7

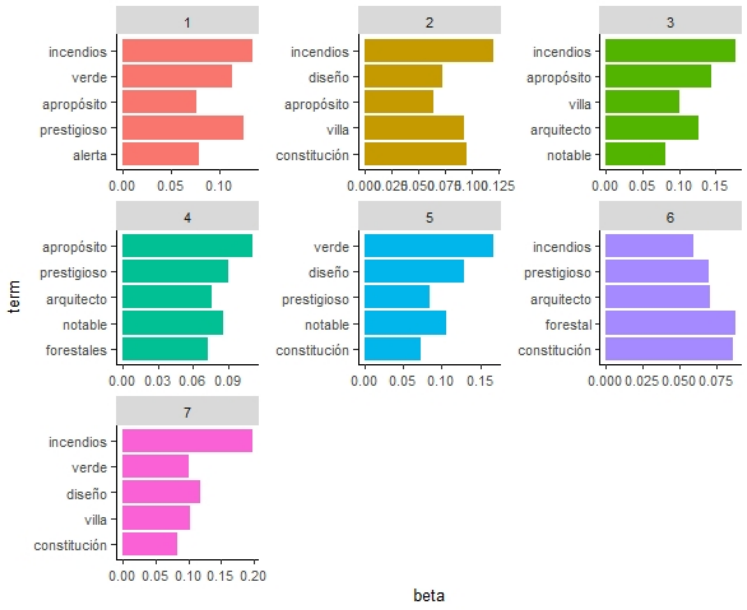


Figura A.11: Coeficiente β tópicos obtenidos el 26/01/2020 - Incendio Santa Juana

Imagen 8

Figura A.12: Coeficiente β tópicos obtenidos entre los días 27-30/01/2020 - Incendio Santa Juana

Apéndice B

BIBLIOGRAFÍA

Bibliografía

Issa Annamoradnejad and Jafar Habibi. A comprehensive analysis of twitter trending topics. In *2019 5th International Conference on Web Research (ICWR)*, pages 22–27. IEEE, 2019.

Rajkumar Arun, Venkatasubramaniyan Suresh, CE Veni Madhavan, and MN Narasimha Murthy. On finding the natural number of topics with latent dirichlet allocation: Some observations. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 391–402. Springer, 2010.

Moumita Basu, Anurag Roy, Kripabandhu Ghosh, Somprakash Bandyopadhyay, and Saptarshi Ghosh. Microblog retrieval in a disaster situation: A new test collection for evaluation. In *SMERP@ ECIR*, pages 22–31, 2017.

Moumita Basu, Saptarshi Ghosh, and Kripabandhu Ghosh. Overview of the fire 2018 track: Information retrieval from microblogs during disasters (irmidis). In *Proceedings of the 10th annual meeting of the Forum for Information Retrieval Evaluation*, pages 1–5, 2018.

Moumita Basu, Kripabandhu Ghosh, and Saptarshi Ghosh. Information retrieval from microblogs during disasters: In the light of irmidis task. *SN Computer Science*, 1(1):61, 2020.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Juan Cao, Tian Xia, Jintao Li, Yongdong Zhang, and Sheng Tang. A density-based method for adaptive lda model selection. *Neurocomputing*, 72(7-9): 1775–1781, 2009.

Elena del Val, Miguel Rebollo, and Vicente Botti. Does the type of event influence how user interactions evolve on twitter? *PLOS one*, 10(5):e0124049, 2015.

Romain Deveaud, Eric SanJuan, and Patrice Bellot. Accurate and effective latent concept modeling for ad hoc information retrieval. *Document numérique*, 17(1):61–84, 2014.

Ingo Feinerer and Kurt Hornik. *tm: Text Mining Package*, 2019. URL https://CRAN.R-project.org/package=tm. R package version 0.7-7.

Jeff Gentry. *twitteR: R Based Twitter Client*, 2015. URL https://CRAN.R-project.org/package=twitteR. R package version 1.1.9.

Saptarsi Goswami, Sanjay Chakraborty, Sanhita Ghosh, Amlan Chakrabarti, and Basabi Chakraborty. A review on application of data mining techniques to combat natural disasters. *Ain Shams Engineering Journal*, 9(3): 365–378, 2018.

Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

Bettina Grün and Kurt Hornik. topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 2011. doi: 10.18637/jss.v040.i13.

Guskin. Hurricane sandy and twitter, 2012. URL https://www.journalism.org/2012/11/06/hurricane-sandy-and-twitter/. PEJ New Media Index.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. *arXiv preprint arXiv:1605.05894*, 2016.

Jessica Li and H Raghav Rao. Twitter as a rapid response news service: An exploration in the context of the 2008 china earthquake. *The Electronic Journal of Information Systems in Developing Countries*, 42(1):1–22, 2010.

Wainer Lusoli, Margherita Bacigalupo, Francisco Lupiáñez-Villanueva, Norberto Nuno Gomes de Andrade, Shara Monteleone, and Ioannis Maghiros. Pan-european survey of practices, attitudes and policy preferences as regards personal identity data management. *JRC Scientific and Policy Reports, EUR*, 25295, 2012.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. Tagging your tweets: A probabilistic modeling of hashtag annotation in twitter. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 999–1008, 2014.

Alexander Mills, Rui Chen, JinKyu Lee, and H Raghav Rao. Web 2.0 emergency applications: How useful can twitter be for emergency response? *Journal of Information Privacy and Security*, 5(3):3–26, 2009.

Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

Nastaran Pourebrahim, Selima Sultana, John Edwards, Amanda Gochanour, and Somya Mohanty. Understanding communication dynamics on twitter during natural disasters: A case study of hurricane sandy. *International journal of disaster risk reduction*, 37:101176, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL https://www.R-project.org/.

RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA, 2015. URL http://www.rstudio.com/.

Bruno Takahashi, Edson C Tandoc Jr, and Christine Carmichael. Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. *Computers in Human Behavior*, 50:392–398, 2015.

Twitter. Q1 2019 earnings report, 2019. URL https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf. Twitter Earnings Report.

Inc. Twitter. Información sobre las api de twitter, 2020. URL https://help.twitter.com/es/rules-and-policies/twitter-api. Centro de Ayuda Twitter.