Evaluación de modelos de aprendizaje no supervisado para el análisis de contenido de tweets generados ante un desastre

Nicolás Jiménez Piñeros, Remberth José Jiménez Vargas

Trabajo de Grado para Optar el título de Ingeniero Industrial

Director

Henry Lamos Díaz

PhD. Física-Matemática

Codirector

Daniel Orlando Martínez Quezada

MSc. Ingeniería Industrial

Universidad Industrial de Santander

Facultad de Ingenierías Físico-Mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2019

Dedicatorias

A Dios, Por haberme permitido llegar a donde me he propuesto y haberme dado salud para lograrlo.

A mi padre, Por creer en mí, y por apoyarme en cada paso que he dado, por enseñarme a pescar en vez de darme el pescado.

Nicolás Jiménez Piñeros

A Dios, por brindarme sabiduría, paciencia y habilidades para sacar el proyecto y la carrera adelante y llenar mi vida de bendiciones.

A mis padres y hermanos Jiménez Vargas por ser mis guías en todo mi proceso de formación personal y profesional, los cuales hoy hacen de mí una persona humilde, responsable y a sus esfuerzos que me permitieron culminar una carrera profesional.

A Nicolás Jiménez, por ser la persona que estuvo apoyándome día a día en todo este proceso de formación, aquella persona que me motivaba a seguir adelante trabajando y estudiando a pesar de los tropiezos que se presentan en el camino, y por ser ese gran amigo que con sus consejos y amistad me permiten culminar este proceso de la forma más agradable.

Remberth José Jiménez Vargas

Agradecimientos

A los profesores Henry Lamos Díaz y Daniel Orlando Martínez Quesada por su generosidad al brindarnos su apoyo con su capacidad y experiencia científica fundamentales para la consecución de este trabajo de investigación.

Al grupo de investigación OPALO, por brindarnos los espacios y herramientas para desarrollar esta investigación.

Tabla de contenido

Lista de tablas

Lista de figuras

Apéndices

(Ver apéndices adjuntos en el CD y pueden visualizarlos en la Base de Datos de la Biblioteca UIS)

Apéndice A. Artículo de Investigación

Resumen

Título del proyecto: Evaluación de modelos de aprendizaje no supervisado para el análisis de contenido de tweets generados ante un desastre[*]

Autores: Nicolás Jiménez Piñeros, Remberth José Jiménez Vargas[**]

Palabras clave: Análisis de contenido, clusters, minería de datos, análisis de texto, aprendizaje automático, Twitter, redes sociales

Se entiende como revolución 4.0 a una etapa en la evolución humana hacia la transición de nuevos sistemas que están construidos sobre la infraestructura digital, trayendo consigo retos tecnológicos que podrían ser usados para resolver problemas y mejorar la calidad de vida de muchas personas. Ahora bien, uno de tantos retos en esta revolución industrial, es el aprovechamiento de la gran cantidad de datos que se generan día a día en la red por cuenta de las redes sociales, datos que al ser procesados e interpretados pueden convertirse en una fuente de información clave para la toma de decisiones en diversos campos de estudio.

En el presente proyecto de investigación aplicando conceptos de minería de texto, se evalúan dos modelos de aprendizaje no supervisado (K-means y LDA), con el lenguaje de programación R y tomando como caso de estudio un conjunto de tweets generados durante los incendios forestales que se presentaron en California (Estados Unidos) a fecha de 16 de noviembre de 2018, bajo el hashtag #california, se comparan dichos modelos permitiendo además de identificar las diferentes temáticas, características y/o patrones de comportamiento de los usuarios al momento de la ocurrencia del desastre, concluir cuál de estos algoritmos proporciona información más amplia y relevante en el tema de investigación.

---

[*] Trabajo de Grado

[**] Facultad de Ingenierías Físico-mecánicas. Escuela de Ingeniería Industrial. Director: Henry Lamos Díaz. Codirector: Daniel Orlando Martínez Quezada

Abstract

Project title: Evaluation of unsupervised learning models for the content analysis of tweets generated in the event of a disaster[*]

Authors: Nicolás Jiménez Piñeros, Remberth José Jiménez Vargas[**]

Keywords: Content analysis, clusters, data mining, text analysis, machine learning, twitter, social networks

Revolution 4.0 is understood as a stage in human evolution towards the transition of new systems that are built on digital infrastructure, bringing with it technological challenges that could be used to solve problems and improve the quality of life of many people. Now, one of the many challenges in this industrial revolution is the use of the large amount of data that is generated on a daily basis on the network by social networks like Twitter, Facebook, Instagram, You Tube, LinkedIn etc. Data that, when processed and interpreted, can become a source of key information for decision-making in various fields of study.

In the current research, applying concepts of text mining, two unsupervised learning models are evaluated (K - means and Latent Dirichlet Allocation), with the programming language R and using a set of tweets as a case study that were generated on November 16th of 2018 during the wildfires occurred in California (United States) under the hashtag #california, It is aimed to compare those models in addition, to identifying the different themes, characteristics and / or behavior patterns in this database, to conclude which of these algorithms provides more extensive and relevant information in the research topic.

---

[*] Bachelor Thesis

[**] Facultad de Ingenierías Físico-mecánicas. Escuela de Ingeniería Industrial. Director: Henry Lamos Díaz. Codirector: Daniel Orlando Martínez Quezada

Introducción

En la actualidad el internet es una herramienta que a muchos, por no decir a la mayoría de las personas les ha facilitado la vida en lo social, económico y científico. La data que se genera en la red de comunicación se presenta de diferentes formas, por un lado los datos de numerosas transacciones que se realizan cada segundo, data generada por sensores en agricultura, logística, entre otros; por otro lado los datos de tipo semi-estructurados como correos electrónicos, tratamientos médicos, imágenes, audio y video, páginas web, tweets, chats, etc. Así que son millones y millones de terabytes de datos que se crean y que circulan a través de la red diariamente, creándose una apremiante necesidad de procesar y convertir dichos datos en información y conocimientos útiles, es allí donde la minería de datos juega un papel muy importante.

La minería de datos consiste en la aplicación de técnicas y procedimientos que buscan extraer patrones mostrando así información útil que puede ser empleada para entender cierto tipo de comportamientos y por supuesto para llegar a tomar decisiones más asertivas. La información y conocimiento adquiridos se pueden utilizar para aplicaciones que van desde análisis de mercado, detección de fraude y retención de clientes, hasta control de producción y exploración científica. (Han, Kamber, & Pei, 2011)

La analítica de grandes volúmenes de datos permite a través de un conjunto de herramientas tecnológicas y algoritmos de diferentes ramas como la estadística, la minería de datos, machine learning, etc, mejorar los procesos de decisión en la logística humanitaria. Por ejemplo, los datos generados por las redes sociales están siendo usada para analizar el impacto social de los procesos de peligro y el comportamiento humano en situaciones de desastre (Murzintcev & Cheng, 2017). Según estadísticas de la ONU los desastres atribuidos a fenómenos climáticos han aumentado cada

año, en promedio 335 por año; lo que representa un 14 % más que en el decenio anterior, causando en los últimos veinte años un promedio anual de 30.000 muertes y más de 4.000 millones de heridos o damnificados. La población más afectada se encuentran en los países más pobres y los daños económicos causados por los desastres son considerablemente más grande que en los países con mayor desarrollo, Margareta Wahlstrom. (20mintos.es, 2015)

Durante y después de un desastre de la naturaleza los usuarios de redes sociales empiezan a generar grandes cantidades de información posteada en Twitter, es aquí donde se ve la posibilidad de manejar de forma correcta esta información para el apoyo a los damnificados y así el correcto procesamiento de esta información hace que el proceso de recuperación de los afectados sea más rápido, y efectivo.

El proyecto se enfocará en la evaluación de dos modelos de aprendizaje no supervisado (K-means y LDA) usando el lenguaje de programación R aplicado a datos de la red social Twitter en los incendios de California que dieron inicio el 8 de noviembre de 2018; y así evaluar estos modelos bajo instancias de benchmarking concluyendo cuál de los dos modelos proporciona más y mejores insights para identificar patrones en este caso de estudio

Es común que durante y después de estos trágicos eventos aparezcan grandes cantidades de información posteada en Twitter, el problema radica en que no se sabe que información es útil o no, para los interesados. El correcto procesamiento de esta información hace que el proceso de recuperación de los afectados sea más rápida, y efectiva.

En este punto es donde vemos la necesidad de utilizar la minería de texto, la cual tiene como objetivo extraer información de texto no estructurado, tales como entidades, reportes, lugares

seguros, comida, etc. y relacionarlas entre sí, es decir, la idea de este proyecto es comprender de forma descriptiva que tanta información pueden proveer las redes sociales cuando un desastre ocurre, y de esta manera en el futuro se puedan usar uno de estos modelos de aprendizaje no supervisado para identificar dicha información que ayude a mitigar o disminuir el impacto generado por dichos desastres.

Tabla 1.
Cumplimientos de objetivos del proyecto

OBJETIVO ESPECIFICOS	CUMPLIMIENTO
⬥ Realizar una revisión de la literatura de técnicas de agrupamiento particional en documentos tipo texto y enfoques de análisis de contenido de redes sociales en desastres.	Capítulo 4
⬥ Consolidar bases de datos de la red social Twitter de un evento catalogado como desastre utilizando la API de Twitter.	Capítulo 6
⬥ Seleccionar y aplicar por lo menos dos técnicas de agrupamiento particional a las bases de datos consolidadas utilizando el lenguaje de programación R.	Capítulo 6
⬥ Comparar los resultados arrojados por los algoritmos mediante instancias de benchmarking	Capítulo 7
⬥ Elaborar un artículo de carácter publicable resumiendo los hallazgos encontrados en el proyecto.	Apéndice A

1. Planteamiento del problema

Por definición, (García Renedo, 2007) desastre como un suceso desafortunado que implica la pérdida o lesión y altera el funcionamiento vital de la sociedad, perjudica la comunidad y el individuo por medio de la quiebra de la estructura social que se traduce en desequilibrio y evidente crisis que amenaza la integridad. Aunque frecuentemente están causados por la naturaleza, los desastres pueden deberse a la actividad humana.

Los desastres naturales han causado más de 8 millones de muertos en todo el mundo, es decir unos 50.000 al año, y han costado 7 billones de dólares desde el comienzo del siglo XX; las inundaciones (38,5% de los daños) y tormentas (20%) representan casi el 60% de esos costes, de acuerdo con el estudio, presentado en Viena durante una reunión de la Unión europea de Geo-ciencias.

En Colombia, desde el año 1900, los terremotos provocaron alrededor del 30% de los fallecimientos debidos a catástrofes naturales. Los sismos causaron el 26% de las pérdidas económicas imputadas a las catástrofes naturales, y las erupciones volcánicas el 1%. (El tiempo, 2016) El director del Instituto de Hidrología, Meteorología y Estudios Ambientales (Ideam), Ómar Franco Torres, señala que un 28% del territorio nacional, es decir un tercio del país, tiene amenaza de inundaciones en 79 municipios. (El país, 2017). Tan solo en 2017, hubo 603.302 personas afectadas por desastres naturales, donde las inundaciones representan el mayor impacto (79%) en la población afectada. (Reliefweb, 2018)

Ahora bien, la gestión de desastres puede definirse como la organización y la gestión de recursos y responsabilidades para abordar todos los aspectos humanitarios de las emergencias, en

particular la preparación, la respuesta y la recuperación a los desastres, a fin de reducir sus efectos. El problema radica en que, durante la materialización de los desastres, no se cuenta con información oportuna para poder actuar de forma inmediata; la información relevante llega cuando el impacto ha sucedido.

Hoy, siglo XXI, La forma de comunicarnos se ha visto afectada gracias a la facilidad que nos brindan las redes sociales para lograr este fin, estas son estructuras sociales que nacen como medios de comunicación para facilitar la difusión de mensajes en canales de información, acortando distancias y tiempos de envío de mensajes.

Esta facilidad que ofrecen las redes sociales al momento de comunicarnos, generando la simplicidad en las personas a comentar y publicar información, percepciones y sentimientos durante la ocurrencia de un desastre, sin embargo, la cantidad de información es tan grande y tan desordenada, que se hace imposible leerla y tomar decisiones en base a ella, es decir, contamos con los recursos, pero no con herramientas que permitan hacer una buena gestión del desastre.

Esta investigación, tiene la intención de utilizar la minería de texto como la herramienta que permite que las grandes cantidades de información generadas por la red social Twitter, sean útiles al momento de tomar decisiones para la gestión de los desastres y la mitigación de sus impactos, evaluando dos algoritmos de aprendizaje no supervisado y concluyendo cuál de los dos genera mejores resultados.

2. Justificación del proyecto

El uso de las redes sociales transforma los estilos de vida, cambia las prácticas y, también, crea nuevo vocabulario, pero todo esto se produce a un ritmo tan acelerado que la cantidad de información que se maneja en estos medios es excesivamente grande y no estructurada, provocando que el análisis de dicha información sea complicado.

Twitter por ejemplo, creado en California en marzo del año 2006. Es una red social donde se comparten mensajes de 280 caracteres, llamados tweets, acompañados de imágenes, videos, menciones, etiquetas y un sinfín de funciones, es utilizada por empresas, políticos, empresarios y personas como medio de comunicación para compartir sucesos y opiniones importantes de gran impacto a la sociedad, desde ese entonces ha tenido una gran acogida a nivel mundial ya que en el primer semestre de 2017 el promedio de usuarios activos de Twitter asciende a más de 328 millones con un crecimiento del 6% respecto al año anterior, según (Forbes, 2017) y donde el 36% de personas entre 18 y 29 años y 23% de las personas entre 30 y 49 años son usuarios de esta plataforma. (York, 2017)

Por otro lado, los desastres son eventos repentinos que afectan y destruyen de cierta manera a una comunidad, según la ONU, los desastres atribuidos en promedio cada año a fenómenos climáticos han sido 335, lo que representa un 14 % más que en el decenio anterior, y han causado en los últimos veinte años un promedio anual de 30.000 muertes y más de 4.000 millones de heridos o damnificados (20 minutos, 2015), el proceso de recuperación después de un desastre es gradual, debido a que se deben atender problemas como la inseguridad y la salud física y mental de los afectados, entre otros.

Es común que durante y después de estos trágicos eventos aparezcan grandes cantidades de información posteada en Twitter, el problema radica en que no se sabe que información es útil o no para los interesados. El correcto procesamiento de esta información hace que el proceso de recuperación de los afectados sea más rápida, y efectiva.

En este punto es donde vemos la necesidad de utilizar la minería de texto, la cual tiene como objetivo extraer información de texto no estructurado, tales como entidades, reportes, lugares seguros, comida, etc. y relacionarlas entre sí, es decir, la idea de este proyecto es comprender de forma descriptiva que tanta información pueden proveer las redes sociales cuando un desastre ocurre, y de esta manera en el futuro se puedan usar uno de estos modelos de aprendizaje no supervisado para identificar dicha información que ayude a mitigar o disminuir el impacto generado por dichos desastres.

3. Objetivos

3.1 Objetivo general

Evaluar dos modelos de aprendizaje no supervisado para el análisis de contenidos de Tweets generados ante un desastre.

3.2 Objetivos específicos

- Realizar una revisión de la literatura de técnicas de agrupamiento particional en documentos tipo texto y enfoques de análisis de contenido de redes sociales en desastres.
- Consolidar bases de datos de la red social Twitter de un evento catalogado como desastre utilizando la API de Twitter.

- Seleccionar y aplicar por lo menos dos técnicas de agrupamiento particional a las bases de datos consolidadas utilizando el lenguaje de programación R.
- Comparar los resultados arrojados por los algoritmos mediante instancias de benchmarking
- Elaborar un artículo de carácter publicable resumiendo los hallazgos encontrados en el proyecto.

4. Revisión de literatura

A pesar de que el nacimiento de las redes sociales es relativamente nuevo y considerando además que la adopción de estas en todo el mundo es un tema aún más reciente, el potencial de estos medios es bastante amplio, por ejemplo, Twitter cuya popularidad no es la más alta con respecto a las demás plataformas; para empresas, agencias y organizaciones empresariales, es el alma de la interacción con las redes sociales tanto así que Twitter se está haciendo cada vez más popular como la mejor opción para el servicio social al cliente. (York, 2017)

La tendencia de los temas en las redes sociales proporciona una fuente importante de información actual sobre eventos en nuestro mundo y también puede proporcionarnos acerca del interés, el perfil y el comportamiento del usuario que cualquier organización puede utilizar para mejorar o comercializar sus productos y servicios. (Sapul, Shiela, Aung, & Jiamthapthaksin, 2017)

En los últimos años, las investigaciones sobre el uso real y potencial de las redes sociales en situaciones de emergencia, desastres y crisis, es un tema que ha generado gran interés. (Alexander, 2014) Las redes sociales se han utilizado ampliamente para la comunicación de crisis durante desastres, y su uso durante eventos extremos ha atraído la atención tanto de investigadores como de profesionales. (Wang & Zhuang, 2017).

La minería de texto y el aprendizaje automático se convierten en una herramienta importante en estas situaciones, debido a que juegan un papel muy importante en cuanto al análisis de información se refiere, gracias a estas técnicas, es posible determinar patrones útiles en el mar de información presentado en estas redes. Ejemplificando de cierta manera los estudios de la combinación de la minería de texto y las redes sociales en las situaciones de desastres encontramos:

En el caso de (Huang, Cervone, & Zhang, 2017) integran la minería de datos y la telecomunicación para analizar eventos de desastres históricos, detección de eventos en tiempo real y el seguimiento de nuevos eventos. La teledetección es un modo de obtener información acerca de fenómenos terrestres y planetarios sin que los instrumentos empleados para adquirir datos estén en contacto directo con dicho fenómeno.

(Huang et al., 2017) sintetizan datos de múltiples fuentes (medios de comunicación, teledetección, Wikipedia y la web), minería de datos espaciales y tecnología de minería de texto para construir una solución arquitectónicamente resistente y elástica para apoyar el análisis de desastres de los acontecimientos históricos y futuros; dos tareas criticas de minería de datos que realizan en la investigación es descubrir los “temas candentes” que se discuten a través de las redes sociales mediante la ejecución de un algoritmo de modelado de temas y descubrir la ubicación de estos temas mediante el algoritmo de agrupamiento. Después de descubrir "temas candentes" a través del modelo LDA[1], las palabras (hashtags) en cada tema se utilizan para coincidir con una lista predefinida de palabras clave relacionadas con un peligro natural, como "huracán",

[1] LDA: Latent Dirichlet allocation

"tormenta", "terremoto", etc. A partir de esto, cuando se detecta una coincidencia, el sistema publicará una alerta para proporcionar advertencias tempranas sobre un posible desastre.

(Namkyung & Junghyae, 2017) en su investigación utilizan los software SNA y UCINet para un análisis de contenido de redes sociales y realizan un análisis de contenido de los informes de situación que la Oficina de Coordinación de Asuntos Humanitarios desarrolló durante sus respuestas a los desastres en Haití 2010 y Japón 2011, con estos análisis se exploraran tres preguntas de investigación: ¿cómo los sistemas de respuesta de Haití y Japón diferían en su estructura?, ¿cómo se implementaba la función de coordinación de la Oficina de Coordinación de Asuntos Humanitarios en diferentes contextos? y ¿qué condiciones específicas facilitaban o impedían las entregas de asistentes humanitarios internacionales a los países afectados y causaban lagunas estructurales entre dos sistemas de respuesta?; se analizaron 16 informes de situación del 12 de enero al 1 de febrero de 2010 para el terremoto de Haití y 16 informes de situación del 12 de marzo al 1 de abril de 2011 para el terremoto y el tsunami en Japón; se identificó y codificó las interacciones (comunicación, la operación conjunta y el intercambio de información / recursos) agente por agente de los informes de situación, al codificar las interacciones, se aplican los procedimientos de codificación específicos definidos por Krippendorff (2004) para responder a las preguntas: ¿quién inició la interacción? (initiat -organización), ¿quién respondió a las solicitudes de colaboración? (organización que responde), y finalmente, ¿cuál fue el contenido de sus interacciones?

Este análisis de contenido crea información estructurada de los datos de las redes sociales y los informes de situación para ser utilizada como datos de entrada para el análisis de red, este análisis revela la estructura de la red que se forma a partir de las interacciones entre las organizaciones en

respuesta a los desastres, por lo tanto, proporciona una perspectiva holística al analizar la dinámica de las organizaciones participantes.

(Murzintcev & Cheng, 2017) desarrollaron sistemas de clasificación informática para filtrar datos relacionados con temas, separarlos en diferentes categorías y visualizarlos en dimensiones espaciales y temporales; se eligieron los siguientes dos grupos de desastres simultáneos: los ciclones tropicales Nari (Filipinas) y Phailin (India) en octubre de 2013; y el ciclón tropical Haiyan (Filipinas) y una inundación en Texas (EE. UU.) En noviembre de 2013. (Murzintcev & Cheng, 2017) utilizan dos modelos de clasificación para el análisis de desastre: Latent Dirichlet Allocation (LDA) y Support Vector Machines (SVM); primero recuperaron mensajes relevantes para los eventos, a través de la API de Twitter y de Internet Archive, en segundo paso lo separaron en varias categorías de acuerdo con los objetivos de su investigación; así mismo, extraen los hashtags de los mensajes recuperados y se utilizan varias heurísticas para filtrar los hashtags no relevantes, esta tarea de filtrado de mensajes se puede formalizar como un problema de clasificación binaria con dos clases, desastre y no desastre. (Murzintcev & Cheng, 2017) prueban la idea de que los hashtags se pueden conectar con eventos relacionados usando información geográfica de bases de datos de desastres. Los experimentos muestran que el método puede encontrar conjuntos de hashtag claramente definidos, incluso en el caso de múltiples eventos simultáneos con impactos similares.

Para (Cheng & Wicks, 2014) el problema de la metodología basada en atributos de datos usando palabras o hashtags radica en que solo siguiendo un selecto grupo de tweets de un conjunto de atributos significa que muchos otros relacionados con el tema se pueden estar perdiendo por completo del estudio, y a esto sumándole la subjetividad debido a que estos son basado en la

percepción y experticia del autor. Con la implementación de un método de estadística de exploración espacio-tiempo (STSS por sus siglas en inglés) (Cheng & Wicks, 2014) buscan una metodología alternativa de detención de eventos y para el análisis utilizan como estudio el accidente del helicóptero en Londres en 2013, y con la ayuda de algoritmos de aprendizaje no supervisado como LDA utilizando R versión 3.0.1 y los paquetes RTextTools y Topicmodels, evalúan si los clúster detectados están relacionados con eventos espacio temporales, es decir clasificar los Tweets en grupos según su contenido.

Finalmente, (Cheng & Wicks, 2014) llegan a la conclusión que STSS es una buena alternativa en cuanto a la detección de eventos de desastres, debido a que la metodología también detecta otros eventos importantes como partidos de fútbol, retrasos de vuelos o trenes a partir de datos de Twitter.

(Bhaskar, 2017) En su investigación propone un método hibrido para localizar y ayudar a las personas afectadas después de un evento de desastre a través de la clasificación de texto en tiempo real, este método hibrido combina la clasificación basada en reglas (RBC), las características lingüísticas (LING) y propone 3 algoritmos de aprendizaje automático para la clasificación de textos: Naïve Bayes, árbol de decisión y Support Vector Machine (SVM).

Los desastres considerados en su caso de estudio incluyen las inundaciones entre India y Pakistán que ocurrieron en septiembre de 2014, una tormenta ciclónica severa llamada HUDHUD en octubre de 2014 y otra tormenta ciclónica severa llamada Nilofar.

Los datos relacionados con los desastres mencionados anteriormente se recopilan por medio de la API de Twitter, estos datos están conformados por 70.817 tweets en total, donde 30.817 tweets

pertenecen India-Pakistán en las inundaciones de Cachemira, 30.000 tweets a HUDHUD y 10.000 tweets a Nilofar; como Twitter solo permite recopilar los datos de los últimos siete días, el resto de los datos se recopilan usando un proveedor externo llamado 'Followthehashtag'.

(Bhaskar, 2017) concluye que la combinación de las reglas, rasgos lingüísticos y SVM lineal mejora el rendimiento del sistema de clasificación de texto produciendo así la más alta precisión para el método hibrido propuesto, este método se puede utilizar para reducir las victimas generadas en un desastre al recibir la información en tiempo real alimentándose de las plataformas de redes sociales.

(Resch, Usländer, Havas, & Usländer, 2017) En su investigación se presenta un enfoque para analizar las publicaciones generadas en las plataformas de redes sociales con el fin de evaluar la huella y el daño causado por desastres naturales combinando técnicas de aprendizaje automático (LDA) con análisis espacial y temporal (auto correlación espacial local para la detección de puntos calientes).

La recolección de datos asociados al desastre se hace a través de la red social Twitter, recopilando 1’012.650 tweets georreferenciados (−123.05°, 37.19°, −121.04°, 38.99°) donde 94.485 tweets fueron publicados el día del evento, para analizar las características del terremoto ocurrido el 24 de agosto de 2014 en Napa (CA, EE.UU).

Después del pre procesamiento de los tweets sin procesar, se aplica una técnica de modelado Latent Dirichlet Allocation (LDA) en forma de cascada, es decir, se aplica LDA a los resultados de la primera iteración en una segunda iteración; de esta forma se extraen los tweets relacionados

con las palabras “terremoto” y “daños”; luego realizan una validación de precisión estadística para investigar la auto correlación espacial local.

Cabe destacar que su investigación se centra en el análisis post-desastres, es decir, los tweets son analizados después de ocurrido el evento, sin embargo, se puede aplicar el campo de la detección de eventos o de alerta temprana, dependiendo de las características de la catástrofe.

(Resch et al., 2017) demuestran que la actividad de twitter aumenta significativamente tan pronto culmina el desastre a pesar de que el evento ocurre a las 3:20 am, donde se espera que la actividad de publicaciones sea baja; también se demuestra que las huellas de terremotos se pueden identificar de manera confiable y precisa como se consiguió en su estudio de caso proporcionando información valiosa sobre la naturaleza y la distribución espacial específica de las áreas afectadas por un terremoto.

(Kibanov, Stumme, Amin, & Lee, 2017) en su investigación buscan identificar las oportunidades que ofrecen las redes sociales para realizar una mejor gestión de los incendios provocados por turberas y las altas concentraciones de neblinas que generan; utilizan la red social Twitter como la principal fuente de datos además de los puntos de acceso y la calidad del aire.

Concentran su estudio en el análisis de casi todos los tweets geo-etiquetados generados en los incendios de turberas ocurridos en la isla Sumatra durante el 2014 ya que la neblina y el impacto generado es la más crítica en esta región en términos de cobertura geográfica.

(Kibanov et al., 2017) tienen como objetivo principal explorar dos temas de investigación: ¿Cómo se relacionan entre sí los eventos relacionados con “incendios” / “neblinas” de turberas y conversaciones en redes sociales sobre temas relacionados con la neblina? Y ¿Cómo capturan los

medios sociales la información situacional sobre las poblaciones afectadas, como los patrones de movilidad, que pueden ser útiles para gestionar la respuesta a desastres?

Para cada tweet de todo el conjunto de datos, independientemente de su tema (neblina, #neblina impacto-neblina, salud-neblina) calcularon la distancia más corta entre él y su punto de acceso más cercano correspondiente, esto determina si otra distribución de distancia de un tema específico de neblina tiene características similares o diferentes. Luego se lleva a cabo el mismo proceso para cada uno de los cuatro conjuntos de datos de neblina.

(Kibanov et al., 2017) demuestran que los usuarios de Twitter reaccionan a las emergencias de neblina y encuentran correlaciones entre los tweets y los hotspots de fuego de turberas; también demuestran que a pesar de las limitaciones que presentan los datos de las redes sociales se pueden utilizar para informar el manejo de desastres en diferentes etapas de una emergencia.

Partiendo de debilidades inherentes de los tweets (textos muy breves y estilos de escritura muy informales) es bastante difícil hacer una investigación del análisis de datos de tweets dando resultados con buen rendimiento y precisión, (Zou & Song, 2016) en su investigación identifican un conducto eficiente para analizar y visualizar los temas candentes de Twitter, Usando modelos de temas (LDA, su extensión CTM) para agrupar los mensajes de Twitter en diferentes temas y luego usar las palabras del tema combinadas con la influencia de los tweets para visualizar gráficamente las tendencias de los temas.

En este artículo, al examinar la sintaxis de Twitter, se propone un esquema de agrupamiento asociado a contextos significativos para un tweet para hacer que los métodos basados en LDA se ejecuten correctamente. Mientras tanto, para identificar la precisión del método propuesto, se hace

un estudio comparativo durante cada paso del proceso completo de análisis de datos. Analizan los resultados de rendimiento antes y después de presentar el esquema de agrupación para el muestreo de datos, el rendimiento antes y después del uso del peso o ponderación TF * IDF para datos, es decir comparan la precisión de CTM, LDA, CTM con TF * IDF y LDA con TF * IDF. Mostrando así desde el resultado, LDA -TF * IDF con esquema de agrupación se identifica como un conducto eficiente para encontrar temas candentes en los datos de Twitter, ya que muestra el mejor rendimiento en la investigación.

(Su et al., 2017) en su investigación tienen como objetivo argumentar la necesidad de que los expertos en comunicación confíen en las herramientas emergentes para el análisis de contenido aprovechando las fortalezas de la codificación humana y los algoritmos inteligentes; en este sentido, utilizan la energía nuclear y la nanotecnología como temas centrales rastreando la expresión de opinión en Twitter antes y después del desastre de Fukushima Daiichi utilizando el método de análisis de contenido HK; el análisis de los tweets se realizó utilizando la plataforma de análisis y monitoreo de redes sociales Crimson Hexagon ForSight; (Su et al., 2017) examinan sistemáticamente los cambios en el sentimiento y el tema de los tweets relacionados con energía nuclear después del desastre con el fin de determinar el grado en que era consistente con lo que se podría esperar después de un evento de desastre de tan alto perfil, además de esto comparan la energía nuclear con nanotecnología demostrando cómo el método híbrido empleado aquí puede rastrear el sentimiento público a través de los problemas al tiempo que introduce pocos falsos positivos u otros sesgos.

(Fersini, Messina, & Pozzi, 2017) en su investigación tienen como objetivo inferir información relevante de la gran cantidad de datos publicados en la red social Twitter, buscando identificar

mensajes de alto valor relacionados con desastre naturales, específicamente terremotos; las palabras clave utilizadas para la extracción de 1500 tweets italianos a través de la API de Twitter son: terremoto, temblor, sismo, enjambre sísmico, réplica y magnitud respecto a las preguntas "Quién", "Qué", "Dónde" y "Cuándo"; primero, para abordar las preguntas Quién y Qué se basan en un enfoque de clasificación por conjuntos llamado promedio del modelo Bayesiano (Bayesian Model Averaging); seguido, para abordar las preguntas Dónde y Cuándo se basan en un modelo probabilístico de predicción estructurada llamado campos aleatorios condicionales (Conditional Random Fields); una vez se aplican estos modelos, se realiza un filtro para identificar mensajes de alto valor que contienen información de conciencia situacional útil para la toma de decisiones, finalmente esta información se visualiza en un mapa utilizando varios tipos de perspectivas.

(Au, 2017) su investigación tuvo como objetivo evaluar el lenguaje y la configuración de los tweets relacionados con los terremotos en Rieti, Norcia - Provincia de Perugia (Italia) y Vrancea (Rumania); aplican el método directo de análisis de relaciones simbióticas para encontrar asociaciones entre las palabras: earthquake, terremoto, disaster, morto, cutremur, demolito, storm, seism, victim; las cuales utilizan como palabras claves para la identificación y extracción de tweets recopilados a través de la API de Twitter; la extracción de los datos se realizó con el método estándar de Sprizer entre el 25 de agosto y el 2 de noviembre de 2016, los datos representan 3.951.042 tweets recopilados, 1.407.791 tweets únicos y un tamaño de base de datos de 0,327 GB; (Au, 2017) utiliza la minería de datos (algoritmo EM, FP-Growth, SEAR, a priori, SVM, PageRank, bayesianos, CART) para diagnosticar patrones y grupos de datos con el fin de identificar el comportamiento de las personas ante el suceso de un desastre además de establecer el efecto y la magnitud del evento; los resultados de su investigación muestran que hay

asociaciones específicas de palabras clave y hashtags en tweets relacionados con terremotos y que la dinámica de las ocurrencias de estas palabras clave están correlacionadas.

(Sapul et al., 2017)**, c**omparan el rendimiento de los resultados de los algoritmos de clustering (k-means y CLOPE) y modelado de temas (LDA) para determinar cuál puede realizar el descubrimiento de temas de tendencia de forma más efectiva usando como caso de estudio tweets de APEC 2016 (Asia-Pacific Economic Cooperation), un foro que reúne 21 países de Asia y el pacífico, y donde su objetivo es promover el libre comercio, un crecimiento económico sostenible y la prosperidad en toda la región.

La metodología de trabajo consta de 6 pasos de pre-procesamiento para reunir y preparar hashtag y palabras como vectores destacados de los tweets (Conocimiento del dominio, recopilación de datos, pre-procesamiento de datos, selección y extracción de características, minería de datos e interpretación y evaluación). Luego, los algoritmos de agrupación para atributos numéricos y atributos categóricos, es decir, k-means y CLOPE, y el popular algoritmo de modelado de temas, la asignación latente de Dirichlet (LDA) se aplica para descubrir patrones significativos.

Los tres tipos de tweets (respuesta, Re-tweet y tweets normales o nuevos) fueron considerados en el estudio. La lista predefinida de temas se identifica y se basa en los objetivos de la cumbre APEC 2016. Estos temas predefinidos son: humano, negocios, seguridad, economía y presidente. Los conjuntos de tweets recopilados incluyen los siguientes 16 campos: Fecha del Tweet, Número de Seguidos, Nombre de pantalla o Nombre de usuario, Re-tweets por tweet, Nombre completo, Número de favoritos o similar por tweets, Texto del Tweet o Título del Tweet, Verificado o no,

Tweet ID, membresía de usuario, plataforma de aplicación, ubicación, número de seguidores, biografía, imagen de perfil e imagen.

Se utilizó una regla de búsqueda para recopilar tweets que contienen los términos #APEC, #APEC20I6, APEC y APEC20I6. Los tweets en inglés fueron los únicos considerados en el estudio debido a la barrera del idioma, y se usó la herramienta de programación R para eliminar signos de puntuación, URL, marcador "RT", y espacios en blanco.

Al comienzo del experimento, se identificaron 10.668 tweets recopilados de los cuales 8.271 eran Re-tweets. Se utiliza un conteo de frecuencia mayor o igual a diez como valor de umbral para cada conjunto de datos. Después de que se identificó el tamaño del umbral, se eliminaron los duplicados para crear tres conjuntos de datos con sus respectivos números de términos: Hashtag (82), Palabras (651) y palabras más Hashtags. (733). Para la comparación se usaron cinco, diez y veinte términos para k igual a 5, donde k es el número de conglomerados. El resultado muestra que, si agregamos más términos, el modelo puede descubrir más patrones, y al usar palabras más Hashtags da más significado que usar solo Hashtag o palabras.

(Hagras, Hassan, & Farag, 2017), en su documento analiza los resultados del uso de la técnica de modelado de temas LDA de asignación de Dirichlet latente para detectar tweets relacionados con el conjunto de datos extraídos de la transmisión de Twitter durante la crisis del tsunami en Japón en 2011. Los Tweets se seleccionan en función de su contenido, donde cada tweet en el conjunto de datos incluye la palabra "tsunami" o el hashtag "# tsunami".

Se utilizaron 3 criterios cualitativos relacionados con el desastre con el fin de filtrar los tweets (¿El tweet está relacionado con un tsunami o un desastre natural? / ¿El tweet informa algún evento

sobre tsunami? / ¿El tweet informa alguna pérdida en vidas o daños?), debido a que, en el momento de la recopilación de estos, se notó que algunos de los tweets eran genéricos e irrelevantes. Solo se eligieron los tweets que satisfacen al menos uno de los criterios para su posterior procesamiento.

Se recolectó un conjunto de entrenamiento de 6700 tweets y un conjunto de prueba de 196 tweets, ambos elegidos manualmente de un conjunto de 700.000 tweets. En los datos de prueba, la perplejidad más baja de 60.594 se alcanzó en ~ K = 2. Se diseñó un algoritmo de evaluación para probar la efectividad del algoritmo LDA. Detectó tweets con un valor de umbral de 0.1 diferencia entre las probabilidades del tema, lo que resulta en un 76% de precisión con una detección exitosa.

Las técnicas simples de aprendizaje supervisado como árboles y Nave Bayes o SVM no son soluciones óptimas para la investigación, debido a que el lenguaje humano no está estructurado porque las personas tienen distintos estilos de escritura y no hay un patrón fijo de escritura aparte de las reglas gramaticales (y aun así estas no son muy usadas en las redes sociales). Además, el aprendizaje supervisado dificulta el procesamiento y consume tiempo. Y es así que esta investigación utiliza el algoritmo de modelado de temas LDA sin supervisión para determinar si los temas de tweets están relacionados con un desastre o no.

(Li, Caragea, Caragea, & Herndon, 2018) tienen como objetivo ampliar la investigación realizada por Li et al. (2015) donde proponen un enfoque de adaptación de dominio utilizando el algoritmo de expectation–maximization (EM) con soft-labels para abordar el problema de identificar los tweets relevantes al momento de ocurrir un desastre, (Li et al., 2018) amplían el enfoque de adopción de dominio, proponiendo el uso de autoformación con hard-labels como estudio de clasificación de texto previo, esto con el fin de obtener más información sobre el

comportamiento de los desastres cuando se presentan en mayor número de diferentes tipos y con mayor cantidad de datos a analizar.

(Li et al., 2018) utilizan un conjunto de datos relativamente grande en comparación a la investigación anterior para evaluar los clasificadores de adaptación de dominio basados en Naive Bayes y self-training con hard-labels (NB-ST), para compararlos con los clasificadores de aprendizaje supervisados solo desde la fuente (NB-S) y con clasificadores de adaptación de dominio basados en expectation–maximization (NB-EM); sus resultados muestran que el uso de datos de origen solo con aprendizaje supervisado puede ayudar cuando los desastres de origen y destino son similares, sin embargo, los enfoques de adaptación de dominio serán mejores cuando solo se tienen datos de origen.

Por último, (Gründer-Fahrer, Schlaf, Wiedemann, & Heyer, 2018), investigan la estructura temática y temporal de la comunicación en redes sociales en el contexto de un desastre natural. Toman como caso de estudio la inundación de Europa Central (Alemania y Austria) que ocurre de finales de mayo a principio de julio de 2013 y recogen datos alemanes generados en las plataformas de redes sociales: Twitter y Facebook; automáticamente analizan los mensajes de las redes sociales combinando secciones transversales con perspectivas longitudinales y singulares con perspectivas comparativas. Para este estudio utilizan el algoritmo de aprendizaje no supervisado Latent Dirichlet Allocation (LDA) para la clasificación y el análisis de la información recuperada de dichas plataformas; para Facebook recuperan los datos de páginas o grupos públicos que contienen la palabra “Hochwasser” (inundación) o “Fluthilfe” (ayuda de inundación); el grupo de datos recogidos fue de 36.377 mensajes de 264 páginas y grupos públicos; para Twitter toman los datos utilizados en la investigación del proyecto de QuOIMA (QuOIMA 2011), estos datos se tomaron

filtrando sesenta y cinco (65) hashtags provenientes de investigaciones de Twitter sobre una muestra de mensajes relevantes y de términos relacionados con los escenarios de desastres utilizados por el Bundesheer austríaco, así como veintinueve (29) nombres de cuentas públicas elegidas manualmente conectadas a la gestión de desastres y ayuda a la inundación (ibid.); los datos resultantes utilizados para Twitter comprendió los 353.982 tweets.

El estudio demuestra que las técnicas de aplicabilidad de los métodos no supervisados para la extracción de información estructurada a partir de las redes sociales Twitter y Facebook en relación con la gestión de desastres son adecuadas y concluye que las dos plataformas de medios sociales Facebook y Twitter, en lugar de competir en la parte idéntica del espectro funcionan como piezas complementarias para la gestión del desastre; el modelado de esta investigación se ha convertido en parte de un prototipo de software de análisis de medios de comunicación social que está actualmente bajo consideración de autoridades alemanas para uso futuro de la gestión de desastres en su país.

Claro está que la minería de texto no ha sido utilizada solo para las redes sociales, Un claro ejemplo es la necesidad de comprender tendencias e innovaciones novedosas en la literatura científica, (Arrivillaga, Greenleaf, Hawthorn, & Alvarado, 2016), en su investigación presentan un producto de datos para la exploración y filtración de un gran corpus de citas de varias bases de datos comerciales distintas, en otras palabras, desarrollan un prototipo de producto de datos que aplica un conjunto de herramientas analíticas a un corpus organizado temporalmente de literatura científica, con el fin de proporcionar una visión de alto nivel de los temas e ideas de investigación contenidos en el corpus.

El objetivo final es una interfaz de usuario altamente interactiva para la exploración intuitiva de corpus en el front-end, respaldada por la ingestión de datos, la fusión, la inferencia y las capacidades de detección de novedades en el back-end.

La alta dimensionalidad de los datos textuales, la reducción de dimensionalidad y el resumen son requisitos principales para un análisis exploratorio efectivo, las principales herramientas analíticas usadas son el modelado de temas (en este caso LDA) y la detección de tendencias. El resultado de los modelos de temas latentes identifica grupos de términos estadísticamente coherentes y proporciona un medio de comparación de documentos. El análisis de tendencias no paramétricas se usa para detectar tendencias temporales en la prevalencia de tema y término.

Se logra una mayor reducción de la información mediante el uso de un algoritmo de detección de tendencias no paramétricas supervisadas desarrollado originalmente en el contexto de las redes sociales (Twitter), con el fin de sugerir términos de posible interés para el usuario en función de su probabilidad incorporando tendencias significativas.

La similitud de los documentos se calcula utilizando la distancia de Hellinger, que es una distancia euclidiana en un espacio de peso temático transformado, y así las consultas de documentos más cercanas se pueden implementar de manera eficiente utilizando una estructura de datos de árbol kd.

Como conclusión final del análisis de la revisión de la literatura, se pudo observar un alto uso del algoritmo LDA para detectar conjuntos de datos relacionados, bien sea en desastres como en otros campos, y debido a esto, el grupo investigador se plantea a implementar este algoritmo en el trabajo de investigación.

5. Marco teórico

En este capítulo, se abordarán las diferentes temáticas de mayor relevancia a tener en cuenta para entender los conceptos y/o teorías que se abordarán en este proyecto de investigación.

5.2. Gestión de desastres

La gestión de desastres puede definirse como la organización y la gestión de recursos y responsabilidades para abordar todos los aspectos humanitarios de las emergencias, en particular la preparación, la respuesta y la recuperación a los desastres, a fin de reducir sus efectos. (Federación Internacional de Sociedades, n.d.)

Como se puede observar en la figura 1, la gestión de los desastres se puede dividir en 3 etapas. La etapa de mitigación y preparación, donde se toman medidas para reducir al mínimo posible la pérdida de vidas humanas y daños materiales. La etapa de respuesta durante el desastre, que consiste en las acciones que se implementan en la situación de desastre para salvar vidas, aliviar el sufrimiento y reducir perdidas económicas. Por último la etapa de recuperación, que son todas aquellas actividades que están orientadas a la restauración y reparación de daños, y también incluye los esfuerzos para reducir los factores de riesgo en próximos desastres.

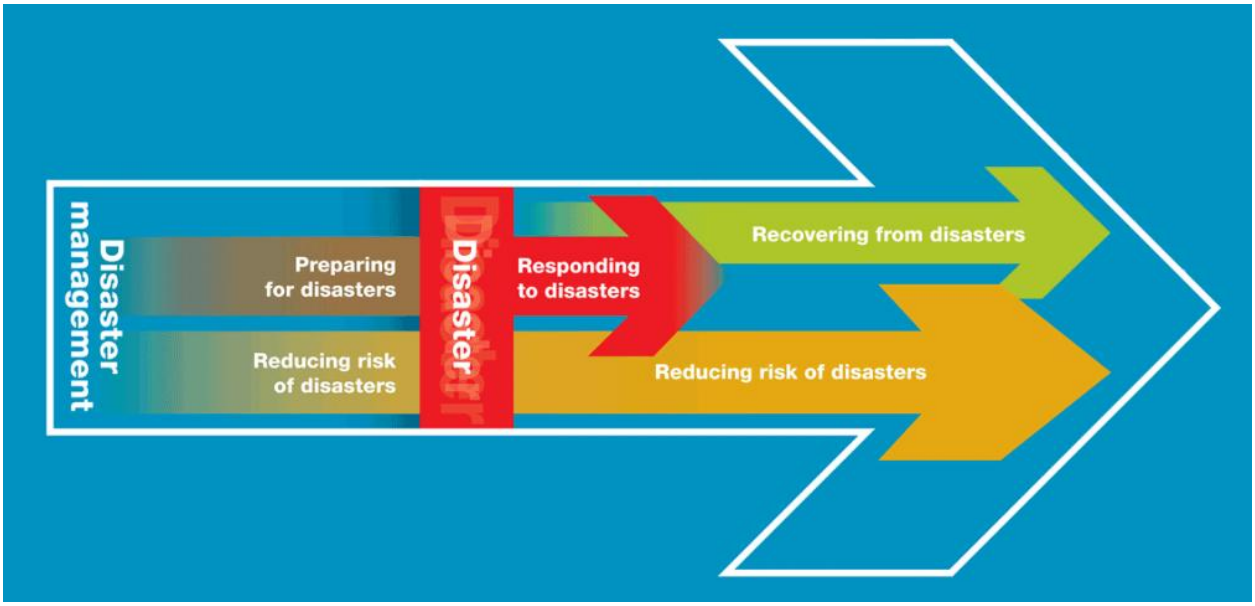

Figura 1. Etapas en la gestión de desastres, Adaptado de Federación Internacional de Sociedades, n.d

5.1. Proceso KDD

Antes de iniciar con el proceso KDD el usuario que realiza el estudio debe tener claro el objetivo final así como desarrollo y entendimiento del dominio de la aplicación, pues en este punto intervienen factores como: conocer los cuellos de botella del dominio, saber qué partes son susceptibles de un procesado automático y cuáles no, cuáles son los objetivos, los criterios de rendimiento exigibles, para qué se usarán los resultados obtenidos, compromisos entre simplicidad y precisión del conocimiento extraído.

Los pasos principales del proceso iterativo del KDD son (Ver Figura 2):

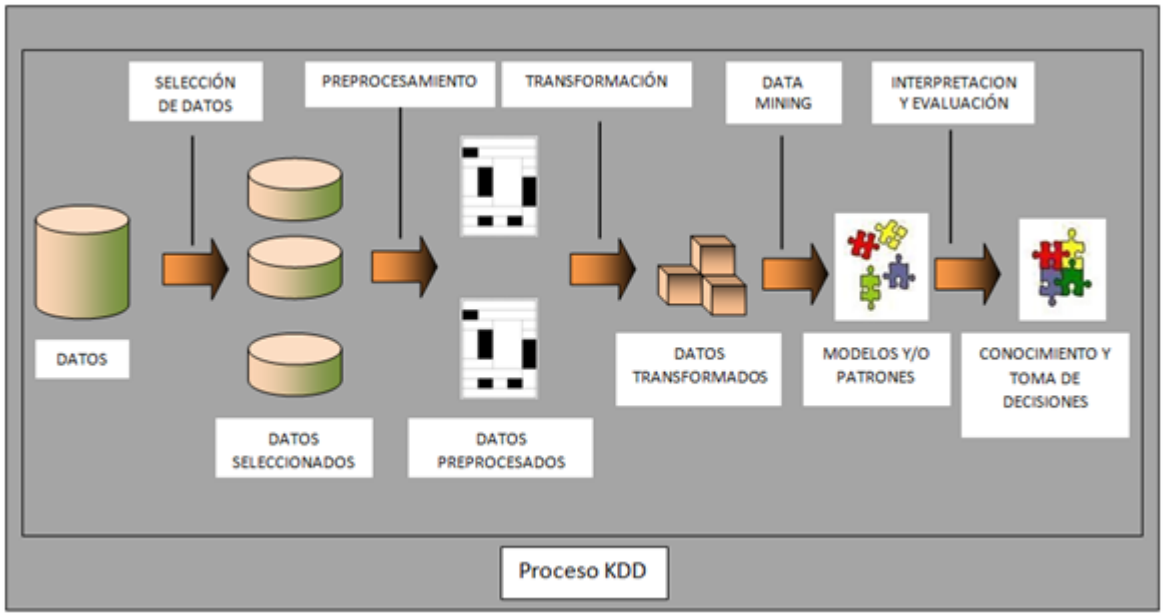

Figura 2. Proceso KDD. Adaptado de Beltrán Martínez 2014

1. Datos: En esta etapa se determina la fuente y el tipo de información a utilizar; los datos relevantes para el análisis se extraen de la o de las fuentes de datos seleccionando el subconjunto de variables sobre los que se realizará el descubrimiento, para esto se debe considerar la homogeneidad de los datos, su variación a lo largo del tiempo, estrategia de muestreo y grados de libertad.

2. Pre procesado: El pre procesado se basa en la preparación y limpieza de los datos extraídos desde las fuentes de datos en un formato manejable necesario para las fases posteriores, en esta etapa se realiza la eliminación de ruido, estrategias para manejar valores ausentes y la normalización de los datos.

3. Transformación y reducción: Consiste en el tratamiento preliminar de los datos, generación y reducción de variables y proyección de los datos en espacios de búsqueda en los que sea más fácil encontrar una solución, teniendo como resultado una estructura de datos

apropiada para las siguientes fases; esta etapa requiere un buen conocimiento del problema y una buena intuición ya que de la buena aplicación de esta etapa depende del éxito o fracaso de la minería de datos.

4. Minería de datos: En esta etapa se selecciona y realiza el modelamiento para minería de datos en donde métodos inteligentes son aplicados con el fin de extraer patrones previamente desconocidos, potencialmente útiles y comprensibles que están contenidos y ocultos en los datos, dependiendo del objetivo de KDD se selecciona el tipo de sistema: clasificación, regresión, agrupamiento de conceptos (clustering), detección de desviaciones, etc.

5. Interpretación del conocimiento: Se identifican los patrones obtenidos basándose en algunas medidas de interés del conocimiento, con posibilidad de iterar de nuevo desde el primer paso; la obtención de resultados aceptables dependerá de la definición de las medidas del conocimiento que permitan filtrar de forma automática; existen técnicas de visualización para facilitar la valoración de los resultados o búsqueda manual de conocimiento útil entre los resultados obtenidos.

Muchas veces los pasos que constituyen el proceso de KDD no están claramente diferenciados, pequeños cambios en una parte pueden afectar fuertemente el resto del proceso; sin quitarle importancia a las fases del proceso KDD, se puede decir que la minería de datos es parte fundamental del proceso y en la que más esfuerzo se realiza. (Beltrán Martínez, 2014)

5.3. Técnicas de pre procesamiento de textos

Es necesario observar que los textos, desde una perspectiva informática, son colecciones no estructuradas de palabras. Un analista de minería de texto por lo general comienza con un conjunto de textos de entrada altamente heterogéneos. Así que el primer paso es la importación de estos textos en un entorno informático, en nuestro caso R Simultáneamente es importante organizar y estructurar los textos para poder acceder de manera uniforme, a este paso se le conoce como pre procesamiento de los datos. Existen diferentes técnicas de pre procesamiento, entre ellas:

5.3.1. Stop Word. Las Stop Words o palabras vacías son las palabras más comunes que se encuentran en cualquier lenguaje natural que tiene muy poco o ningún contexto semántico significativo en una oración. Simplemente lleva la importancia sintáctica que ayuda en la formación de la oración. Como una operación de pre procesamiento, debe eliminarse para facilitar la tarea adicional y acelerar la tarea central en el procesamiento de texto. (Jaideepsinh K & Jatinderkumar R, 2016)

5.3.2. Stemming. El Stemming es un método para reducir una palabra a su raíz. Stemming aumenta el recall que es una medida sobre el número de documentos que se pueden encontrar con una consulta.

Cuando queremos buscar una palabra en documentos, tal vez nos interese utilizar este método para recuperar más documentos relacionados con la palabra. Por ejemplo si buscamos “biblioteca” a secas, nos devolverá los documentos que contengan esta palabra. Pero si aplicamos Stemming, la raíz seria “bibliotec” y nos devolvería

documentos que contengan además de “biblioteca” los que tienen “bibliotecario”. (Blanco & Sanz, 2016)

5.3.3. Document Term Matrix (DTM). Es una matriz donde las filas son los documentos, Tweets, Textos y las columnas son los términos o palabras (Token)

Para el cálculo de la matriz documento termino, se pueden usar tres tipos de medidas, el peso TF (Term-Frequency o frecuencia de termino), el peso IDF (Inverse document frequency o frecuencia inversa de documento) o la ponderación TF-IDF (Term frequency – Inverse document frequency o frecuencia de término – frecuencia inversa de documento).

- ***Factor TF:*** Frecuencia de Aparición de un Término, Es la suma de todas las ocurrencias, es decir, el número de veces que aparece un término en un documento permitiendo determinar su capacidad de representación. A este tipo de frecuencia de aparición también se la denomina "Frecuencia de aparición relativa" por que atañe a un documento en concreto y no a toda la colección. (Blázquez Ochando, 2012)

La frecuencia de aparición de un término (n) en un documento (D) es la suma de las ocurrencias de dicho término.

- ***Factor IDF:*** Frecuencia Inversa del Documento para un Término, Es el coeficiente que determina la capacidad discriminatoria del término de un documento con respecto a la colección. Es decir, distinguir la homogeneidad o heterogeneidad del documento a través de sus términos, esto significa que cuanto menor sea la cantidad de documentos, así como la frecuencia absoluta de aparición del término, mayor será su factor IDF y a

la inversa, cuanto mayor sea la frecuencia absoluta relativa a una alta presencia en todos los documentos de la colección, menor será su factor discriminatorio.

El factor IDF es único para cada término de la colección. Esto significa que su cálculo, véase ecuación 1, el IDF de un término (n) se realiza aplicando el logaritmo en base 10 de N (Número total de documentos) dividido entre la "Frecuencia de documentos para un término $df_{(n)}$" (o lo que es lo mismo el número de documentos de la colección en los que aparece el término (n) dado). Al valor resultante se le suma 1 para corregir los valores para los términos con IDF muy bajos (Aunque esta variación depende del sistema de recuperación). (Blázquez Ochando, 2012)

$$IDF_{(n)} = \log\left(\frac{N}{df_{(n)}}\right) \qquad Ecuación\ (1)$$

- ***Ponderación TF-IDF:*** Ponderación frecuencia de término – frecuencia inversa de documento, Es el producto de su frecuencia de aparición en dicho documento (TF) y su frecuencia inversa de documento (IDF) (Ver ecuación 2)

$$TF - IDF_{(n,D)} = Tf_{(n,D)}\ x\ IDF_{(n)} \qquad Ecuación\ (2)$$

5.3.4. Ejemplo del cálculo de pesos Por ejemplo, Supongamos estos tres documentos (cada frase es un documento) (Blanco & Sanz, 2016)

- Documento 1: Juega al fútbol.
- Documento 2: Le gusta el baloncesto.
- Documento 3: Mira un partido de fútbol.

La siguiente matriz tiene como filas a los documentos, y una columna por cada palabra diferente que hay en el total de documentos (vocabulario). En la figura 3 se representa la frecuencia de la palabra en el documento para cada palabra. De esta manera el documento no1 tiene las palabras “juega”, “al” y “futbol” por lo que la fila uno tiene valor 1 para esas palabras y 0 para

	juega	al	futbol	le	gusta	el	baloncesto	mira	un	partido	de
1	1	1	1	0	0	0	0	0	0	0	0
2	0	0	0	1	1	1	1	0	0	0	0
3	0	0	1	0	0	0	0	1	1	1	1

Figura 3. Transformar un documento a valores numéricos, Modificado de Blanco & Sanz, 2016

el resto.

En la tabla 2 se muestra el cálculo del peso IDF usando la ecuación 1.

Tabla 2.
Ejemplo del cálculo IDF

TERMINO	N	IDF
Juega		0,4771
Al		0,4771
Futbol		0,1761
Le		0,4771
Gusta		0,4771
El	3	0,4771
Baloncesto		0,4771
Mira		0,4771
Un		0,4771
Partido		0,4771
de		0,4771

Modificado de Blanco & Sanz, 2016

Finalmente, para el cálculo TF-IDF, y usando la ecuación, la matriz DTM con peso TF-IDF quedaría (Ver tabla 3)

Tabla 3.
Cálculo TF-IDF

	Juega	**al**	**futbol**	**le**	**gusta**	**El**	**baloncesto**	**mira**	**un**	**partido**	**de**
D1	0,4771	0,4771	0,3522								
D2				0,4771	0,4771	0,4771	0,4771				
D3			0,3522					0,4771	0,4771	0,4771	0,4771

5.4. Ejemplo conceptual en R de pre procesamiento de Tweets

Este ejemplo Tomado de (Santana, 2016) es basado en tweets copiados de la cuenta pública de @amazon, el proceso general sigue estos pasos:

1. Cargar Datos: Para este ejemplo, los datos se cargan de un archivo .csv en Dropbox, el cual tiene dos columnas: el texto y la clase a predecir. (Ver Figura 4):

texto	clase
BREAKING TECH Amazons Might Spend Billions More On Its HighRisk Kindle Experime	NEG
BREAKING TECH Amazons Might Spend Billions More On Its HighRisk Kindle Experime	NEG
AMZN Bestselling Diary of a Wimpy Kid Series Coming to Kindle http://tco/4B9jvD4f	NEG
Amazons Might Spend Billions More On Its HighRisk Kindle Experiment AMZN AAPL G	NEG
Amazons Might Spend Billions More On Its HighRisk Kindle Experiment AMZN AAPL G	NEG
Amazon: Targeting Texas Instruments Wireless Unit: Could http://tco/vGPGIDk4 Inc	POS
RT jeffmello: Tech News: Amazons Might Spend Billions More On Its HighRisk Kindle E	POS
Amazons Might Spend Billions More On Its HighRisk Kindle Experiment AMZN AAPL G	POS
Check this out on AMZN: Real Made up Monsters http://tco/S4gd7pS2 my new hilaric	POS

Figura 4. Paso 1, Cargar Datos, modificado de Santana 2016

2. Crear Corpus: El corpus es un conjunto de documentos, que pueden ser artículos periodísticos, noticias, currículos, tweets, chat, o cualquier colección de textos que se vaya a utilizar para predecir/clasificar. En este ejemplo se usan tweets, y la columna "texto" del .csv será el corpus.

   Una forma de crear el corpus en R es usando la librería tm de la siguiente forma:

   - Crear un objeto Vector Source que será el origen de los datos que luego se utilizan para crear un corpus volátil o VCorpus

- Crear un objeto VCorpus que es un corpus que se guarda en memoria, con lo cual es volátil. En este objeto se considera cada observación (tweet para este ejemplo) como un documento. El origen de datos para crear un Vcorpus siempre será un VectorSource. (Ver Figura 5)

Figura 5. Paso 2, Crear Corpus. Modificado de Santana 2016

3. Limpiar Corpus: Luego que se tiene el VCorpus, se procede "limpiar" el corpus (Ver figura 6) de la siguiente forma:

- Sustituir los signos de puntuación por espacios (replace punctuation)
- Eliminar los números (remove numbers)
- Eliminar el doble espacio (strip whitespace)
- Convertir en minúscula todas las palabras (tolower)
- Sustituir algunas palabras abreviadas (strip replace all fixed)

- Se transforma en documento plano (plaintext document). Esto para cuando se usan funciones que no retornan un Text Documents.

- Eliminar los sufijos de las palabras usando el algoritmo Porter Stemmer (stem document).

- Eliminar palabras sin significados, como pronombres, preposiciones, etc. usando el Stop Words o lista de palabras que trae la librería tm (remove words).

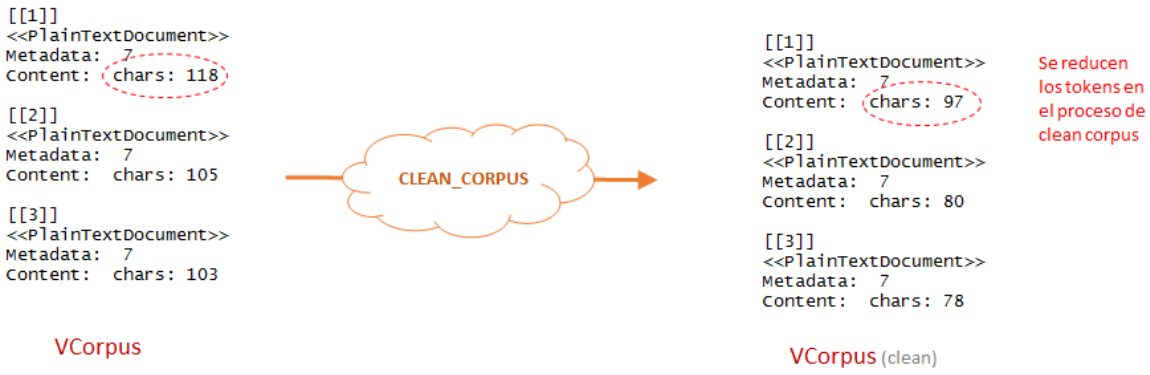

Figura 6. Paso 3, Limpiar corpus, Modificado de Santana 2016

4. Crear Document Term Matrix: Con la librería tm se pueden crear dos objetos para analizar el documento (Ver Figura 7):

- Term Document Matrix (TDM) donde las filas son los términos o palabras (token) y las columnas los documentos o tweets. Este objeto se usa para analizar frecuencia y asociación de palabras, es decir para "entender" el corpus.

- Document Term Matrix (DTM) donde las filas son los documentos o tweets y las columnas son los términos o palabras (token). Para este ejemplo se usa la DTM para crear una matriz de datos que luego se usa para modelar el clasificador. En la creación del DTM se eliminan las palabras que solo tienen una ocurrencia, así como las que

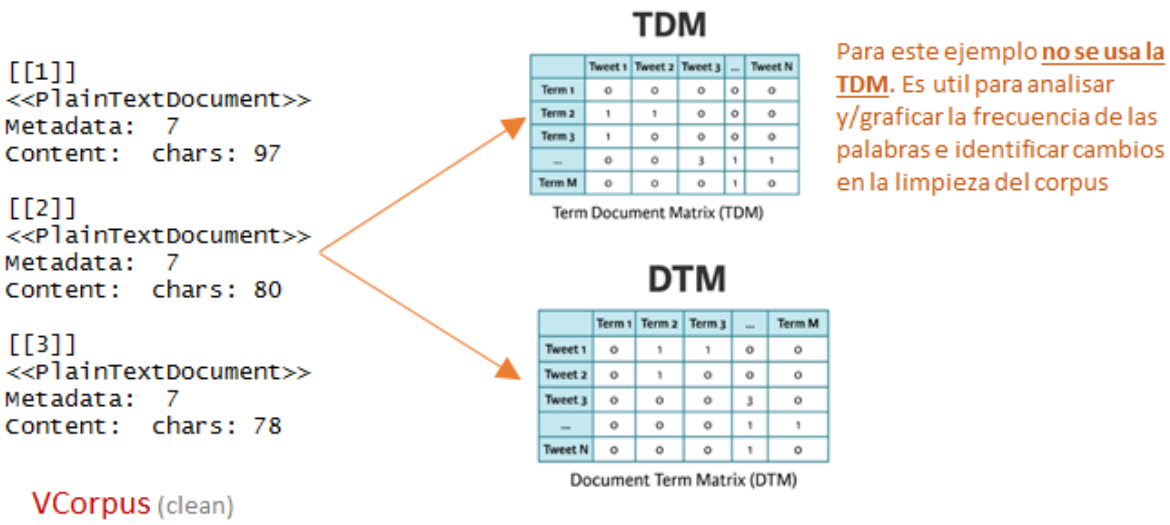

Figura 7. Paso 4, Crear Document Term Matrix, Modificado de Santana 2016

tienen pocas letras.

5. Se crea una matriz: Se usa la DTM para crear la matriz a usar (Ver figura 8)

Figura 8. Large Matrix. Modificado de Santana 2016

5.5. Clustering

El proceso de agrupar un conjunto de objetos físicos o abstractos en clases de objetos similares se denomina Clustering. Un clúster es una colección de objetos de datos que son similares entre sí dentro del mismo clúster y son diferentes a los objetos en otros clústeres. Un grupo de objetos de datos se puede tratar colectivamente como un grupo y, por lo tanto, se puede considerar como una forma de compresión de datos. (Han et al., 2011) Además, no depende de una taxonomía predefinida: la decisión de si un texto pertenece a un grupo u otro se hace dinámicamente y se basa en el contenido del conjunto de documentos. (mean cloud, n.d.)

No debe confundirse con clasificación o categorización de textos, que consiste en asignar a un solo texto una o más categorías de una taxonomía predefinida. La creación de un modelo de clasificación requiere entrenar un motor con textos preclasificados manualmente o definir una serie de reglas para cada categoría (lo que se conoce como aprendizaje supervisado) (mean cloud, n.d.)

5.6. Aprendizaje Automático

El aprendizaje automático trata de la construcción de programas que utilizando la experiencia sean capaces de mejorar automáticamente su rendimiento. Este campo ha recibido la influencia de otros muchos campos como la estadística, la inteligencia artificial, la biología y la teoría de la información, entre otros. (Ledezma, 2004)

Las técnicas de aprendizaje automático nos permiten establecer modelos utilizando datos de ejemplo o experiencias pasadas. De este modo se podrán identificar ciertos patrones o regularidades en los datos y así podremos construir buenas aproximaciones al problema, dos de los métodos de aprendizaje automático más ampliamente adoptados son Aprendizaje Supervisado y Aprendizaje no Supervisado.

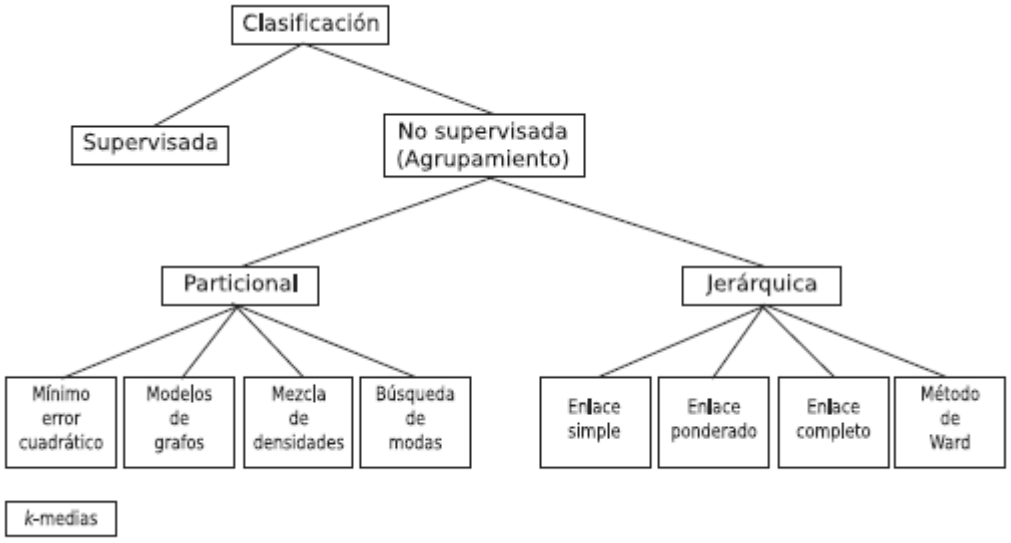

Figura 9. Árbol de tipos de métodos de clasificación. Modificado de (Palma & Marín, 2008)

5.6.1. Aprendizaje Supervisado. Se refiere a los procesos en donde se dispone de información tanto de entrada (Espacio de instancias) como de salida (Espacio de

Salida), en otras palabras, al algoritmo escogido se le proporcionan una serie de ejemplos con sus correspondientes etiquetas, es decir, que todos los ejemplos han sido clasificados "a priori". De esta forma en el proceso de aprendizaje, el algoritmo compara su salida actual con la etiqueta del ejemplo para luego realizar los cambios que sean necesarios.

Los valores de los atributos pueden ser también nominales o continuos. Si estos valores pertenecen a un número definido de clases (por ejemplo atributo sexo es categórico) se dice que es una tarea de clasificación y si el valor deseado es continuo. (Por ejemplo los atributos tamaño, peso o edad son numéricos) la tarea es una regresión. (Ledezma, 2004)

5.6.2. Aprendizaje no supervisado. La idea común de todos ellos es tratar de representar los datos de entrada del sistema de forma que esta representación refleje la estructura estadística que los define. En otras palabras, tratar de encontrar patrones o características que sean significativas en los datos de entrada. En este caso no se dispone de ninguna salida con la que comparar el rendimiento del método, por lo tanto, la única información que tendrá el sistema es la que proporcionen los datos de entrada. (Palma & Marín, 2008)

5.7. Algoritmos de agrupamiento

El proceso de agrupamiento (Clúster analysis en inglés) no es más que la organización de una colección de elementos en un conjunto de grupos homogéneos. Dichos elementos están representados por un vector de valores de atributos, es decir, son puntos de algún espacio

multidimensional. Estos valores también se suelen denominar atributos, componentes o simplemente variables. Intuitivamente, dos elementos pertenecientes a un agrupamiento válido deben ser más parecidos entre sí que aquellos que estén en grupos distintos. Partiendo de esta idea se desarrollan las técnicas de agrupamiento. Estas técnicas dependen de cómo sean los datos de partida, de qué medidas de semejanza se estén utilizando y de qué clase de problemas se estén resolviendo. (Palma & Marín, 2008)

Es importante distinguir entre clasificación supervisada (dentro de la cual se encuentra el análisis discriminante) y clasificación no supervisada (dentro de la cual se incluyen las técnicas de agrupamiento), en el primer caso, se posee la información de a qué clase pertenece cada elemento, y lo que se desea es determinar cuáles son los atributos que intervienen en la definición de las clases y qué valores de los mismos determinan estas. En el segundo caso, no tenemos ninguna información acerca de organización de los elementos en grupos o clases y, por lo tanto, el objetivo es encontrar dicha organización en base a la proximidad entre los elementos. (Palma & Marín, 2008)

Existen dos grandes grupos de algoritmos de agrupamiento, el agrupamiento Jerárquico y el agrupamiento no Jerárquico o particional, y serán explicados continuación.

5.7.1. Algoritmos de Agrupamiento Jerárquico. Los algoritmos de agrupamiento jerárquico tienen por objetivo agrupar Clústeres para formar un nuevo o bien separar alguno ya existente para dar origen a otros dos, de tal forma que, si sucesivamente se va efectuando este proceso de aglomeración o división, se minimice alguna distancia o bien se maximice alguna medida de similitud. Los métodos jerárquicos se subdividen

en Aglomerativos y disociativos. Cada una de estas categorías presenta una gran diversidad de variantes.

5.7.2. Algoritmos de agrupamiento no jerárquico o particional. Dado D, un conjunto de datos de n objetos, y k, el número de clústeres a formar, un algoritmo de partición organiza los objetos en k particiones ($k \leq n$), donde cada partición representa un clúster. Los clústeres se forman para optimizar un criterio de partición objetivo, como una función de disimilitud basada en la distancia, de modo que los objetos dentro de un clúster sean "similares", mientras que los objetos de diferentes clústeres sean "diferentes" en términos de los atributos del conjunto de datos. (Han et al., 2011)

5.8. Algoritmo K-means

Entre el grupo de algoritmos particionales es uno de los más conocidos y de los más simples, este divide la base de datos dada en k grupos de forma a priori, es decir, el algoritmo k-means toma el parámetro de entrada k, y divide un conjunto de n objetos en k clústeres para que la similitud resultante del intra clúster sea alta, pero la similitud entre clústeres sea baja. La similitud de proximidad se mide con respecto al valor medio de los objetos en un grupo, que se puede ver como centroide o centro de gravedad del grupo. (Han et al., 2011)

El algoritmo k-means funciona de la siguiente manera.

En primer lugar, selecciona aleatoriamente k de los objetos, cada uno de los cuales representa inicialmente una media o centro del grupo. Para cada uno de los objetos restantes, un objeto se asigna al clúster al que es más similar, según la distancia entre el objeto y la media del clúster. Luego calcula la nueva media para cada grupo. Este proceso se repite hasta que la función de

criterio converge. Normalmente, se utiliza el criterio de error cuadrado, definido como (Ver ecuación 3):

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} |p - m_i|^2 \quad Ecuación\ (3)$$

Donde E es la suma del error cuadrado para todos los objetos en el conjunto de datos; p es el punto en el espacio que representa un objeto dado; y mi es la media del grupo Ci (tanto p como m_i son multidimensionales). En otras palabras, para cada objeto en cada grupo, la distancia desde el objeto a su centro de grupo se cuadra, y las distancias se suman. Este criterio intenta hacer que los clústeres k resultantes sean tan compactos y tan separados como sea posible.

El problema del empleo de estos esquemas es que fallan cuando los puntos de un grupo están muy cerca del centroide de otro grupo, también cuando los grupos tienen diferentes tamaños y formas. (Han et al., 2011)

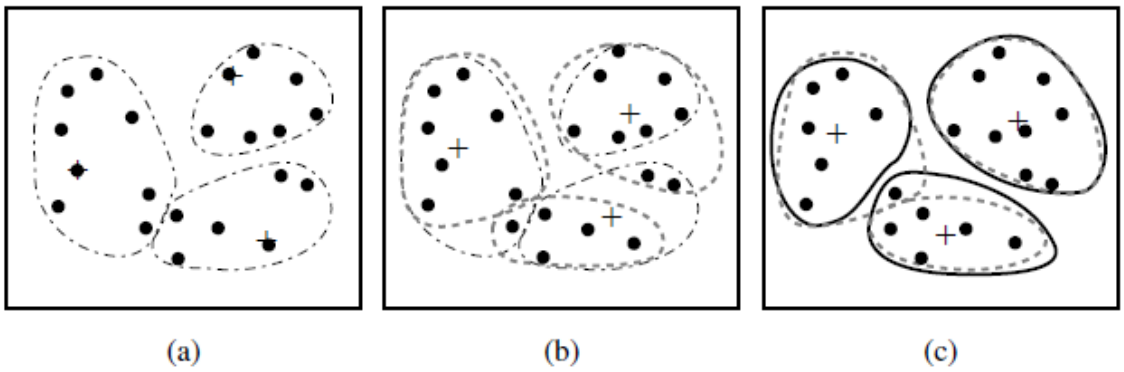

Figura 10. Agrupación de un conjunto de objetos basado en el método k-means. Modificado de libro Data mining Han 2011 Nota: La media de cada grupo está marcada con un +

5.8.1. Ejemplo del funcionamiento de K means. El proceso de K-medias inicialmente determina el número de grupos K y se asume el centroide o centro de los grupos; para determinar estos centroides se tienen dos alternativas:

- Tomar de forma aleatoria K objetos como centroides iniciales
- Tomar los primeros K en secuencia

A partir de ese momento, el algoritmo ejecuta los siguientes pasos hasta que no se muevan del grupo, a esto se le llama criterio de convergencia, los pasos son:

- Determinar el o los centroides iniciales de acuerdo al número de clústeres esperado
- Determinar la distancia de cada objeto con relación a los centroides
- Agrupar los objetos con base en la distancia mínima.

El proceso se representa en el siguiente diagrama de flujo (Ver figura 11):

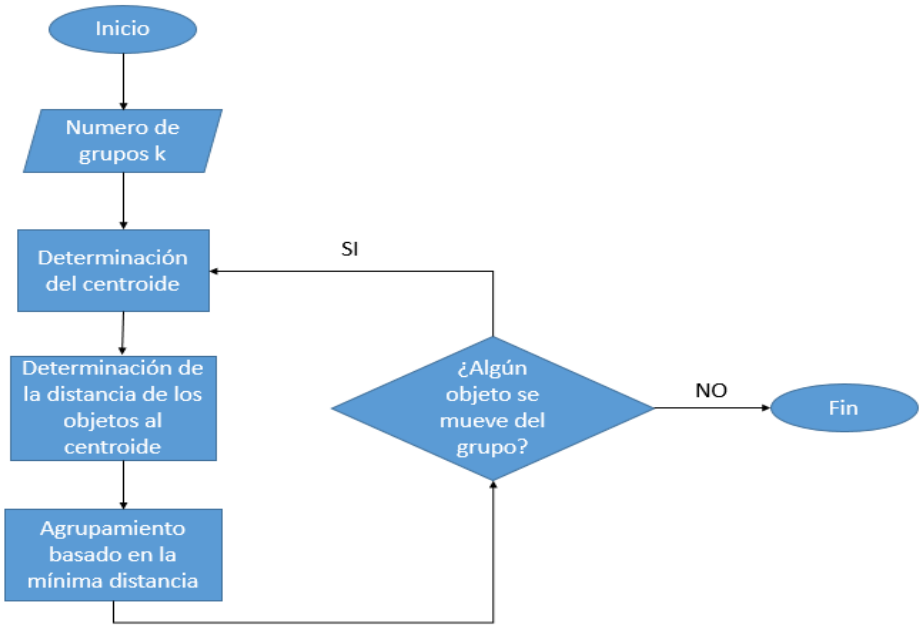

Figura 11. Diagrama de flujo proceso K-means. Modificado de (Teknomo, n.d.)

A continuación, se muestra cómo funciona la iteración, este ejemplo fue adoptado de (Teknomo, n.d.):

Supongamos que tenemos cuatro tipos de materiales con dos atributos: peso y dureza, (Ver Tabla 4)

Tabla 4.
Datos ejemplo K means

MATERIAL	PESO	DUREZA
A	2	2
B	4	1
C	1	3
D	3	4

Cada material puede ser representado como un punto en un espacio coordenado (Ver Figura 12)

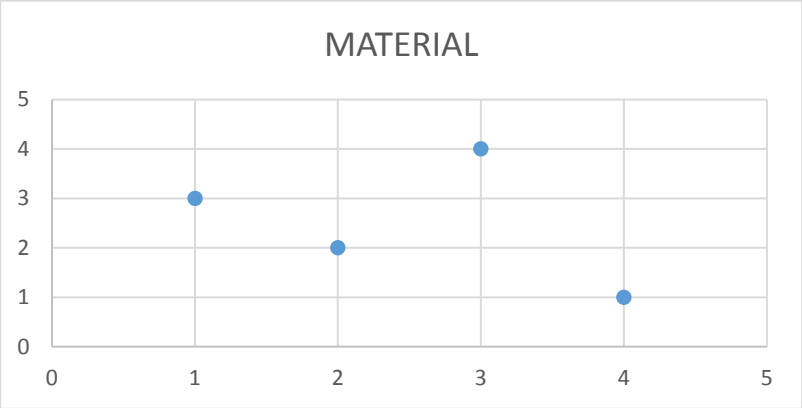

Figura 12. Representación del material, ejemplo k means.

Determinamos los centroides de forma aleatoria: C1= (2,2) C2= (4,1) (Ver Figura 13)

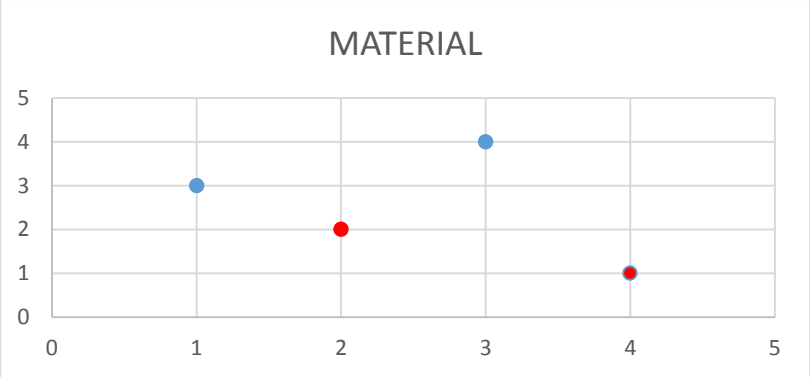

Figura 13. Representación centroides de forma aleatoria, ejemplo K means

Ahora calculamos la distancia de cada objeto respecto a los centroides utilizando la distancia euclidiana:

$$Dist_{XY} = \sqrt{\sum_{k=1}^{m} (X_{iK} - X_{jK})^2} \qquad Ecuación\ (4)$$

Cálculo de las distancias:

$$D_{A,\,C1} = ((2-2)^2 + (2-2)^2\ )\text{^}(1/2) = 0$$

$$D_{B,\,C1} = ((4-2)^2 + (1-2)^2\ )\text{^}(1/2) = 2{,}24$$

$$D_{C,\,C1} = ((1-2)^2 + (3-2)^2\ )\text{^}(1/2) = 1{,}41$$

$$D_{D,\,C1} = ((3-2)^2 + (4-2)^2\ )\text{^}(1/2) = 2{,}24$$

$$D_{A,\,C2} = ((2-4)^2 + (2-1)^2\ )\text{^}(1/2) = 2{,}24$$

$$D_{B,\,C2} = ((4-4)^2 + (1-1)^2\ )\text{^}(1/2) = 0$$

$$D_{C,\,C2} = ((1-4)^2 + (3-1)^2\ )\text{^}(1/2) = 3{,}61$$

$$D_{D,\,C2} = ((3-4)^2 + (4-1)^2\ )\text{^}(1/2) = 3{,}16$$

Después de esta primera iteración obtenemos la siguiente matriz:

$$Do = \begin{pmatrix} 0 & 2{,}24 & 1{,}41 & 2{,}24 \\ 2{,}24 & 0 & 3{,}61 & 3{,}16 \end{pmatrix}$$

Dada la matriz se asigna cada objeto al grupo en el cual tiene el mínimo error de partición de cada elemento en cada grupo, es decir, donde la distancia es minina, siguiendo esto el material A se asigna al centroide 1, el material B al centroide 2, etc..

$$Go = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Al tener los miembros de cada grupo, volvemos a calcular los centroides:

$$C1 = \left(\frac{2+1+3}{3}\ , \frac{2+3+4}{3}\ \right) = \left(\frac{6}{3}\ , \frac{9}{3}\right) = \ (2\ , 3)$$

$C2 = \left(\frac{4}{1}\ , \frac{1}{1}\right) = \ (4, 1)$, obtenemos lo siguiente:

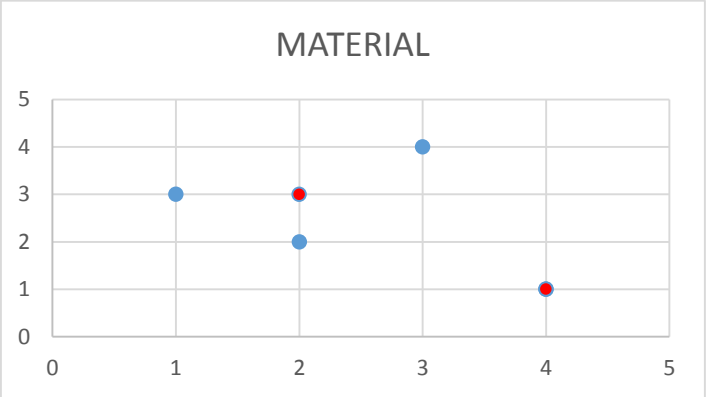

Figura 14. Representación de centroides finales, ejemplo K means.

Nuevamente calculamos la distancia respecto a los nuevos centroides:

$$D1 = \begin{pmatrix} 1 & 2{,}82 & 1 & 1{,}41 \\ 2{,}83 & 0 & 3{,}61 & 3{,}16 \end{pmatrix}$$

Se asigna nuevamente cada objeto al respectivo centroide:

$$G1 = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

En este punto los objetos quedaron asignados a los mismos centroides, obteniendo la misma clasificación que la asignación anterior, por lo tanto, se concluye que los centroides se han estabilizado y no es necesario continuar iterando.

5.9. Latent Dirichlet Allocation (LDA)

La asignación latente de Dirichlet (LDA, por sus siglas en inglés) es un método particularmente popular para ajustar un modelo de tema. Trata cada documento como una mezcla de temas, y cada tema como una mezcla de palabras. Esto permite que los documentos

se "superpongan" entre sí en términos de contenido, en lugar de separarse en grupos discretos. (Silge & Robinson, n.d.)

Es un modelo generativo que permite que conjuntos de observaciones puedan ser explicados por grupos no observados que explican por qué algunas partes de los datos son similares, es decir, se trata de un modelo probabilístico, ya que dice que un documento se crea mediante la selección de los temas y las palabras de acuerdo a las representaciones probabilísticas del texto natural.

La probabilidad inherente en los modelos de selección de cada palabra se deriva del hecho de que el lenguaje natural nos permite utilizar múltiples palabras diferentes para expresar la misma idea. Expresar esta idea en el modelo LDA, sirve para crear un documento sin un corpus, lo que se podría determinar cómo una distribución de temas. Para cada palabra del documento que se está generando, se escoge un tema de una distribución de Dirichlet de temas. A partir de ese tema, se elige una palabra seleccionada al azar basada en otra distribución de probabilidad condicionada en ese tema. Esto se repite hasta que el documento se ha generado. (Benitez Andradez, José; Valvuena Lopéz, 2011)

En LDA, cada documento puede verse como una mezcla de varias categorías. Esto es similar a Probabilistic Latent Semantic Analysis (PLSA), excepto que en LDA se asume que la distribución de categorías tiene una distribución a priori de Dirichlet. La clave en LDA es que las palabras siguen una hipótesis de bolsa de palabras o, más bien que el orden no importa, que el uso de una palabra es ser parte de un tema y que comunica la misma información sin importar dónde se encuentra en el documento, este supuesto es necesario para que las probabilidades sean intercambiables y permitan una mayor aplicación de métodos matemáticos. Las palabras

que aparecen con menos frecuencia en los documentos únicos, pero son comunes en muchos documentos diferentes probablemente son indicativo de que existe un tema común entre los documentos. Cuando se genera un resumen, la capacidad de recoger los matices de los temas del documento permiten que la información más relevante sea incluida con menos posibilidades de repetición y dar así un resumen mejor. (Benitez Andradez, José; Valvuena Lopéz, 2011)

El modelo de tópicos LDA asume que en un corpus existe una "estructura oculta de tópicos" y una Estructura de "variables observadas". Lo que hace LDA es que atrapa dicha intuición en un "modelo de variables ocultas". En un modelo de variables ocultas se practica y postula una estructura oculta en los datos observados en el corpus correspondiente. Luego de obtener dicha estructura, se aprende de esta misma usando "probabilidades a posteriori" lo que se denomina un modelo probabilístico generativo. (Chandía Sepúlveda, 2016)

5.9.1. Proceso generativo de LDA. El objetivo del LDA en este paso es modelar el proceso generativo real por uno "sintético", que se aproxime al real, y tratar de encontrar parámetros para este, que se ajusten bien (o lo mejor posible) a los datos. (Cenditel, 2016)

El proceso supone que un documento se genera como mezclas de palabras de tópicos con cierta probabilidad. En concreto, el LDA asume el siguiente proceso generativo para un corpus:

1. Se establece el vocabulario a usar.

2. Se determinan un número (fijo) K de tópicos, con su respectiva distribución de palabras (Distribución multinomial).

3. Se establece el número M de documentos que tendrá el corpus.

4. Para cada uno de los M documentos:

    4.1. Se establece el número N de palabras que tendrá el documento (Por ejemplo de acuerdo a una distribución Poisson).

    4.2. Se elige una distribución θ de tópicos para el documento (de acuerdo a una distribución Dirichlet (α) sobre el conjunto fijo de K tópicos).

    4.3. Para cada una de las N palabras:

        4.3.1. Se selecciona un tópico Zn. (de acuerdo a una distribución multinomial (θ))

        4.3.2. Se selecciona una palabra del tópico (de acuerdo a la distribución de palabras en el tópico establecida en el paso 2).

Este proceso define una distribución de probabilidad conjunta sobre ambas, las variables ocultas y las variables observadas. El análisis de los datos se construye usando la distribución de probabilidad conjunta para calcular la distribución condicional de las variables ocultas dadas y las variables observadas. Esta distribución condicional es lo que en estadística bayesiana se llama distribución a posteriori. (Cenditel, 2016).

5.10. Selección del número óptimo de Cluster's

Una pregunta que surge a la hora de estudiar los algoritmos de clustering es, ¿Cómo hallar el número óptimo de Clusters? A continuación se muestran 4 métodos para determinar el número de clusters a la hora de aplicar algoritmos de este tipo. (Moya, n.d.)

5.10.1. Elbow method. Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferentes números de clusters (desde 1 a N clusters), siendo la inercia la suma de las distancias al cuadrado de cada objeto del cluster a su centroide:

$$Inercia = \sum_{i=0}^{N} ||x_i - \mu||^2 \qquad Ecuación\ (5)$$

Una vez obtenidos los valores de la inercia tras aplicar el K-means de *1* a *N* clusters, representamos en una gráfica lineal la inercia respecto del número de clusters. En esta gráfica se debería de apreciar un cambio brusco en la evolución de la inercia, teniendo la línea representada una forma similar a la de un brazo y su codo. El punto en el que se observa ese cambio brusco en la inercia nos dirá el número óptimo de clusters a seleccionar para ese data set; o dicho de otra manera: el punto que representaría al codo del brazo será el número óptimo de clusters para ese data set.

Como se puede apreciar en los resultados obtenidos, el método del codo devuelve unos valores muy coherentes y se ajusta a los resultados esperados. Es posible que al aplicar este método para otro conjunto de datos no se aprecie "el codo" o incluso se observen dos o más codos (o cambios bruscos en la evolución de la inercia). En ese caso habría

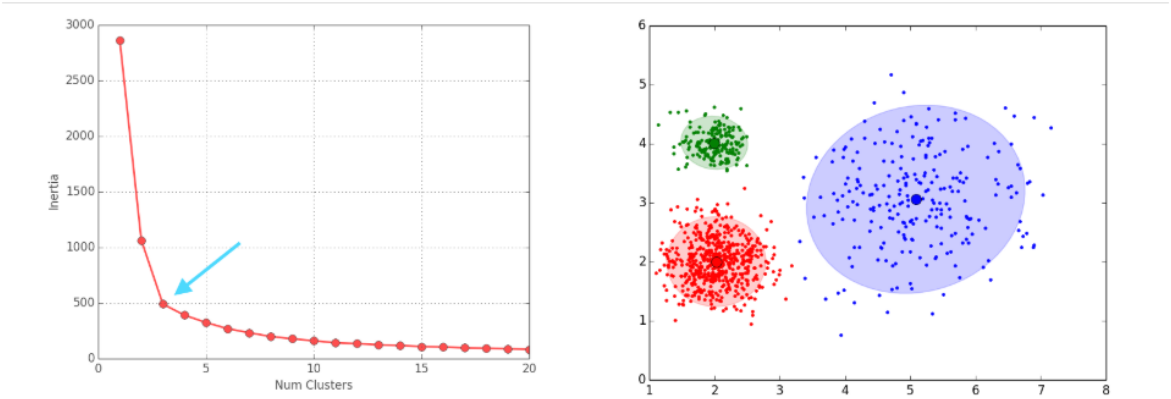

Figura 15. Elbow Method, Modificado de (Moya n.d)

que estudiar más en detalle o con otras técnicas el número óptimo de clusters a seleccionar.

5.10.2. Silhouette method. El método de silueta promedio calcula la silueta promedio de las observaciones para diferentes valores de k. El número óptimo de conglomerados k es el que maximiza la silueta de promedios sobre un rango de valores posibles para k (Kaufman y Rousseeuw, 1990).

El algoritmo es similar al método del codo y se puede calcular de la siguiente manera:

- Calcular el algoritmo de agrupamiento (por ejemplo, k-means clustering) para diferentes valores de k. Por ejemplo, variando k de 1 a 10 clusters.

- Para cada k, calcular la silueta promedio de las observaciones (avg.sil).

- Trazar la curva de avg.sil de acuerdo con la cantidad de conglomerados k.

- La ubicación del máximo se considera como el número apropiado de racimos.

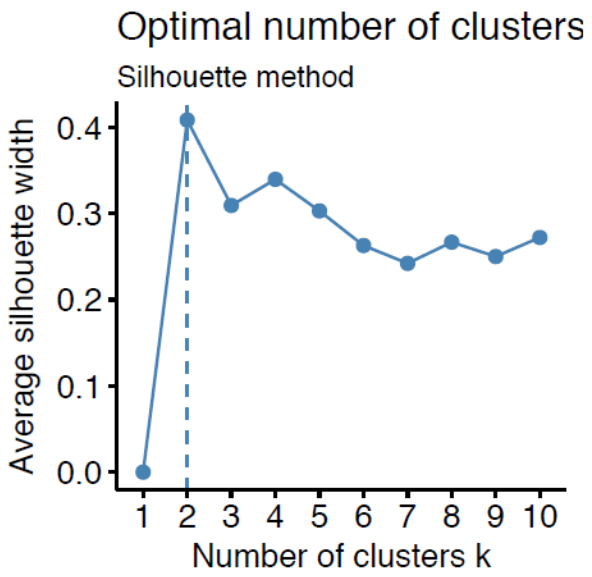

Figura 16. Silhouette Method. Modificado de (Moya n.d)

5.10.3. Gap. (brecha); similar al método del codo, cuya finalidad es la de encontrar la mayor diferencia o distancia que hay entre los diferentes grupos de objetos que vamos formando para representarlos en un dendrograma. Para ello vamos cogiendo las distancias que hay de cada uno de los enlaces que forman el dendrograma y vemos cual es la mayor diferencia que hay entre cada uno de estos enlaces. El script que se muestra a continuación calcula estas diferencias (para alguno de los 3 data sets de los que

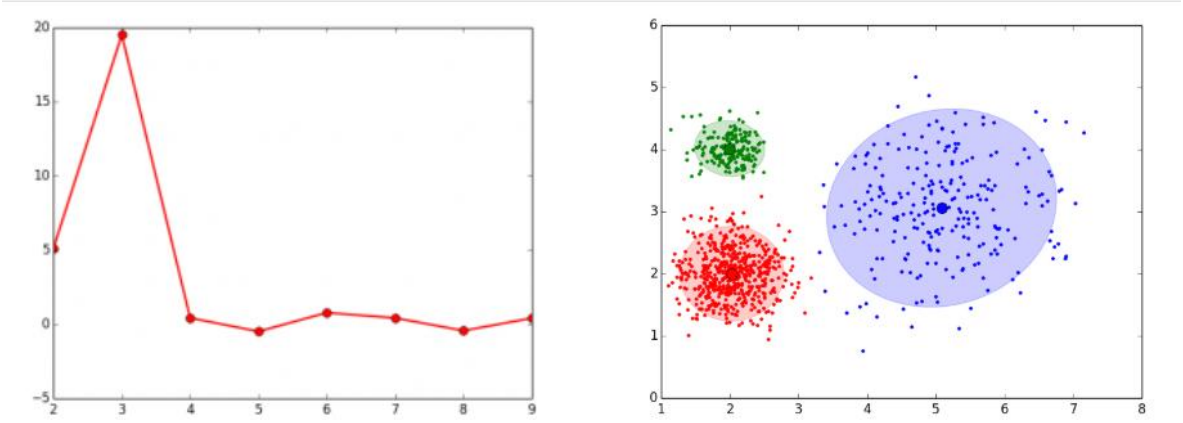

Figura 17. Método GAP, Modificado de (Moya n.d)

disponemos) y las representa en una gráfica lineal, siendo el punto máximo, el número de Clusters óptimo para ese data set.

5.10.4. Máxima verosimilitud. Es un método habitual para ajustar un modelo y estimar sus parámetros, y funciona probando con diferentes valores de k (# de grupos) cuál de ellos produce la máxima verosimilitud (log-likehood) dado los datos.

6. Caso de Estudio

Para la evaluación de los dos modelos de aprendizaje no supervisado se eligió una base de datos de mensajes que se dieron durante un desastre natural, posteriormente usando una API que es

nuestro caso es la de Twitter se recuperaron los tweets generados por los usuarios. (Ver numeral 6.1)

Antes de programar los algoritmos, es muy importante hacer una limpieza o pre-procesamiento de la base de datos (Ver numeral 6.2) Consecuentemente, una vez preparados los datos se programan y ejecutan los algoritmos K-means y LDA donde los resultados se verán en el numeral 6.3

6.1. Evento de desastre: Incendios forestales California (Estados Unidos)

Para recolección de la base de datos se hizo una revisión periódica en páginas de internet de aviso de desastres naturales como (El CRED[1]). Una vez detectado el desastre natural se procede a hacer una revisión con palabras clave en Twitter con el fin de revisar que tan comentado es este desastre en esta plataforma, si el desastre no es muy comentado, continuamos con la revisión de los desastres, hasta encontrar uno con una tendencia considerable en la red social, todo esto con el fin de consolidar una base de datos con una cantidad considerable de tweets.

En este proyecto se usó como caso de estudio los incendios de California (Camp fire, Woolsey fire y Hill fire).

En total hubo 89 muertos, donde Camp fire en el condado de Butte, con 86 muertos, más de 18733 estructuras destruidas fue el incendió más grande y con mayor devastación de los 3 desastres, las otras 3 muertes fueron causadas por el incendio de Woolsey (en los Ángeles) donde se presentaron daños a unas 1500 estructuras; por su parte en el incendio de Hill ubicado en el

[1] Centre for Research on the Epidemiology of Disasters

condado de Ventura no presentó muertos y hubo 4 edificios destruidos. (San Francisco Chronicle, 2018)

Inicialmente se identificaron los hashtags #Campfire, #Woolseyfire, #Hillfire y #california, pero debido a la capacidad reducida de procesamiento del computador a usar, no fue posible usar todos los hashtags, y al final, se tomó solo el hashtag #california, porque era el hashtag más global y que abarcaba información de los tres incendios.

La descarga de los tweets se hizo el 16 de noviembre de 2018 basados en el hashtag #california donde se obtuvo un total de 17301 tweets, los cuales 13504 eran en Inglés y el resto en otros idiomas como español, italiano y francés.

6.2. Preparación de los datos (limpieza o pre-procesamiento de la base de datos)

Para el primer paso en la preparación de los datos, se seleccionan únicamente los tweets en inglés (13504 tweets), debido a que es el idioma predominante.

Seguidamente, se toma solo la información texto, luego se eliminan los hashtags (#) las etiquetas (@) y los links adjuntos al texto, así como los signos de puntuación, los números y se convierte todo el texto en minúscula, y finalmente en este paso se eliminan los tweets repetidos y los tweets vacíos, quedando así un total de 8019 tweets.

En la siguiente etapa se construye el corpus y se procede a hacer la tokenización del mismo, pero antes se hace una nueva limpieza del texto, ahora se eliminan palabras que no agregan valor a la investigación (stop words) y además se hace stemming con el fin de que las palabras con misma raíz se junten en una sola y no se genere ruido, una vez esto se procede con la construcción

del Document Term Matrix (DTM) (DTM con ponderación TF-IDF para k-means y DTM con peso TF para LDA) y se convierte en un espacio vectorial (data frame) para iniciar el proceso de análisis. Es importante tener en cuenta, que al momento de la eliminación de los stop words y el stemming, algunos tweets pueden quedar vacíos, es por esta razón que se eliminan los documentos vacíos de la DTM y se toman en cuenta solo los documentos que contienen palabras, de esta manera la base de datos totalmente limpia queda reducida a un total de 8011 tweets.

6.3. Ejecución de los algoritmos y Resultados

En esta sección se describen los pasos para la ejecución de los algoritmos K-means y LDA así como los respectivos resultados de cada uno.

6.3.1. K-means. Antes de ejecutar el algoritmo, es necesario hallar el número de clusters en los que se dividirá la base de datos, en este caso, se usó el método de silueta, arrojando un valor de k=4 (Ver figura 18). Se ejecuta el algoritmo K means utilizando la distancia Manhattan y la variación del algoritmo K-means Mc-Queen, (empezando desde 50 conjuntos aleatorios, hasta un máximo de 100) arrojando la siguiente distribución de los tweets (Ver tabla 5)

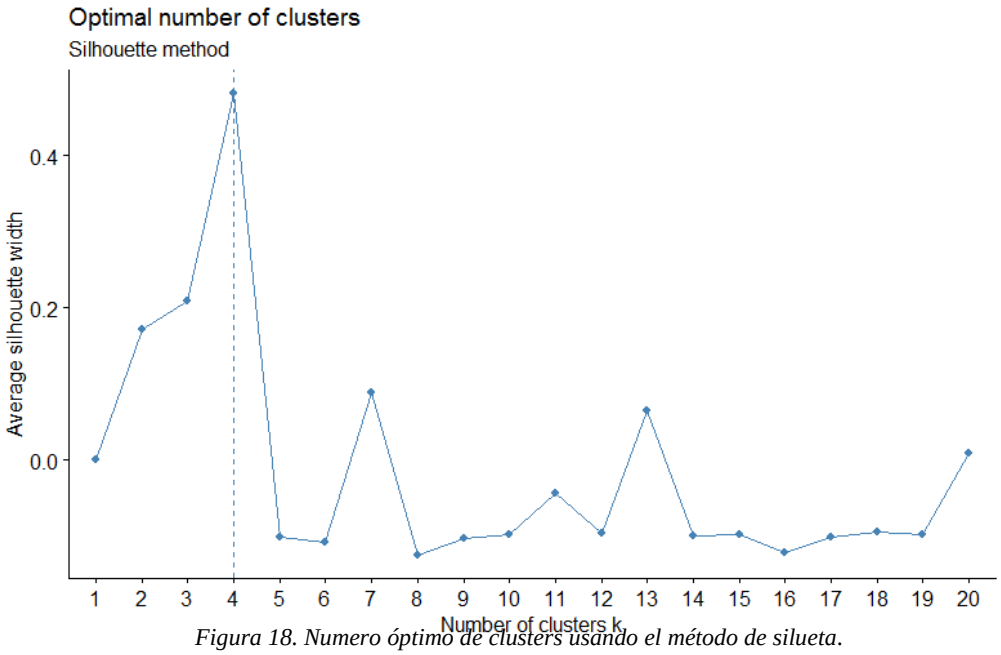

Figura 18. Numero óptimo de clusters usando el método de silueta.

Tabla 5.
Distribución de tweets K-means

Clúster	Cantidad
1	908
2	2965
3	2765
4	1373

6.3.2. LDA. Así mismo como para el algoritmo k-means, se necesita encontrar el número de grupos en los que se va a clasificar la base de datos, (con lo cual se usa el método máxima verosimilitud) Arrojando un valor de 18 grupos (ver figura 19)

Es importante aclarar que para la ejecución del algoritmo LDA se usa DTM con pesos de frecuencia de termino (TF), no como en K-means que se usan pesos TFIDF, Se ejecuta el algoritmo LDA con su variante Gibbs, y como resultado arroja la siguiente distribución de los tweets en los 18 grupos dados. (Como LDA es un modelo de tipo probabilístico, los documentos (tweets) están asignados basados en que tema tienen mayor probabilidad de pertenecer). (Ver tabla 6)

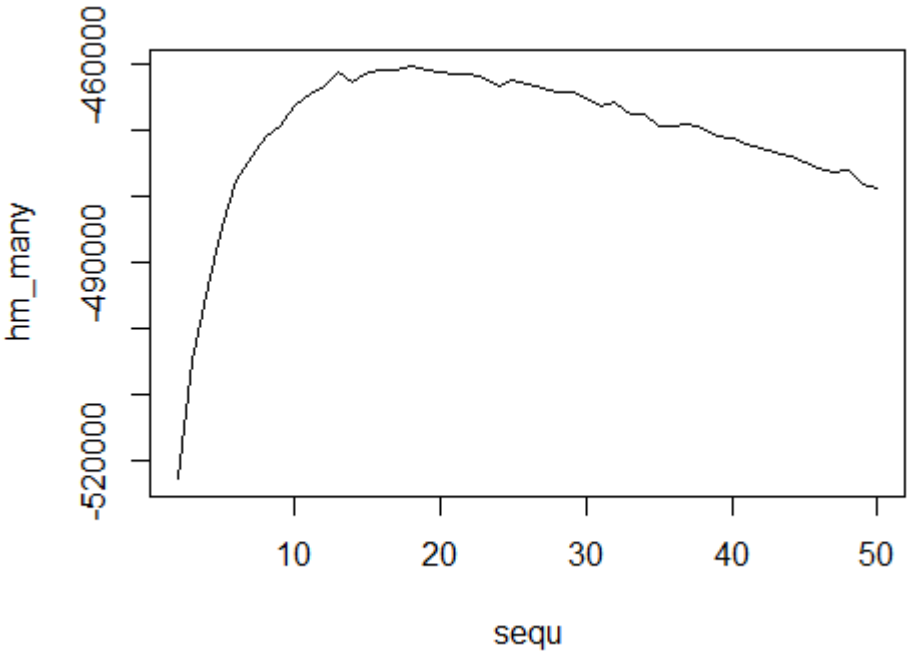

Figura 19. Número óptimo de clusters, usando el método máxima verosimilitud

Tabla 6.
Distribución de tweets LDA

Clúster	Cantidad
1	798

Clúster	Cantidad
2	821
3	632
4	597
5	531
6	461
7	531
8	396
9	379
10	309
11	330
12	423
13	271

Clúster	Cantidad
14	305
15	268
16	251
17	220
18	488

7. Análisis de resultados y comparación

El objetivo de este capítulo es identificar patrones que permitan entender el contenido de la base de datos mediante el análisis de los clúster producto del resultado de cada algoritmo, y finalmente comparar estos mediante un cuadro comparativo o donde bajo ciertos criterios se determinará que algoritmo bridó mejores resultados.

7.1. Análisis de clusters K-MEANS

7.1.1. Cluster 1. En el cluster 1 se encuentran agrupados 908 tweets (el más pequeño de los grupos), donde se puede observar que las palabras más repetidas son Beach (playa), live (vivir), rock (atribuyéndose a el género musical), sunset (atardecer) entre otros. (Ver figura 20), lo que nos lleva a interpretar que en este grupo se encuentran tweets que se refieren a cosas, experiencias que las personas quieren compartir sobre la vida en California, no

parece tener relación al desastre en particular, y esto puede ser debido a que el hashtag usado (#california) se puede considerar muy general, y no muy explícito hacia el desastre.

Algunos ejemplos del tipo de tweets que se pueden encontrar en este grupo son:

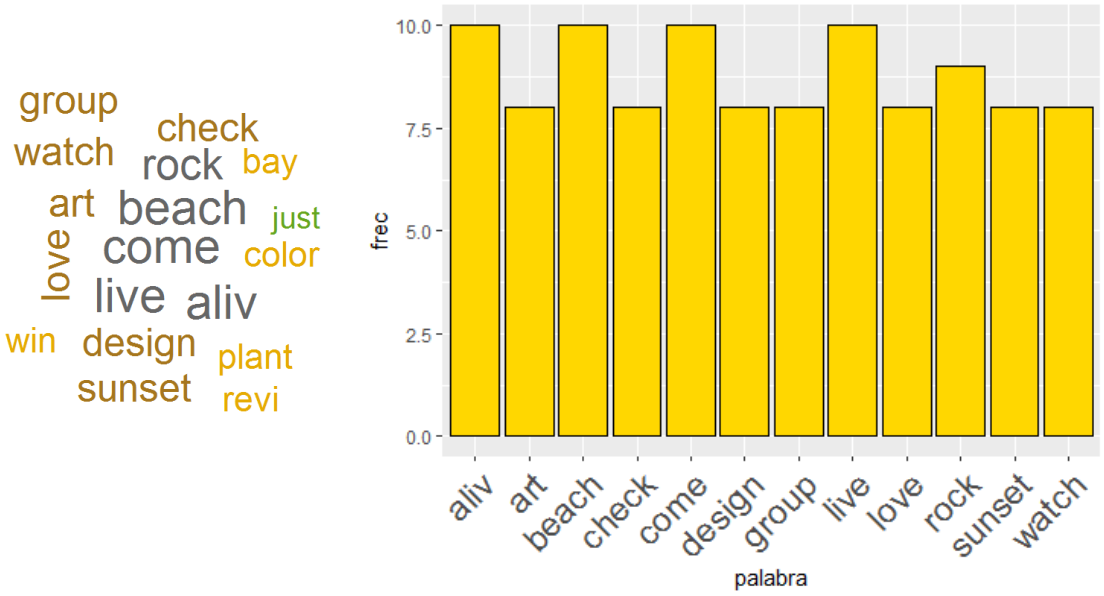

Figura 20. Cluster 1. Algoritmo kmeans

- "Watching the sunrise in the east Venice beach boardwalk CA[1]" // *"Mirando el amanecer en el este de la del paseo marítimo de la playa de Venecia CA"*

- "Having a little workout and fun at the famous gym Venice muscle beach, in Venice beach boardwalk CA" // *"Tener un poco de ejercicio y diversión en el famoso gimnasio Venice muscle beach, en el paseo marítimo de la playa de Venice CA"*

- "Everyone knows the coast is beautiful but add a five star resort is a fantastic spa and a scenic golf course and its downright irresistible" // *Todo el mundo sabe que la costa*

[1] CA: Diminutivo de California

es hermosa, pero agregarle un resort de cinco estrellas es un fantástico spa y un pintoresco campo de golf y es absolutamente irresistible.

7.1.2. Cluster 2. Para el cluster 2 con 2965 tweets (el grupo con más cantidad de tweets), en la Figura 21, se observa que las palabras que más destacan son fire (fuego/incendio), people (personas), help (ayuda), need (necesitar), etc.

Estas palabras nos indican que los tweets de este grupo hablan acerca de los incendios que se estaban presentando en California, más específicamente manifestando la necesidad de ayuda a las personas, y donde se puede corroborar en los ejemplos que a continuación se presentan.

- Send the troops on the border to #CALIFORNIA to help with the victims, They're badly needed THERE-MUCH more than sitting around waiting for a caravan // *Envíen las tropas de la frontera a CALIFORNIA para ayudar con las víctimas, ellos son mucho más necesarios que estar sentados esperando un carnaval.*

- California wildfires: Town of Paradise will need 'total rebuild // *Incendios California: El pueblo de Paradise necesitará total reconstrucción.*

- Allyn Pierce[1] lives in #Paradise, #California. Pierce is a nurse and intensive care unit manager at his local hospital so he helps others daily and is likely a hero to some. // *Ally Pierce vive en Paradise, California. Pierce es un enfermero y jefe de la unidad de*

[1] Ally Pierce: Es un enfermero que ayudó a evacuar a los pacientes de un hospital conduciendo su camioneta a través de las llamas

cuidados intensivos in su hospital local, el ayuda a otros diariamente y es probablemente un héroe para algunos.

- The #Campfire, the most deadly and destructive wildfire in #California history, has destroyed hundreds of homes, put thousands of lives at risk and separated pets from their owners// *Campfire, el incendio forestal más mortal y destructivo en la historia de California, ha destruido cientos de hogares, ha puesto miles de vidas en riesgo y separado mascotas de sus dueños.*

- HEY #Resistance do YOU think a nuke would only kill #Gunowners fallout is NOT selective, looks like we need rules of engagement for #CivilWar you #California people are nuts, you elect morons like this, psst he would murder YOU too. // *"Hey #resistencia, ¿Ustedes creen que una bomba nuclear solo mata a los portadores de armas?, NO es selectivo, parece que necesitamos reglas de compromiso para la #GuerraCivil, ustedes los californianos están locos, Eligen personas como esta, posdata, él también te mataría"*

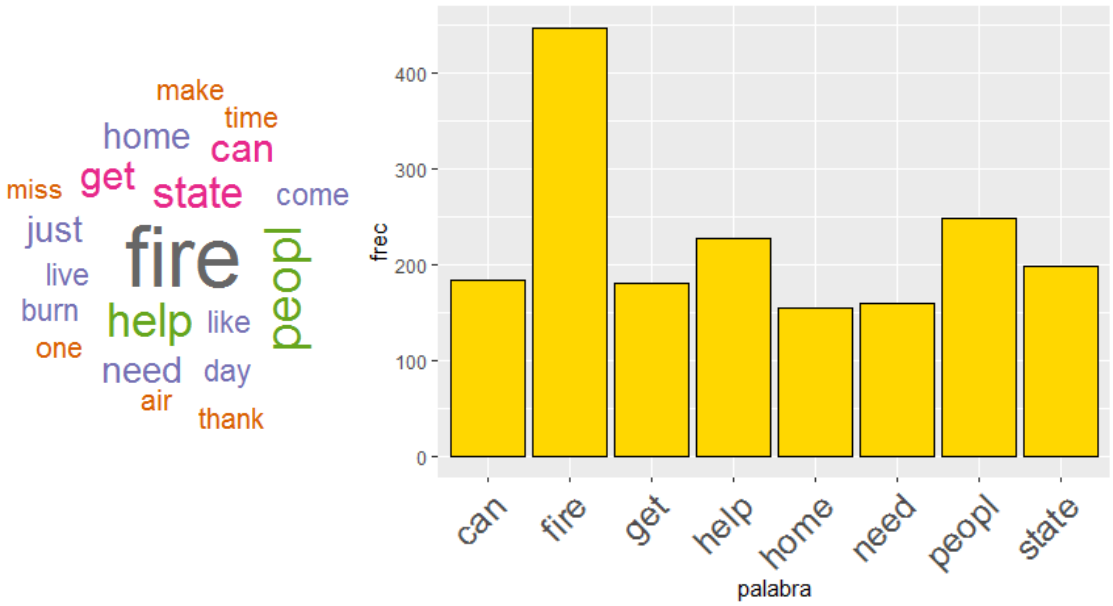

Figura 21. Cluster 2. Algoritmo Kmeans

Es importante destacar que en este grupo también se presentan tweets sin contenido alusivo al incendio (contrario a lo que la frecuencia de palabras hacen alusión en un primer vistazo), como se muestra en el último ejemplo, el cual habla acerca de una opinión acerca de las bombas nucleares y la crítica a la inclinación política de las personas en california.

7.1.3. Cluster 3. En cuanto al cluster 3 con 2765 tweets y observando la figura 22, se puede notar la presencia de gran parte de las palabras que se evidencian en el grupo anterior, tales como fire, get, need, llevándonos a inferir que este grupo tiene un comportamiento parecido al cluster 2.

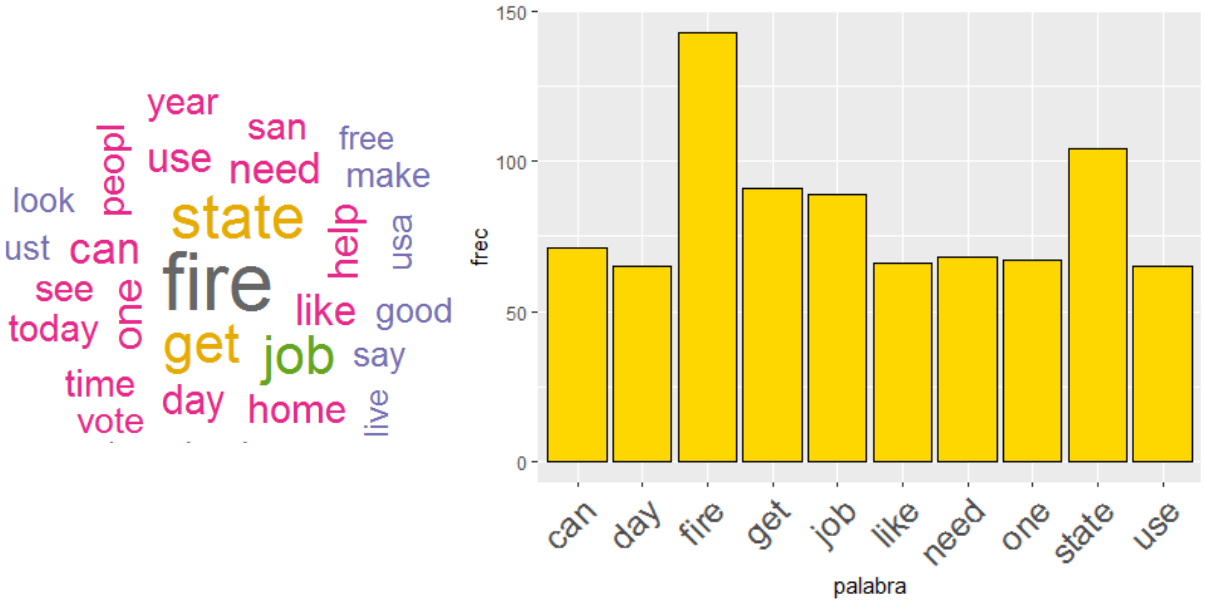

Figura 22. Cluster 3. Algoritmo Kmenas

Algunos ejemplos de tweets que pertenecen a este grupo son:

- Southern #California residents faced with the loss of lives and homes in the #WoolseyFire also are grappling with the destruction of a vast swath of public lands that are popular destinations for hikers, horseback riders and mountain bikers. // *Los residentes del sur de California que se enfrentan a la pérdida de vidas y casas en el #WoolseyFire también están lidiando con la destrucción de una vasta franja de tierras públicas que son destinos populares para excursionistas, jinetes y ciclistas de montaña.*

- What are #California mayors doing to help #EndCAHomelessness? Join us for a live-streamed conversation Friday at noon. // *¿Qué están haciendo los alcaldes de #California para ayudar a ayudar personas sin hogar de CA? Únase a nosotros para una conversación transmitida en vivo el viernes al mediodía.*

- Communist #California has been a constant nuisance to the rest of America and has maliciously interfered with #POTUS[1] #Trump's efforts to #MAGA[2] // *El comunista #California ha sido una molestia constante para el resto de los Estados Unidos y ha interferido maliciosamente con #POTUS # los esfuerzos de Trump para #MAGA*

- I'm convinced bill johnson of #bethel #Church in redding #california is the next david koresh, and its #followers are the somewhere between the manson family and branch davidians.... #lunacy #Heretic #false #Gospel #Jesus help us // *Estoy convencido de que Bill Johnson de la #iglesia #Bethel en Redding California es el próximo David Koresh, y sus # seguidores son el lugar entre la familia Manson y los Davidianos de rama ... #locura #Herética #falso #Evangelio #Jesús ayúdanos*

- JOB: Mountain View CA USA - Sales Assistant Associate - Sales Assistant position with great potential for: Sales Assistant position with great potential for career and financial growth. Significant opportunity to as much JOBS #SOUTHERN #CALIFORNIA // *TRABAJO: Mountain View, CA, EE. UU., Asociado Asistente de ventas: puesto de Asistente de ventas con un gran potencial para: puesto de Asistente*

[1] POTUS: President of the United States (Presidente de los Estados Unidos)
[2] MAGA: Make America Great Again (Hacer América grande de nuevo)

de ventas con un gran potencial de crecimiento profesional y financiero. Oportunidad significativa para tanto TRABAJOS #SOUTHERN #CALIFORNIA

- Job Laguna hills CA USA customer service representative dispatcher coordinator customer service re customer service representative coordinator for our residential hvac division coordinates technicians for call o jobs // *Job Laguna hills CA Representante de atención al cliente de EE. UU. Coordinador de despacho de atención al cliente re coordinador de atención al cliente de nuestra división de hvac residencial coordina a técnicos para llamadas o trabajos*

Como se puede percatar en los ejemplos presentados, existen tweets acerca de ofertas de trabajo en california (relacionados con la palabra job en la figura 20) y tweets con tendencias políticas lo cual no tiene relación si el grupo habla en su gran mayoría de un desastre natural.

7.1.4. Cluster 4. En el cluster 4 con 1373 tweets, tiene un comportamiento muy parecido con los clusters 2 y 3, donde las palabras más frecuentes (ver figura 23) son fire, people, help, pero aparecen palabras como death (muerte), toll (cifra), burn (quemar) un vocabulario que no tenía tanta relevancia en los grupos anteriores, pero que tiene relación directa con el contenido de los tweets de los grupos 2 y 3. Algunos ejemplos de los contenidos de los tweets presentes en este clúster son:

- California fire service veteran flees Paradise as it burns from the Camp Fire. The inferno has torched 130,000 acres, and is responsible for at least 48 deaths. // *El veterano del servicio de bomberos de California huye del Paraíso mientras se quema*

del Camp Fire. El infierno ha incendiado 130,000 acres y es responsable de al menos 48 muertes.

- The number of people unaccounted for in the #CampFire zone of Northern #California more than doubled on Thursday as authorities said the death toll in the deadliest wildfire in California history rose by seven, to 63. // *La cantidad de personas desaparecidas en la zona #CampFire del norte #California aumentó a más del doble el jueves, ya que las autoridades dijeron que la cifra de muertos en el incendio forestal más mortífero en la historia de California aumentó en siete, a 63.*

- North #California fire: death toll now at 71, more than 1,000 missing. It looks this is the most worst #US #wildfire I ever have followed. But, why keeps my stomach sending a feeling it is about one or more arsons where property speculation is the motive? // *Incendio en el norte de California: número de muertos ahora en 71, más de 1,000 desaparecidos. Parece que este es el peor incendio forestal de los Estados*

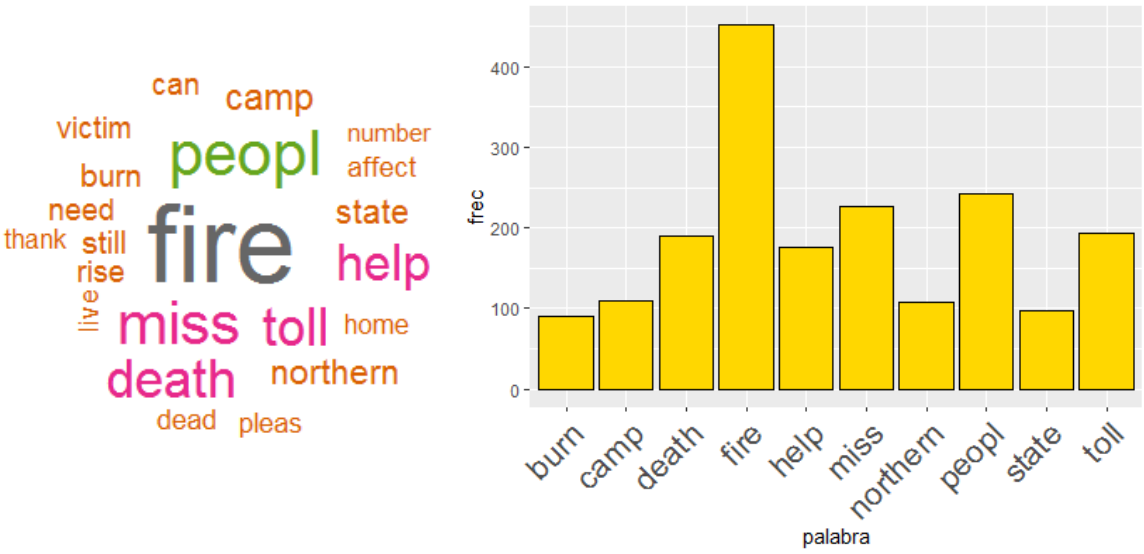

Figura 23. Cluster 4. Algoritmo Kmeans

Unidos que he seguido. Pero, ¿por qué mi estómago me envía una sensación de que se trata de uno o más de los incendios provocados donde la especulación de la propiedad es el motivo?

7.2.Análisis de clústers LDA.

Para el análisis de los clusters de LDA se dividieron en 2 partes, los grupos que agregan valor en la identificación de patrones claves para entender la base de datos, y los grupos triviales, o que no aportan valor para dicho fin.

7.2.1. Clusters Relevantes. Se define como clúster relevante, a los grupos cuyas palabras más frecuentes no son tan comunes, y por contrario son claves para inferir el contenido o tema de los tweets de ese grupo. Los clusters clasificados como relevantes por el grupo investigador son los clusters 1, 2, 4, 5, 7, 11, 16, 17 y 18 relacionados a continuación.

7.2.1.1. El clúster 1 por su parte cuyas palabras más frecuentes (Ver figura 24) son air (aire), quality (calidad), smoke (humo), bad (malo), fire (incendio) indica que los tweets de este

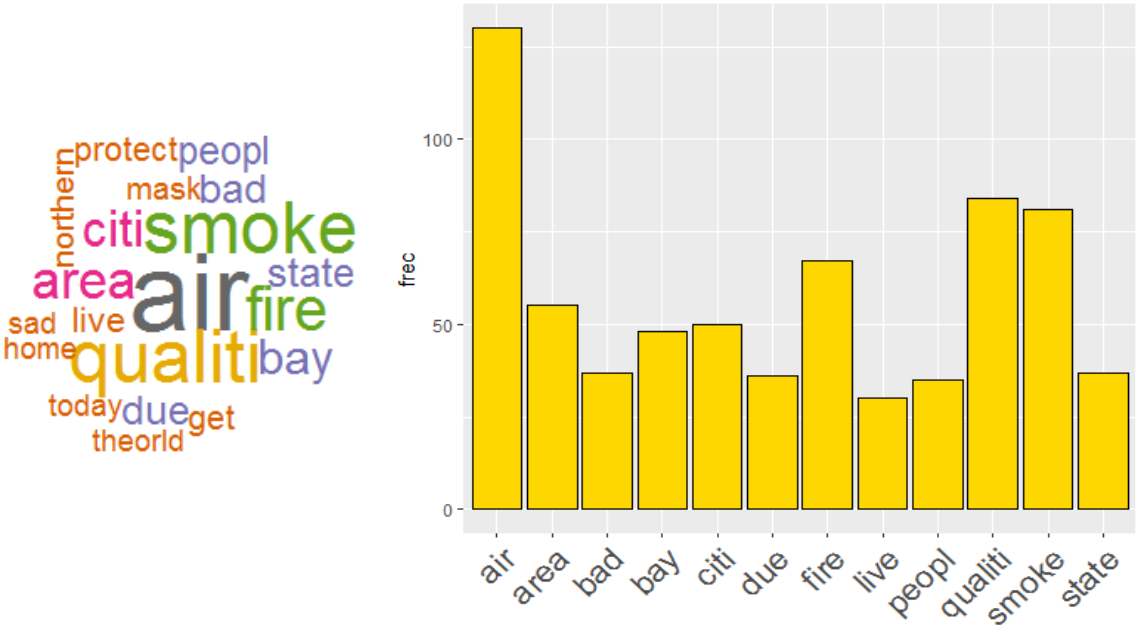

Figura 24. Cluster 1. Algoritmo LDA

grupo en su mayoría, hablan acerca de la mala calidad del aire debido a incendio en ese estado. Y es que según (Turkewitz & Richtel, 2018) Los incendios forestales que han asolado gran parte de California están presentando a los residentes un nuevo peligro: el aire tan espeso por el humo se encuentra entre los más sucios del mundo, lo cual se evidencia en los ejemplos de los tweets pertenecientes a este grupo:

- HAZARDOUS AIR WARNING: Authorities say there are 600 people now unaccounted for in California as the smoke from the devastating wildfires is closing schools and reaching hazardous levels. // ADVERTENCIA DE AIRE PELIGROSO: Las autoridades dicen que hay 600 personas desaparecidas en California, ya que el humo de los devastadores incendios forestales está cerrando escuelas y llegando a niveles peligrosos.

- My dust mask turned brown after being outside for a few hours today. No, this isn't makeup. Makeup would be on the INSIDE of the mask too. Keep that in mind if you're in NorCal like me. // *Mi tapabocas se volvió marrón después de estar afuera por unas horas hoy. No, esto no es maquillaje. El maquillaje también estaría en el INTERIOR del tapabocas. Tenlo en cuenta si estás en el norte de California como yo.*

- Just looked up the air quality index of Northern #California and compared it to cities around the world. Right now we are worse than countries like China and Philiippines // *Acabo de buscar el índice de calidad del aire del norte de California y comparar esto con ciudades alrededor del mundo. Ahora mismo, nosotros estamos peor que ciudades como China y Pilipinas.*

7.2.1.2. Por otro lado en el cluster 2 se puede inferir que el contenido de los tweets es acerca agradecimientos, pensamientos y/u oraciones por los bomberos y/o afectados con el

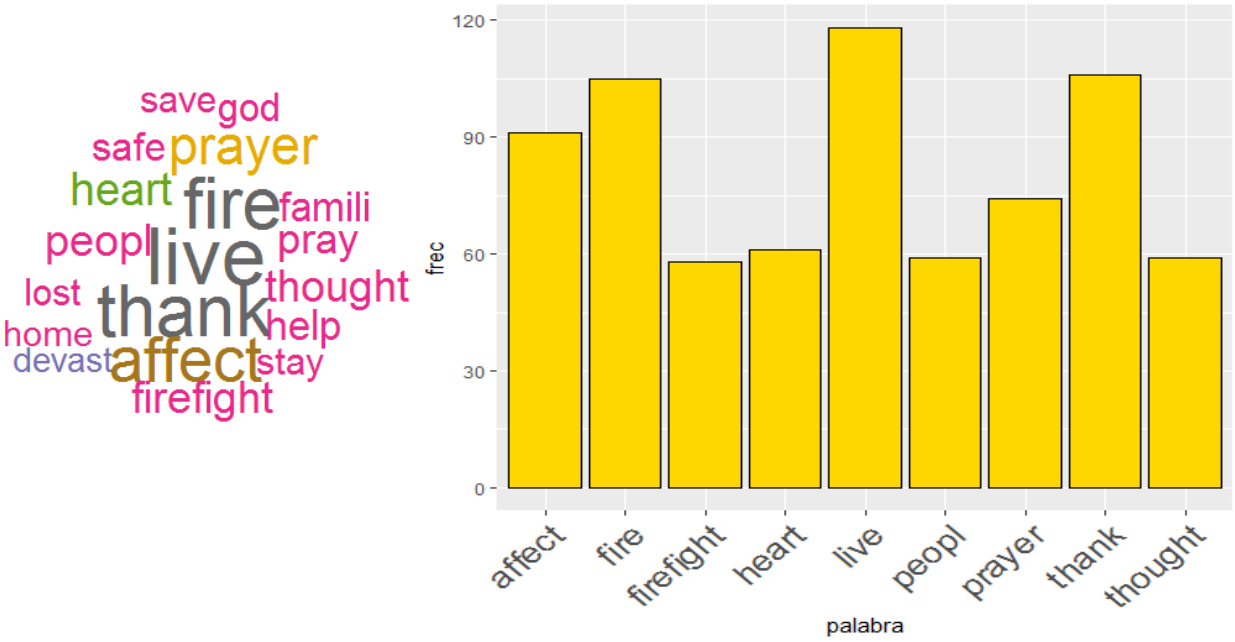

Figura 25. Cluster2. Algoritmo LDA

incendio. (ver figura 25).

Cabe resaltar, como se observa en los tweets ejemplo a continuación, no todos los tweets presentes en este grupo son alentadores, por el contrario existen algunos con contenido de rabia y frustración.

- At a Diwali candle lighting event at the #whitehouse, #presdonaldtrump began by recogizing lives lost in the California #wildfires, thanking firefighters and pledging the full suppot of the federal government. // *En el evento de iluminación de velas Diwali en la #casablanca, #presdonaldtrump comenzó por reconocer las vidas perdidas en los #incendiosforestales de California, agradeciendo a los bomberos y prometiendo el apoyo total del gobierno federal.*

- I can assure you that the California wildfires ARE NOT a blessing from God. Therefore, they have to be a punishment for the sinful lives of the many. // *Les puedo asegurar que los incendios forestales de California NO SON una bendición de Dios. Por lo tanto, tienen que ser un castigo para las vidas pecaminosas de muchos.*

- Thoughts and Prayers continue to be with the people of California // *Pensamientos y oraciones estan con las personas de California*

- We are happy to report #Charlie, rescued by a #California #Firefighter has been reunited with his family. Thank You to all the first responders and firefighters who risk their lives to rescue those in need including the animals affected by the fires. // *Nos complace informar que #Charlie, rescatado por un #Bombero de #California se ha reunido con su familia. Gracias a todos los socorristas y bomberos que arriesgan sus vidas para rescatar a los necesitados, incluidos los animales afectados por los incendios.*

7.2.1.3. En cuanto al cluster 4, se puede observar claramente en la figura 26 que el contenido de este grupo es mayormente político, deduciendo que hay discusiones acerca de votos para demócratas y republicanos en el estado de California, lo cual se puede corroborar dado que las elecciones legislativas de los estados unidos que se presentaron el pasado 6 de noviembre y donde los demócratas recuperaron el control de la cámara de representantes y los republicanos mantienen por su parte el senado (The Guardian, 2018)

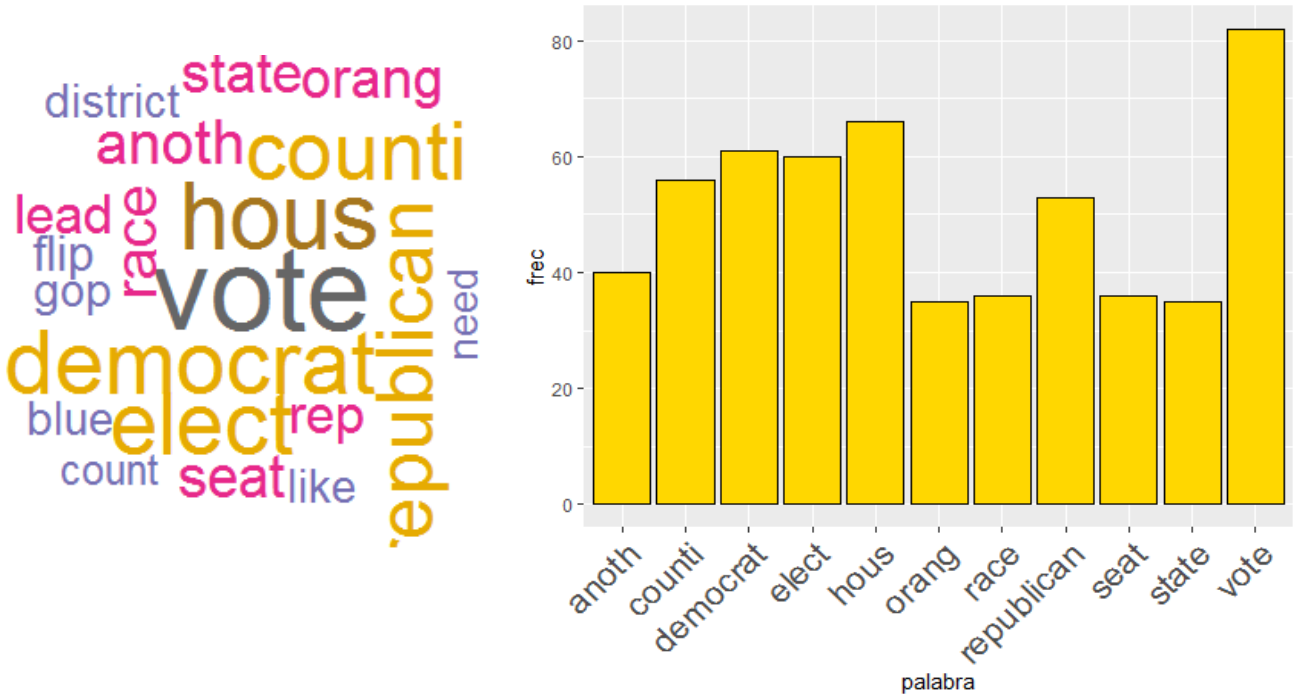

Figura 26. Cluster 4. Algoritmo LDA

Algunos ejemplos de los tweets en este grupo son:

- #California #Politics #Government #Elections --> Republicans Lose More Ground in SoCal House Races As Counting Continues // *#California #Politica #Govierno #Elecciones; Los republicanos pierden más terreno en la casa del sur de California y la cuenta continua*

- No, slow vote counts in #California don't mean fraud. To suggest otherwise is irresponsible // *No, el conteo lento de votos en #California no significa fraude. Sugerir lo contrario es irresponsable.*

- #California elects trash like this? Lol. And they diminish those who voted for @POTUS // *#California elige basura como esta? Jajaja Y disminuyen a los que votaron por @POTUS.*

- Proof democrats stole the house and are trying to steal the senate and have already stolen some senate seats // Prueba de Demócratas robaron la casa y están tratando de robar el Senado y ya han robado algunos asientos del Senado

7.2.1.4. Al ver la figura 27 se puede deducir que el cluster 5 contiene tweets que hablan de las playas de San Francisco y San Diego, dado que la cantidad de repeticiones de la palabra

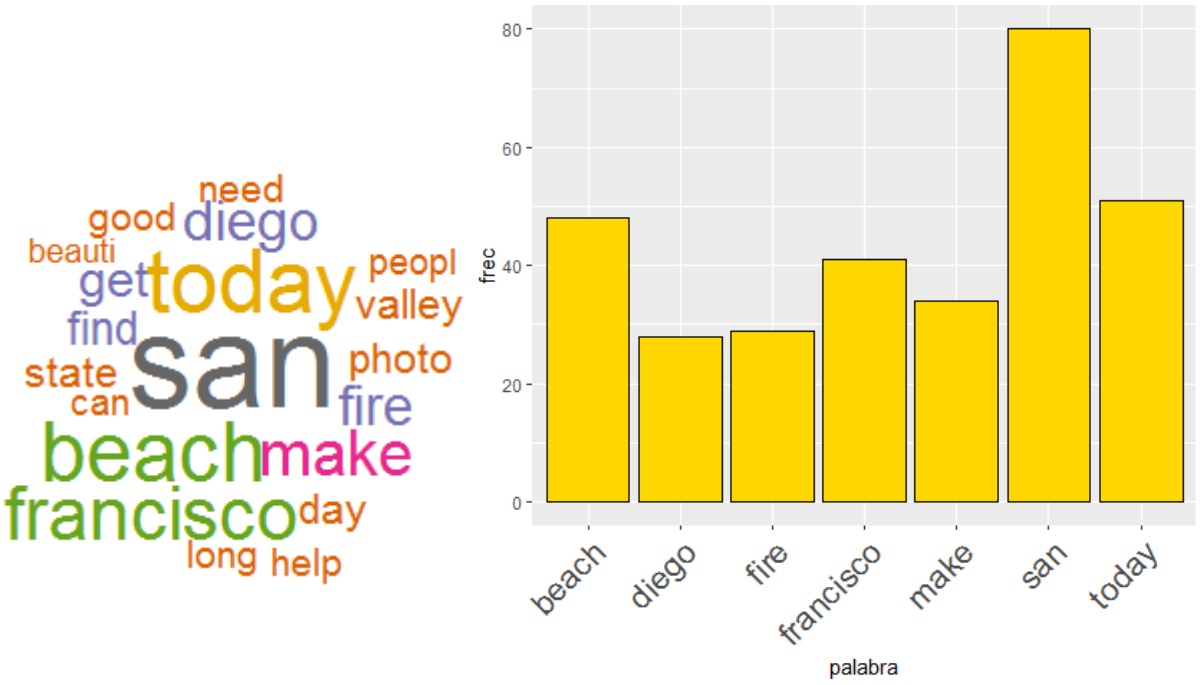

Figura 27. Cluster 5. Algoritmo LDA

fire (incendio) se repite muy poco (cerca de 30 veces) se puede inferir que en este grupo en su gran mayoría los tweets son acerca del día a día en estos lugares y en California en general. Lo cual se puede ratificar con los ejemplos presentados a continuación:

- I'm so thrilled I got to meet #lianemoriarity tonight! Her #booktour for #nineperfectstrangers made only 1 stop in #california: not San Francisco, not Los Angeles - it was in San Diego! Thanks to @warwicksbooks, the #universityofsandiego for making this wonderful evening happen! // *¡Estoy tan emocionada que pude conocer a #lianemoriarity esta noche! Su #booktour para #nineperfectstrangers hizo solo una parada en #california: no en San Francisco, ni en Los Ángeles, ¡fue en San Diego! ¡Gracias a @warwicksbooks, a la #universidaddesandiego por hacer que esta maravillosa noche suceda!*

- @DaveEdlund in the engine room of the Queen Mary, Long Beach, CA. She transported 750,000 Americans to the European Theater // *@DaveEdlund en la sala de máquinas del Queen Mary, Long Beach, CA. Ella transportó a 750,000 estadounidenses al teatro europeo.*

- I got tired eyes san francisco california // *Tengo los ojos cansados San Francisco California*

- Newport Beach CA - Visit Balboa Fun Zone #travel #California // *Newport Beach CA – Visita Balboa Fun Zone #viaje #California.*

7.2.1.5. Detallando la figura 28 se puede inferir que el cluster 7 es acerca de los incendios en California, pero en gran parte son tweets acerca de ayuda (help) a las víctimas, algunos pueden ser manifestando la necesidad de esta, otros agradeciéndola u apoyándola. Otro dato importante es que la ayuda no es solo para personas (people) también se puede observar la palabra anim (el Stemming de la palabra animal).

Algunos ejemplos de tweets que aparecen en el cluster 7 se muestran a continuación, vale la

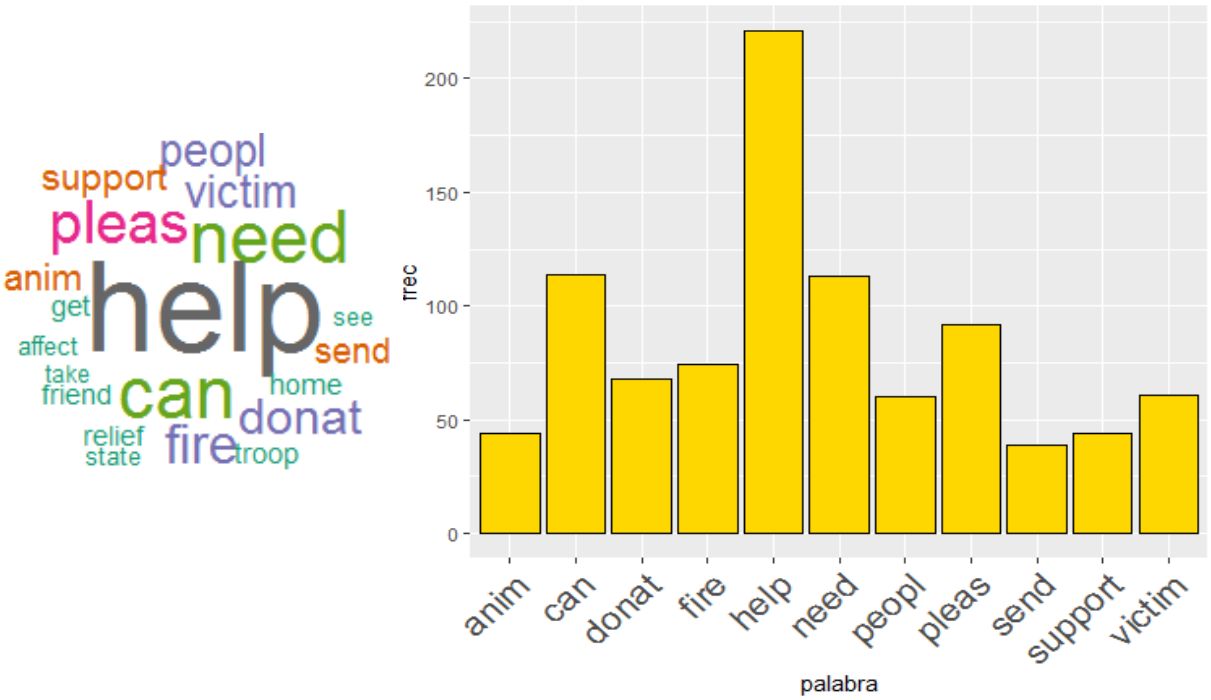

Figura 28. Cluster 7. Algoritmo LDA

pena resaltar, la presencia de tweets donde comparten información de cómo ayudar a las víctimas de esta tragedia.

- Here's how to help the victims as wildfires rage across #California. Via @TIME https://t.co/JVYMB2lAn7/3 // *Aquí se explica cómo ayudar a las víctimas, ya que los incendios forestales se propagan a través de #California. A través de @TIME https://t.co/JVYMB2lAn7/3*

- As the #fires in #California continue to rage we will be sharing links to #nonprofits and #funds where you can donate to help the animals and people in need. Starting with the IAFF[1] disaster relief for firefighters. We thank you for your continued service: https://t.co/D6JuwC7gIt https://t.co/QEuWzd4vZF/3 // *A medida que los # incendios en #California continúan avanzando, estaremos compartiendo enlaces de #sinanimodelucro y #cuentas donde puede donar para ayudar a los animales y las personas necesitadas. Comenzando con el alivio de desastres IAFF para los bomberos. Le agradecemos su servicio continuo: https://t.co/D6JuwC7gIt https://t.co/QEuWzd4vZF/3*

- The right way to donate to the victims of the California wildfires https://t.co/JVYMB2lAn7/3 // *La forma correcta de donar a las víctimas de los incendios forestales de California https://t.co/JVYMB2lAn7/3*

- How to help victims of the California wildfires: https://t.co/mv3VlhlcaM // *Cómo ayudar a las víctimas de los incendios forestales de California: https://t.co/mv3VlhlcaM*

[1] IAFF: International Association of Fire Fighters (Asociación Internacional de Bomberos)

7.2.1.6. En el cluster 11, las palabras más frecuentas que se observan en la figura 29 muestran una tendencia a hablar acerca de ofertas de trabajo en California, en cuanto a servicio al cliente, energía, ventas, entre otros.

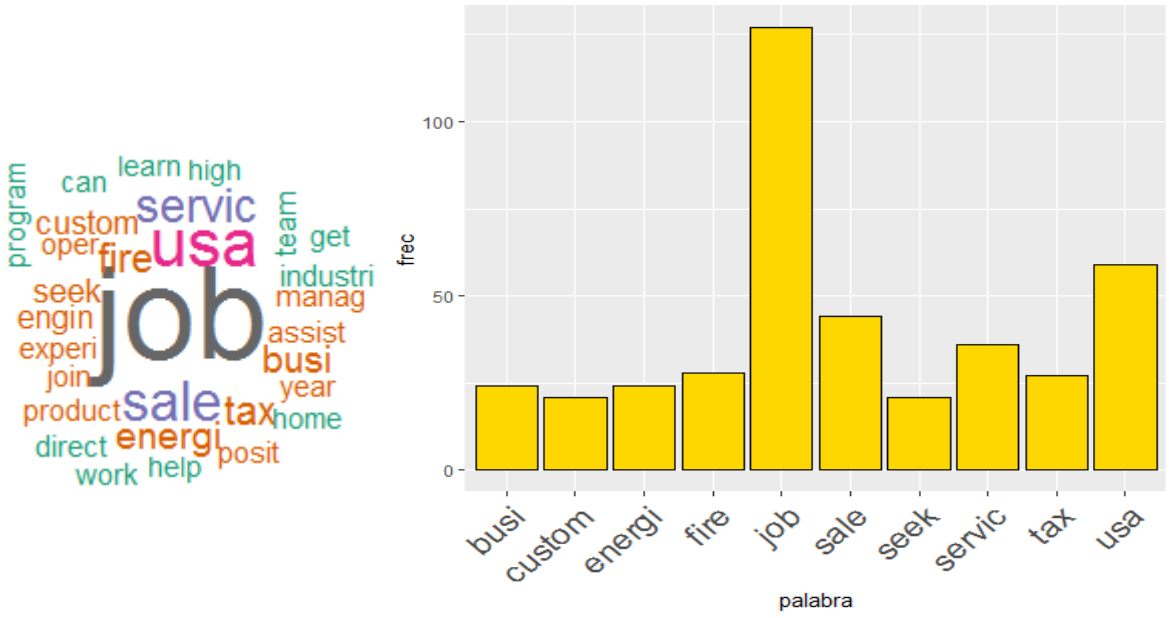

Figura 29. Cluster 11. Algoritmo LDA

Al momento de revisar ejemplos de los tweets pertenecientes a este cluster podemos observar que también la palabra trabajo (job) es usada para manifestar la necesidad de recontruir algunas zonas afectadas, como se puede apreciar en el último ejemplo.

Ejemplos:

- JOB: Anaheim CA USA - Full Time - Electronics Sales -- Paid Training - E.V Solutions is a private sales a: E.V Solutions is a private sales and marketing firm that has immediate hire opportunities for entry level individua.. JOBS #SOUTHERN #CALIFORNIA // *TRABAJO: Anaheim, CA, EE. UU. - Tiempo completo - Ventas de productos electrónicos - Capacitación pagada - E.V Solutions es una empresa privada de ventas: E.V Solutions es una empresa privada de ventas y mercadotecnia que tiene oportunidades de*

contratación inmediata para personal de nivel de entrada .. TRABAJOS #SOUTHERN #CALIFORNIA

- JOB: Mountain View CA USA - Sales Assistant Associate - Sales Assistant position with great potential for: Sales Assistant position with great potential for career and financial growth. Significant opportunity to as much a… JOBS #SOUTHERN #CALIFORNIA // *TRABAJO: Mountain View, CA, EE. UU., Asistente asociado de ventas: puesto de Asistente de ventas con un gran potencial para: puesto de Asistente de ventas con un gran potencial de crecimiento profesional y financiero. Oportunidad significativa tanto como ... TRABAJOS #SUTHERN #CALIFORNIA*

- @EsmeAlaki @SenSanders @RepRoKhanna Sales tax needs to increase to 25% for most goods; same as Norway and Denmark. Only food and essentials have lower tax. And add the service and labor tax that Newsom wants. Increasing taxes will take care of everyone in // *@EsmeAlaki @SenSanders @RepRoKhanna El impuesto a las ventas debe aumentar hasta el 25% para la mayoría de los bienes; Igual que Noruega y Dinamarca. Sólo los alimentos y los productos básicos tienen impuestos más bajos. Y añada el servicio y el impuesto laboral que quiere Newsom. El aumento de los impuestos se hará cargo de todos en*

- The director of the US emergency agency says a California town ravaged by wildfire will need a \"total rebuild\" job that will take several years.\" https://t.co/5QU10G140P #California #wildfire #Paradise #destruction" // *El director de la agencia de emergencia de EE. UU. Dice que una ciudad de California devastada por un incendio forestal*

necesitará un trabajo de "reconstrucción total" que durará varios años. Https://t.co/5QU10G140P #California #wildfire #Paradise #destruction "

7.2.1.7. El cluster 16 muestra un tema bastante interesante, debido a que observando la figura 30, se puede inferir que los tweets más relevantes de este grupo son acerca del problema migratorio que tiene estados unidos en la frontera (border) con México

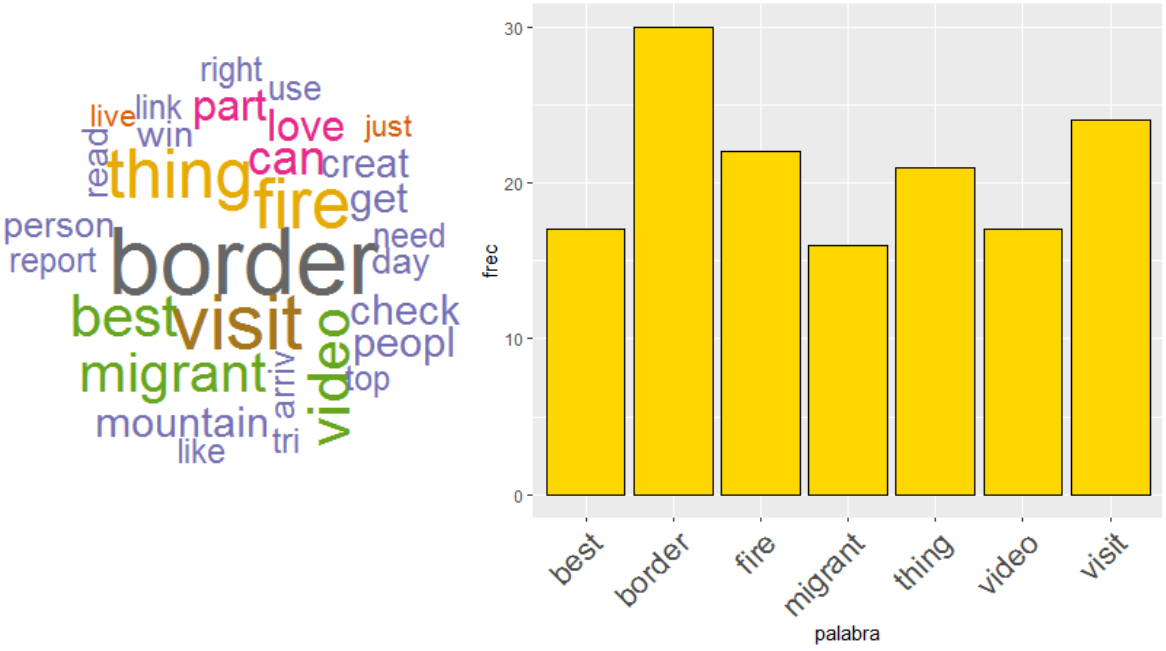

Figura 30. Cluster 16. Algoritmo LDA

A continuación se presentan varios ejemplos de los tweets de este clúster, donde se puede observar que varios de estos tweets son solicitudes al gobierno de enviar la ayuda pertinente a california en vez de concentrar sus esfuerzos en la lucha migratoria con México.

- Greek police beat migrants at border, send them back to Turkey naked, locals say https://t.co/bMSa5yzRB9 #news #business #california https://t.co/urvmbkwIZA // *La policía griega golpea a los migrantes en la frontera, los envía de vuelta a Turquía*

desnudos, los lugareños dicen https://t.co/bMSa5yzRB9 #news #business #california https://t.co/urvmbkwIZA

- @realDonaldTrump @POTUS is 100% correct about @jerrybowngov's neglect of the forests in #California contributing to the devastation of the #CaliforniaFires. And 90% of migrants in the caravans; illegals here, now, are not h… // *@realDonaldTrump @POTUS es 100% correcto acerca de la negligencia de @jerrybowngov de los bosques en #California que contribuye a la devastación de los #CaliforniaFires. Y el 90% de los migrantes en las caravanas; Los ilegales aquí, ahora, no son h ..*

- Dear @realDonaldTrump Please send the troops on the border to #CALIFORNIA to help with the victims of #wildfires! They're badly needed THERE-MUCH more than sitting around waiting for a \CARAVAN.\" @SecretaryZinke @Interior" // *Estimado @realDonaldTrump ¡Envíe las tropas en la frontera a #CALIFORNIA para ayudar con las víctimas de los #incendiosforestales! Son muy necesarios MUCHO más que quedarse esperando a un \ CARAVAN. \ "@SecretaryZinke @Interior"*

7.2.1.8. En la figura 31, se puede apreciar que las palabras que sobresalen y que ayudan a identificar patrones en el clúster 17 son change (cambio), climate (clima), lo que nos muestra que un claro indicio de que este grupo de tweets en cierta parte trata sobre como el cambio climático tiene relación con el desastre. Algunos ejemplos de los tweets

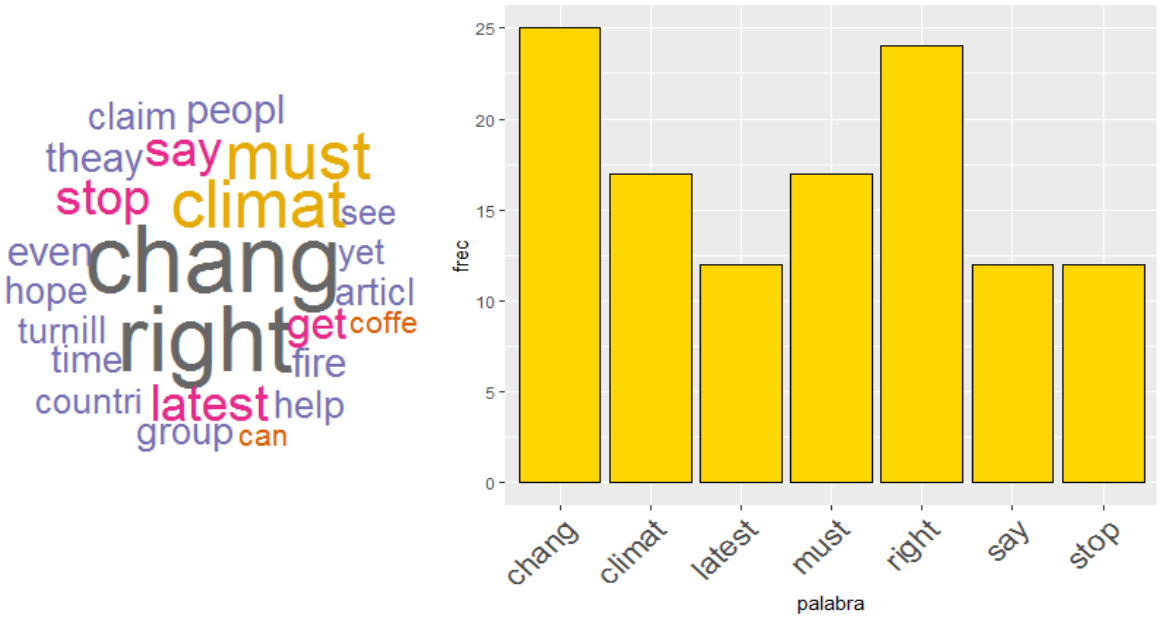

Figura 31. Cluster 17 Algoritmo LDA

encontrados en este clúster son:

- Some corners of the right-wing internet are already cooking up a grand conspiracy theory to explain the fires in #California: airborne laser attacks meant to clear the way for high-speed rail. https://t.co/vqAIA7XB9C // *Algunos rincones del internet de la derecha ya están preparando una gran teoría de la conspiración para explicar los incendios en #California: ataques con láser en el aire, destinados a despejar el camino para el ferrocarril de alta velocidad. https://t.co/vqAIA7XB9C*

- @realDonaldTrump @JerryBrownGov will meet tomorrow in a show of solidarity for suffering Californians from the #WoolseyFire and other hardships plaguing Californians Governor Brown Don't Go into a long-winded speech about climate change // *@realDonaldTrump @JerryBrownGov se reunirán mañana en una muestra de solidaridad por los californianos que sufren desde el #WoolseyFire y otras dificultades que aquejan al gobernador de California. Gobernador Brown no participe en un discurso prolongado sobre el cambio climático*

- What's Behind The California Wildfires? Is It More Than Climate Change? Are Corporations Involved? https://t.co/brmtYv4N0k // *¿Qué hay detrás de los incendios forestales de California? ¿Es más que el cambio climático? ¿Están involucradas las corporaciones? https://t.co/brmtYv4N0k*

7.2.1.9. Finalmente, en cuanto al clúster 18, como se evidencia en la figura 32, se puede analizar que es un grupo basado en cifras de personas muertas y desaparecidas por los incendios ocurridos en el norte de california.

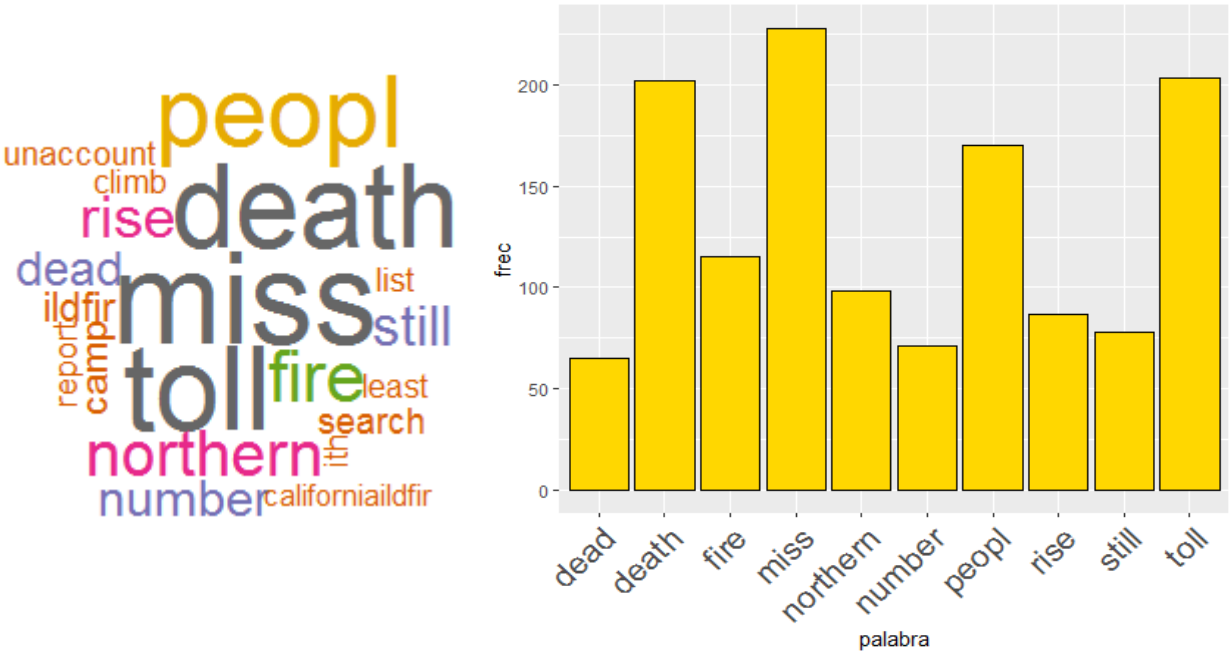

Figura 32. Cluster 18. Algoritmo LDA

Ejemplos de tweets presentados en este cluster:

- #Trump headed to #California to see firsthand the grief and devastation from the deadliest US wildfire in a century. Authorities confirmed a new death toll of 71 and said they were trying to locate 1,011 people even as they stressed that not all are believed missing. https://t.co/Bb9hdjChFl // *#Trump se dirigió a #California para ver de primera mano el dolor y la devastación de los incendios forestales más mortíferos de los Estados Unidos en un siglo. Las autoridades confirmaron un nuevo número de muertos de 71 y dijeron que estaban tratando de localizar a 1.011 personas, incluso cuando destacaron que no todos se creen desaparecidos. https://t.co/Bb9hdjChFl*

- Hundreds of people are missing in #California's #CampFire. With each day that goes by, their families refuse to let hope wane. https://t.co/Z0GF8TrEWM // *Cientos de personas están desaparecidas en #CampFire de #California. Con cada día que pasa, sus familias se niegan a dejar que la esperanza se desvanezca. https://t.co/Z0GF8TrEWM*

- Camp fire death toll rises to 71 with more than 1,000 missing https://t.co/OsvuElpXaZ // *El número de muertos en Camp fire aumenta a 71 y faltan más de 1.000 https://t.co/OsvuElpXaZ*

- Crews search for #California fire victims as list of missing passes 600 #fires https://t.co/GBv5mgyILC https://t.co/TkbpkL7XUm // *Los equipos de búsqueda de víctimas de incendios de #California como la lista de pases perdidos 600 #fires https://t.co/GBv5mgyILC https://t.co/TkbpkL7XUm*

7.2.2. Clusters no relevantes. Se define como clusters no relevantes, a los grupos cuyas palabras más frecuentes no permiten inferir acerca del contenido de los tweets, por lo tanto dichos grupos demuestran una falta de identidad.

Los clusters del algoritmo LDA que no agregan valor son los clúster 3, 6, 8, 9, 10, 12, 13, 14 y 15. Dichos grupos se pueden observar en las figuras 33, 34, 35, 36, 37, 38, 39, 40 y 41 donde se encuentran sus respectivas nueves de palabras e histogramas de frecuencia

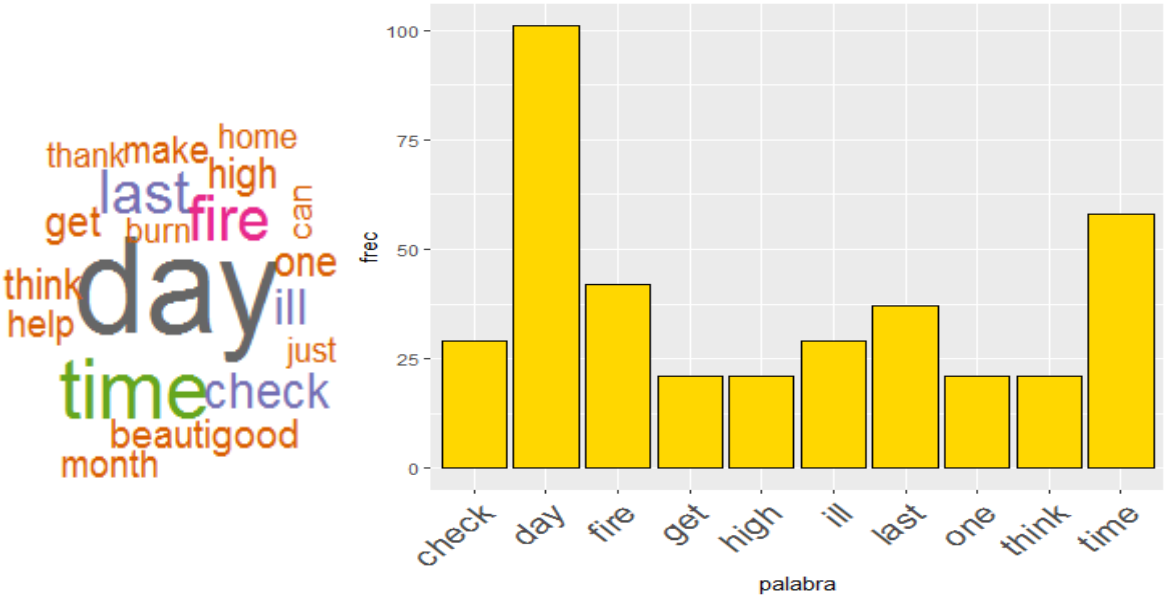

Figura 33. Cluster 3. Algoritmo LDA

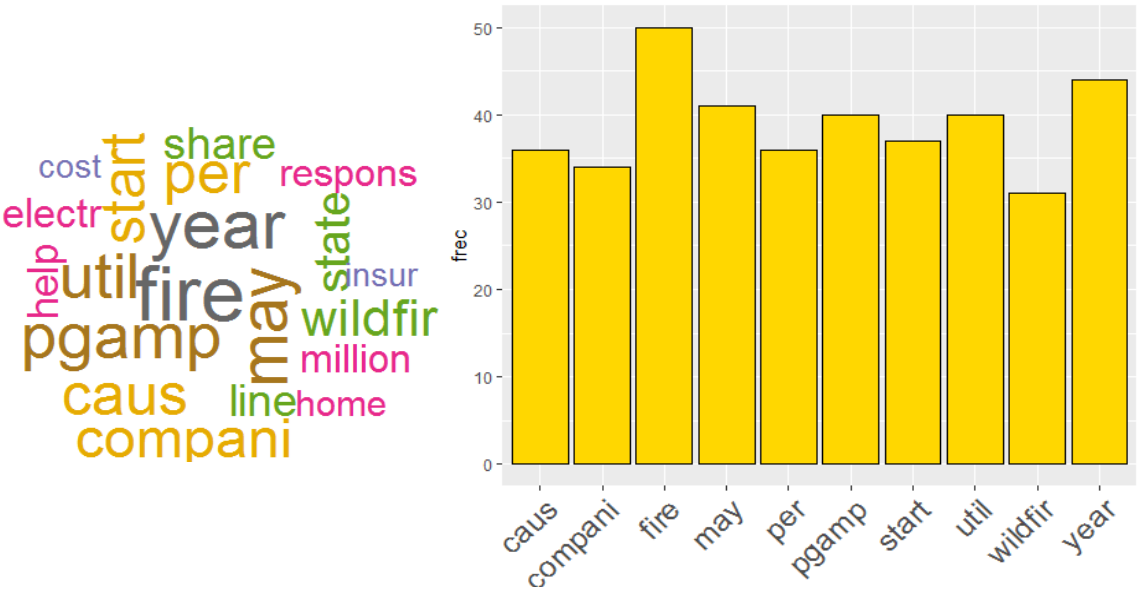

Figura 34. Cluster 6. Algoritmo LDA

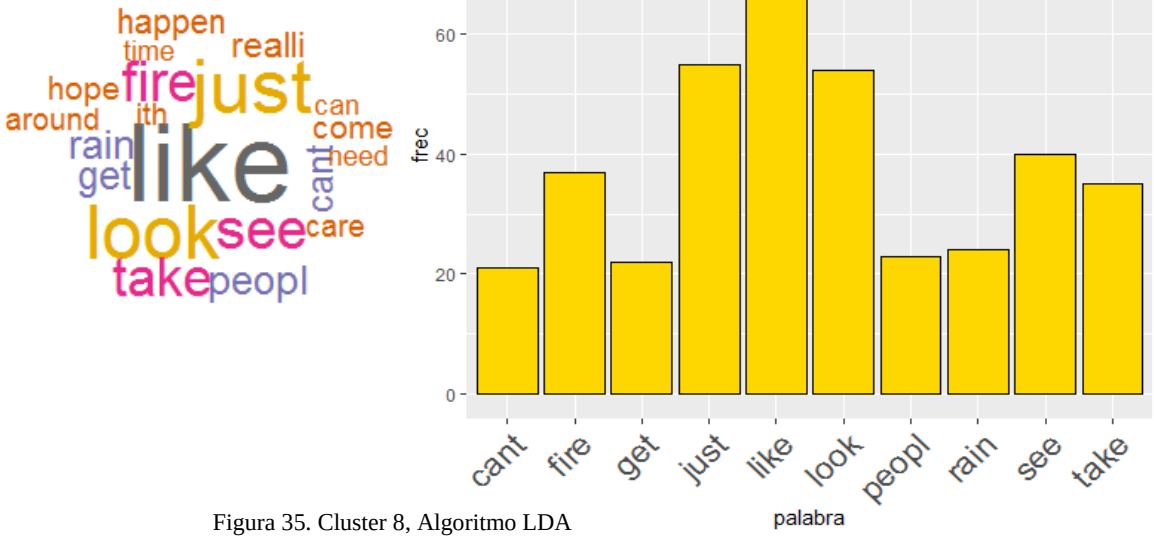

Figura 35. Cluster 8, Algoritmo LDA

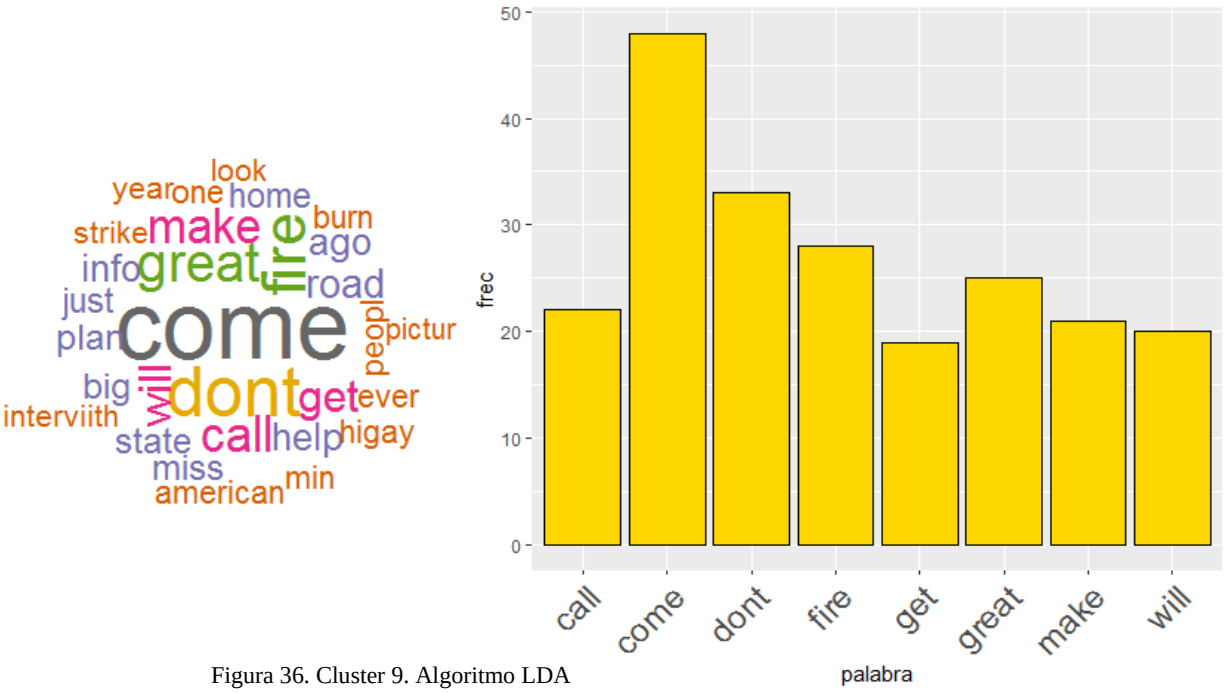

Figura 36. Cluster 9. Algoritmo LDA

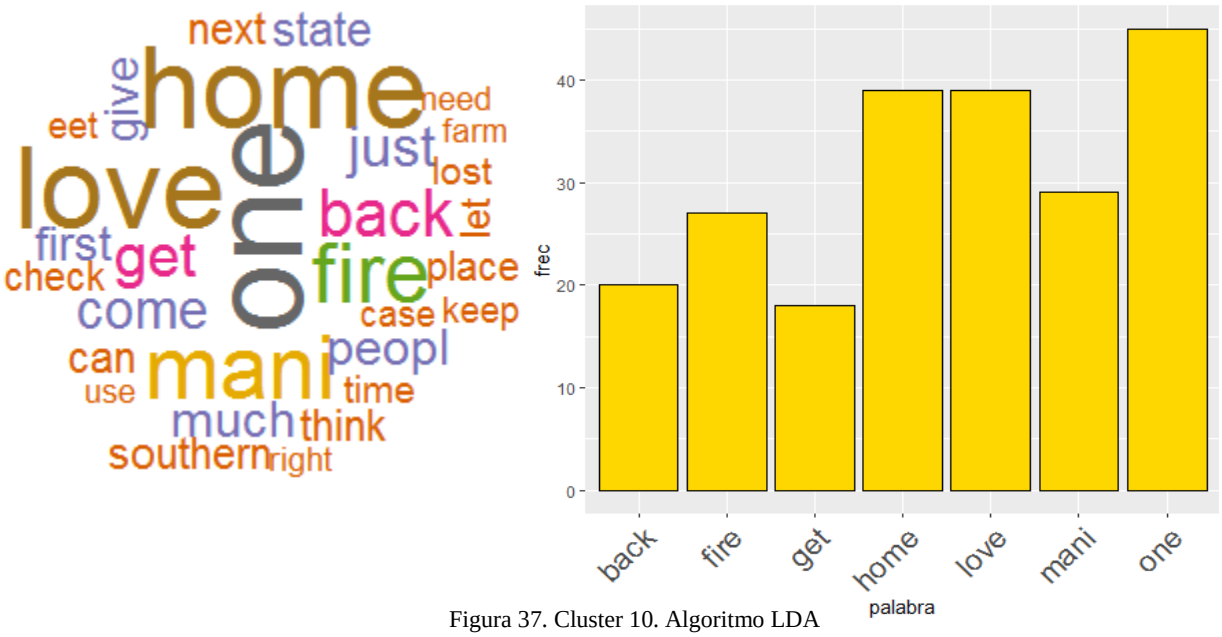

Figura 37. Cluster 10. Algoritmo LDA

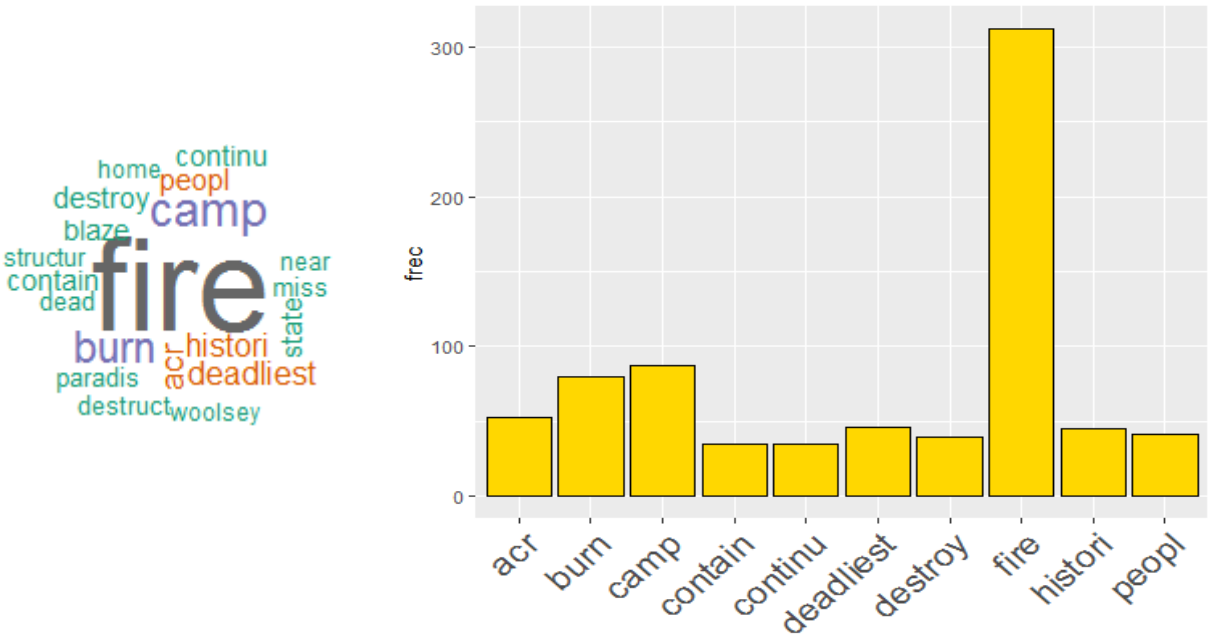

Figura 39. Cluster 12. Algoritmo LDA

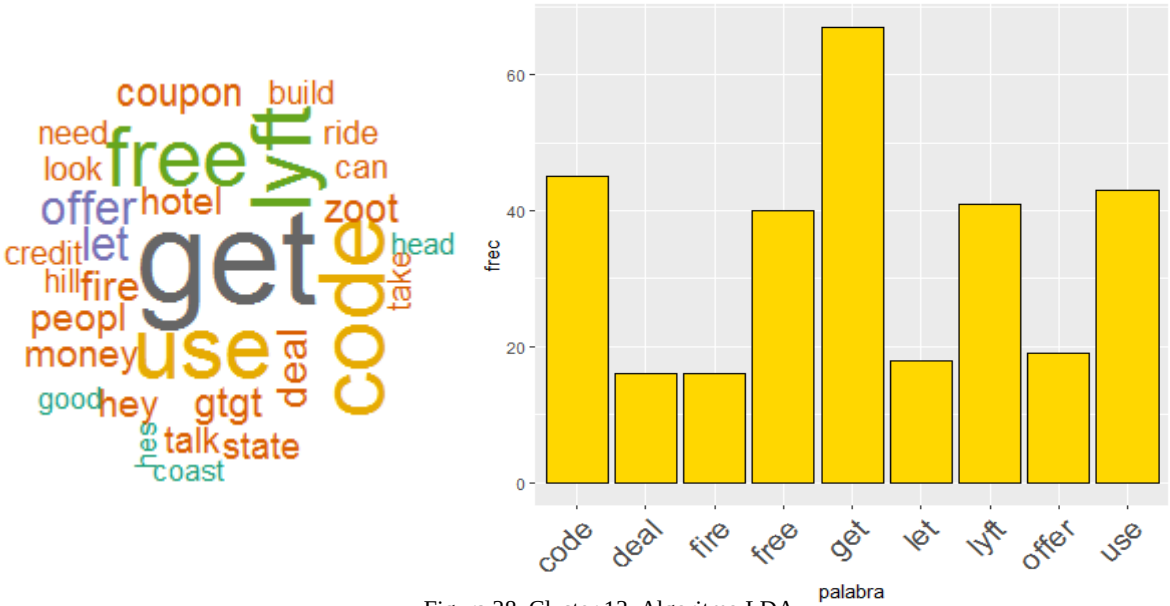

Figura 38. Cluster 13. Algoritmo LDA

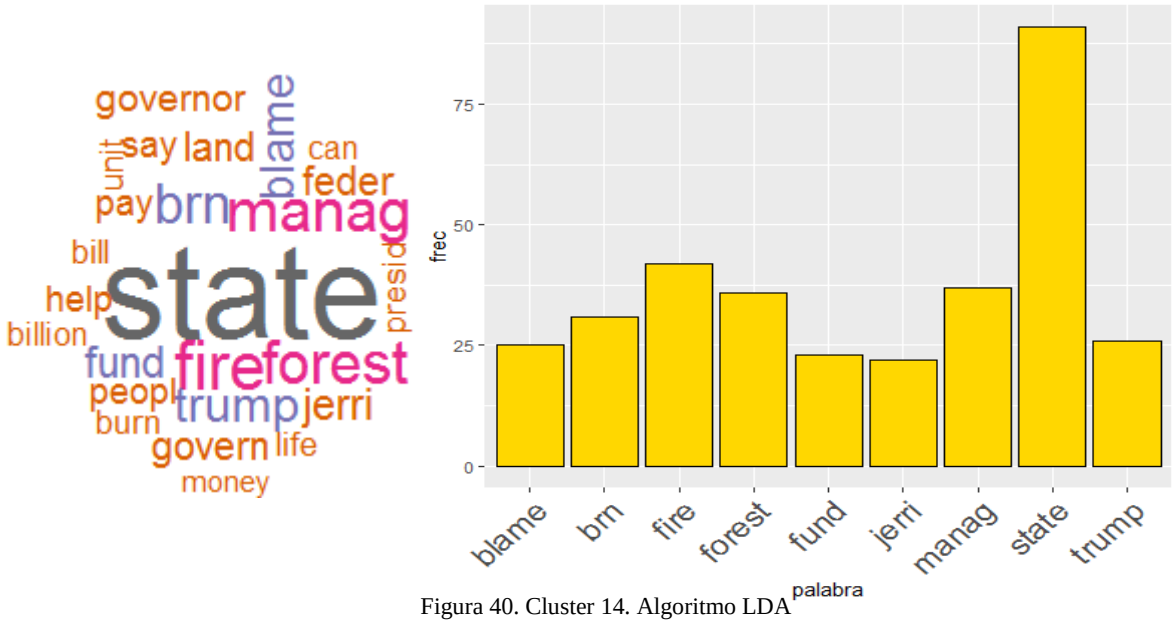

Figura 40. Cluster 14. Algoritmo LDA

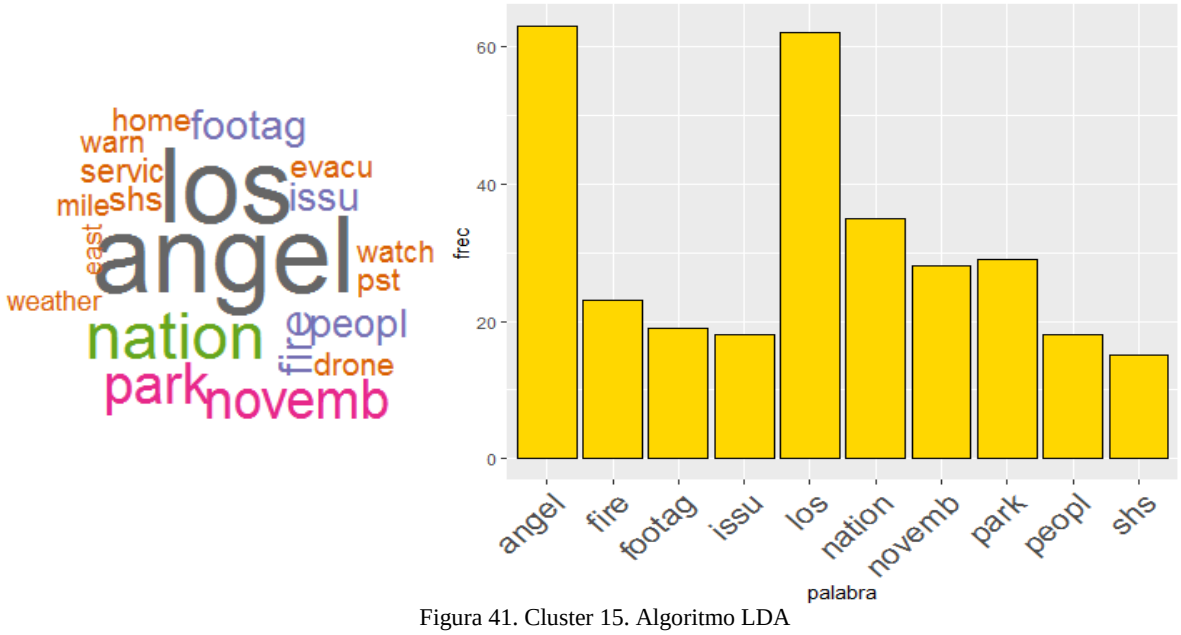

Figura 41. Cluster 15. Algoritmo LDA

Los códigos realizados bajo el lenguaje de programación R para la ejecución del proyecto se encuentran disponibles en el siguiente link:

https://github.com/nicolasjimenezp/Base-de-datos-California

7.3. Cuadro Comparativo

En la tabla 7 se encuentra una comparación conceptual bajo ciertos criterios que para el grupo investigador definen las importancias cualitativas en un algoritmo.

Para la ejecución de los algoritmos se usó un computador con memoria RAM de 6 GB, con procesador Intel core i5-3210M (de dos núcleos) con una velocidad por núcleo de 2,5 GHz. Bajo estas condiciones, el algoritmo que menos tiempo computacional tardó fue el algoritmo LDA, con aproximadamente 2 horas y 30 minutos, 66,66% menos tiempo (5 horas menos) que lo que tardó por su parte el algoritmo k-means.

El número de clusters en que un algoritmo clasifica la base de datos es un punto de partida que permite al investigador identificar el número de temas que están implícitos en dichos datos, y es al momento del análisis donde se puede corroborar cuales de estos grupos verdaderamente ayudan al investigador a entender y obtener patrones de dicha información, respecto a este punto, k-means clasificó la base de datos en 4 grupos, de los cuales solo se pudo identificar que los tweets hablaban de 3 temas en particular: la vida en California, la necesidad de ayuda y por último información sobre cifra de muertos y desaparecidos. Por lo tanto, de los 4 grupos que

inicialmente el algoritmo desarrolló solo 3 de ellos fueron importantes para el entendimiento de la base de datos.

Tabla 7.
Comparación entre modelos de aprendizaje no supervisado

CRITERIO	ALGORITMO K-MEANS	ALGORITMO LDA
Tiempo computacional	Aprox. 7,5 horas	Aprox. 2,5 horas
Numero de temas en los que se dividió	4	18
Temas que se pudieron identificar.	3	9
Coherencia dentro de los clústers	Media	Media

Por su parte el algoritmo LDA clasificó en 18 grupos, de estos fue posible identificar que en la base de datos aparecían de 9 temas, acerca de la calidad del aire, agradecimientos, política, vida en California, necesidad de ayuda y como ayudar, trabajos, migrantes, cambio climático y por último cifra de muertos y desaparecidos. Por lo tanto, de los 18 grupos que inicialmente el

algoritmo desarrolló inicialmente solo 9 fueron importantes para el entendimiento de la base de datos.

La coherencia dentro de los grupos se define como que tan congruentes están discriminados los tweets dentro del grupo, donde ambos algoritmos muestran algunas debilidades, dado que en ambos algoritmos se puede observar que dentro de los grupos existen tweets que no se relacionan con el tema de que dicen tratar. Esto se puede atribuir al modo de escritura que es informal (debido a la limitación en el número de caracteres a usar en Twitter). Además, en el caso del algoritmo LDA, dado que es un algoritmo de tipo probabilístico, esta clasificación de los documentos (tweets) es dada por la mayor probabilidad de ocurrencia en dicho tópico, por lo que en ocasiones aparecen tweets que no están directamente relacionados con el tema de dicho tópico.

8. Conclusiones

El objetivo de este proyecto de investigación se enfocaba en la evaluación de dos modelos de aprendizaje no supervisado tomando como fuente primaria tweets que fueron generados ante los incendios forestales presentados en California a principios de Noviembre 2018. Para lo cual se implementaron: En primer lugar el algoritmo K-means el cual divide la base de datos en *k* grupos de *n* objetos midiendo la similitud de proximidad con respecto al valor medio de los objetos en un grupo (centroide o centro de gravedad del grupo). Y en segundo lugar el algoritmo LDA (Latent Dirichlet Allocation) el cual trata cada documento (tweet) como una mezcla de temas y cada tema como una mezcla de palabras, y donde el orden de las palabras en los documentos no importa, dado que el uso de una palabra es ser parte de un tema y que comunica la misma información sin importar dónde se encuentra en el documento; se trata de un modelo probabilístico, es decir, cada documento puede verse como una mezcla de varias categorías y asume que la distribución de clusters tiene una distribución a priori de Dirichlet

La revisión de la literatura permitió la apertura a conceptos básicos indispensables para el entendimiento del tema de investigación, así como también para la identificación de técnicas y tecnologías actuales en el campo del Big data.

En cuanto a la red social Twitter, se puede destacar la presencia masiva de ruido en las publicaciones, esto puede deberse también a las restricciones en cuanto al número de caracteres disponibles en una publicación (280) lo que genera un uso incorrecto del idioma y por consecuente una dificultad mayor a la hora del análisis en los resultados.

En lo que respecta a la fase experimental, R al ser un software de uso gratuito permite elaborar estos estudios desde cualquier lugar, además, el lenguaje de programación R es muy amigable con el usuario, dado que presenta opciones de ayuda y librerías que permiten una programación mucho más ágil.

Para el análisis, los agrupamientos de los tweets, permitieron identificar y extraer de una manera práctica información, patrones y tendencias de sobre lo que estaba ocurriendo en esa ventana de tiempo en California. Para el caso del Algoritmo K-means dividió la base de datos en 4 grupos, los cuales sirvieron para identificar y/o explicar tan solo 3 temas de la base de datos, lo cual no es de gran utilidad, si de buscar patrones que sirvan para entender un conjunto masivo de datos se refiere. Por su parte el algoritmo LDA, fue considerablemente más rápido en tiempo de ejecución comparado con el algoritmo K-means, y dividió la base de datos en 19 grupos, permitiendo identificar 9 temas, mostrando que es una herramienta bastante útil a la hora de identificar patrones en cantidades masivas de documentos.

Finalmente, el uso de técnicas de minería de texto en tweets publicados ante un desastre, permiten visualizar de manera rápida las diferentes opiniones, inquietudes, canales para ayudar y para ser ayudado, información que puede ser de gran ayuda en la etapa de respuesta durante el desastre y en la etapa de recuperación en la gestión de los desastres, debido a que con esta información, el tomador de decisiones puede enfocarse en el clúster que más elementos aporta para dar distribuir información de manera más oportuna durante y después de los desastres

.

9. Recomendaciones

Para complementar o extender este trabajo de investigación, considerando los resultados obtenidos, se recomienda hacer para el análisis de los resultados un análisis de sentimientos, lo cual serviría para entender de mejor manera la postura que los usuarios de Twitter desean mostrar al momento de postear en un evento catalogado como desastre.

Por otro lado podría ser interesante la comparación de estos algoritmos, pero esta vez usando lenguajes de programación diferentes, y usando diferentes tamaños de bases de datos, pudiendo comparar los resultados obtenidos de los diferentes software y los tiempos de ejecución obtenidos en cada uno.

Referencias Bibliográficas

20mintos.es. (2015). La ONU cifra en 335 los desastres naturales al año por fenómenos climáticos y 30.000 sus muertos. Retrieved March 22, 2018, from https://www.20minutos.es/noticia/2611725/0/desastres-naturales-clima/aumento-mundo-cambio/climatico-relacionados/

Alexander, D. E. (2014). Social Media in Disaster Risk Reduction and Crisis Management. *Science and Engineering Ethics*, *20*(3), 717–733. https://doi.org/10.1007/s11948-013-9502-z

Arrivillaga, J., Greenleaf, D., Hawthorn, M., & Alvarado, R. (2016). Revealing the landscape: Detecting trends in a scientific corpus. *2016 IEEE Systems and Information Engineering Design Symposium, SIEDS 2016*, 292–297. https://doi.org/10.1109/SIEDS.2016.7489317

Au, I. R. N. (2017). Analysis of data volumes circulating in SNs after the occurrence of an earthquake Main features of Social Media data, *20*(3), 286–298.

Beltrán Martínez, B. (2014). Minería de datos. *Benemérita Universidad Autónoma de Puebla*, *1*(5), 67. Retrieved from http://bbeltran.cs.buap.mx/NotasMD.pdf

Benitez Andradez, José; Valvuena Lopéz, J. (2011). Motores de Búsqueda Web: Estado del arte en Probabilistic Latent Semantic Analysis y en Latent Dirichlet Allocation aplicado a problemas de acceso a la información en la Web. Retrieved March 19, 2018, from https://www.jabenitez.com/personal/MASTER/MOTORES_DE_BUSQUEDA_WEB/TAREAS/MBW-TareaExtraFinal-JoseAlbertoBenitezAndrades-y-JuanAntonioValbuenaLopez.pdf

Bhaskar, V. (2017). Author ’ s Accepted Manuscript Mining Crisis Information : A Strategic

Approach for Detection of People at Risk through Social Media Analysis. *International Journal of Disaster Risk Reduction*. https://doi.org/10.1016/j.ijdrr.2017.12.002

Blanco, E.-J., & Sanz, H. (2016). Algoritmos de clustering y aprendizaje automático aplicados a Twitter. *Universidad Politécnica de Catalunya*. Retrieved from https://upcommons.upc.edu/bitstream/handle/2117/82434/113257.pdf?sequence=1&isAllowed=y

Blázquez Ochando, M. (2012). Técnicas avanzadas de recuperación de información: Frecuencias y pesos de los términos de un documento. Retrieved January 17, 2019, from http://ccdoc-tecnicasrecuperacioninformacion.blogspot.com/2012/11/frecuencias-y-pesos-de-los-terminos-de.html

Cenditel. (2016). *Guía Teorica del Proyecto Modelado de Tópicos Indice general*. CENDITEL.

Chandía Sepúlveda, B. (2016). APLICACIÓN Y EVALUACIÓN LDA PARA BRUNO CHANDÍA SEPÚLVEDA Profesor Guía : Rodrigo Alfaro Arancibia.

Cheng, T., & Wicks, T. (2014). Event detection using twitter: A spatio-temporal approach. *PLoS ONE*, *9*(6), 1–10. https://doi.org/10.1371/journal.pone.0097807

El país, C. (2017). Colombia, un país expuesto a desastres como el de Mocoa por vulnerabilidad al cambio climático. Retrieved March 18, 2018, from http://www.elpais.com.co/colombia/un-pais-expuesto-a-desastres-como-el-de-mocoa-por-vulnerabilidad-al-cambio-climatico.html

El tiempo. (2016). Consecuencias de Desastres naturales desde 1900 - Archivo Digital de Noticias de Colombia y el Mundo desde 1.990. Retrieved March 18, 2018, from http://www.eltiempo.com/archivo/documento/CMS-16568384

Federación Internacional de Sociedades. (n.d.). Gestión de desastres - IFRC. Retrieved October 28, 2018, from https://www.ifrc.org/es/introduccion/disaster-management/gestion-de-desastres/

Fersini, E., Messina, E., & Pozzi, F. A. (2017). Earthquake management: a decision support system based on natural language processing. *Journal of Ambient Intelligence and Humanized Computing*, *8*(1), 37–45. https://doi.org/10.1007/s12652-016-0373-4

García Renedo, M. (2007). *Psicología y desastres : aspectos psicosociales*. Universitat Jaume I.

Gründer-Fahrer, S., Schlaf, A., Wiedemann, G., & Heyer, G. (2018). *Topics and topical phases in German social media communication during a disaster*. *Natural Language Engineering* (Vol. 24). https://doi.org/10.1017/S1351324918000025

Hagras, M., Hassan, G., & Farag, N. (2017). Towards natural disasters detection from twitter using topic modelling. *Proceedings - 2017 European Conference on Electrical Engineering and Computer Science, EECS 2017*, 272–279. https://doi.org/10.1109/EECS.2017.57

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. (M. Kamber, Ed.), *Elsevier* (Second, Vol. 12). ELSEVIER. https://doi.org/10.1007/978-3-642-19721-5

Huang, Q., Cervone, G., & Zhang, G. (2017). A cloud-enabled automatic disaster analysis system of multi-sourced data streams: An example synthesizing social media, remote sensing and Wikipedia data. *Computers, Environment and Urban Systems*, *66*, 23–37. https://doi.org/10.1016/j.compenvurbsys.2017.06.004

Kibanov, M., Stumme, G., Amin, I., & Lee, J. G. (2017). Mining social media to inform peatland fire and haze disaster management. *Social Network Analysis and Mining*, *7*(1), 1–19.

https://doi.org/10.1007/s13278-017-0446-1

Ledezma, A. I. (2004). *Aprendizaje Automatico En Conjuntos De Clasificadores Heterog´E Neos Y Modelado De Agentes*. Universidad Carlos III de Madrid. Retrieved from http://e-archivo.uc3m.es

Li, H., Caragea, D., Caragea, C., & Herndon, N. (2018). Disaster response aided by tweet classification with a domain adaptation approach. *Journal of Contingencies and Crisis Management*, *26*(1), 16–27. https://doi.org/10.1111/1468-5973.12194

mean cloud. (n.d.). Text Clustering | MeaningCloud. Retrieved January 17, 2019, from https://www.meaningcloud.com/products/text-clustering

Moya, R. (n.d.). Selección del número óptimo de Clusters - Jarroba. Retrieved September 7, 2018, from https://jarroba.com/seleccion-del-numero-optimo-clusters/

Murzintcev, N., & Cheng, C. (2017). Disaster Hashtags in Social Media. *ISPRS International Journal of Geo-Information*, *6*(7), 204. https://doi.org/10.3390/ijgi6070204

Namkyung, O., & Junghyae, L. (2017). Activation and variation of the United Nation ' s cluster coordination model : a comparative analysis of the Haiti and Japan disasters, (May 2015), 37–41. https://doi.org/10.1080/13669877.2015.1017826

Palma, J., & Marín, R. (2008). *Inteligencia artificial: técnicas, métodos y aplicaciones*. España: McGrawHill.

Reliefweb. (2018). Colombia: Desastres Naturales 2017 a 4 de abril de 2018 - Colombia | ReliefWeb. Retrieved December 20, 2018, from

https://reliefweb.int/report/colombia/colombia-desastres-naturales-2017-4-de-abril-de-2018

Resch, B., Usländer, F., Havas, C., & Usländer, F. (2017). Combining machine-learning topic models and spatiotemporal analysis of social media data for disaster footprint and damage assessment. *Cartography and Geographic Information Science*, *00*(00), 1–15. https://doi.org/10.1080/15230406.2017.1356242

San Francisco Chronicle. (2018). California Fire Map: 2018 Wildfire Tracker for Northern, Central and Southern California. Retrieved January 17, 2019, from https://projects.sfchronicle.com/2018/fire-tracker/

Santana, E. (2016). Data Mining con R: Text Mining con Twiter. Retrieved April 8, 2018, from http://apuntes-r.blogspot.com.co/2016/11/text-mining-con-twiter.html#more

Sapul, M., Shiela, C., Aung, T. H., & Jiamthapthaksin, R. (2017). Trending topic discovery of Twitter Tweets using clustering and topic modeling algorithms. *Proceedings of the 2017 14th International Joint Conference on Computer Science and Software Engineering, JCSSE 2017*. https://doi.org/10.1109/JCSSE.2017.8025911

Silge, J., & Robinson, D. (n.d.). Text Mining with R. Retrieved September 7, 2018, from https://www.tidytextmining.com/

Su, L. Y. F., Cacciatore, M. A., Liang, X., Brossard, D., Scheufele, D. A., & Xenos, M. A. (2017). Analyzing public sentiments online: combining human- and computer-based content analysis. *Information Communication and Society*, *20*(3), 406–427. https://doi.org/10.1080/1369118X.2016.1182197

Teknomo, K. (n.d.). Ejemplo numérico de Agrupamiento K-medias. Retrieved September 7, 2018,

from https://people.revoledu.com/kardi/tutorial/kMean/EjemploNumerico.htm

The Guardian. (2018). Democrats take control of House but Republicans tighten grip on Senate | US news | The Guardian. Retrieved January 17, 2019, from https://www.theguardian.com/us-news/2018/nov/06/midterm-elections-2018-exit-polls-voters

Turkewitz, J., & Richtel, M. (2018). Air Quality in California: Devastating Fires Lead to a New Danger - The New York Times. Retrieved January 30, 2019, from https://www.nytimes.com/2018/11/16/us/air-quality-california.html

Wang, B., & Zhuang, J. (2017). Crisis information distribution on Twitter: a content analysis of tweets during Hurricane Sandy. *Natural Hazards*, *89*(1), 161–181. https://doi.org/10.1007/s11069-017-2960-x

York, A. (2017). Social Media Demographics for Marketers. Retrieved March 18, 2018, from https://sproutsocial.com/insights/new-social-media-demographics/#twitter

Zou, L., & Song, W. W. (2016). LDA-TM: A two-step approach to Twitter topic data clustering. *Proceedings of 2016 IEEE International Conference on Cloud Computing and Big Data Analysis, ICCCBDA 2016*, 342–347. https://doi.org/10.1109/ICCCBDA.2016.7529581