Aplicación de técnicas de agrupamiento (clustering) para el análisis estadístico de tendencias en twitter basado en el lenguaje de programación R.

Víctor Alfonso Sanabria Ruiz

Trabajo de grado para optar el título de

Ingeniero Industrial

Director:

Henry Lamos Díaz

PH.D. Física-Matemáticas

Codirector:

Daniel Orlando Martínez Quezada

Ingeniero Industrial

Universidad Industrial de Santander

Facultad de Ingeniería Físico-Mecánicas

Escuela de Estudios Industriales y Empresariales

Bucaramanga

2017

DEDICATORIA

A Dios por orientar mi camino día a día y llenarme de tantas bendiciones y fortalezas en mi vida

A mi papá por darme tantas lecciones de vida, que con su humildad y sencillez me hizo una persona de bien y que desde el cielo me guía, ilumina mis pasos y cuida de mí.

A mi mamá, por ser la amiga y compañera que me ha ayudado a crecer, gracias por estar siempre conmigo en todo momento. Gracias por la paciencia que has tenido, por cada uno de los consejos sabios. Gracias mamá por estar al pendiente durante toda esta etapa.

A mis hermanos, Carlos, Jairo, Yanet, Ludy y Naty que con su amor me han enseñado a salir adelante, gracias por la paciencia, por preocuparse por su hermano menor, gracias por compartir sus vidas, pero sobre todo gracias por ese apoyo incondicional.

Víctor Sanabria

AGRADECIMIENTOS

A mí director Henry Lamos Díaz por brindarme la oportunidad, su conocimiento y orientación profesional.

A mi codirector Daniel Orlando Martínez Quezada por la orientación, la paciencia y el tiempo dedicado durante el desarrollo del proyecto.

A la familia Ruiz Florián por todo su apoyo porque en los momentos de dificultad y alegrías de mi vida estuvieron presentes, gracias familia por darme el ejemplo de amor y humildad.

A Alejandra, Paola y Debora por esa linda amistad, apoyo incondicional y momentos vividos junto a ellas

A los compañeros del grupo OPALO, por su apoyo y enseñanzas; que permitieron la realización de éste trabajo.

Al grupo OPALO por abrirme las puertas en la investigación, brindándome las herramientas necesarias para la realización del proyecto creciendo como profesional.

Víctor Sanabria

Tabla de contenido

Lista de Tablas

Lista de Figuras

Apéndices

Apéndice A. Artículo de Investigación, ver apéndice adjunto en el CD y puede visualizarlos en base de datos de la Biblioteca UIS

Resumen

Título del proyecto: Aplicación de técnicas de agrupamiento (clustering) para el análisis estadístico de tendencias en twitter basado en el lenguaje de programación R[*].

Autor: Víctor Alfonso Sanabria Ruiz[**]

Palabras clave: Aprendizaje Automático, Agrupamiento, Minería de Texto, Redes Sociales.

DESCRIPCIÓN

En las últimas décadas, el uso de técnicas de aprendizaje automático no supervisado en aplicaciones de redes sociales se ha visto de manera positiva en la comunidad científica ya que permite el descubrimiento de conocimiento a partir de datos sin una intervención previa. Este tipo de aplicaciones normalmente se encuentra asociadas a un marco tradicional de análisis de texto el cual consta de cuatro fases consecutivas: definición de corpus, pre-procesamiento, representación y descubrimiento de conocimiento. En la primera se define los documentos objeto de estudio los cuales son conocidos como corpus, el pre-procesamiento da una forma al corpus que permite analizar con métodos estadísticos, la representación que consiste en la transformación del corpus de documentos a un espacio vectorial para ser procesados en la fase de descubrimiento de conocimiento generando modelos de aprendizaje automático como los de agrupamiento. En el presente trabajo se evaluaron diferentes variantes del algoritmo k-means en una base de datos de prueba. Además, un caso de estudio para el análisis de texto es presentado, en este se definió como corpus los tweets del usuario de un periódico local en una ventana de tiempo de dos meses, utilizando representaciones de TF-IDF con el fin de aplicar un algoritmo de agrupamiento k-means que permitieron identificar tendencias características, junto a análisis descriptivos adicionales se lograron identificar índices de impacto a lo largo del tiempo.

---

[*] Trabajo de grado.

[**] Facultad de Ingenierías Fisicomecánicas. Escuela de Estudios Industriales y Empresariales. Director: PhD. Henry Lamos Díaz, Codirector: Ing. Industrial Daniel Orlando Martínez Quezada

Abstract

Project title: Application of clustering techniques for the statistical analysis of trends in twitter based on the programming language R*.

Author: Víctor Alfonso Sanabria Ruiz**

Keywords: Machine Learning, Clustering, Text Mining, Social Networks

DESCRIPTION

In the last decades, the use of unsupervised machine learning techniques in social networks applications has been seen in a positive way for the scientific community since it allows the knowledge discovery from data without prior intervention. This type of applications is usually associated to a traditional text analysis framework which consists of four consecutive phases: definition of corpus, preprocessing, representation and knowledge discovery. In the first phase, the documents that are object of study are defined, which ones are known as corpus, the pre-processing gives a structure to the corpus that allows to analyze with statistical methods, the representation that consists in the transformation of the corpus of documents to a vector space to be processed in the knowledge discovery phase generating machine learning models such as clustering. In the present work we evaluated different variants of the k-means algorithm in a test database. Moreover, a case study for text analysis is presented, in which the user's tweets of a local newspaper were defined as a corpus in a two-month time window, using a TF-IDF representation in order to apply a k-means algorithm that allowed to identify characteristic trends, along with additional descriptive analyzes, were able to identify index of impact over time.

* Degree Project.

** Faculty of Physicomechanical Engineering. Industrial and Business School. Director: PhD. Henry Lamos Díaz, Codirector: Ing. Daniel Orlando Martínez Quezada

Introducción

La presente investigación se refiere al aprendizaje automático no supervisado la cual tiene una serie de técnicas para que una máquina pueda aprender y generalizar comportamientos a partir de información suministrada y que debido al uso masivo de redes sociales tales como Twitter, Facebook e Instagram es posible el acceso a información relevante, sujeto a la volatilidad de la misma. Tendencias como la minería de texto o el manejo de grandes datos que buscan dentro de sus líneas de investigación técnicas para extraer conocimiento sin una estructura definida. Las técnicas de agrupación de documentos de texto presentan relevancia en áreas como: mercadeo, gestión de desastres, biomédicos, aplicaciones académicas, de seguridad, entre otros. (Won, D., Song, B. M., & McLeod, D. 2006).

La agrupación es un enfoque ampliamente usado en minería de datos para el descubrimiento de eventos importantes en un conjunto de documentos de texto. (Luo, C., Li, Y., & Chung, S. M. 2009). Siendo una técnica importante que puede ser usada para la organización automática de documentos en grupos relacionados. Además, tiene un papel importante en la búsqueda de grandes bases de datos, particularmente en las áreas de clasificación de documentos y estrategias de presentación de resultados (Aljaber, B., Stokes, N., Bailey, J., & Pei, J. 2010), donde la minería de texto entra como un enfoque al descubrimiento de patrones interesantes y nuevos conocimientos a partir de un conjunto de datos, su objetivo es descubrir tendencias, desviaciones y asociaciones en una gran cantidad de información textual, en forma general se puede decir que la minería de

texto es el proceso encargado del descubrimiento de conocimientos que no existe explícitamente. (Montes, M., & de Lenguaje Natural, G. L. 2014).

El análisis de texto se rige a una serie de etapas como seleccionar los datos óptimos a procesar, en la segunda etapa los datos son sometidos a su respectivo pre procesamiento, enseguida los textos se transforman a algún tipo de representación estructurada o semi-estructurada que facilite su posterior análisis, mientras que en la cuarta etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevos conocimientos; a continuación en la figura 1 muestra el proceso de análisis de datos para su interpretación y consolidación.

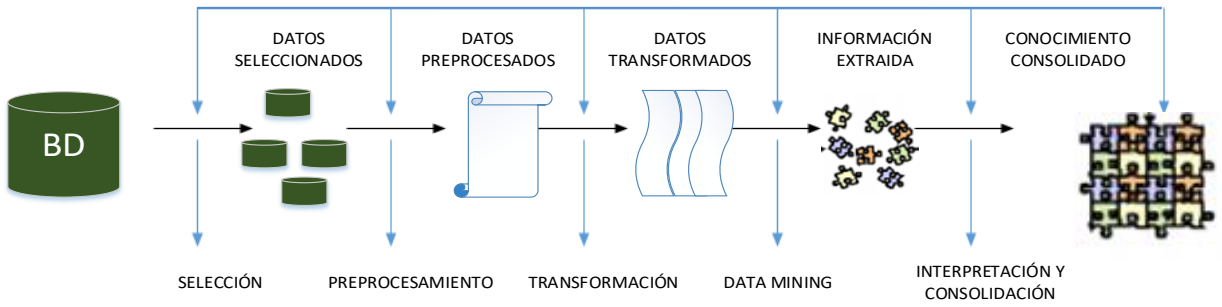

Figura 1. Metodología para el descubrimiento de conocimiento en base de datos. Modificado de: Gil, N. L. Q., &

La materia prima del presente proyecto es la información que se encuentra en las Redes Sociales con el propósito de realizar un análisis exhaustivo mediante técnicas clustering, teniendo este dos finalidades, por un lado, presentar una revisión de la literatura sobre las diferentes técnicas de agrupamiento de textos utilizadas en el área de aprendizaje automático, y por otra parte aplicar las técnicas encontradas a instancias de la literatura y datos de la red social Twitter con el fin de hallar patrones de comportamiento de los usuarios en temas específicos.

Tabla 1.
Cumplimiento de objetivos del proyecto

Objetivos Específicos	Cumplimiento
• Efectuar una revisión bibliográfica del estado del arte actual de técnicas de agrupamiento en documentos tipo texto.	Capítulo 4
• Seleccionar e identificar las técnicas de transformación de datos no estructurales a datos estructurales.	Capítulo 5
• Implementar y validar los algoritmos de clustering mediante instancias del benchmarking.	Capítulo 6
• Realizar un análisis de datos con la finalidad de establecer relaciones y características de los individuos.	Capítulo 7
• Con base a los resultados obtenidos en la investigación realizada, elaborar un artículo de carácter publicable.	Apéndice A

1. Planteamiento del Problema

En los mercados cambiantes de la actualidad con ya más de una década, del surgimiento de internet, hacen que actualmente los usuarios tengan la posibilidad de generar su propio contenido y compartirlo públicamente con mayor facilidad. En este auge, las redes sociales han cobrado gran popularidad, en particular la plataforma de Twitter la cual permite a sus usuarios compartir mensajes de texto de máximo 140 caracteres con sus familiares, amigos y seguidores. Diariamente se publican más de 500 millones de mensajes, comúnmente llamados tweets.

Las tecnologías de la información han experimentado crecimientos espectaculares desde los años 50, a un ritmo en el que la potencia de la informática crece exponencialmente todos los años. A este crecimiento natural de la informática le ha acompañado la gran cantidad de volúmenes de datos que se manejan. Encontrar y monitorear la opinión de muchos usuarios es una tarea difícil ya que identificar información relevante y extraerla de forma resumida es un procedimiento arduo para ser realizado manualmente. (Domínguez, Y. 2007).

Debido a este mercado cambiante de las redes sociales surge la utilización de técnicas de agrupamiento (clustering) ya que ha sido ampliamente tratada en diversos campos. El agrupamiento es una técnica común en el área de análisis estadístico y se ha ido extendiendo, por su aplicabilidad, a otros campos como el reconocimiento de patrones, minería de datos, análisis de imágenes o aprendizaje automático.

Tradicionalmente, las técnicas de agrupamiento se han clasificado en dos tipos fundamentales: las técnicas jerárquicas y las particionales. Las primeras corresponden con análisis recursivos de agrupamiento en los datos para ir obteniendo de una manera paulatina una jerarquía de pertenencia de grupos. En la segunda clasificación, se pretende obtener un único nivel de subconjuntos, donde cada uno de ellos recoge un comportamiento homogéneo con respecto al conjunto total.

Por lo tanto, para mejorar el conocimiento de los usuarios a través de la generación de contenido en redes sociales, se realiza *analytics* mediante *Machine Learning,* realizando evaluaciones de las diferentes variantes del algoritmo k-means en una base de datos de prueba. Además, un caso de estudio real utilizando datos de la red social twitter en un conjunto de técnicas denominadas clustering describiendo brevemente los conceptos básicos de este campo siguiendo una estructura definida de los documentos objeto de estudio los cuales son conocidos como corpus, el pre-procesamiento da una forma al corpus que permite ser analizada con métodos estadísticos, la representación que consiste en la transformación del corpus de documentos a un espacio vectorial para ser procesados en la fase de descubrimiento de conocimiento generando modelos de aprendizaje automático como los de agrupamiento. Con todo ello se ha conseguido mejorar las capacidades adquiridas en la etapa de estudio previo a partir de la puesta en práctica de los anteriores conocimientos adquiridos así mismo el desarrollo del trabajo servirá para conocer el potencial de la API de twitter.

2. Justificación del Proyecto

Hoy en día muchas de las organizaciones dirigidas a sectores industriales poseen problemas a la hora de tomar decisiones con base en la búsqueda de estrategias competitivas, que logren aumentar la innovación, productividad y a su vez generar cambios constantes en el producto, manejo de la publicidad, estrategias de promoción, definir las necesidades del cliente y satisfacerla al máximo; lo que actualmente es una de las desventajas en las organizaciones Colombianas, el tener grandes cantidades de datos e información almacenada y guardada de forma incorrecta, lleva a que muchos de estos datos y especificaciones no sean utilizados debidamente, con el fin de crear estrategias que contribuyan a la formación de industrias competitivas e innovadoras. (Echeverri, L. A., Retamoza, A. M. P., de la Rosa, M. O., Barros, I. V., Álvarez, D. D. O., & Guerrero, E. C. 2013).

El crecimiento en el volumen de datos generados por diferentes sistemas y mediciones de actividades cotidianas en la sociedad es un factor que tiene influencia en la necesidad de modificar, optimizar y concebir métodos y modelos de almacenamiento y tratamiento de datos que suplan las falencias que presentan las bases de datos y los procesos de KDD. (Leal, E. J. 2016).

Debido al problema de tratamiento de datos se tiene en cuenta la evolución de las redes sociales para el respectivo análisis donde han surgido multitud de herramientas destinadas a la analítica; y en el caso de Twitter esto no es una excepción. Entre la información que esta ofrece al ser algo más parecido a un medio de comunicación donde la mayoría de la información es pública

basándose en la transmisión de contenidos que a los propios usuarios interesan, generando así nuevas tendencias a escala global. (Guy Kawasaki & Peg Fitzpatrick 2014)

Los criterios más importantes que se tienen en cuenta en el presente proyecto son la eficiencia de los algoritmos y su fiabilidad, es por ello que se ve la importancia de realizar la presente investigación con el fin de observar las técnicas de agrupamiento que ofrece un análisis de las bases de datos con información selectiva, obteniendo resultados relevantes, información veraz y precisa, para la predicción de comportamientos de patrones y tendencias, hacer pronósticos, encontrar relaciones, generar redes de tendencias en twitter, que sirvan de apoyo en la toma de decisiones y explotar de forma eficiente el lenguaje de programación R.

3. Objetivos

3.1 Objetivo General

Aplicar técnicas de análisis de clusters que permitan clasificar documentos en bases de datos del benchmarking.

3.2 Objetivos Específicos

- Efectuar una revisión bibliográfica del estado del arte actual de técnicas de agrupamiento en documentos tipo texto.
- Seleccionar e identificar las técnicas de transformación de datos no estructurales a datos estructurales.
- Realizar un análisis de datos con la finalidad de establecer relaciones y características de los individuos.
- Implementar y validar los algoritmos de clustering mediante instancias de benchmarking.
- Con base en los resultados obtenidos en la investigación realizada elaborar un artículo de carácter publicable.

4. Revisión de la Literatura

La primera vez que se escuchó el término base de datos fue en un congreso celebrado en California en 1963, (Korth, H. F., Silberschatz, A., Sudarshan, S., & Pérez, F. S. 1993). Los orígenes de las bases de datos se remontan a la antigüedad donde ya existían bibliotecas y toda clase de registros. Además, también se utilizaban para recoger información sobre las cosechas y censos. Sin embargo, su búsqueda era lenta y poco eficaz y no se contaba con la ayuda de máquinas que pudiesen reemplazar el trabajo manual.

Para manipular automáticamente las bases de datos grandes, surge la minería de datos que desde los años sesenta cuando en esos momentos los estadísticos utilizaban los términos de data fishing, data Mining o data archaerlogy; partiendo de esto en los años ochenta se empezó a hablar de los términos KDD que por sus siglas en inglés significa proceso de extracción de conocimiento a partir de datos del cual la minería de datos forma parte. A partir de ese año se fueron creando varias empresas dedicadas a prestar servicios relacionados con la minería de datos (Braga, L.P., Valencia, L.I, & Carvajal, S.S 2009), teniendo su origen en los sistemas de información cuya finalidad era recopilar información sobre un tema determinado para tomar decisiones.

El almacenamiento de datos se ha convertido en una tarea rutinaria de los sistemas de información de las organizaciones. Esto es aún más evidente en las empresas de la nueva economía, la telefonía, el marketing directo, etc. Los datos almacenados son un tesoro para las organizaciones, es donde se guardan las interacciones pasadas con los clientes, la contabilidad de

sus procesos internos y representan la memoria de la organización, pero con tener memoria no es suficiente debiendo pasar a la acción inteligente sobre los datos para extraer la información que almacenan. (Aluja Banet, T. 2001).

La tecnología de Internet actual y su creciente demanda necesita el desarrollo de tecnologías de minería de datos más avanzadas para interpretar la información y el conocimiento de los datos distribuidos por todo el mundo. En este siglo la demanda continuará creciendo, y el acceso a grandes volúmenes de datos traerá la mayor transformación para la sociedad. Por tanto, el desarrollo de tecnologías de minería de datos continuará siendo una importante área de estudio. (Gilbert, K., Sánchez, R. R., & Santos, J. C. R. 2006).

En al ámbito empresarial y el mundo de los negocios, durante la última década del pasado siglo y los primeros años de este se hablaba de Business Intelligence (BI) para hacer referencia al conjunto de estrategias y herramientas que una empresa tenía a su disposición para poder analizar los datos de su organización. Con el BI se hacían previsiones y análisis.

Las herramientas y conceptos que hoy se agrupan bajo la denominación Big Data que está relacionado con lo que se ha conocido como minería de datos, un campo de las ciencias de la computación que intenta descubrir patrones en grandes volúmenes de datos. La minería de datos, al igual que el Big Data, utiliza los métodos como el aprendizaje automático y la Estadística para analizar los patrones en las bases de datos con las que trabaja.

Pero también es importante comprender que además de los datos estructurados, aquellos otros que provienen de fuentes de información conocidas y que, por tanto, son fáciles de medir y analizar a través de los sistemas tradicionales, empezamos a poder y querer manejar datos no estructurados (Tascón, M. 2013).

La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones previamente desconocidos e interesantes, tales como grupos de registros de datos, es por ello que la tecnología puede generar nuevas oportunidades de negocios, brindando las siguientes capacidades: (Allegue Lorenzo, Á., & Rueda Silva, A. C. 2016).

Predicción de tendencias y comportamientos: Este tipo de modelo, toma como base una hipótesis definida por el usuario y posteriormente se efectúa una prueba de validación contra los datos, para evaluar la calidad de estos modelos, tales como simplicidad e interoperabilidad etc.

Modelos de descubrimiento: Este tipo de modelos intentan encontrar relaciones y patrones de comportamiento en el conjunto de datos para ofrecer conocimiento sobre un problema concreto. (Corso, C. L., & Alfaro, S. L. 2009).

A continuación, se presenta una recopilación de las investigaciones relacionadas con la minería de texto para conocer las contribuciones previas y actuales tomadas como referencia para la ejecución del tema propuesto.

(Jain, A. K., Murty, M. N., & Flynn, P. J. 1999). Abordaron el tema clustering, donde presentan métodos de agrupamiento desde una perspectiva de reconocimiento de patrones estadísticos, con el objetivo de proporcionar consejos útiles y referencias a conceptos fundamentales accesibles a la amplia comunidad de profesionales de la agrupación. Presentaron una taxonomía de las técnicas de agrupamiento e identificando temas transversales y avances. También describieron algunas aplicaciones importantes de algoritmos de agrupación tales como segmentación de imágenes, reconocimiento de objetos y recuperación de información.

(Liu, H., & Huang, S. T. 2003). Realizan un estudio acerca de las técnicas de agrupamiento que se utilizan comúnmente para identificar grupos naturales de datos no etiquetados. Sin embargo, las etiquetas de las clases no siempre corresponden a la agrupación natural de los datos, así técnicas utilizadas para el análisis exploratorio de datos, pueden capturar la estructura natural de los mismos.

(Ramos, J. 2003, December). En este trabajo, examinaron los resultados de la aplicación Término Frecuencia inversa del documento (TF-IDF) para determinar qué palabras en un corpus de documentos podrían ser más favorables para su consulta. Como el término implica, TF-IDF calcula valores de cada palabra en un documento a través de proporción inversa de la frecuencia de la palabra en un documento en particular al porcentaje de documentos en los que aparece la palabra. Altos números de TF-IDF implican una relación con el documento en el que aparecen, sugiriendo que, si esa palabra apareciera en una consulta, el documento podría ser de interés para el usuario. Proporcionamos pruebas de que este simple Algoritmo que clasifica eficientemente palabras relevantes que puede mejorar la recuperación de consultas.

(Cardona, A., Nigro, N., Sonzogni, V., & Storti, M. 2006). **“**Realiza un estudio en el año 2006 en donde analizan diferentes metodologías y criterios para realizar análisis de agrupamiento sobre datos multivariados. El análisis de agrupamiento tiene por objetivo formar grupos de elementos, de manera tal que los pertenecientes a un mismo grupo sean parecidos entre sí y distintos a los miembros de los restantes grupos. Se describen consideraciones para los dos grandes tipos de métodos: jerárquicos y de partición. Los primeros proveen una estructura de grupos a diferentes niveles de granularidad según su nivel de similitud, mientras que los segundos dividen el conjunto muestral en grupos internamente homogéneos. En el caso de los métodos jerárquicos, se analiza en detalle las diferentes medidas de asociación y distancia utilizadas por el método, así como también el ligamiento usado para recalcular las distancias”.

(Pascual, D., Pla, F., & Sánchez, S. 2007). Para el año 2007 realizan un estudio sobre diferentes técnicas de agrupamiento, las que se encuentran dentro del campo de estudio del reconocimiento de patrones, haciendo hincapié en las técnicas basadas en densidad. Mostraron el estudio comparativo de tres algoritmos K-means, CURE y DBSCAN, empleando bases de datos reales y artificiales.

(Porras, J. C. C., Laverde, R. M., & Diaz, J. R. 2008). Hablan sobre la gran cantidad de datos y el elevado volumen de información que se tienen actualmente que ha hecho necesario contar con técnicas automáticas que permitan indagar, organizar y extraer información implícita presente en las enormes bases de datos que contienen información bibliográfica, económica, genética, información derivada de experimentos e investigaciones y muchos otros tipos de información la

cual extraer de forma manual resulta prácticamente imposible a medida que va creciendo el tamaño de las bases de datos.

(Aliguliyev, R. M. 2010). Hablan sobre la agrupación de los documentos de texto donde es un problema central en la minería de texto que se define como la división de un conjunto de documentos en grupos de acuerdo a sus temas o contenidos principales. La agrupación de documentos tiene muchos propósitos, incluyendo la recuperación de información, la generación de un resumen, la extracción automática de tema, navegar por las colecciones de documentos, organización de la información en formato digital las bibliotecas y los temas de detección.

(Kuna, H., García Martínez, R., & Villatoro, F. (2009). busca como establecer una taxonomía relacionada con la calidad de los datos, analizando los procesos de explotación de información que mejor aplican a la identificación de patrones de pistas de auditoria, se explorarán esas procesos analizando las ventajas y desventajas de cada una en su estudio, demostrando con esto que la explotación de información es un elemento fundamental de un proceso más amplio que tiene como objetivo el descubrimiento de conocimiento en grandes bases de datos.

(González, D. P. 2010). Realiza una investigación en el 2010 basada en el aprendizaje no supervisado basándose en conjunto de datos mostrando la importancia para el conocimiento del comportamiento de una población, de la cual sólo se tiene una cantidad N de sus elementos. Al estudiar el proceso de división en clases, muestra de que cada técnica está diseñada para realizar una clasificación de tal modo que cada grupo sea lo más homogéneo y lo más diferente de los

demás como sea posible. El resultado de cada método de agrupamiento dependerá del algoritmo en concreto, del valor de los parámetros y de la medida de similaridad / disimilaridad adoptada.

Inspirado en el estudio de la minería de datos, su artículo se basó en organizar datos en agrupaciones sensibles donde es uno de los modos más fundamentales de comprensión y aprendizaje. Por ejemplo, un esquema común de clasificación científica coloca a los organismos en un sistema de taxones de rango: dominio, reino, filo, clase, etc. El análisis de clúster es el estudio formal de métodos y algoritmos para agrupar o agrupar características intrínsecas percibidas o similitud. El análisis de clústeres no utiliza etiquetas de categorías que marcan objetos con identificadores anteriores, es decir, etiquetas de clase. La ausencia de información de categoría distingue el agrupamiento de datos (aprendizaje no supervisado) de la clasificación o el análisis discriminante (aprendizaje supervisado). El objetivo de la agrupación es encontrar la estructura en los datos y por lo tanto es de naturaleza exploratoria. La agrupación tiene una larga y rica historia en una variedad de campos científicos. Uno de los algoritmos de agrupamiento más populares y sencillos, K-means, se publicó por primera vez en 1955. A pesar de que K-means fue propuesta hace más de 50 años y miles de algoritmos de agrupación han sido publicados desde entonces, K-means es hoy día ampliamente utilizado. (A. K. 2010).

(Kalogeratos, A., & Likas, A. 2011). En el año 2011 estudiaron la importancia del agrupamiento de documentos que es un enfoque de aprendizaje no supervisado para separar automáticamente documentos similares de un corpus en el mismo grupo, llamado cluster, y documentos diferentes entre grupos. La cual se analiza la idea de que, a pesar de que los centroides son los prototipos de agrupamiento óptimo con respecto a ciertos objetivos (Por ejemplo, basándose en la similitud de

coseno), su óptima podría también convertirse en un inconveniente en la baja calidad de datos (por ejemplo, valores atípicos, ruido). Especialmente, a medida que el número de objetos de datos es menor comparado con la complejidad de un Clustering (es decir, número de clusters, dimensionalidad), los centroides se convierten en representantes de grupos menos apropiado.

(Bouras, C., & Tsogkas, V. 2012). En el 2012 se centran en la agrupación de los datos, que en general, han sido ampliamente investigados por la comunidad científica en los últimos 20 años. Especialmente para la agrupación de documento, se han propuesto una gran variedad de técnicas. Un objetivo importante del agrupamiento de documentos es el de mejorar los resultados de los sistemas de recuperación de información en términos de precisión y sensibilidad. Esta a su vez conduce a servir mejor los resultados filtrados y adecuados a sus usuarios, ayudando en esencia, el proceso de toma de decisiones.

(Anchalia, P. P., Koundinya, A. K., & Srinath, N. K. 2013, June). Para el año 2013 en su artículo, dieron a conocer la importancia del K-Means Clustering el cual es un método utilizado para clasificar conjuntos de datos estructurados o no estructurados. Este es común y eficaz para clasificar los datos debido a su simplicidad y capacidad para manejar voluminosos conjuntos de datos. Se acepta el número de clusters y el conjunto inicial de centroides como parámetros. La distancia de cada elemento en los conjuntos de datos se calcula con cada uno de los centroides del respectivo racimo. El elemento se asigna entonces al clúster con el que la distancia del artículo es la menor. El centroide del grupo a la cual se ha asignado el elemento se vuelve a calcular. Uno de los métodos más importantes y comúnmente utilizados Agrupar los elementos de un conjunto de datos utilizando K-Means Clustering es calculando la distancia entre el punto y la media elegida.

Esta distancia suele ser la distancia euclidiana, aunque hay otras definiciones de cálculo de distancia existentes. Esta es la métrica más común para la comparación de puntos.

(Jiménez, C. M. 2014). En el año 2014 analiza su estudio en las estructuras más conocidas para el tratamiento de los big data, la estructura de un conjunto de datos, de este tipo de información está empezando a ser un requisito innegable para la supervivencia de muchas empresas y organizaciones. Como consecuencia de ello han surgido en los últimos años términos como big data, Mapreduce, Hadoop o computación en la nube. Así, la demanda de los llamados "científicos de datos" está creciendo exponencialmente. Plantea en su artículo una introducción divulgativa a todos estos términos y analiza las estructuras más conocidas para el tratamiento de los big data.

(Delgado Calle, C. 2015). En su investigación en el 2015, se basó en el modelo de identificación de Meta-Topics a través de análisis semántico de conjuntos de datos extraídos de twitter, centra su estudio en obtener las publicaciones de los gustos de los usuarios a los que sigue el sujeto y organizarlas de acuerdo a un tema general que comprenda varios temas como deportes, videojuegos, cine, etc. dependiendo del usuario. Para ello, se llevó a cabo un análisis semántico de las publicaciones (o tweets) extraídas de dichos usuarios, para su posterior análisis, de modo que se pudiera encontrar una relación entre los tweets.

(Cotelo, J. M., Cruz, F., Ortega, F. J., & Troyano, J. A. 2015). En su artículo, la cual tiene como propósito mostrar cómo es posible sacar partido de la información estructurada y no estructurada que proporciona la red social Twitter. Beneficiando la resolución de tareas relacionadas con la red social: la recuperación de tweets y la clasificación de tweets; en ambos casos prueban técnicas

que hacen uso tanto del contenido de los mensajes como de toda la estructura que los rodea, los textos están acompañados de mucha información estructurada que relaciona a múltiples entidades de distinta naturaleza, en el entorno de Twitter hay muchos objetos (tweets, hashtags, usuarios, palabras, ...) y relaciones entre ellos (co-ocurrencia, menciones, re-tuiteos, ...) que ofrecen innumerables posibilidades de procesado e información estructurada que proporciona Twitter.

En su trabajo (Martín Morales, S. 2016), presenta un análisis basado en el sentimiento (positivo, negativo o neutro) de un conjunto de textos obtenidos de Twitter en la diversidad de plataformas gratuitas en la web en las que puedes registrarte y compartir estados, fotos, pensamientos, etc. Gracias a ello, estas plataformas recogen millones de datos de todas las personas, y una vez procesados reportan información muy útil, entre otras muchas cosas, para estudios de mercado.

5. Marco Teórico

5.1 Aprendizaje Automático

Es la rama de la inteligencia artificial cuyo objetivo es desarrollar técnicas que permitan a las computadoras aprender. De forma más concreta, se trata de crear programas capaces de generalizar comportamientos a partir de una información suministrada en forma de ejemplos. Es, por lo tanto, un proceso de inducción del conocimiento.

El aprendizaje automático se encarga de diseñar y desarrollar algoritmos que permitan a las máquinas ser más eficientes y realizar tareas sin supervisión humana. El aprendizaje automático

trata de producir de manera automática modelos, como pueden ser reglas o patrones, de una serie de datos iniciales.

Usando modelos computacionales tales que, de forma automática, les permitan resolver nuevos problemas o mejorar su comportamiento en problemas ya vistos, En general, 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. (A. I. L., Mur, R. A., & de Miguel, M. A. S. 2004).

5.2 Enfoques de Aprendizaje

5.2.1 Aprendizaje supervisado. Con el aprendizaje supervisado nos referimos a todas aquellas aplicaciones o procesos en los que se dispone de información tanto de los valores de entrada del sistema como de los valores de salida deseados. De manera global, dos de los problemas típicos en el aprendizaje supervisado son el de clasificación y el de regresión. En clasificación, los valores deseados corresponden con las etiquetas de cada caso (información cualitativa), mientras que, en regresión, la información de salida es el valor real a estimar (información cuantitativa). Gallardo (Campos, M. 2009).

El aprendizaje supervisado consiste en un tipo de aprendizaje automático en donde al algoritmo que se utiliza 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.

5.2.2 Aprendizaje no supervisado Aprendizaje no supervisado es un método de Aprendizaje Automático donde un modelo es ajustado a las observaciones. Se distingue del Aprendizaje Supervisado por el hecho de que no hay un conocimiento a priori. En el aprendizaje no supervisado, un conjunto de datos de objetos de entrada es tratado. Así, el aprendizaje no supervisado típicamente trata los objetos de entrada como un conjunto de variables aleatorias, siendo construido un modelo de densidad para el conjunto de datos. (Espino, A. I. L., Mur, R. A., & de Miguel, M. A. S. 2004).

5.2.3 Aplicación de Algoritmos de clasificación supervisada. El objetivo de las técnicas de clasificación es la asignación de objetos a uno de varios grupos bien definidos, por un lado, está la clasificación no supervisada, trata a la clasificación como el descubrimiento de las clases de un determinado problema. Es decir que contamos con un conjunto de elementos descritos por un conjunto de características, sin conocer a que clase pertenece cada uno de ellos.

En cambio, en la clasificación supervisada enfoca el problema de clasificación de otra manera, es decir parte de un conjunto de elementos descrito por un conjunto de características donde conocemos la clase al cual pertenece. A este concepto se suele denominar conjunto de datos de entrenamiento o conjunto de aprendizaje. La clasificación supervisada ha sido aplicada en numerosos ámbitos como el diagnóstico de enfermedades, la concesión o rechazo de créditos bancarios.

Otro concepto fundamental en el ámbito de los métodos de clasificación son los diversos criterios para la evaluación de los clasificadores. Es decir, estimar la bondad de un clasificador, se conoce como proceso de validación, y esto nos permite efectuar una medición sobre la capacidad de predicción del modelo generado a partir de un clasificador.

Una alternativa de verificar o medir la bondad del clasificador es la matriz de confusión. Una matriz de confusión nos permite visualizar mediante una tabla de contingencia la distribución de errores cometidos por un clasificador. Esta matriz de confusión para el caso de dos clases tiene la siguiente apariencia. (Corso, C. L. 2009).

Tabla 2.
Matriz de confusión.

	REAL	
PREDICCIÓN	POSITIVO	NEGATIVO
POSITIVO	**TP**	**FP**
NEGATIVO	**FN**	**TN**

Nota: Aplicación de algoritmos de clasificación supervisada, Adaptado de: Universidad Tecnológica Nacional, Facultad Regional Córdoba. 2009

- TP (Verdaderos positivos): instancias correctamente reconocidas por el sistema.
- FN (Falsos negativos): instancias que son positivas y que el sistema dice que no lo son.
- FP (Falsos positivos): instancias que son negativas pero el sistema dice que no lo es.
- TN (Verdaderos negativos): instancias que son negativas y correctamente reconocidas como tales.

5.2.4 Métricas de la matriz de confusión. Ayuda a evaluar la "calidad" de un modelo de clasificación supervisado.

- **Accuracy:** es la proporción del número total de predicciones que son correctas.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (1)$$

- **Precisión:** es la proporción de los casos predichos positivos que son correctos.

$$P = \frac{TP}{TP+FP} \quad (2)$$

- **Recall:** es la proporción de casos positivos que fueron identificados correctamente.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

- **F-measure:** es una métrica que toma en cuenta tanto el recall como la precisión.

$$F = \frac{2(P*Recall)}{(P+Recall)} \quad (4)$$

5.2.4.1 Ejemplo de aplicación de la Matriz de Confusión. Los datos de este ejemplo consisten en información de 4601 mensajes de correo electrónico, En un estudio para tratar de predecir si el correo electrónico era correo basura, o "Spam". Los datos son públicos Disponible en ftp.ics.uci.edu. La variable de respuesta es binaria, con valores de correo electrónico o spam, mediante la matriz de confusión apta para los datos de prueba la tasa global de errores es del 5,5%. (Hastie, T., Tibshirani, R., Friedman, J. 2009)

Tabla 3.
Matriz de confusión de datos de prueba.

	PREDICCIÓN	
REAL	email(1)	spam(0)
email(1)	**58.3% (TP)**	**2.5%(FN)**
spam(0)	**3.0%(FP)**	**36.3%(TN)**

Nota: Adaptado de: *"*The Elements of Statistical Learning Data Mining, Inference, and Prediction" Second Edition, 2009.

Al ver la cantidad en cada celda de la matriz se podrá saber en cuantas ocasiones ha sido exacta la predicción del modelo teniendo en cuenta que cada columna de la matriz representa el número de predicciones de cada clase, mientras que cada fila representa a las instancias en la clase real, siendo la etiqueta 1 positiva y 0 negativa

La primera celda de resultados, que contiene el porcentaje 58.3%, indica el número de verdaderos positivos para el valor 1. Dado que 1 indica que es un email, esta estadística indica que el modelo predijo el valor correcto de email en 58.3%; la celda de verdaderos negativos indica que en un 36.3% el modelo predijo correctamente que de los 4601 mensajes el 36.3% son spam; la celda que contiene 3.0% indica el número de falsos positivos, esta estadística indica que, en un 3.0% el modelo predijo que eran email cuando en realidad son spam; por ultimo La celda que indica el número de falsos negativos, o el porcentaje que el modelo predijo que eran spam cuando en realidad eran email. Sumando los valores de las celdas contiguas diagonalmente, se puede determinar la exactitud total del modelo. Una diagonal indica el número total de predicciones exactas y la otra indica el número total de predicciones erróneas.

Medidas de rendimiento:

Accuracy: esta medida muestra el porcentaje de acierto del clasificador, puede no ser una medida de rendimiento adecuada cuando el número de casos negativos es mucho mayor que el número de casos positivos.

Accuracy = 94.5% de acierto del clasificador

Precisión: mide el porcentaje de respuestas positivas del clasificador que son correctas y que corresponden realmente a muestras positivas.

Precisión = 96% de respuestas que realmente los mensajes son email.

Recall: mide el porcentaje de respuestas positivas reales que son correctas y que predijeron positivas.

Recall = 95% de respuestas que el clasificador predijo que si eran email.

F-measure: es una medida de exactitud de una prueba, que considera tanto la precisión como el recall, para calcular la puntuación de F-measure puede interpretarse como un promedio ponderado de la precisión y el recall, donde un puntaje F-measure alcanza su mejor valor en 1 y el peor en 0.

F-measure= 95%

5.3 Clustering

Una forma de aprendizaje no supervisado es la agrupación (en inglés, clustering) entre los que se pueden señalar COBWEB, EM, K-Means. El proceso de clustering consiste en la división de los datos en grupos de objetos similares, para medir la similitud entre objetos suelen utilizar diferentes formas de distancia: distancias como Euclídea, Manhattan, Maximum, Minkowski, etc. El clustering juega un papel muy importante en aplicaciones de minería de datos, tales como exploración de datos científicos, recuperación de la información y minería de texto, aplicaciones sobre bases de datos en aplicaciones web, (Garre, M., Cuadrado, J., Sicilia, M. A., Rodríguez, D., & Rejas, R. 2007).

El análisis de agrupamiento encuentra clusters de objetos de datos que son similares en un sentido a otro. Los miembros de un conglomerado se parecen más unos a otros de lo que son como miembros de otros grupos. El objetivo del análisis de agrupamiento es encontrar agrupaciones de alta calidad de tal manera que la semejanza entre clústeres sea baja y la similitud entre grupos sea alta.

El agrupamiento es útil para explorar datos. Si hay muchos casos y no hay agrupaciones obvias, los algoritmos de agrupamiento pueden usarse para encontrar agrupaciones naturales. El agrupamiento también puede servir como un útil paso de pre procesamiento de datos para identificar grupos homogéneos sobre los cuales construir modelos supervisados.

La agrupación también puede utilizarse para la detección de anomalías. Una vez que los datos se han segmentado en grupos, es posible que algunos casos no encajan bien en los clústeres. Estos casos son anomalías o valores atípicos.

5.4 Clustering de texto

Agrupar documentos de texto en diferentes grupos de categorías es un paso en la indexación, la recuperación, la gestión y la minería de texto abundante de datos en la web o en sistemas de información corporativos. Agrupamiento de texto puede describirse intuitivamente como hallazgo, dado un conjunto de vectores de datos en un espacio multidimensional.

Clustering de Texto descubre automáticamente la estructura implícita en una colección de documentos, identificando los temas más frecuentes dentro de la colección y distribuyendo los documentos en varios grupos (clusters). Esta distribución se caracteriza por maximizar la similitud entre los elementos de un mismo grupo y a la vez maximizar las diferencias entre los diversos grupos (Jing, L. 2008).

5.4.1 Diferencias entre la clasificación y el clustering de texto. clasificar o categorizar en el texto consiste en asignar a un texto individual una o varias categorías entre una taxonomía previamente definida. Crear 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). MeaningCloud proporciona una funcionalidad de categorización a través de su Clasificación de Texto, que ofrece diversos modelos de clasificación estándar predefinidos (p. ej.: IPTC para noticias, IAB para contenidos web) y también la posibilidad de que

el usuario pueda crear modelos a medida mediante las herramientas de personalización del producto.

Por el contrario, el **clustering** se ejecuta generalmente sobre un conjunto de documentos a la vez para distribuirlos en varios grupos atendiendo a sus similitudes. Y no parte de una taxonomía predefinida, sino que la decisión sobre qué textos van a un grupo y qué textos van a otro se toma dinámicamente en función de los contenidos del conjunto de documentos. Por lo tanto, el clustering no requiere la definición previa de una taxonomía ni el consiguiente entrenamiento o definición de reglas, en un enfoque que se conoce como aprendizaje no supervisado.

Clasificación y clustering son dos enfoques complementarios. La clasificación es apropiada cuando se conoce a priori la estructura que se va a dar al conjunto de documentos y es necesario analizar documentos individuales. El clustering requiere analizar un conjunto de documentos a la vez (y el resultado cambia si se altera el conjunto), pero ofrece el potencial de **descubrir la estructura implícita y los temas significativos** que emergen del contenido de los propios documentos.

5.4.2 Aplicación de clustering de texto. El clustering está especialmente indicado en aquellas aplicaciones donde se trata de detectar relaciones entre varios textos, de distribuirlos dinámicamente en agrupaciones naturales o de descubrir los temas más relevantes que emergen de sus contenidos y expresarlos en sus propios términos. En particular, en el importante campo del análisis de la Voz del Cliente o la gestión de la Experiencia del Cliente, el clustering se aplica allí donde se requiere descubrir la “nueva voz” de esos clientes.

- **Seguimiento y análisis de medios (sociales y tradicionales):** Detección de contenido duplicado, identificación de plagios, noticias relacionadas.
- **Recuperación de la información y sistemas de recomendación:** Agrupación de resultados de búsqueda, ayuda a la navegación, sugerencia de información relacionada, recomendación de contenidos y productos.
- **Análisis de minería de opiniones:** Descubrimiento de temas no predefinidos ni previstos en encuestas y reclamaciones (que permita una gestión más proactiva y una respuesta más eficaz); agregación y descripción utilizando "sus propias palabras"; análisis de la voz del cliente, empleado, ciudadano, etc.; gestión de ideas.
- **Organización de documentos:** Estructuración de colecciones de documentos y expedientes en función de los temas implícitos que emergen de forma natural de los propios contenidos y no de taxonomías externas
- **Aplicaciones de Seguridad:** Muchas empresas y gobiernos utilizan la Minería de Textos para el seguimiento y análisis de fuentes en línea de texto sin formato, como las noticias de Internet, blogs, etc. para fines de seguridad nacional. También está involucrado en el estudio del texto cifrado / descifrado.
- **Biomédicos:** Se refiere a la Minería de Texto aplicado a los textos y la literatura del dominio de la biología molecular y biomedicina. Es un campo de investigación bastante reciente en el borde del procesamiento del lenguaje natural, la bioinformática, la informática médica y la lingüística computacional.
- **Marketing:** Está empezando a utilizar en la comercialización, y más concretamente, en análisis de gestión de relaciones con clientes, se aplican para mejorar los modelos de análisis predictivo para la pérdida de clientes.

- **Aplicaciones académicas**: El tema de la Minería de Textos es de importancia para publicadores que tengan grandes bancos de datos que requieran de indexación. Esto es el caso en particular para disciplinas científicas en las que hay una gran cantidad de información muy específica en forma de texto escrito. (Clustering de Texto. (s.f.).

5.4.3 Beneficios del clustering de texto.

- Identificar “hechos” y datos puntuales a partir del texto de los documentos.
- Agrupar documentos similares (clustering).
- Determinar el tema o temas tratados en los documentos mediante la categorización automática de los textos.
- Identificar los conceptos tratados en los documentos y crear redes de conceptos.
- Facilitar el acceso a la información repartida entre los documentos de la colección, mediante la elaboración automática de resúmenes, y la visualización de las relaciones entre los conceptos tratados en la colección.
- Visualización y navegación de colecciones de texto.

5.5 Pre-procesamiento de datos.

El pre procesamiento y la limpieza de datos son tareas importantes que normalmente se deben llevar a cabo para que el conjunto de datos se pueda usar de forma eficaz para el aprendizaje automático, sin datos de calidad, no hay calidad en los resultados de la minería de datos.

¿Por qué pre procesar y limpiar datos?

Por el volumen de datos y la tasa de crecimiento de los datos no estructurados siendo superior al de los datos estructurados. Por ejemplo, twitter genera 12 Terabytes de información cada día, de acuerdo con Gartner (empresa consultora y de investigación de las tecnologías de la información con sede en Stamford, Connecticut, Estados Unidos), la tasa anual de crecimiento de datos es del 40 a 60 por ciento, pero para los datos no estructurados en empresas, la tasa de crecimiento puede llegar al 80 por ciento (Juan Vidal. 2014).

Se recopilan datos del mundo real de varios orígenes y procesos y pueden contener irregularidades o datos dañados que comprometen la calidad del conjunto de datos. Los problemas de calidad de datos más habituales que surgen son:

- **Incompletos**: en los datos no hay atributos o contienen valores faltantes.
- **Ruidosos**: los datos contienen registros erróneos o valores atípicos
- **Incoherentes**: los datos contienen discrepancias o registros en conflicto.
- eliminación de palabras vacías, son palabras que se consideran como no descriptivo dentro de un enfoque de bolsa de palabras. Ellas comprenden típicamente preposiciones, artículos, conectores, etc. (Hotho, A., Staab, S., & Stumme, G. 2003).

Los datos de calidad son un requisito previo para los modelos predictivos de calidad. Para evitar la "entrada y salida de elementos no utilizados" y mejorar la calidad de los datos y, por tanto, el rendimiento del modelo, es fundamental llevar a cabo una pantalla de mantenimiento de datos para

detectar problemas de datos al principio y decidir acerca de los pasos de limpieza y pre procesamiento de datos correspondientes (Santiago Cortez, S. D. 2016)

5.5.1 Importancia de la etapa de limpieza

- Asegura la calidad de los datos que vamos a procesar.
- Evita la información no veraz o errónea.
- Ahorra costes de espacio en disco al eliminarse la información duplicada.
- Agiliza las consultas por la ausencia de datos repetidos o inservibles.
- Ayuda a tomar decisiones estratégicas correctas.

5.5.2 Stop-words. Las stop-words son palabras comunes que no tienen significado relevante en un sistema de recuperación. Son una parte del lenguaje natural con el que un minero de texto se encontrará, Estas son generalmente palabras de alta frecuencia que no dan ninguna información adicional. La eliminación de stop words reduce la dimensionalidad del espacio de términos. Las stop-words o palabras vacías son palabras en documentos de texto como artículos, preposiciones, pro-sustantivos, etc. que no da el significado a los documentos, no se miden como palabras clave en aplicaciones de minería de texto (Vijayarani, S., Ilamathi, M. J., & Nithya, M. 2015).

5.5.3 Transformar un documento a valores numéricos. La transformación de un documento, que contiene palabras, a un vector de números con el cual los algoritmos pueden trabajar se realiza de la siguiente manera. Supongamos estos tres documentos (cada frase es un documento):

1. Juega al futbol.

2. Le gusta el baloncesto.

3. Mira un partido de futbol.

La siguiente matriz tiene como filas a los documentos, y una columna por cada palabra diferente que hay en el total de documentos (vocabulario). La idea es poner la frecuencia de la palabra en el documento para cada palabra. De esta manera el documento no1 tiene las palabras "juega", "al" "futbol" por lo que la fila uno tiene valor 1 para esas palabras y 0 para el resto, (Blanco-Hermida Sanz, E. J. 2016).

	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 2. Procesamiento de lenguaje. Modificado de: Blanco-Hermida Sanz, E. J,

5.6 Representaciones.

5.6.1 Frecuencias y pesos de los términos de un documento Este modelo entiende que los documentos pueden expresarse en función de unos vectores que recogen la frecuencia de aparición de los términos en los documentos. Los términos que forman esa matriz serían términos no vacíos, es decir, dotados de algún significado a la hora de recuperar información y, por otro lado, estarían almacenados en formato "stemmed" (reducidos los términos a una raíz común, tras un procedimiento de aislamiento de la base que agruparía en una misma entrada varios términos).

También es posible considerar una frecuencia de aparición de los términos mayor, para denotar un documento como más idóneo para resolver la consulta del usuario. Pero en todo caso, es necesario estimar cuál es el valor de cada término de cara a la recuperación, representación y discriminación de los contenidos en el corpus documental, ya que son muchos más factores los que entran en juego.

La ponderación de los términos es el proceso que tiene como finalidad conocer la importancia de los términos para representar un documento y permitir su posterior recuperación. Esto implica que se debe determinar el poder de resolución de los términos de la colección, o lo que es lo mismo, la capacidad de los términos para representar el contenido de los documentos en la colección, que permitan identificar cuáles son relevantes o no ante la consulta del usuario. Al valor e índice que es capaz de determinar este extremo se le denomina "peso del término" o "ponderación del término" y su cálculo implica determinar la "Frecuencia de aparición del término TF" y la "Frecuencia inversa del documento para un término IDF".

5.6.2 Factor TF: Frecuencia de Aparición de un Término. El factor TF es la suma de todas las ocurrencias o el número de veces que aparece un término en un documento. 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.

Tabla 4.
Características del factor TF

Factor TF	
Denominación	TF = Frecuencia de aparición del término
Descripción	Es la frecuencia de aparición de un término a lo largo de un documento. Dicho de otra forma, el número de veces que este se repite en el documento, lo que permite determinar su capacidad de representación.
Finalidad	Representativa
Casos	Frecuencia de aparición TF baja. Representatividad elevada. Frecuencia de aparición TF media. Frecuencia de aparición TF alta. Muy baja representatividad.

Nota: Frecuencias y pesos de los términos de un documento, adaptado de: Ochando M. 2013.

Su cálculo se efectúa una vez el texto del documento ha sido normalizado, según los procesos de depuración mencionados en artículos anteriores. Posteriormente se lleva a cabo el conteo de las veces que el término aparece presente en el documento.

$$Tf_{ij} = como\ la\ frecuencia\ del\ termino\ T_i\ en\ un\ documento\ d_j. \quad (6)$$

5.6.3 Factor IDF: Frecuencia Inversa del Documento para un Término. El factor IDF de un término es inversamente proporcional al número de documentos en los que aparece dicho término. 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.

Tabla 5.
Características del Factor IDF

	Factor IDF
Denominación	Inverse Document Frequency = Frecuencia Inversa del Documento para un término
Descripción	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.
Finalidad	Discriminatoria
Casos	Poder discriminatorio bajo. El término es genérico y aparece en la mayoría de los docs. Poder discriminatorio medio. Poder discriminatorio alto. El término es especializado y aparece en pocos docs.

Nota: Frecuencias y pesos de los términos de un documento adaptado de: Ochando M. 2014

$$idf(t) = log \frac{N}{df(t)} \quad (7)$$

N = Número total de documentos.

df(t) = Es la frecuencia de documentos que contienen el termino t.

Un ejemplo de aplicación de la fórmula del factor IDF, es la que se muestra a continuación:

Se muestra 4 documentos, donde después de hacer la respectiva limpieza, se representa la matriz de frecuencia.

D1: los planetas giran alrededor del sol.

D2: las agujas del reloj giran.

D3. Las peonzas giran al igual que giran los planetas.

D4: los planetas y el sol son astros.

Tabla 6.
Frecuencia de documentos.

	Planetas	**Giran**	**Alrededor**	**Sol**	**Agujas**	**Reloj**	**Peonzas**	**Igual**	**Astros**
D1	1	1	1	1	0	0	0	0	0
D2	0	1	0	0	1	1	0	0	0
D3	1	2	0	0	0	0	1	1	0
D4	1	0	0	1	0	0	0	0	1

Tabla 7.
Ejemplo de cálculo IDF

Cálculo IDF		
Término	**N**	**IDF**
Planetas		0.124
Giran		0.124
Alrededor		0.602
Sol		0.301
Agujas	4	0.602
Reloj		0.602
Peonzas		0.602
Igual		0.602
Astros		0.602

5.6.4 Peso o Ponderación TF-IDF. El peso de un término en un documento es el producto de su frecuencia de aparición en dicho documento (TF) y su frecuencia inversa de documento (IDF)

$$\mathrm{TF}-\mathrm{IDF}(\mathrm{td}) = \mathrm{TF} * \mathrm{IDF}(\mathrm{td}) \quad (8)$$

Tabla 8.
Ejemplo de cálculo TF-IDF

CALCULO DEL PESO TF-IDF				
Término	**D1**	**D2**	**D3**	**D4**
Planetas	0.124	0	0.124	0.124
Giran	0.124	0.124	0.248	0
Alrededor	0.602	0	0	0
Sol	0.301	0	0	0.301
Agujas	0	0.602	0	0
Reloj	0	0.602	0	0
Peonzas	0	0	0.602	0
Igual	0	0	0.602	0
Astros	0	0	0	0.602

Los pesos obtenidos son denotativos de la importancia del término en cada documento y servirá a la postre para calcular otros valores indispensables para la recuperación de información en los distintos modelos booleano, vectorial y probabilístico (Ochando, D. M. 2013).

5.7 Stemming

El stemming es una representación para reducir una palabra a su raíz. Stemming aumenta la sensibilidad la cual es una medida sobre el número de documentos que se pueden encontrar con una consulta.

Los métodos de stemming intentan construir las formas básicas de las palabras, es decir, quitar el plural de Sustantivos, de los verbos, u otros afijos. Un stem es un grupo natural de palabras con

igual (o muy similar) significado. Después del proceso de derivación, cada palabra está representada por su raíz.

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 devolverá documentos que contengan además de "biblioteca" los que tienen "bibliotecario" (Hotho, A., Nürnberger, A., & Paaß, G. 2005, May).

Stemming ha sido ampliamente utilizado como una técnica fundamental en la recuperación de la información, especialmente en el contexto de los motores de búsqueda web, en la minería de texto y el análisis del sentimiento como una técnica de pre procesamiento, y para las bases de datos cuando la construcción de grandes índices en los documentos. Las razones principales son la reducción de la memoria y las demandas informáticas que son necesarias para manejar grandes cantidades de Datos. Es así como las palabras originales de los stems sólo se pueden restaurar de un diccionario. Esto es Especialmente relevante para grandes cantidades de textos cortos (como Twitter), (Feinerer, I. 2010, December).

Tabla 9.
Terminaciones y stems

C	terminaciones	Word Stems	S
C1	Experimental		
C2	Experiment	experi	S1
C3	Experience		
C4	Experiences		
C5	Adhere	Adher	S2

Nota: adaptado de Frecuencias y pesos de los términos de un documento: Ochando M. 2013

5.8 K-means

Dentro de los algoritmos principales se encuentran el algoritmo K-means que fue utilizado por primera vez por James (MacQueen, J. 1967, June). En 1967 algoritmo de aprendizaje sin supervisión que resuelve el conocido problema de agrupación. El procedimiento sigue una manera sencilla y fácil de clasificar un determinado conjunto de datos a través de un cierto número de clusters.

El nombre de K-means viene porque representa cada uno de los clusters por la media (o media ponderada) de sus puntos, es decir, por su centroide. La representación mediante centroides tiene la ventaja de que tiene un significado gráfico y estadístico inmediato. Cada cluster por tanto es caracterizado por su centro o centroide, que se encuentra en el centro o el medio de los elementos que componen el cluster. Kmeans es traducido como K-medias y se realiza en 4 etapas. (Cambronero, C. G., & Moreno, I. G. 2006),

Paso 1: Elegir aleatoriamente K objetos que forman así los K clusters iniciales. Para cada cluster k, el valor inicial del centro es = xi, con la xi únicos objetos de Dn pertenecientes al cluster.

Paso 2: Reasigna los objetos del cluster. Para cada objeto x, el prototipo que se le asigna es el que es más próximo al objeto, según una medida de distancia, (habitualmente la medida euclidiana).

Paso 3: Una vez que todos los objetos son colocados, recalcular los centros de K cluster. (los baricentros)

Paso 4: Repetir los pasos 2 y 3 hasta que se alcance un determinado criterio de parada.

Figura 3. Pasos del algoritmo Kmeans. Modificado de: Larranaga, P., Inza, I., & Moujahid, A. Tema 14. Clustering.

Tal y como puede verse en la Figura 3, en el algoritmo propuesto por McQueen se comienza considerando los k primeros elementos del fichero de casos como los k centroides iniciales, o dicho de forma equivalente como conglomerados con un único elemento. A continuación, y siguiendo el orden establecido en la figura 3, cada uno de los objetos se va asignando al conglomerado con centroide más próximo, con la característica que al efectuar cada asignación se recalculan las coordenadas del nuevo centroide.

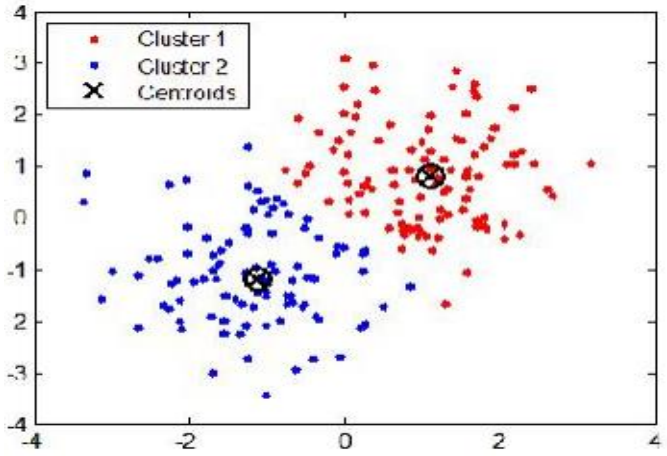

Figura 4. Representación del agrupamiento del K-MEANS. Modificado de: Cambronero, C. G., & Moreno, I. G, 2006.

5.8.1 Variaciones del algoritmo K-means. Estas variaciones ofrecidas por el lenguaje de programación R son funciones que realiza el proceso K-Means, de acuerdo con diferentes algoritmos. Estos algoritmos se describen a continuación:

5.8.1.1 Algoritmo Lloyd/Forgy. Dado un conjunto de k centros Z, para cada centro z en Z, V(z) denote su vecindad. Ese es el conjunto de puntos de datos para los cuales z es el vecino más cercano. Cada etapa del algoritmo de Lloyd/Forgy mueve cada punto central z al centroide de V (z) y luego actualiza V (z) recalculando la distancia de cada punto a su centro más cercano. Estos pasos se repiten hasta la convergencia. Tenga en cuenta que el algoritmo de Lloyd/Forgy puede

quedar atrapado en soluciones localmente mínimas que están lejos de lo óptimo. Por esta razón es común considerar la heurística basada en la búsqueda local, en la cual los centros se intercambian dentro y fuera de una solución existente (generalmente al azar). Este intercambio sólo se acepta si disminuye la distorsión media, de lo contrario se ignora.

5.8.1.2 Algoritmo de Hartigan. Dado n objetos con p variables medidos en cada objeto x (i, j) para i = 1,2, ..., n; J = 1,2, ..., p; K-means asigna cada objeto a uno de los K grupos o clusters para minimizar la suma de cuadrados dentro del clúster:

$$Sum(k) = \sum_{i=0}^{n} \sum_{j=0}^{p} (x(i,j) - x(k,j))^2 \quad (9)$$

Donde x (k, j) Es la variable media j de todos los elementos del grupo K.

Además de la matriz de datos, se requiere una matriz K x p que dé los centros de agrupamiento inicial para los grupos de K. A continuación, los objetos se asignan inicialmente al clúster con la media de clúster más cercana. Dada la asignación inicial, el procedimiento consiste en buscar iterativamente la partición K con una suma de cuadrados localmente óptima dentro de un grupo de puntos móviles de un grupo a otro.

5.8.1.3 Algoritmo de MacQueen. Este algoritmo trabaja moviendo repetidamente todos los centros de agrupación a la media de sus respectivos conjuntos de Voronoi que se basa fundamentalmente en la proximidad. (Dasgupta, S. 2013)

5.8.2 Ejemplo 2 K-means numérico con dos variables. Supóngase que se tiene cuatro tipos de medicina con dos atributos (peso e índice ph) como se muestra en la siguiente tabla.

Tabla 10.
Muestra de cuatro tipos de medicina con dos atributos

MEDICINA	PESO (gramos)	INDICE PH
A	1	1
B	2	1
C	4	3
D	5	4

El número de grupos k es k=2

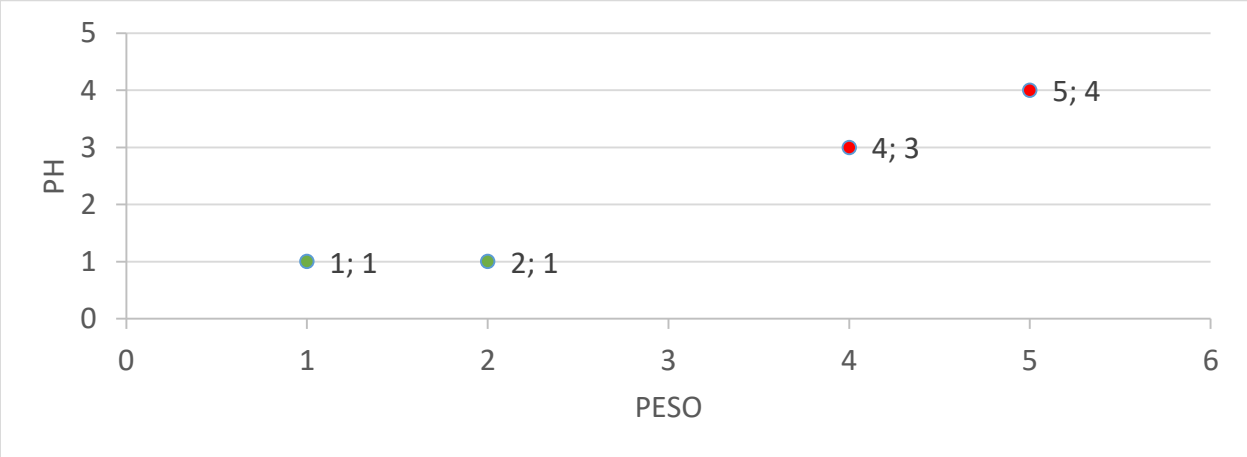

Figura 5. Representación gráfica para determinar el centroide

Sean C1 y C2 los centroides de los k-grupos. Si se eligen las medicinas A y B como los primeros centroides que quedan en los puntos (1,1) y (2,1).

Se calcula la distancia de cada objeto a los centroides utilizando la distancia euclidiana:

$$D(i,j) = (\sum_{j=1}^{i}(Xij - X(i)j)^2)\text{^}(1/2)) \quad (10)$$

Cálculo de Distancias

D (C, C1) = $\sqrt{((4-1)^2 + (3-1)^2)} = 3{,}61$

D (D, C1) = $\sqrt{((5-1)^2 + (4-1)^2)} = 5$

D(C, C2) = $\sqrt{((4-2)^2 + (3-1)^2)} = 2{,}83$

D (D, C2) = $\sqrt{((5-2)^2 + (4-1)^2)} = 4{,}24$

D (C2, C1) = $\sqrt{((2-1)^2 + (1-1)^2)} = 1$

Matriz

$$D_0 = \begin{bmatrix} 0.00 & 1.00 & 3.61 & 5.00 \\ 1.00 & 0.00 & 2.83 & 4.24 \end{bmatrix}$$

Se asigna entonces cada objeto a cada grupo teniendo en cuenta el mínimo error de participación de cada elemento en cada grupo.

$$G_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}$$

La medicina A se asigna al grupo 1 con centroide en (1,1) y las medicinas B, C, D al grupo 2 con centroide en (2,1).

Primera Iteración:

Conociendo los miembros de los grupos se vuelve a calcular el centroide de cada grupo y la matriz de distancia, como el grupo 1 tiene un solo miembro el centroide de ese grupo permanece igual, el grupo 2 tiene ahora 3 miembros por tal motivo se procede a calcular un nuevo centroide.

$$C2 = \left(\frac{2+4+5}{3}\ ,\ \frac{1+3+4}{3}\right) = \ (3{,}67;\ 2{,}67)$$

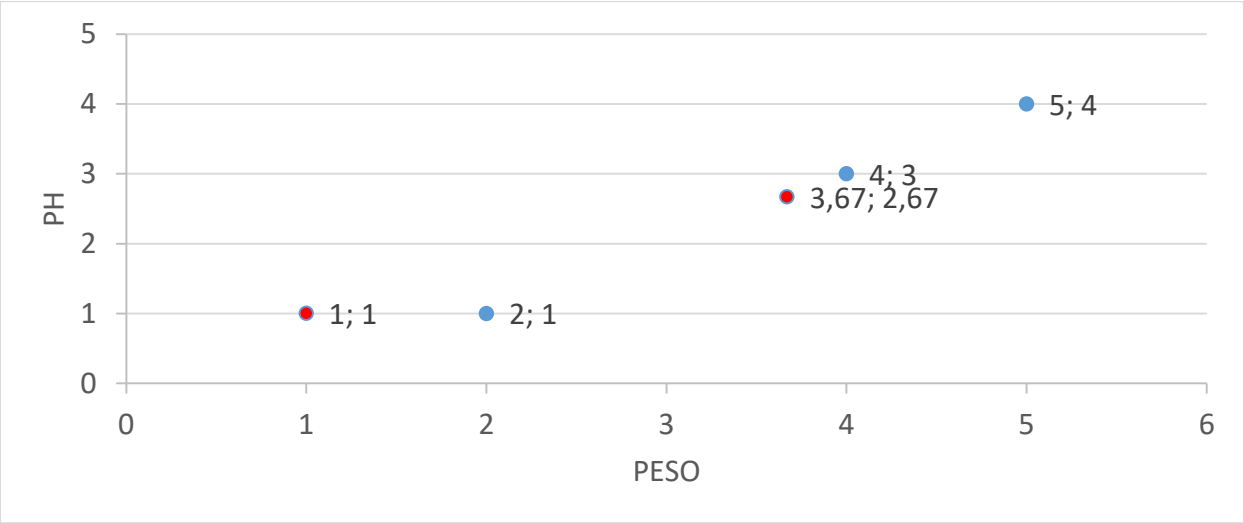

Figura 6. Representación gráfica con los nuevos centroides.

Se recalculan las distancias con el nuevo centroide

$D(C,C1) = \sqrt{((3,67-1)^2 + (2,67-1)^2)} = 3,14$

$D(C,C2) = \sqrt{((3,67-2)^2 + (2,67-1)^2)} = 2,36$

$D(D,C1) = \sqrt{((3,67-4)^2 + (2,67-3)^2)} = 0,47$

$D(D,C2) = \sqrt{((3,67-5)^2 + (2,67-4)^2)} = 1,89$

Matriz

$$D_1 = \begin{bmatrix} 0.00 & 1.00 & 3.61 & 5.00 \\ 3.14 & 2.36 & 0.47 & 1.89 \end{bmatrix}$$

Se vuelve asignar cada objeto a cada grupo teniendo en cuenta el mínimo error de participación de cada elemento en cada grupo

$$G_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Segunda iteración:

Se recalculan de nuevo los centroides.

$C1 = \left(\frac{1+2}{2} + \frac{1+1}{2}\right) = (1{,}5\ ;\ 1)$

$C2 = \left(\frac{4+5}{2} + \frac{3+4}{2}\right) = (4{,}5\ ;\ 3{,}5)$

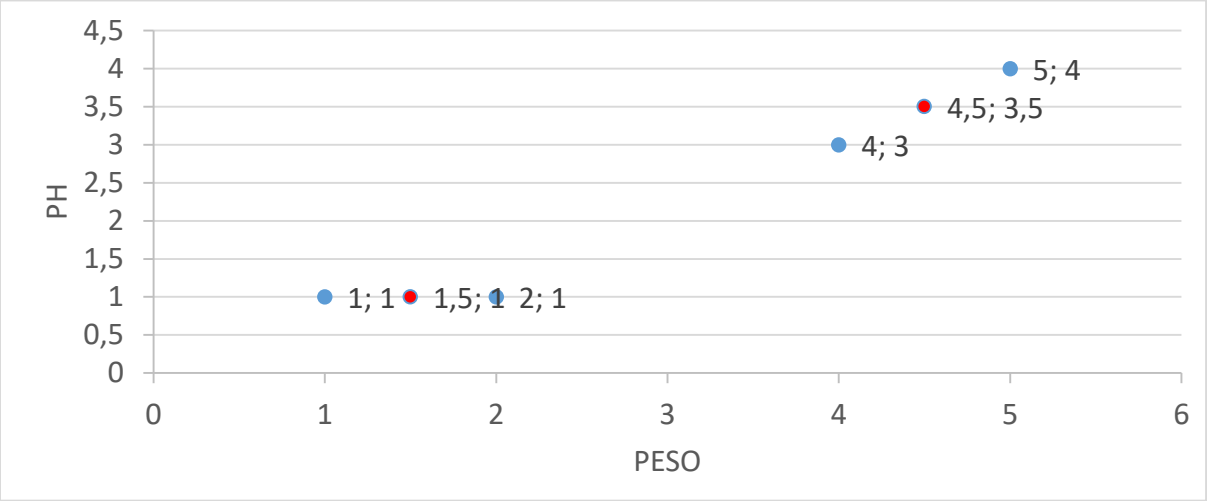

Figura 7. Representación gráfica de centroides.

Se recalculan las distancias con los nuevos centroides.

D (C1, A) = $\sqrt{((1{,}5-1)^2 + (1-1)^2)} = 0{,}5$

D (C2, B) = $\sqrt{((4{,}5-1)^2 + (3{,}5-1)^2)} = 4{,}3$

D (C1, B) = $\sqrt{((1{,}5-2)^2 + (1-1)^2)} = 0{,}5$

D (C2, B) = $\sqrt{((4{,}5-2)^2 + (3{,}5-1)^2)} = 3{,}5$

D (C1, C) = $\sqrt{((1{,}5-4)^2 + (1-3)^2)} = 3{,}20$

D (C2, C) = $\sqrt{((4{,}5-4)^2 + (3{,}5-3)^2)} = 0{,}71$

D (C1, D) = $\sqrt{((1{,}5-5)^2 + (1-4)^2)} = 4{,}61$

D (C2, D) = $\sqrt{((4{,}5-4)^2 + (3{,}5-3)^2)} = 0{,}71$

Matriz

$$D_2 = \begin{bmatrix} 0.50 & 0.50 & 3.20 & 4.61 \\ 4.30 & 3.54 & 0.71 & 0.71 \end{bmatrix}$$

Se vuelve asignar cada objeto a cada grupo teniendo en cuenta el mínimo error de participación de cada elemento en cada grupo.

$$G_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

El cálculo de centroides se ha estabilizado y no es necesario continuar iterando

5.8.3 Medidas de distancia

5.8.3.1 Distancia euclidiana. La distancia euclidiana calcula la raíz de la diferencia cuadrática Entre coordenadas de un par de puntos u objetos.

$$\mathrm{Dist}_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2} \qquad (11)$$

***5.8.3.2 Distancia Manhattan*.** La distancia entre dos puntos es la suma de las diferencias (absolutas) de sus coordenadas.

$$\mathrm{Dist}_{XY} = |X_{ik} - X_{jk}| \qquad (12)$$

5.8.3.3 Distancia máximum. Se calcula como la magnitud absoluta de la diferencia entre coordenadas de un par de objetos.

$$\mathrm{Dist}_{XY} = max_k |X_{ik} - X_{jk}| \qquad (13)$$

5.8.3.4 Distancia Minkowski. Distancia de Minkowski es la distancia métrica generalizada.

Tenga en cuenta que cuando p = 2, la distancia se convierte en la distancia euclidiana. Cuando p = 1 se convierte en distancia de manhattan. La distancia máximum es una variante de la distancia Minkowski donde P = ∞ (tomando un límite). Esta distancia se puede usar tanto para variables ordinales y cuantitativas.

$$\mathrm{Dist_{XY}} = \left(\sum_{K=1}^{d} |\mathrm{X_{ik}} - \mathrm{X_{jk}}|^{\frac{1}{p}}\right)^{p} \quad (14)$$

Las métricas de distancia se utilizan para encontrar objetos de datos similares que conducen a desarrollar algoritmos robustos para la minería de datos funcionalidades como la clasificación y la agrupación.

Los algoritmos necesitan asumir que el número de grupos (Clusters) se conoce a priori. Un paso importante en la agrupación es seleccionar una métrica de distancia, que determinará cómo es la similitud de dos elementos

5.8.4 Elbow Method. Este método utiliza los valores de la inercia obtenidos tras aplicar el K-means a diferente número 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:

$$\boldsymbol{Inercia} = \sum_{\boldsymbol{i=0}}^{\boldsymbol{N}} \|\boldsymbol{xi} - \boldsymbol{u}\|^2 \quad (15)$$

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 esa data set; o, dicho de otra manera: el punto que representaría al codo del brazo será el número óptimo de Clusters para ese conjunto de datos.

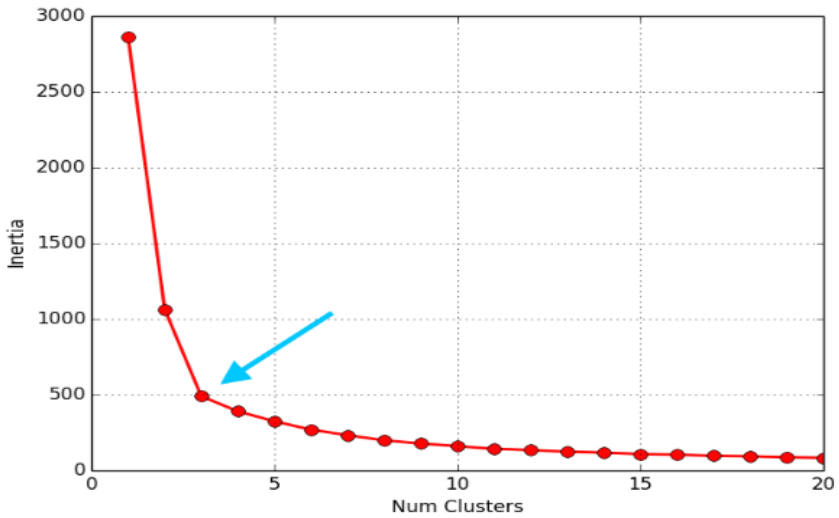

Figura 8. Cantidad de Óptima de Clúster. Modificado de: Ricardo Moya, (2016).

5.8.5 Ejemplo 2 Kmeans numérico en R. Si se quiere optimizar una segmentación de clientes con k-mean, una posible técnica es comparar las 4 variaciones del algoritmo incluidos en R para k-means y elegir la que tenga mejores resultados para los datos analizados. A esta técnica le dicen validación cruzada de métodos.

Las 4 variaciones del algoritmo son: Lloyd, Forgy, MacQueen y Hartigan-Wong. Para compararlas puede usarse la "distancia intracluster", que es la suma de las distancias entre los centroides. El algoritmo que tenga la mayor "distancia intracluster" sería el ganador ya que sería la mejor separación de grupos.

Las diferencias entre los 4 algoritmos se centran en la forma de elegir los centroides iniciales, y la forma que usan para la asignación de cada individuo al nuevo cluster

Tabla 11.
Clientes.

	NOMBRE	EDAD	MONTO CONSUMO
1	JUAN	55	19
2	PEDRO	30	56
3	MARIA	80	11
4	ISABEL	78	57
5	DIEGO	98	34
6	LUIS	87	72
7	LUCIA	46	42
8	FRANCISCA	38	43
9	ALBERTO	49	23
10	GARCIA	54	98
11	SOTO	92	17
12	VICTOR	28	24
13	ESTEBAN	19	12
14	JOSE	29	48
15	BETO	46	56

Nota: ejemplo para una segmentación de clientes. Adaptado de Santana E. (2014)

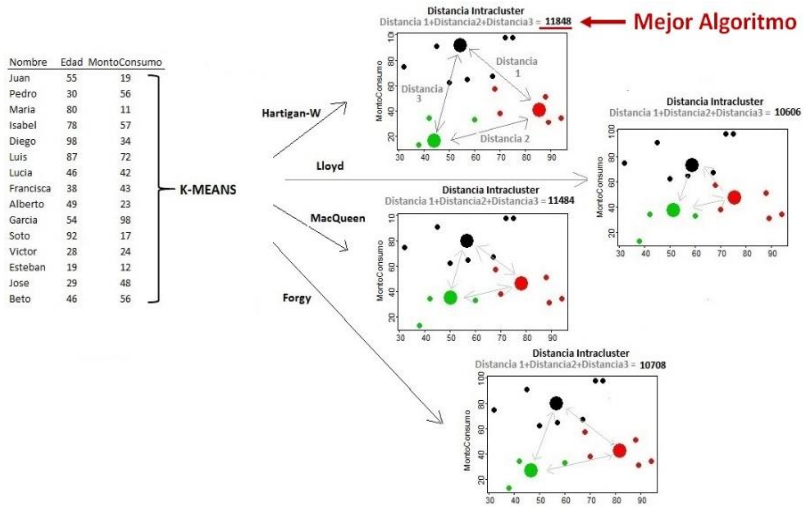

Figura 9. "Comparación de algoritmos intracluster". Modificado de: Santana E. 2014.

El resultado sería una segmentación realizada por el algoritmo ganador, para este caso es "Hartigan-Wong", quedando de la distribución así:

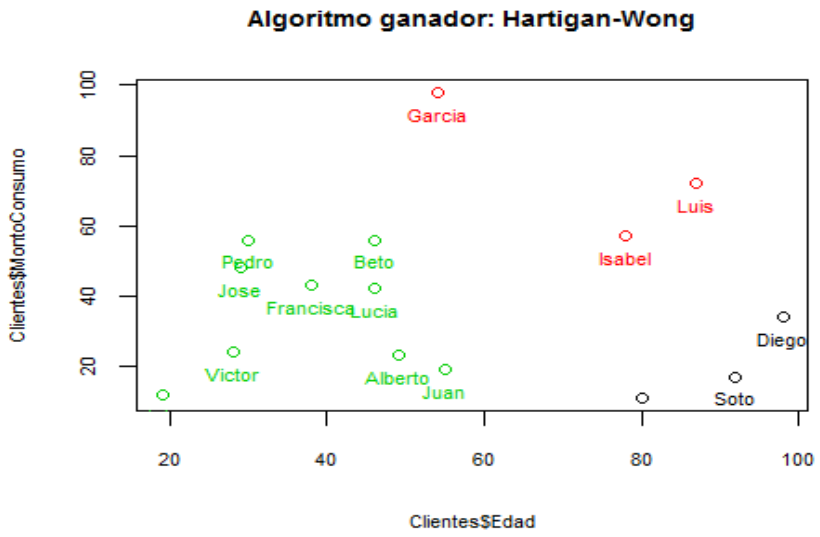

Figura 10. "Segmentación realizada por el algoritmo ganador". Modificado de: Santana E. 2014

5.8.6 Ejemplo 3 Tf-Idf y K-means en R. En la siguiente documentación, desglosaremos cuatro pequeños documentos para ilustrar la idea.

D1: The sky is blue

D2: The sun is bright today

D3: The sun in the sky is bright

D4: We can see the shining sun, the bright sun.

La primera parte de la fórmula tf (t, d) es simplemente calcular el número de veces que cada palabra apareció en cada documento. Por supuesto, al igual que con los métodos de minería de

texto comunes: stop- words, los signos de puntuación se eliminarán de antemano y las palabras se convertirán en casos más bajos, y así realizar la matriz de frecuencia.

Muchos conjuntos de datos de minería de texto existentes están en forma de una clase DocumentTermMatrix (del paquete tm), al crear un vector muestra la inspección y matriz Tf y Tf-Idf de los datos.

Seguidamente se aplica el método Elbow (del paquete factoextra), la cual ayudará a escoger el número de grupos óptimo con el que divide el conjunto de muestras, así pues, el valor de K óptimo será aquel en el que el gráfico dibuja un pico o cambio brusco en la evolución.

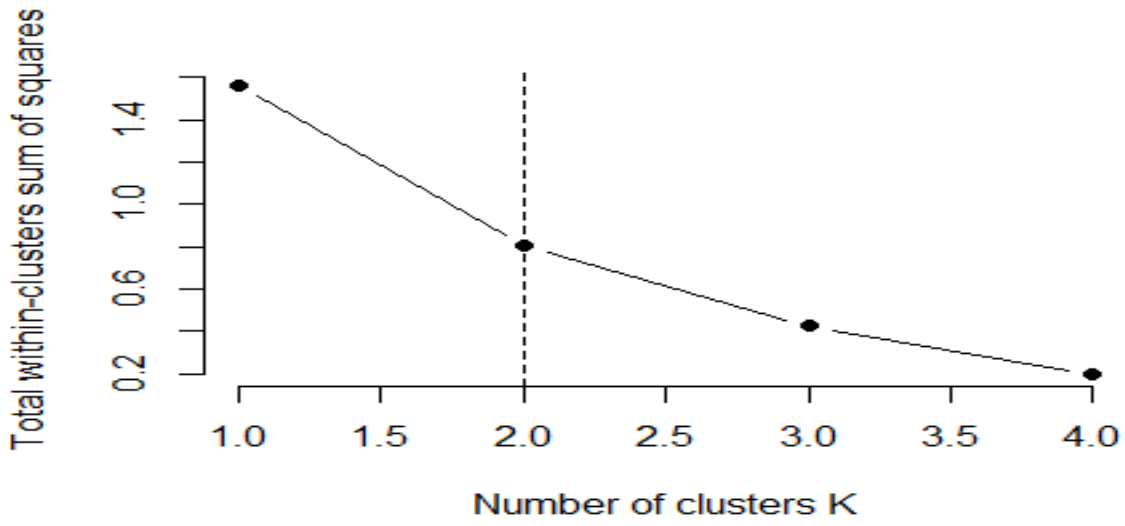

Figura 11. Cantidad de Óptima de Clúster

Por último, se halla el K-means (del paquete tm), el cual consiste en dividir el espacio en el que se representan los datos en K grupos. A cada muestra se le asigna el grupo más cercano mediante el cálculo de la distancia euclídea, entre la muestra y el centro de gravedad de cada cluster.

6. Resultados

Con el fin de evaluar las diferentes combinaciones de las variantes del algoritmo k-means, (Hartigan, Lloyd, Forgy y MacQueen) y las distancias, (Euclidiana, Maximum, Manhattan, Minkowski); se propone evaluar estas con una base de datos del benchmark obtenidas de un conjunto de entrenamiento para la recopilación de spam de SMS recopilados para la investigación de spam de teléfonos móviles, la cual contiene una colección compuesta por 1630 mensajes en inglés, etiquetados según su legitimidad de spam o ham, sea spam mensajes no deseados y ham mensajes significativos.

Proporción Ham y Spam

Ham = 1423 = 0.87

Spam = 207 = 0.12

Siendo ham positivo (2) y spam negativo (1).

6.1 Prueba de datos.

De acuerdo al conjunto de datos que contiene etiquetas se quiso verificar que tanto coincide la clasificación natural (obtenida por el clustering) y la real (dada si es spam o ham), lo que en realidad se sacó de este análisis fue cual se comportó mejor donde una de las medidas fue el Accuracy que se planteó soportado además con la suma de cuadrado entre clusters Betweenss.

6.1.1 Benchmarking de algoritmos y distancias mediante las medidas de desempeño.

Tabla 12.
Métrica Accuracy.

	MEDIDAS DE DISTANCIA					
Algoritmos	Euclidiana	Maximum	Manhattan	Minkowski (p=3)	Minkowski (p=4)	Minkowski (p=5)
Hartigan	0.7037	0.7515	0.6485	0.7411	0.746	0.7479
Lloyd	0.6982	0.7509	0.6485	0.7374	0.7442	0.7472
Forgy	0.6982	0.7509	0.6485	0.7374	0.7442	0.7472
MacQueen	0.7037	0.7509	0.6485	0.7411	0.746	0.7472

Tabla 13.
Métrica F-measure

	MEDIDAS DE DISTANCIA					
Algoritmos	Euclidiana	Maximum	Manhattan	Minkowski (p=3)	Minkowski (p=4)	Minkowski (p=5)
Hartigan	0.8256947	0.858044	0.778849	0.8511988	0.8544303	0.85563
Lloyd	0.8218682	0.857643	0.778849	0.8487632	0.8532206	0.85523
Forgy	0.8218682	0.857643	0.778849	0.8487632	0.8532206	0.85523
MacQueen	0.8256947	0.857643	0.778849	0.8511988	0.8544303	0.85523

En la tabla 12 se observa que mediante la métrica Accuracy, que indica el porcentaje de acierto del clasificador; en los algoritmos no se ve una diferencia significativa que por lo contrario en las distancias si se observa que hay diferencias, donde la distancia máximum es la mejor ya que arrojó el mejor porcentaje de acierto del clasificador siendo la más alta con respecto a las otras.

El F-measure es una medida de exactitud de una prueba, que considera tanto la precisión como el recall, en la tabla 13 muestra una relación con Accuracy que indica que si el porcentaje es alto el F-measure también, pero no todo se queda en los verdaderos negativos si no que muestra un

porcentaje de precisión del 0.8580 la cual no es tan malo ya que un puntaje F-measure alcanza su mejor valor en 1 y el peor en 0.

Es importante aclarar que estas medidas de desempeño no son buenas con relación a la clasificación no supervisada ya que la proporción de spam da mayor que la clasificación supervisada, y que lo que se busca en realidad es ver que tan parecidas son, sacando como conclusión que es mejor utilizar una clasificación trivial.

6.1.2 Benchmarking de algoritmos y distancias en la suma de cuadrados entre cada cluster.

En la tabla 14 muestra la comparación de las 4 variaciones del algoritmo Lloyd, Forgy, MacQueen y Hartigan-Wong con las diferentes distancias, donde para compararlas se usa la "distancia intracluster", que es la suma de las distancias entre los centroides. El algoritmo que tenga la mayor "distancia entrecluster" sería el ganador ya que sería la mejor separación de grupos, la cual los resultados muestran que la mejor separación de grupos es 2582072 con el algoritmo Hartigan y distancia Manhattan.

Tabla 14.
Betwweenss, suma de cuadrados entre clúster.

Algoritmos	**Euclidiana**	**Maximum**	**Manhattan**	**Minkowski (p=3)**	**Minkowski (p=4)**	**Minkowski (p=5)**
Hartigan	2002477	2115367	2582072	2029510	2062207	2080264
Lloyd	2000520	2115362	2581996	2025578	2062207	2080262
Forgy	2002249	2117513	2580142	2029179	2062067	2080262
MacQueen	2002477	2115362	2582072	2029510	2062207	2080262

7. Procesamiento de lenguaje natural (caso de estudio)

7.1 Datos de twitter

Twitter es una popular red social fundada en San Francisco en 2006. Su principal característica es que los mensajes tienen un límite de 140 caracteres, son los llamados "tweets". Se estima que tiene más de 800 millones de usuarios en todo el mundo y que genera 65 millones de tweets al día (Shiels, M. 2011).

Twitter realiza algunas acciones simples sobre los mensajes recibidos en nuestra línea de tiempo, tales como reenviar el mensaje de otro (Retweets), responder/citar a alguien o escribir un mensaje privado todo esto para darle más visibilidad a cada tweet, el motivo por los cuales se eligió Twitter sobre otra red social es muy sencillo: es una red social donde el contenido es público y accesible. Además, ofrece una API sencilla y amigable de utilizar y twitter la pone a disposición permitiendo hacer uso de su servicio Web para acceder a los datos de su servidor. Se puede hacer búsquedas, crear nuevos tweets, recuperar los tweets de un usuario, etc, permitiendo que la interfaz facilite la interacción humano-software.

Empleando la técnica de minería de texto se hará un análisis descriptivo de un conjunto de datos obtenidos de la red social Twitter y gracias al API que permite incursionar en el núcleo de twitter, se obtuvo 3160 tweets de meses comprendidos entre el 17/01/2017 al 17/03/2017. Del usuario del Periódico de noticias de Bucaramanga, Santander y el mundo - Vanguardia Liberal. **“vanguardiacom”,** con 241 mil seguidores**.**

Para recuperar los documentos de la base de datos propuesta, uno de los métodos más elementales fue comprobar la presencia o ausencia de las palabras que forman la cadena de consulta del usuario en cada documento, la ponderación de los términos y el proceso que tiene como finalidad conocer la importancia de los términos para representar un documento y permitir su posterior recuperación a esta forma de representar los tweets se identifica como tf o frecuencia de términos.

Sin embargo, aquellas características que fueron muy frecuentes en el conjunto de todo el tweet hacen que no sean tan relevantes al momento de identificarlos. Por lo tanto, se incorporó un factor de frecuencia inversa de documento que atenúa el peso de las características que ocurren con mucha frecuencia en la colección de tweets e incrementa el peso de las características que ocurren pocas veces.

7.2 Elbow Method

Se determina el número de clusters en el conjunto de datos ya procesados de twitter del usuario “vanguardiacom” por la regla Elbow Method.

Se aplicó clustering a la base de datos de “vanguardiacom” de twitter, para ver qué tipo de información se podría obtener. Se utilizó la distancia manhattan y la variación del algoritmo K-means Hartigan clasificada como la mejor de acuerdo al benchmarking, enseguida se ejecutó el Elbow Method punto en el que se observa en la figura 12 ese cambio brusco en la inercia que dirá el número óptimo de Clusters a seleccionar para esta base de datos, donde arrojó un resultado de

4 clusters, veremos los términos de cada uno de ellos para hacernos una idea de que tema es más relevante en cada uno de los grupos.

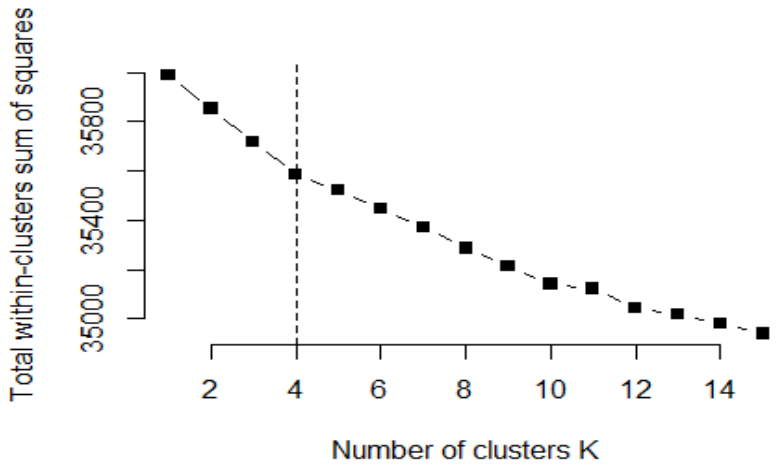

Figura 12. Cantidad Óptima de Clúster de los datos procesados.

7.3 Análisis General

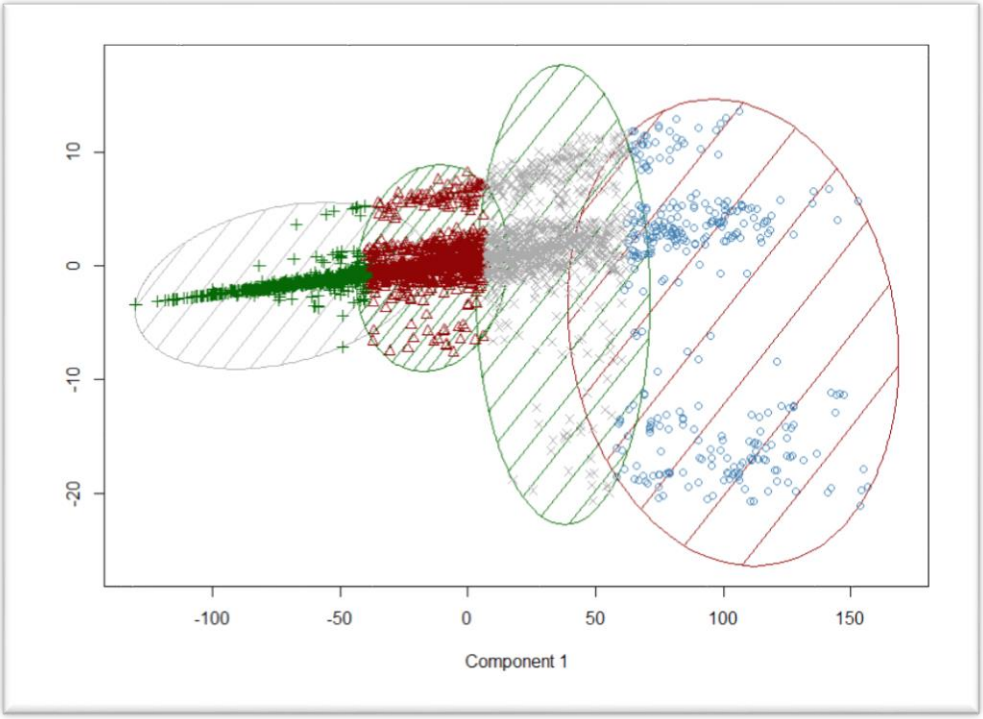

Figura 13. PCA Documento

Se tiene una segmentación de k-means con 11159 variables, donde se representó en un gráfico de dispersión y la cual se utilizó una técnica llamada "*Análisis de Componentes Principales (PCA)*" para resumir todas las variables en solo 2, que representaron el componente 1 y 2, asignando un color diferente y símbolo a cada cluster.

La distribución de los tweets en cada uno de los cluster, es la siguiente:

Tabla 15.
Distribución de Tweets.

clusters	cantidad
1	725
2	393
3	1116
4	926

Esta reducción de dimensionalidad busca la proyección según la cual los datos queden mejor representados en términos de mínimos cuadrados, en la figura 13 muestra una variabilidad del 83.37% indicando con esto que del 100 % de información se está perdiendo un 16.63%, pero de igual forma es una variabilidad muy buena para la cantidad de variables reducidas.

En la figura 13 se observa el comportamiento del método k-means, viendo que entre grupos no hay documentos en común o sea no se traslapan; de igual forma se llega a un análisis donde se puede ver que en el cluster de color rojo y gris no hay documentos dispersos lo que quiere decir que en esos clusters puede haber temas muy relacionados entre sí, para el cluster de color verde se evidencia que hay poca dispersión entre ellos pero aún puede seguir teniendo temas muy relacionados y por último el cluster de color azul muestra una dispersión importante queriendo

decir que hay una variedad de temas y que no están muy relacionados, también se puede ver en la tabla 15 que el cluster 2 tiene una baja cantidad de documentos y por eso las frecuencias para éste son bajas en comparación con los otros.

Es importante decir que se encontró similitudes entre los clusters de color rojo y verde, gris y azul porque pueden existir temas muy comunes entre ellos; por lo contrario, el color verde y el azul, rojo y gris, son muy diferentes teniendo probablemente distintos temas para el análisis.

7.4 Análisis de histograma

Twitter trabaja como una herramienta que tiene la capacidad de proveer grandes volúmenes de opiniones claves y tendencias para la evaluación e información durante un evento, por tal motivo se muestra en el siguiente ejemplo la búsqueda realizada en los dos meses, al usuario vanguardiacom en los últimos 3160 tweets, sacando información de frecuencia tanto de Retweets como Favoritos.

Los retweets permiten reenviar un tweet lanzado por otro usuario que generalmente se le sigue, dicho mensaje se visualiza en la line de tiempo, se torna muy interesante y se busca que todos sus seguidores de twitter tengan acceso a él. En la figura 14 se observa que más de 700 tweets no tuvieron retweets, más de 800 tweets tuvieron al menos 1 retweet, también se puede evidenciar que hay dos tendencias importantes en tweets marcados con 98 y147 retweets los cuales fueron:

“Jorge Robledo denuncia que Banco Agrario hizo préstamos irregulares a Odebrecht”

"Los sobrevivientes del Chapecoense alzaron la Copa Sudamericana"

Una de las características más interesantes de **Twitter**, y probablemente de las menos usadas, son los **Favoritos**. Al igual cómo los mensajes que envías a Twitter terminan definiendo quien eres, los mensajes que marcas como favoritos tienen un efecto similar. En determinado momento puede ser mayor la impresión que das con estos mensajes, muchas veces seleccionamos los mejores tweets basados en nuestros intereses personales para recordar momentos o eventos especiales. Todo mensaje que marcas como favorito es público y puede ser visto en tu perfil, en la figura 15 muestra la frecuencia de favoritos que hubo en los dos meses teniendo relevancia un tweet marcado con 342 favoritos el cual fue, "Los sobrevivientes del Chapecoense alzaron la Copa Sudamericana" recordando la triste tragedia del avión de Chapecoense.

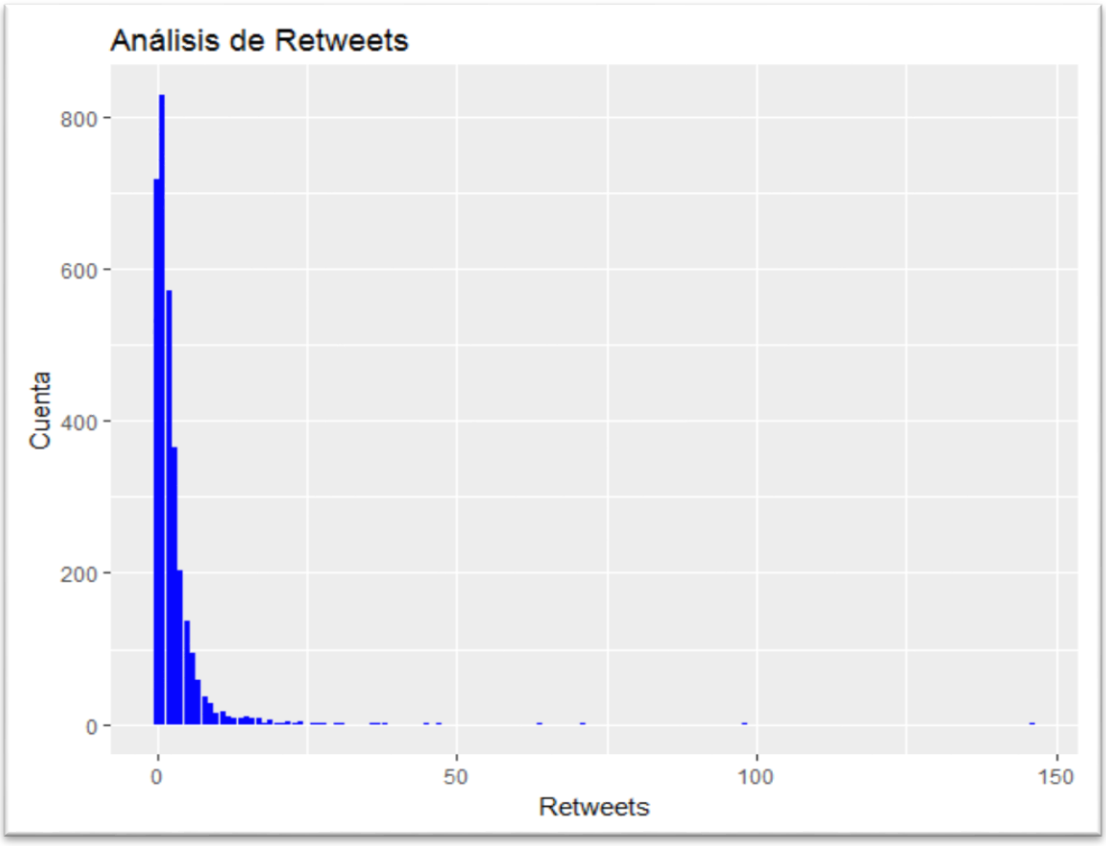

Figura 14. Análisis de Retweets.

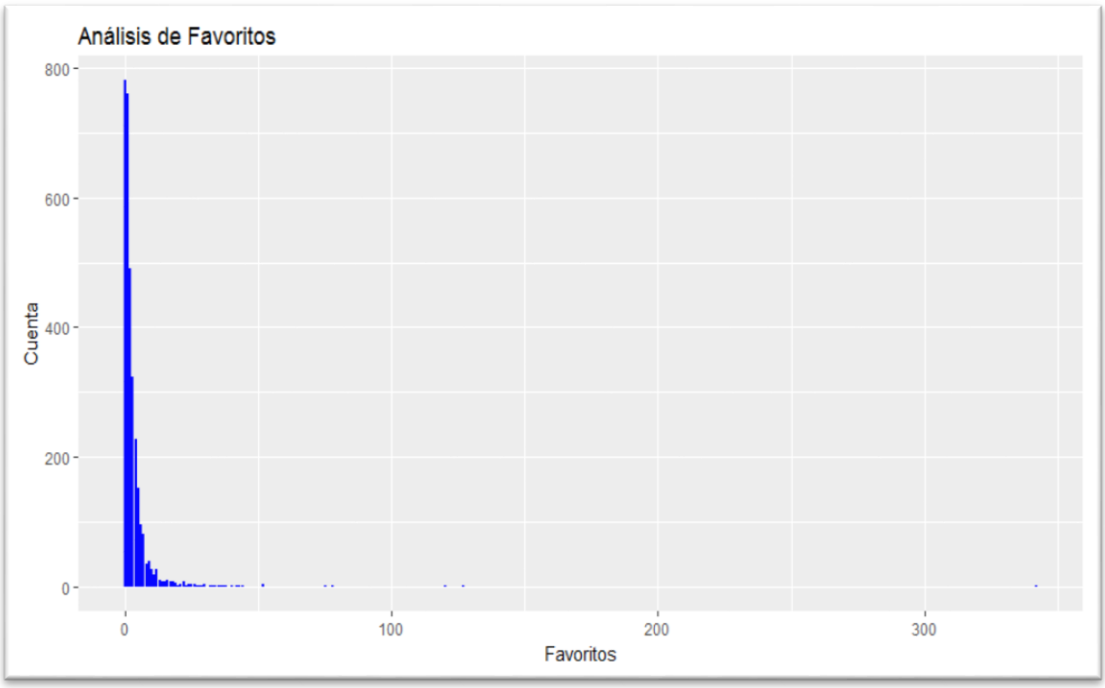

Figura 15. Análisis de Favoritos.

7.5 Análisis de Tweets

7.5.1 Cluster 1

Figura 16. Nube de palabras cluster 1.

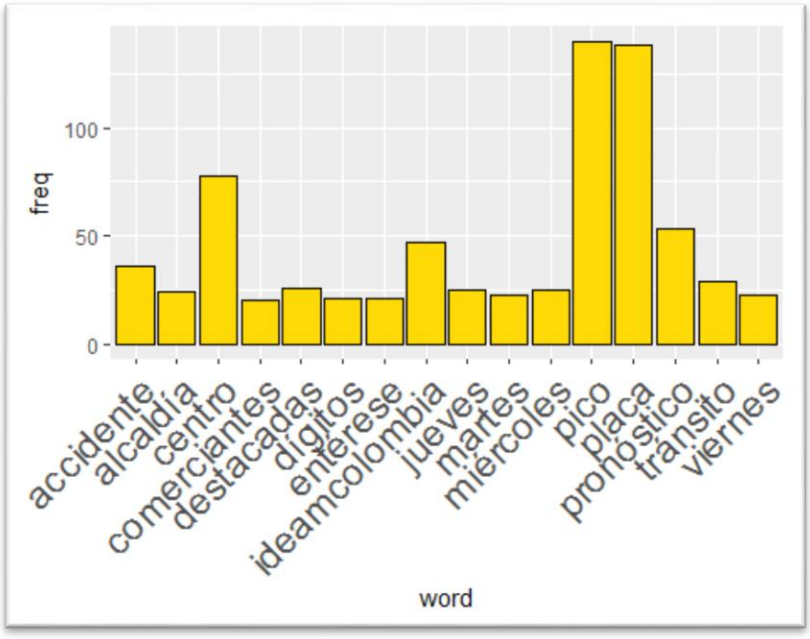

Figura 17. Frecuencia de términos cluster 1.

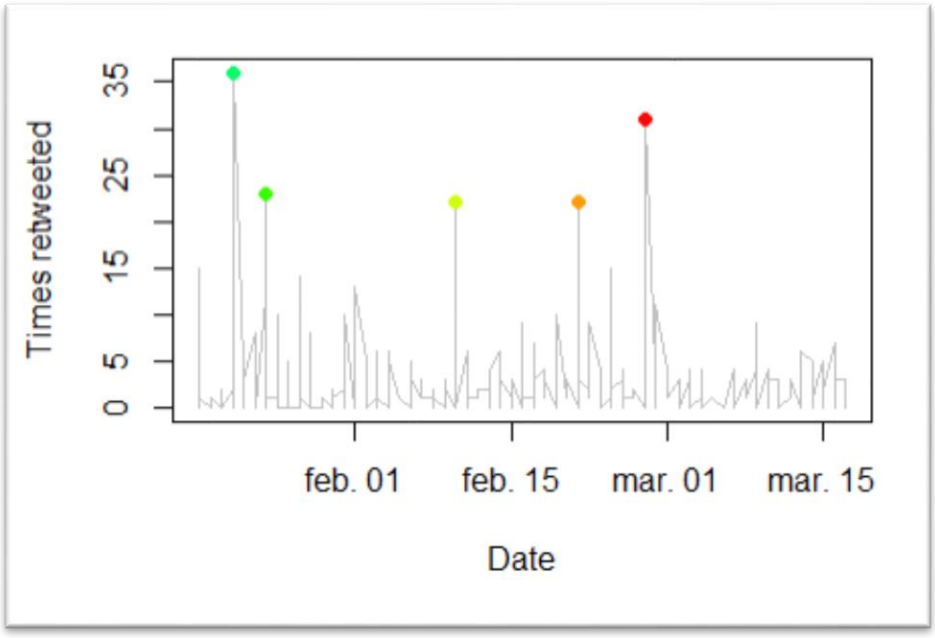

Figura 18. Gráfico de Retweets cluster 1.

El resultado arrojado en el cluster 1 muestra que los temas fuertes encontrados fueron los cambios en el pico y placa, teniendo una frecuencia por encima de 120 y la palabra centro con una frecuencia por encima de 50, palabras destacadas como se pueden ver en la figura 16, que marcaron una tendencia importante en este cluster ya que para estas fechas abarcaron temas como la aplicación del nuevo pico y placa en toda la ciudad, donde vanguardia uso este medio para comunicar todos los cambios que se generaron en la ciudad, se destaca que la palabra dígitos va muy relacionada con estas palabras en tweets como:

- Pico y Placa este viernes en #Bucaramanga: dígitos terminados en 9 y 0. Siga las noticias en Vanguardia.
- Pico y Placa este jueves en #Bucaramanga: dígitos terminados en 7 y 8. Siga las noticias en

Otras palabras principales que se puede observar en la figura 17 y que no son muy relevante son palabras como: accidente, ideamcolombia, pronóstico, tránsito, que crean un nivel de importancia teniendo en cuenta que informa a los Bumangueses sobre eventos que se presentan a causa de accidentes de tránsito e inundaciones o accidentes a causa de las lluvias.

En la figura 18, indica la frecuencia de retweets en dos meses, dejando ver que para la segunda quincena de enero y la segunda quincena de febrero el gráfico muestra una periodicidad de picos altos marcados como retweets, también indica que hay 5 tweets de mayor relevancia en esos dos meses, mostrando que en esos 5 tweets hubo más de 20 retweets marcados como tendencias en el cluster 1; es importante decir, que estos 5 tweets no tienen relación con el tópico principal del cluster.

7.5.2 Cluster 2

Figura 19. Nube de palabras cluster 2.

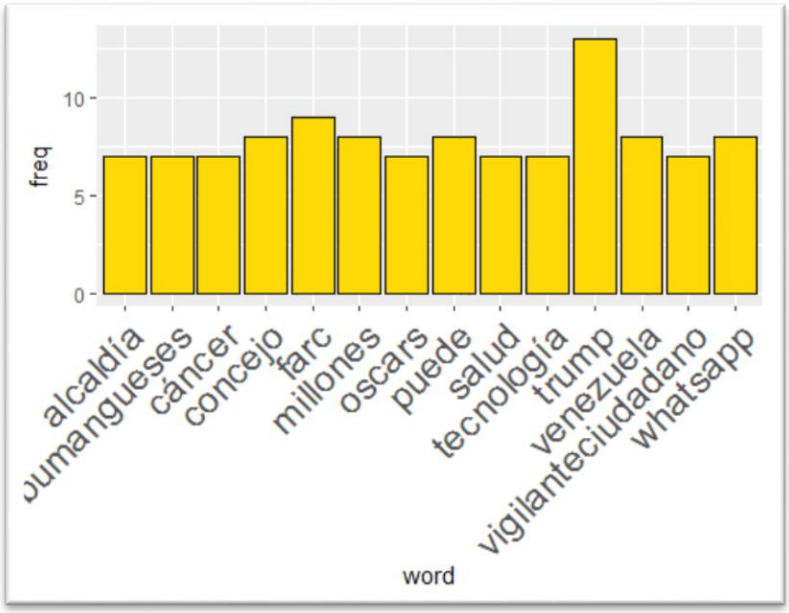

Figura 20. Frecuencia de términos cluster 2.

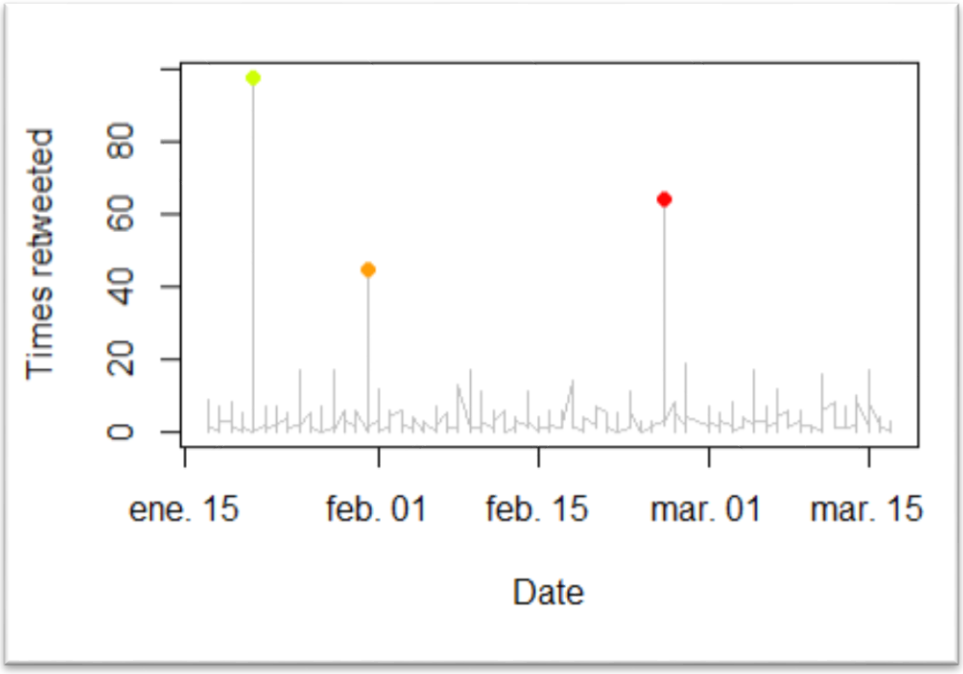

Figura 21. Gráfico de Retweets cluster 2.

En el cluster 2 muestra un término con una alta frecuencia como la palabra trump, teniendo una relevancia en política ya que en la actualidad a nivel mundial se habla de los diferentes cambios en lo que respecta al poder político social, en EEUU que está fomentando fuertemente el actual

gobierno, teniendo un personaje atípico como Trump. Al informar sobre este tema crea la expectativa sobre los cambios que se puedan presentar en el mundo siendo EEUU una potencia mundial.

Otro tema político que se está viviendo en el momento es el de Venezuela, ya que es otro término de una frecuencia importante y que el objetivo es informar sobre la crisis económica que sufre el vecino país fenómeno que ha afectado en los últimos años y que ha generado una denunciada crisis humanitaria en Venezuela. Esta situación se da en productos con precios regulados, como alimentos (leche, diversos tipos de carne, pollo, café, arroz, aceite, harina precocida, mantequilla, entre otros), productos de primera necesidad (papel higiénico, aseo personal), medicinas (para tratar el cáncer entre otros).

En la figura 20 muestra el término Whatsapp de gran preeminencia donde el objetivo principal es comunicar y alertar a los Bumangueses de los peligros que existen entre los que utilizan la tecnología, ya que a través de cadenas que circulan en este medio engañan, mal informan y causan mucho daño a los usuarios.

Un término de muy poca frecuencia mostrada en la figura 20 como cáncer, pero de gran importancia, siendo significativo para los ciudadanos ya que es un tema que afecta a toda la población y lo que se busca es informar a toda la ciudadanía de los diferentes casos que se están presentando en nuestra región, con el objetivo que las personas tomen conciencia que esta enfermedad detectada a tiempo se puede curar y a veces por falta de desconocimiento e información terminan en una enfermedad sin salida.

En cuanto a la figura 21 se observa una tendencia importante con 98 retweets al tweet “Los sobrevivientes del Chapecoense alzaron la Copa Sudamericana” y que aun viviendo el dolor de la tragedia la gente sigue mostrando sus condolencias y solidaridad a las víctimas del accidente aéreo.

Otra tendencia marcada con 45 retweets fue “La Corte estudia una demanda para incluir las corridas de toros y las riñas de gallos como delito de maltrato animal…” manifestando con esto el apoyo a esta iniciativa para acabar con las corridas de toros y peleas de gallos.

De acuerdo con estas tendencias significativas mostradas en la figura 21 es importante decir que no hay relación con respecto a los tópicos principales en la frecuencia de términos de nube de palabras e histograma; se puede decir, que estos temas no fueron tan comunes, pero si relevantes, teniendo un impacto interesante en los seguidores.

7.5.3 Cluster 3

Figura 22. Nube de palabras cluster 3.

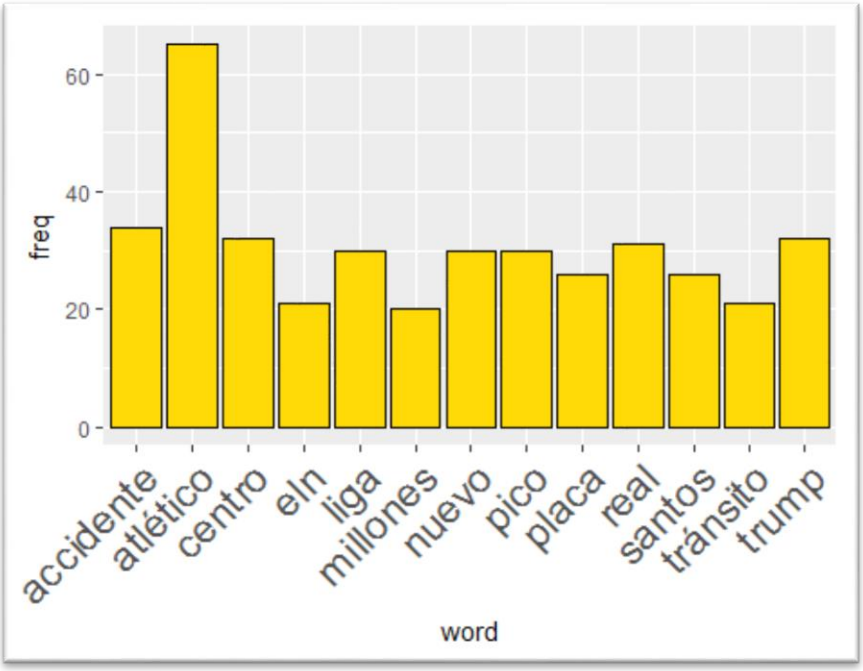

Figura 23. Frecuencia de términos cluster 3.

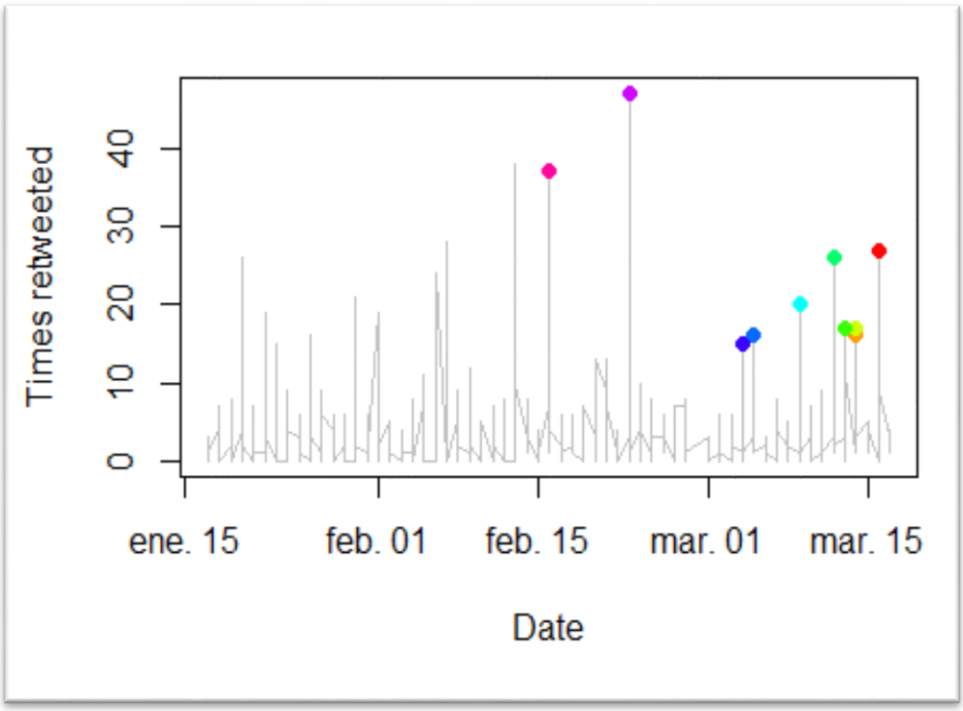

Figura 24. Gráfico de Retweets cluster 3.

Para el cluster 3 en la figura 22 y 23 indican una frecuencia de términos importantes en temas de futbol como atlético, real, Madrid y liga; dando información a contenidos como el atlético

Bucaramanga que en el inicio de la liga colombiana no empezó de la mejor forma y con malos resultados desataron el descontento de los hinchas llegando al punto de hacer protestas y manifestaciones de rechazo para la expulsión del técnico; también nos arroja una variedad de información sobre los equipos del Real Madrid, Real Bucaramanga y el Real Santander de acuerdo a cada liga donde juegan, exaltando marcadores, posiciones y lideratos.

Otra frecuencia importante son palabras como pico, placa y centro, la cual es un tema enfocado a la dirección de tránsito y la Alcaldía de Bucaramanga diseñaron un esquema de la restricción vehicular en el centro de Bucaramanga causando gran controversia entre los diferentes sectores del comercio de dicha zona, utilizando el twitter como un medio de comunicación frente a los cambios que refiere el tema, siguiendo día a día las anomalías que se presentaron.

Otros términos importantes como la paz y el eln no tuvieron gran trascendencia en este cluster, pero muestra tweets relevantes con los atentados del eln que pondrían en duda el proceso de paz con esta guerrilla.

Finalmente se puede observar en la figura 24 que en los dos meses hubo una actividad de retweets marcando tendencias importantes; que en la primera quincena del mes de marzo hay 8 tweets que tuvieron más de 15 retweets, en contándose temas de política como el siguiente tweet: “Caravana de seguridad de Didier Tavera transitó en contravía por carril de Metrolínea” con 26 retweets, esto indica que la gente sigue estando en desacuerdo con actividades y acciones ilegales por parte de funcionarios de entidades del estado.

Otro tweet relevante fue “tembló en varias ciudades de #Colombia a las 10:38 a.m. con una magnitud de 5.1” muestra la gran importancia de las aplicaciones de mensajería instantánea las cuales son claves para comunicarse ante cualquier eventualidad de sismo, cómo pedir ayuda o simplemente informar del movimiento telúrico que azotó a Santander.

Para la segunda quincena del mes de febrero como revela la figura 24 muestra una tendencia importante en dos tweets que tuvieron más de 35 retweets los cuales fueron los siguientes: “Senador Robledo pidió a procuraduría sancionar a Fiscal General por escándalo Odebrecht” y “Vea la denuncia de corrupción que hizo Claudia López al PAE de la Gobernación de #Santander” esto evidencia que compartir y estar informados por las redes sociales de todo los actos de corrupción reflejan la indignación en la sociedad, transformándolo esto en presión social, manifestaciones pacíficas y rechazo a todo corrupto.

7.5.4 Cluster 4

Figura 25. Nube de palabras cluster 4.

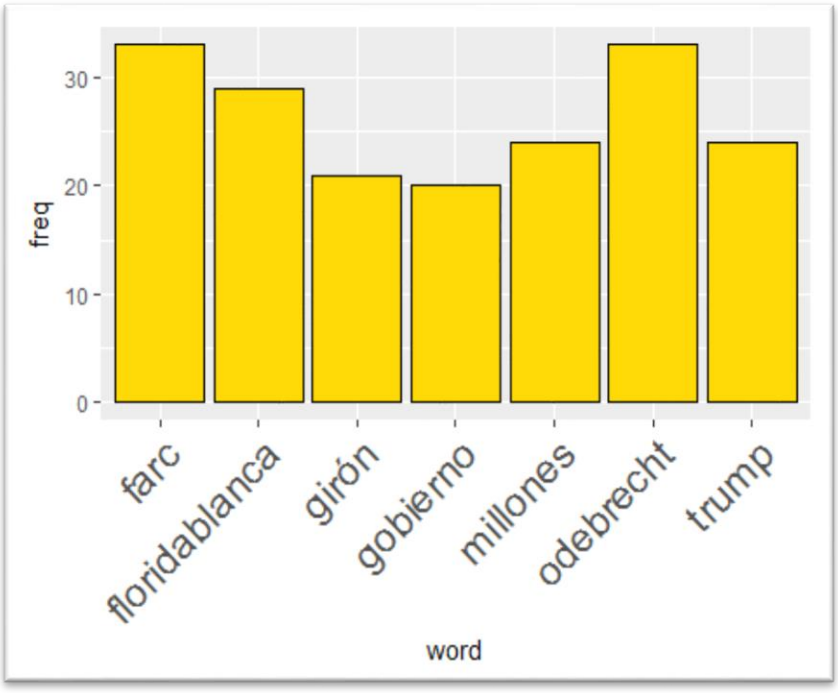

Figura 26. Frecuencia de términos cluster 4.

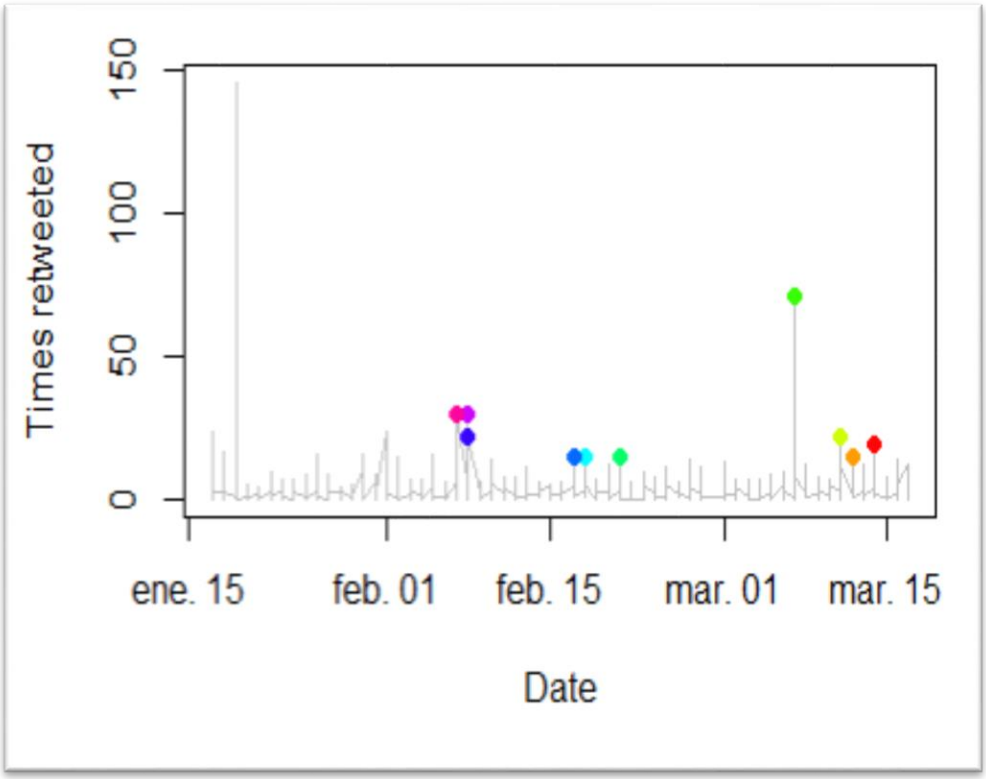

Figura 27. Gráfico de Retweets cluster 4.

En el cluster 4 muestra un término farc de gran relevancia la cual busca informar la evolución que ha llevado el proceso de paz con participación activa del pueblo en iniciativas en reformas políticas, económicas y sociales, fomentando una gran expectativa: ¿qué va a pasar con el proceso de paz? A raíz de todo esto el ELN abrió una puerta importante hacia los diálogos con el gobierno, creando esperanza y controversias en los colombianos.

El escándalo de corrupción orquestado por Odebrecht otro tema de gran envergadura el cual explotó en Colombia a finales del 2016, la sombra que se extiende tras la constructora brasileña es larga, tanto en el tiempo como en cuanto a los países a los que afectó, y ha desatado una ola de tensiones políticas que aún golpea a la mayor parte de Latinoamérica. Siendo tendencia el 19 de enero de 2017 con 146 retweets en el tweet "Jorge Robledo denuncia que Banco Agrario hizo préstamos irregulares a Odebrecht" como se puede ver en la figura 27.

Otro de los tweets con mayor relevancia fue "#EnDesarrollo Hinchas del Atlético Bucaramanga realizan un plantón frente a las oficinas del club…" con 71 retweets mostrando y dando a conocer el inconformismo a los directivos por los malos resultados del equipo en lo que lleva la liga y exigiendo resultados satisfactorios.

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

8. Conclusiones

En el presente proyecto se realizó una revisión de variantes de K-means (algoritmo de clustering particional) donde se realizó una aplicación real utilizando datos de la red social twitter, todo esto enmarcado dentro del área del aprendizaje automático y más concretamente en un conjunto de técnicas denominadas técnicas de agrupamiento (clustering) por lo cual se ha descrito brevemente los conceptos básicos de este campo, algunas de las técnicas de uso más comunes y sus aplicaciones más conocidas, así como la segmentación de los seguidores mediante la utilización de técnicas Machine Learning y extrayendo tendencias a partir de métodos de procesamiento de lenguaje natural.

En cuanto a la implementación, se desarrolló minería de texto, el algoritmo k-means con el que se segmento los datos de estudio, el método Elbow con el que se obtuvo el número de clusters óptimo, representación como la tf-idf, y análisis de tendencias en twitter. Con todo ello se ha conseguido mejorar los conocimientos adquiridos en la etapa de estudio previo a partir de la puesta en práctica de las instrucciones impartidas. El desarrollo del trabajo también ha servido para conocer las posibilidades del análisis en Twitter y el potencial de la API.

En lo que respecta a la red social Twitter, nos ofrece una API fácil de utilizar pero que viene limitada. Hubiese sido más fácil trabajar con un acceso que no estuviese tan restringido, pero no ha supuesto grandes inconvenientes. Se ha visto que los tweets están llenos de ruido, la gente escribe con muchas faltas de ortografía. Al tener un límite de 140 caracteres, estamos un poco

condicionado a la hora de trabajar con algoritmos que hacen uso de procesamiento de lenguaje natural o análisis de texto.

Los agrupamientos aplicados a Twitter, permitieron extraer ciertas informaciones para ver de qué hablan, ya que permite sacar información, tendencias importantes sobre un acontecimiento o tema en discusión. Al ser una colección de temas seguramente el algoritmo trabajado se comportó de una manera eficiente creando clusters con mejor calidad. Aún más, el clustering fue una ayuda para ver por encima los temas que se hablan en los tweets que se trabajaron; en la fase experimental se han conseguido resultados interesantes como la extracción de los temas más frecuentes y la gran cantidad de seguidores desde el punto de vista del mercado objetivo de la cuenta.

Por lo tanto, se han cumplido los objetivos marcados, aunque todavía hay mucho margen de mejora y unas posibilidades enormes en materia de minería de texto para propuestas a un trabajo futuro, algunas de éstas se discuten en la sección 9.

9. Recomendaciones.

A partir del estado actual de la implementación resultante de este trabajo de investigación, se desprenden diferentes líneas de trabajo sobre las cuales se podrían continuar.

Hay que ver como eliminar todo el ruido de los tweets que al final impide trabajar con comodidad. Se podría utilizar etiquetado gramatical para realizar un análisis de las palabras más profundo. Además, habría que poder corregir la variación de la calidad del contenido en lo que

respecta a la jerga de uso común en internet. Así mismo, hacer un amplio análisis del contenido en sí, existe una extensa gama de información de contenido disponible. Por ejemplo, Twitter permite a los usuarios Símbolo "#", llamado hashtag, para marcar palabras clave o temas en un Tweet.

Un problema detectado en la investigación fue la sensibilidad de tiempo ya que los diferentes usuarios de twitter publican noticias e información varias veces al día, pueden querer comunicarse Instantáneamente con amigos, al enviar una consulta a Twitter, los resultados devueltos son sólo varios minutos de antigüedad. Además de comunicarse y compartir entre sí, los usuarios publican comentarios sobre eventos recientes, tales como nuevos productos, películas, deportes, juegos, Campañas políticas, etc. El gran número de actualizaciones en tiempo real de información es abundante, lo que proporciona muchas oportunidades de detección y monitoreo de un evento, esto puede llevar a una investigación relevante sobre como analizar los tweets en tiempo real.

Por otro lado, se ha constatado que el número de usuarios que utilizan la geolocalización en sus tweets el usuario lo restringe por una política de protección de datos, por lo que al analizar las cuentas en particular a partir de este criterio se convierte en algo no representativo. De todos modos, el hecho de utilizar estos datos geográficos de twitter puede ser importantes para algunas aplicaciones como análisis de población, mercadeo, político, etc., teniendo gran relevancia este factor.

Referencias Bibliográficas

Aljaber, B., Stokes, N., Bailey, J., & Pei, J. (2010). Document clustering of scientific texts using citation contexts Retrieval, *13(2), 101-131. Information.*

Aluja Banet, T. (2001). La minería de datos, entre la estadística y la inteligencia artificial.

Aliguliyev, R. M. (2010). Clustering Techniques and Discrete Particle Swarm Optimization Algorithm for Multi-Document Summarization. *Computational Intelligence, 26(4), 420-448.*

Allegue Lorenzo, Á., & Rueda Silva, A. C. (2016). *Seguimiento automatizado de variables de control en pacientes crónicos.*

Anchalia, P. P., Koundinya, A. K., & Srinath, N. K. (2013, June). MapReduce design of K-means clustering algorithm. *In 2013 International Conference on Information Science and Applications (ICISA) (pp. 1-5). IEEE.*

Araya, M. S. J. A. Z., Brenes, L. R. T. V., & Acuña, L. M. F. (2012) Aprendiendo estadística con R.

Bouras, C., & Tsogkas, V. (2012). A clustering technique for news articles using WordNet. *Knowledge-Based Systems, 36, 115-128*.

Blázquez Ochando, Manuel. "Técnicas avanzadas de recuperación de información: procesos, técnicas y métodos." (2013).

Blanco-Hermida Sanz, E. J. (2016). Prototipo de clustering orientado motor de búsqueda (Bachelor's thesis, Universitat Politècnica de Catalunya).

Braga, l.p., Valencia, l.i, & Carvajal, S.S. Introducción a la minería de datos, Sao Pablo, 2009 Cambronero, C. G., & Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. *Inteligencia en Redes de Comunicación, Universidad Carlos III de Madrid.*

Cardona, A., Nigro, N., Sonzogni, V., & Storti, M. (2006). Análisis numérico de diferentes criterios de similitud en algoritmos de clustering. *Mecánica Computacional, 25, 993-1011*.

Cambronero, C. G., & Moreno, I. G. (2006). Algoritmos de aprendizaje: knn & kmeans. *Intelgencia en Redes de Comunicación, Universidad Carlos III de Madrid.*

Cotelo, J. M., Cruz, F., Ortega, F. J., & Troyano, J. A. (2015). Explorando Twitter mediante la integración de información estructurada y no estructurada. *Procesamiento del Lenguaje Natural, 55, 75-82.*

Corso, C. L., & Alfaro, S. L. (2009). Algoritmos de Data Mining aplicados la enseñanza basada en la Web.

Corso, C. L. (2009). Aplicación de algoritmos de clasificación supervisada usando Weka. *Córdoba: Universidad Tecnológica Nacional, Facultad Regional Córdoba.*

Clustering de Texto. (s.f.). Obtenido de MeaningCloud: https://www.meaningcloud.com/es/productos/clustering-de-texto

Dasgupta, S. (2013). Algorithms for minimally supervised learning. *Online, November, 9.*

Delgado Calle, C. (2015). Utilización de técnicas de clustering para mejorar la detección de meta-topics en conjuntos de datos extraídos de Twitter.

Echeverri, L. A., Retamoza, A. M. P., de la Rosa, M. O., Barros, I. V., Álvarez, D. D. O., & Guerrero, E. C. (2013). Minería de datos como herramienta para el desarrollo de estrategias de mercadeo B2B en sectores productivos, afines a los colombianos: una revisión de casos. *Sotavento MBA, (22), 126-136.*

Espino, A. I. L., Mur, R. A., & de Miguel, M. A. S. (2004). Aprendizaje automático en conjuntos de clasificadores heterogéneos y modelado de agentes. *Universidad Carlos III de Madrid, Departamento de Informática.*

Feinerer, I. (2010, December). Analysis and algorithms for stemming inversion. *In Asia Information Retrieval Symposium (pp. 290-299). Springer Berlin Heidelberg.*

Gallardo Campos, M. (2009). Aplicación de técnicas de clustering para la mejora del aprendizaje.

Garre, M., Cuadrado, J., Sicilia, M. A., Rodríguez, D., & Rejas, R. (2007). Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software. *Revista Española de Innovación, Calidad e Ingeniería del Software, 3(1), 6-22.*

Gilbert, K., Sánchez, R. R., & Santos, J. C. R. (2006). Minería de datos: Conceptos y tendencias. Inteligencia artificial: *Revista Iberoamericana de Inteligencia Artificial, 10(29), 11-18.*

González, D. P. (2010). Algoritmos de agrupamiento basados en densidad y variación de clusters (Doctoral dissertation, Universitat Jaume I, Departament de Llenguatges i Sistemes Informàtics).

Guy Kawasaki and Peg Fitzpatrick, (2014). *The Art of Social Media: Power Tips for Power Users*, Penguin Books,

Hernández Leal, E. J. (2016). Aplicación de técnicas de análisis de datos y administración de Big Data ambientales (*Doctoral dissertation, Universidad Nacional de Colombia-Sede Medellín)*

Hotho, A., Nürnberger, A., & Paaß, G. (2005, May). A brief survey of text mining. *In Ldv Forum (Vol. 20, No. 1, pp. 19-62).*

Hotho, A., Staab, S., & Stumme, G. (2003). *Text clustering based on background knowledge. Technical report 425, 36.*

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern recognition letters, 31(8), 651-666.*

ain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR), 31(3), 264-323.*

Jiménez, C. M. (2014). “Bid data". Un nuevo paradigma de análisis de datos. *In Anales de mecánica y electricidad (Vol. 91, No. 6, pp. 10-16). Asociación de Ingenieros del ICAI.*

Jing, L. (2008). Survey of text clustering. Department of Mathematics, *the University of Hong Kong, HongKong, China, ISBN, 7695-1754.*

Juan_Vidal. (2014). *DataPrix*. Obtenido de Big data: Gestión de datos no estructurados: http://www.dataprix.com/blog-it/big-data/big-data-gestion-datos-no-estructurados

Kalogeratos, A., & Likas, A. (2011). Document clustering using synthetic cluster prototypes. *Data & Knowledge Engineering, 70(3), 284-306.*

Korth, H. F., Silberschatz, A., Sudarshan, S., & Pérez, F. S. (1993). *Fundamentos de bases de datos (No. 005.7406 005.7406 K85f2E2v). McGraw-Hill.*

Kuna, H., García Martínez, R., & Villatoro, F. (2009). Procedimientos de la explotación de información para la identificación de datos faltantes, con ruido e inconsistentes. *In XI Workshop de Investigadores en Ciencias de la Computación.*

Liu, H., & Huang, S. T. (2003). Evolutionary semi-supervised fuzzy clustering. *Pattern Recognition Letters, 24(16), 3105-3113.*

Luo, C., Li, Y., & Chung, S. M. (2009). Text document clustering based on neighbors. *Data & Knowledge Engineering, 68(11), 1271-1288.*

MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. *In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).*

Martín Morales, S. (2016). Análisis de información proveniente de redes sociales como Twitter (Bachelor's thesis).

Montes, M., & de Lenguaje Natural, G. L. (2014). Minería de texto: Un nuevo reto computacional. Laboratorio de Lenguaje Natural, *Centro de investigación en computación, instituto politécnico nacional.*

Pascual, D., Pla, F., & Sánchez, S. (2007). Algoritmos de agrupamiento. Método Informáticos Avanzados.

Porras, J. C. C., Laverde, R. M., & Díaz, J. R. (2008). Técnicas de lógica difusa aplicadas a la mineria de datos. *Scientia et Technica, 3(40), 1-6.*

Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning.

Santiago Cortez, S. D. (2016). Microsoft Azure. Obtenido de Tareas para preparar los datos para el aprendizaje automático mejorado: https://docs.microsoft.com/es-es/azure/machine-learning/machine-learning-data-science-prepare-data

Sarduy Domínguez, Y. (2007). El análisis de información y las investigaciones cuantitativa y cualitativa. *Revista cubana de salud pública, 33(3), 0-0.*

Shiels, M. (2011). Twitter co-founder Jack Dorsey rejoins Company. *BBC News, 28(39, 20011)*

Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three Different Distance Metrics. *International Journal of Computer Applications, 67(10).*

Tascón, M. (2013). Introducción: Big Data. Pasado, presente y futuro. Telos: *Cuadernos de comunicación e innovación, (95), 47-50.*

Vijayarani, S., Ilamathi, M. J., & Nithya, M. (2015). Preprocessing Techniques for Text Mining-An Overview. *International Journal of Computer Science & Communication Networks, 5(1), 7-16.*

Won, D. Song, B. M., & McLeod, D. (2006). An approach to clustering marketing data. In Proceedings of the 2nd International Advanced Database Conference.