Prototipo de Sistema de Exploración y Generación de Herramientas de Análisis para Datos de Twitter

Tatiana Novoa Triana

Julio 2020

Director de Proyecto

Ixent Galpin

Maestría en Ingeniería y Analítica de Datos

Universidad de Bogotá Jorge Tadeo Lozano

Bogotá D.C.

UTADEO
UNIVERSIDAD DE BOGOTÁ JORGE TADEO LOZANO

Tabla de Contenido

Lista de Figuras

Lista de Tablas

1. Resumen

Internet y las redes sociales han permitido a las personas comunicarse y expresarse libremente acerca de cualquier tema que se comparta por dichos medios. Con el pasar del tiempo y debido a la popularidad de muchas de estas redes sociales se empieza a observar que la información que se genera cada minuto es de tal magnitud que la tecnología misma para almacenarla y procesarla se empieza a ver limitada y se empieza a requerir de nuevas técnicas para ello, a este alto volumen de información es a lo que se le conoce como *Big Data*. Una de estas redes sociales es Twitter, aquí las personas pueden expresar sus opiniones sobre cualquier tema, incluso empresas usan esta red para conocer que piensan sus clientes acerca de los productos o servicios.

Este proyecto busca desarrollar una herramienta que permita generar algunos de los tantos análisis que se pueden generar a partir de esta información, entre ellos está el análisis de sentimiento, dicho análisis consiste en clasificar las opiniones de las personas en positivas, negativas o neutrales acerca del tema que se desea indagar. Para poder cumplir el objetivo y desarrollar la herramienta se utilizarán herramientas de código abierto, como es Spark para procesamiento paralelo de datos con Clústers, así como también Django, el cual es un framework de Python para desarrollo ágil de entornos web.

Palabras Clave: Twitter, Análisis de Sentimientos, Clusterización de Palabras, Nube de Palabras, Frecuencia de Palabras, Spark, Clúster, Django, *Big Data*

2. Abstract

Internet and social networks have allowed people to communicate and express themselves freely about any topic that is shared by users. Over time and due to the popularity of many of these social networks, we find that the information generated every minute of the day is of such magnitude that the technology to store and process seems to be limited and, as such, new techniques are required for such high volumes of information known as Big Data. One of these social networks is Twitter, there people can express their opinions on any topic, even companies use this network to know what their customers think about products or services.

This project seeks to develop a tool that allows generating some of the many analyses that can be generated from this information, among which is sentiment analysis. This analysis consists in classifying the opinions of people in positive, negative or neutral about the topic user want to investigate. In order to fulfill the objective and develop the tool, open-source libraries will be used, such as Spark for parallel data processing with clusters, as well as Django, which is a Python framework for agile development of Web environments.

Keywords: Twitter, Sentiment Analysis, Words Clustering, Word Cloud, Word Frequency, Spark, Cluster, Django, Big Data

3. Introducción

Actualmente la cantidad de datos estructurados y no estructurados que se generan a diario en redes sociales como Facebook, Twitter, Instagram, así como en toda la Web es enorme. El avance tecnológico ha hecho cada vez más fácil que las personas de todo el mundo puedan expresar abiertamente sus opiniones y compartan sus puntos de vista sobre cualquier tema, ya sea político, económico, social, cultural, entre otros.

La plataforma Hootsuite es una plataforma de gestión de redes sociales que realiza anualmente un estudio del uso de las redes sociales alrededor del mundo. Entre los años 2014 y 2020 se identifica un crecimiento del 33,8% a nivel mundial del tiempo usado en redes sociales como se puede identificar en la Figura 1.

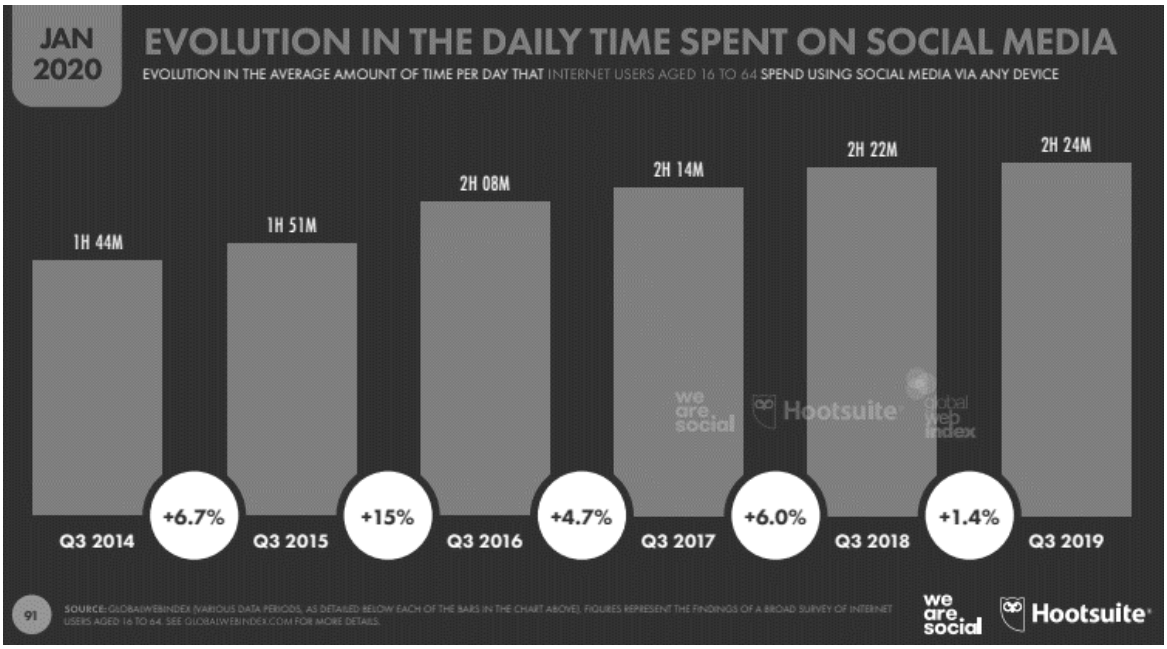


Figura 1. Evolución del tiempo diario usado en redes sociales *(Kemp, 2020)*.

Este crecimiento exponencial de la información ha implicado un gran reto en cuanto a la forma y técnicas que se deben utilizar para almacenar y procesar dicha información de la manera más eficiente posible, es decir, que implique el menor costo en tiempo y capacidades computacionales posible. Dada esta situación, algunos frameworks como Hadoop y Spark fueron implementados para hacer este procesamiento de datos más sencillo. Hadoop por ejemplo pone a disposición dos componentes como son, HDFS (Hadoop Distributed File System) utilizado para el almacenamiento de la información, pues almacena los datos en bloques de memoria y lo distribuye en Clústers, y MapReduce para toda la parte de procesamiento.

Centrándonos en Twitter, por ejemplo, donde cada minuto se generan millones de *tweets* acerca de diversos temas, analizar uno por uno podría ser una tarea bastante tediosa. Dado el caso, una compañía desee conocer la opinión de sus clientes acerca de sus productos o simplemente se quiera conocer la opinión de las personas acerca de una temática en particular, rastrear cada opinión es bastante complicado. En esta situación, algunas técnicas como análisis de sentimientos, frecuencia de palabras y nubes de palabras juegan un papel importante.

El objetivo del análisis de sentimiento es identificar la polaridad y la actitud de un grupo de usuarios frente a un tema específico. Este proceso involucra varias ramas de la computación como es el procesamiento del lenguaje natural, minería de texto y aprendizaje automático. Para realizar el análisis de sentimiento de *tweets* se debe superar algunos inconvenientes dados por las diferencias en la forma en que los usuarios pueden expresarse por medio de la red social, esto incluye palabras propias y particulares de la jerga, errores de ortografía, palabras que no ofrecen significado como son los artículos, pronombres, preposiciones entre otras (también llamadas *stop words* en inglés). Para solucionar esto, algunos lenguajes de programación como R y Python

ofrecen librerías para Hadoop y Spark que permiten procesar esta información de manera masiva, realizar todo el proceso de limpieza de la información y generar los análisis correspondientes.

Para llevar a cabo este trabajo, se usará Hadoop HDFS para almacenar la información extraída a partir de la API de Twitter, haciendo uso de un nodo maestro y dos esclavos. Adicionalmente, se utilizará librerías de Spark para Python para realizar todo el proceso de limpieza y preparación de los datos, así como también todo el análisis de sentimiento y generación de nubes de palabras. Finalmente, para la visualización se utilizará Django, de esta manera se ofrecerá al usuario un entorno Web amigable para la consulta y exploración de los análisis realizados.

4. Objetivos

4.1. Objetivo General

Diseñar y desarrollar un prototipo de una aplicación Web que permita exploración y generación de herramientas de análisis escalable de los datos generados en la red social Twitter.

4.2. Objetivos Específicos

- Configuración de clúster haciendo uso de herramientas Hadoop y Spark.
- Exploración y configuración de servicio para obtención de datos de Twitter.
- Desarrollo de prototipo de aplicación Web para la generación de herramientas de análisis (Frecuencia de Palabras, análisis de Sentimientos y Clusterización de palabras entre otros) de las consultas realizadas.
- Evaluación de resultados mediante dos casos de estudio.

5. Marco Teórico

Con el fin de tener un entendimiento global del contexto en el cual se realiza el proyecto, es necesario comprender diversos conceptos. De esta manera, en el siguiente capítulo se estudiarán ciertas definiciones correspondientes a tecnologías de Twitter, Clúster y los análisis realizados como análisis de sentimiento y clusterización utilizadas entre otros que son necesarios para el desarrollo del proyecto.

5.1. Redes Sociales

Una Red Social como su nombre indica hace referencia a los vínculos o interacciones sociales de cualquier tipo (amistosas, amorosas, comerciales, etc.) que se generan entre grupos de personas. A partir de 1991 cuando que se hizo pública *World Wide Web* (www) surgieron nuevos modelos de comunicación que nos permiten compartir dichos vínculos haciendo uso de sistemas de información y rompiendo las barreras impuestas por el tiempo y el espacio (López Zapico, 2020).

Las redes sociales han adquirido un fuerte impacto social en el mundo de manera que se han convertido en una herramienta indispensable en las nuevas generaciones, ya que por medio de ellas se ofrecen funcionalidades para compartir ideas, vivencias, noticias, opiniones, imágenes y que posibilita la publicación de información y servicios que facilitan la cotidianidad de las personas.

El funcionamiento de las redes sociales se basa en que todos los usuarios de estas crean un perfil y establecen una red de contactos con los cuales tienen algún interés en común y de esta manera permite compartir contenidos de texto o multimedia.

Las redes sociales cuentan con características que las identifican y permiten también describir más de sus funcionalidades como se pueden identificar en la Tabla 1.

Tabla 1. Características de Redes Sociales

Característica	Descripción
Entretenimiento	Seguir el contenido de marcas, personas públicas o empresas que genere algún interés comercial o alguna afinidad con los usuarios.
Información	A través de las publicaciones de notas o contenidos informativos que generen comunicación o interés a los usuarios.
Contactos Personales	Buscar y conectarse con familiares o amigos.
Contactos Profesionales	Hacer contactos profesionales de tu sector o profesión que se podría materializar fuera de línea.
Comunidades Online	Para las empresas con el tiempo permite crear una comunidad en torno a tu marca y se puede poner a disposición productos o servicios.
Publicidad	Permite realizar publicidad online e identificar a que publico se quiere llegar.
Branding	Colocar anuncios de empresas o marcas para que se puedan ver a través de redes sociales sin que los usuarios necesariamente la sigan.

Existen diferentes tipos de redes sociales que habilitan diferentes funcionalidades. la Tabla 2 muestra los tipos de redes sociales y algunos ejemplos de cada una de ellas.

Tabla 2. Tipos de Redes Sociales

Tipo	Descripción
Horizontales	Son abiertas para cualquier tipo de usuario o finalidad. Ejemplo: Facebook, Instagram o Twitter
Verticales	Los usuarios tienen puntos en común y sirven para una o varias finalidades concretas. Ejemplo: LinkedIn, Tripadvisor o Spotify

Hoy en día las redes sociales se enfrentan a aspectos positivos y negativos que afectan la forma en que se distribuye la información y agudiza los efectos que dicha información tiene en cada una de las personas. En la Tabla 3 encontramos algunos de estos aspectos.

Tabla 3. Aspectos positivos y negativos de las redes sociales

Aspectos Positivos	Aspectos Negativos
Rapidez de distribución de información	Exhibicionismo
Conocimiento de información de interés	Exceso de vanidad
Facilidad para realizar contacto con otras personas	Fragilidad de privacidad
Acceso a todo tipo de contenidos	Propagación de noticias falsas
Autopromoción a través de perfiles	Pérdida de tiempo
Generación de Entretenimiento	Exceso de publicidad
Promoción y ventas de productos online	Errores costosos

A continuación, profundizaremos en la red social sobra la cual se va a enforcar el desarrollo de este proyecto.

5.1.1. Twitter

Twitter es una red social de tipo horizontal basada en *microblogging* que consiste en comunicarnos por medio de mensajes cortos a los cuales en dicha herramienta denominamos *tweets* (Abella García, 2015).

Sin embargo, como lo indica (Carballar, 2011) una de las particularidades que tiene Twitter es que tiene la capacidad de compartir el contenido publicado por medio de una cuenta, con cualquier otra persona del mundo que tenga interés de verla. Esta característica hace que se comporte más como un blog en donde no se necesita ningún tipo de suscripción para ver su contenido.

Desde el año 2007 esta herramienta inició con la búsqueda de compartir mensajes entre un grupo de personas haciendo uso de SMS y ha venido evolucionando hasta convertirse en una de las aplicaciones más usadas en donde puedes seguir a personas o empresas para compartir sus ideas, opiniones o productos con el resto del mundo. Así mismo como lo indica (Fainholc, 2011) “Twitter como herramienta de comunicación directa presenta un crecimiento descomunal, que

como es previsible, se constituye en una caja de resonancia de y en la vida social, cultural, política y económica de un país o región".

Esto último también conlleva a que la información publicada en estas cuentas corresponde a un interés particular de la persona detrás de la plataforma que es quien publica los *tweets*. Por este motivo se han generado diferentes prácticas como creación de perfiles falsos o viralización de noticias falsas que han generado diferentes estudios relacionados con la influencia de estas plataformas en los diferentes ámbitos sociales y políticos del mundo (Waxman, 2017). Estas prácticas son las que nos obligan a realizar un mejor análisis y entendimiento en conjunto y no solo basarnos en la información publicada en una sola cuenta.

5.2. Clústers

Los Clústers computacionales se han venido convirtiendo en una de las herramientas más populares para el procesamiento de grandes volúmenes de datos (o lo que llamamos *Big Data*). Esto se debe a que corresponden a redes de equipos de computación (de bajo o alto costo) interconectados entre sí y que permiten dividir el procesamiento de los datos paralelamente en los diferentes equipos que estén configuradas dentro en la misma red (Petrocelli, 2017).

Un Clúster es un grupo de equipos independientes que ejecutan una serie de aplicaciones de forma conjunta y aparecen ante clientes y aplicaciones como un solo sistema. Los Clústers permiten aumentar la escalabilidad, alta disponibilidad y fiabilidad de múltiples niveles de red (Olea, 2004).

Existen Clústers de diferentes tipos de acuerdo con las necesidades para las cuales se requieran y de acuerdo con esto también se definen las características de sus componentes como se puede observar en la Tabla 4.

Tabla 4.Tipos de Clúster

Tipo	Descripción
Alto Desempeño	Resolver problemas que requieren de mucho procesamiento
Alta Disponibilidad	Mantener aplicaciones en pleno funcionamiento el mayor tiempo posible
Balance de Carga	Hacer que cada equipo de cómputo reciba o entienda una solicitud

Es importante establecer cuál es la necesidad inicial para identificar los componentes ya que la tarea se torna ardua si no se tiene el software adecuado y una arquitectura de hardware eficiente (Piccoli, 2011).

Para este proyecto vamos a profundizar en los Clúster de alto de desempeño ya que es el tipo de Clúster que se configura para este proyecto. Este tipo de Clúster cuenta con un componente de hardware y otro de software que permite distribuir las tareas para que se procesen paralelamente. A continuación, describiremos las características de cada uno de estos componentes.

5.2.1. Hardware Clúster

De manera general un Clúster se define como un conjunto de equipos de cómputo interconectadas a través de alguna tecnología de red. Se caracterizan por estar formadas por estaciones de trabajo de alto rendimiento, conectadas a través de redes de alta velocidad (Pérez Hernández, 2003).

El Clúster está compuesto por un nodo maestro y varios nodos esclavos. El nodo maestro es aquel que se encarga de administrar los recursos y adicionalmente recibir y distribuir las tareas entre los diferentes nodos esclavo que tenga el Clúster a través de una red privada.

Los nodos esclavos son aquellos que se encargan de realizar la ejecución y devolver el resultado de las tareas que el nodo maestro distribuye a través de la red privada. En la Figura 2 podemos encontrar la arquitectura que tiene un Clúster de alto desempeño.

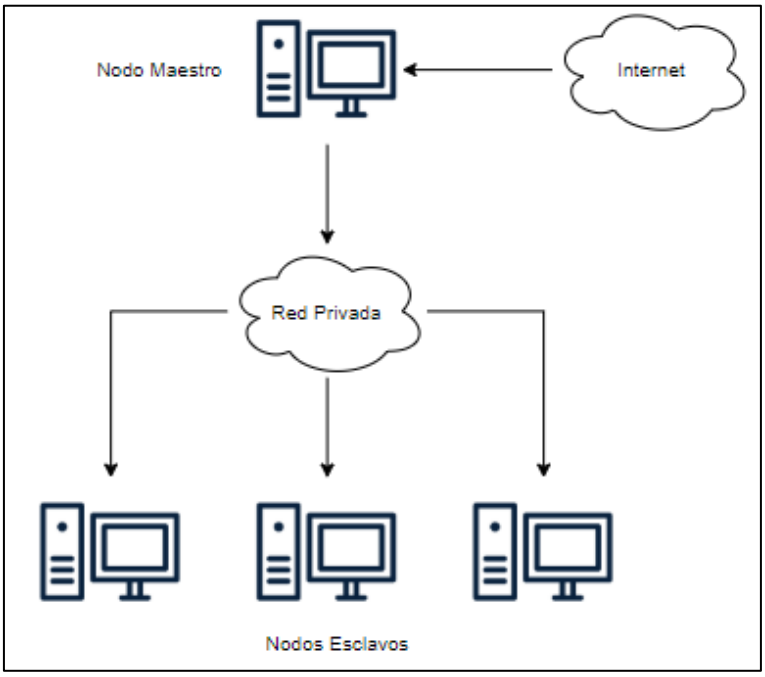


Figura 2. Arquitectura clúster alto desempeño

Para el proceso de comunicación entre cada uno de los nodos se requiere que todos los nodos se encuentren en la misma red y que el nodo maestro tenga comunicación y permisos para administrar, controlar y modificar los recursos de una maquina con los nodos esclavos usando mecanismos de autenticación por medio una llave que cifra los mensajes enviados entre los equipos. Esto se realiza mediante el protocolo SSH o *Secure Shell* que habilita las conexiones seguras a través de Internet (Hostinger, 2019). En la Figura 4 podemos ver el flujo que utiliza el Protocolo SSH para el cifrado de mensajes.

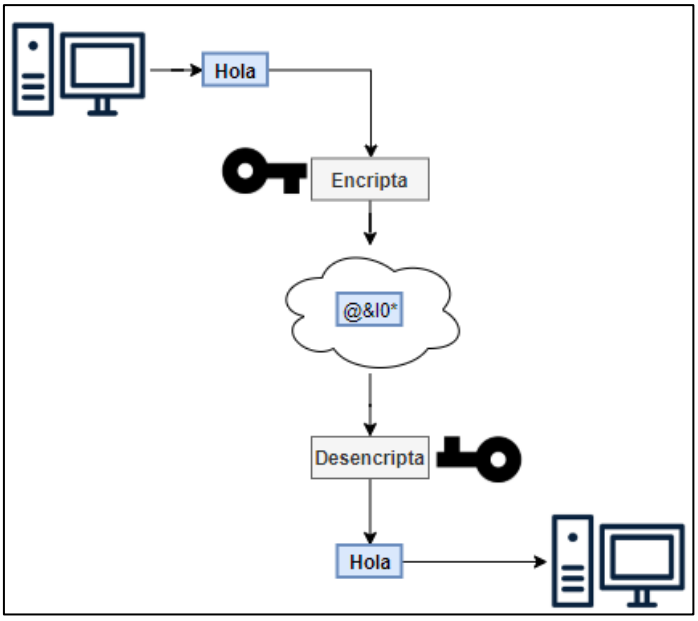


Figura 3. Protocolo SSH o Secure Shell

5.2.2. Software Clúster

Existen diferentes componentes de software que permiten la configurar y administrar los procesos que se ejecutan a través del clúster, sin embargo, para los fines de este proyecto nos centraremos en dos de las más utilizadas para procesamiento de grandes volúmenes de datos.

5.2.2.1.Hadoop

Apache Hadoop es un framework *open-source* perteneciente a Apache Software Foundation creado por Doug Cutting y está basado en el sistema de archivos de Google y MapReduce, el cual permite el procesamiento distribuido de grandes volúmenes de datos a través de Clústers. El framework de Hadoop nace con el fin de solucionar uno de los problemas que se enfrenta al trabajar con *Big Data*, como lo es el procesamiento de estos grandes volúmenes de datos de una manera eficiente, debido a que las tecnologías y base de datos tradicionales no tienen las capacidades para enfrentar este problema de una manera rápida y a la vez permita optimizar el uso

de los recursos y capacidades disponibles. Para solucionar este inconveniente, Hadoop pone a disposición clústers, los cuales son capaces de almacenar y procesar algoritmos paralelamente.

Para ilustrar el inconveniente un poco mejor, se puede exponer el ejemplo dado por el autor (White, 2009): Antiguamente, un disco duro en 1990 podía almacenar 1370 MB de datos y tenía una velocidad de lectura de aproximadamente 4.4MB/s, es decir se podía leer todo el disco en unos 5 minutos. Unos 20 años después, se tienen discos de mínimo 1 terabyte con una velocidad de lectura de unos 100MB/s, es decir se podría leer la totalidad de los datos en más o menos 2 horas y media. Como se puede observar el tiempo de lectura aumentara de manera proporcional al volumen de datos, de esta manera para optimizar esto se propone acceder la información de forma paralela teniendo réplicas de la información y a la vez que sea tolerable a fallos sin tener perdida de la información, es ahí donde se introduce el sistema de archivos de Hadoop HDFS (Hadoop Distributed File System).

HDFS es un sistema de archivos diseñado para almacenar archivos de gran tamaño en clústers. Esta construido bajo la idea que la forma más eficiente de procesar datos es realizando solo una escritura y realizar múltiples lecturas (*write-once, read-many pattern*). Este sistema es altamente tolerable a fallas debido a que se configura la cantidad de replicaciones de la información que se almacena.

Un Clúster de HDFS tiene dos tipos de nodos, un maestro (namenode) y uno o varios esclavos (datanodes). El maestro se encarga básicamente de administrar la estructura del sistema de archivos y la metadata de todos los archivos y directorios de dicha estructura, dicha información se almacena directamente en el disco (Borthakur, 2019). Adicionalmente el maestro es quien sabes que datanode tiene asignado cual bloque de un archivo en específico. El otro componente es el

datanode o esclavo, estos se encargan de almacenar y devolver los bloques solicitados por el cliente, adicionalmente reportan al maestro (namenode) que bloques están almacenando.

Otra de las situaciones que se debe solucionar al tener los datos distribuidos en diferentes nodos es volverlos a unir, para resolver esto Hadoop provee un modelo de programación llamado MapReduce. Este paradigma está conformado por dos fases principales, *Map* y *Reduce*. El proceso de Mapeo (*Map*) recibe como entrada un par llave-valor y realiza los procesos correspondientes para generar nuevos pares llave-valor, y la tarea de reducción (*Reduce*) se encarga de combinar los pares generados en cada nodo para generar las salidas correspondientes.

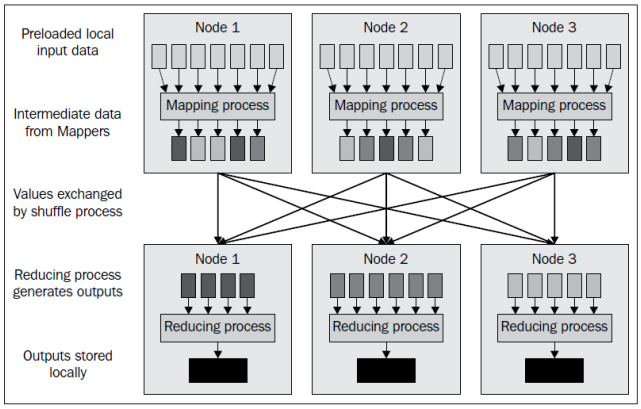


Figura 4. MapReduce (Prajapati, 2013)

5.2.2.2.Spark

Apache Spark es un sistema distribuido y altamente escalable para análisis de datos en memoria, el cual provee la capacidad de desarrollar aplicaciones en lenguajes de programación conocidos como, por ejemplo, Java, Scala, Python y R (Frampton, 2015). A diferencia de Apache Hadoop el cual escribe datos en disco, Spark fue optimizado para procesar datos rápidamente haciendo uso de la memoria lo cual puede ser hasta 100 veces más rápido que Hadoop MapReduce. Actualmente, Apache Spark es uno de los proyectos top de Apache con gran apoyo de la comunidad y de empresas como Databricks, IBM y Huawei (Scott, 2015). Spark ofrece múltiples

beneficios, pero se pueden resaltar tres características en especial, la primera consiste en la facilidad de uso gracias a la gran cantidad de API's que se han desarrollado y se encuentran bien documentadas para diferentes propósitos que van desde procesamiento de datos hasta tareas de aprendizaje automático o *machine learning*. Otra de las características que ya se han mencionado es la velocidad, el realizar sus procesos en memoria hace que estos trabajos sean mucho más rápidos que cuando se trabajan en disco, así se realicen de forma distribuida como lo hace Hadoop. Finalmente, encontramos el soporte de múltiples lenguajes de programación entre los cuales se destaca Java, Python, R y Scala.

5.3. Técnicas de Análisis de Datos

Existe una gran variedad de herramientas de análisis de datos que nos permiten obtener un mayor aprendizaje de un conjunto de datos. Sin embargo, para este proyecto profundizaremos en dos que corresponden a lo implementado en la aplicación.

5.3.1. Frecuencia de Palabras

La frecuencia de palabras o análisis cuantitativo de texto es una herramienta nos permite identificar la diversidad de palabras que hay en un texto según su frecuencia de uso y su tipo. Esta técnica fue inicialmente utilizada para desencriptar textos cifrados identificando los patrones de las letras usadas en un texto encriptado y sustituyéndolas con otras con el fin de encontrar el significado del texto.

Para realizar este análisis se requiere inicialmente estandarizar y eliminar aquellas palabras que no le dan un valor agregado al análisis como por ejemplo las palabras conectoras, signos de puntuación, Etc. De esta manera garantizamos que el resultado nos permita identificar las palabras más frecuentes y obtener un mayor nivel de entendimiento del texto que se está analizado.

Existen varias formas de visualizar este tipo de análisis entre las más comunes están los diagramas de barras en donde tenemos un diagrama de dos ejes x, y en donde uno de los ejes contiene las palabras y el otro la frecuencia de dicha palabra.

Otra forma de visualización también muy utilizada es la nube de palabras, que consiste en un gráfico en donde aparecen las palabras más frecuentes en un texto y en donde la medición se identifica dado que el tamaño y en algunos casos el color que tiene la palabra sea proporcional a la frecuencia que tiene la palabra en el texto. Un ejemplo de esto lo podemos ver en la Figura 5.

Figura 5. Ejemplo Nube de Palabras

5.3.2. Análisis de Sentimientos

Un sentimiento corresponde en este caso a la opinión o actitud de un autor acerca de un tema en específico. Análisis de sentimiento o minería de opiniones es entonces, extraer la opinión, sentimiento o preferencias del usuario a partir de un texto subjetivo, en pocas palabras, detectar la polaridad de un texto, esta puede ser positiva, negativa o neutral (Rodrigues, Chiplunkar, & Rao, 2016).

El análisis de sentimientos busca establecer el tono emocional con el cual se escriben los *tweets* obtenidos, esto permite que nos hagamos una idea del sentimiento que generan los *tweets* relacionados con la palabra o frase consultada. Para ello se le aplica la polaridad, a través de la cual se clasifica el mensaje en función de la intención que tenga el autor al realizarlo.

Este tipo de análisis es comúnmente realizado con la información generada a partir de las redes sociales debido a que la mayoría de la información recopilada corresponde a opiniones esto genera un gran reto que implica almacenar y procesar el gran volumen de información generado, en este caso Twitter. Durante la investigación de este tema se encontraron muchos estudios donde se aplica diferentes técnicas de *Big Bata* para realizar análisis de sentimiento para diferentes finalidades.

5.3.3. Clusterización de Palabras

El proceso de clusterización de Palabras consiste principalmente en encontrar grupos cuyos elementos sean similares entre sí y diferentes a los elementos de otros grupos. Dichos grupos son conocidos como clústers (Vicente, 2005).

El algoritmo K-Means corresponde a una de las metodologías de clusterización basado en procesamiento no supervisado, este tipo de algoritmos buscan patrones entre los datos sin tener una predicción u otro tipo análisis de estos. Fue propuesto por MacQueen en el año 1968 y es comúnmente usada para encontrar rápidamente una solución. Este algoritmo es simple, directo y se basa en análisis de varianzas (Villagra, 2009).

El resultado de K-means depende de la cantidad de grupos inicial que se escojan. es por esto que para seleccionar el numero correcto de grupos se debe ejecutar varias veces el proceso con diferentes valores de grupos y con los resultados que estos generen se define cual es el número de Clústers óptimo para el conjunto de datos (Cuell, 2009).

La principal desventaja del algoritmo de K-means es que el clúster resultante es sensible a la selección inicial de los centroides y puede converger a un óptimo local. Por lo tanto, la selección inicial de los centroides dirige el proceso de K-means y la partición resultante está condicionada a la elección esos centroides (Vicente, 2005).

6. Estado del Arte

En el proceso de investigación de este proyecto se encontraron diferentes proyectos relacionados con elaboración de procesos de análisis haciendo uso de diferentes herramientas que podemos ver en la Tabla 5.

Tabla 5. Proyectos Análisis y Procesamiento de Datos

Cita	Herramientas Utilizadas	Fuente de Datos	Tema Analizado	Parametrizable
(Cuell, 2009)	Clusterización de Palabras	Datos Climatológicos	Clima	NO
(Bhangle & Krishnan, 2018)	Minería de Datos Análisis de Sentimientos	Twitter	Deportes	NO
(Dutta, Sharma, Natani, Khare, & Singh, 2017)	Análisis de Sentimientos	Twitter	Industria Aérea	NO
(Karau, Konwinski, Wendell, & Zaharia, 2015)	Hadoop Spark	Twitter	General	NO
(SAURA, 2018)	*Machine Learning* Análisis de Sentimientos	Twitter	Comercial	NO
(Congosto, 2011)	Analítica de Grafos	Twitter	Política	NO
(Abellán, 2012)	Estadística	Twitter	Política	NO
(Nabel, 2010)	Estadística	Twitter	Política	NO
(López-García, 2015)	Estadística	Twitter	Política	NO

En cada uno de estos proyectos se identifica la importancia que tiene el análisis de los datos para diferentes ámbitos investigativos y sociales.

En el caso de (Cuell, 2009) se puede identificar como utiliza datos climatológicos de Canadá para probar una teoría en la cual se pueden realizar análisis con una reducción de datos de procesamiento y realizando agrupaciones sobre los mismos.

Debido al impacto que tienen las redes sociales en el mundo, se hace cada vez más necesario poder entender y analizar dicha información y por eso en varios de los proyectos encontrados se hacen uso de diferentes herramientas de análisis sobre datos de publicaciones de redes sociales.

Los autores (Bhangle & Krishnan, 2018) en su artículo *Twitter Sentimental Analysis on Fan Engagement* exponen como objetivo analizar a partir de diferentes técnicas de *machine learning* como los fans reaccionan hacia diferentes deportes. En su investigación se realiza todo el proceso de minería de texto y preprocesamiento de la información para posteriormente utilizar esta salida como entrada para los algoritmos de máquinas de soporte vectorial y arboles de decisión, de esta manera obtener las clasificaciones en la polaridad de los tweets.

Otros autores como (Dutta, Sharma, Natani, Khare, & Singh, 2017) en su artículo *"Sentimental Analysis for Airline Twitter data"* tienen como objetivo proveer a la industria aérea de un análisis que le permita conocer el grado de satisfacción de sus clientes, esto a partir de la extracción de *tweets*, pre procesado de los mismos y clasificar los sentimientos a través de un algoritmo bayesiano con herramientas como R y Rapidminer.

Así mismo (SAURA, 2018) es otro de los autores que en su artículo *"Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de #BlackFriday"* hace uso de análisis de sentimientos para analizar los *tweets* basados en una de las tendencias "*BlackFriday*" relacionada con un periodo de tiempo en el cual se generan diferentes ofertas en el comercio de diferentes partes del mundo.

(Congosto, 2011) es otro autor que utilizo la herramienta analítica de grafos para realizar análisis de la actividad de la red social frente a las elecciones presidenciales del año 2011 en

España. En este articulo el autor concluye con la importancia de este tipo de análisis para identificar la opinión política en un proceso electoral.

Así como (Congosto, 2011) los artículos de los autores (Abellán, 2012), (Nabel, 2010) y (López-García, 2015) buscaron hacer uso de la información generada de la red social Twitter para realizar análisis de opinión política en los procesos electorales de diferentes países. La diferencia con estos autores es que hicieron uso de herramientas estadísticas para establecer de qué forma impactaban las publicaciones de la red social en las definiciones políticas que se llevaron a cabo en diferentes países.

Varios autores han utilizado diferentes metodologías para realizar procesamiento de datos de Twitter. Este es el caso de (Karau, Konwinski, Wendell, & Zaharia, 2015) quienes en su libro "*Learning spark: lightning-fast big data análisis"* explican de qué manera hacer procesamiento de *tweets* a través de un clúster hadoop. Sin embargo, este tipo de herramientas deben ser ejecutadas por alguien que tenga conocimientos específicos en este tipo de tecnologías, lo que dificulta que personas sin estas habilidades puedan realizar este tipo de análisis fácilmente.

En todos los proyectos expuestos anteriormente se puede identificar la forma en que con diferentes mecanismos se busca generar un mejor entendimiento de la opinión registrada por los usuarios para diferentes objetivos. Hoy en día existe una necesidad cada vez más grande de involucrar análisis de redes sociales en la mayoría de los estudios, proyectos o trabajos en los cuales se requiera obtener un entendimiento de la opinión que se registra en las redes sociales.

Este proyecto pretende generar una herramienta estandarizada que habilite la elaboración de diferentes tipos análisis de la información en tiempo real generada a través de la red social Twitter sobre cualquier ámbito de interés para cualquier usuario. Incorporando el componente del clúster

que permite generar una mayor escalabilidad de manera que se puedan generar procesos que requieran mayor nivel de procesamiento.

7. Desarrollo

Para el desarrollo de este proyecto se realiza una aplicación que permita a los usuarios realizar análisis de *tweets* sobre una palabra o frase que quieran consultar.

7.1. Arquitectura General

La arquitectura utilizada para este proyecto está compuesta por tres elementos que son Aplicación Web, Servicio de Twitter y Clúster. En la Figura 6 se puede identificar la arquitectura de la aplicación utilizada para el desarrollo.

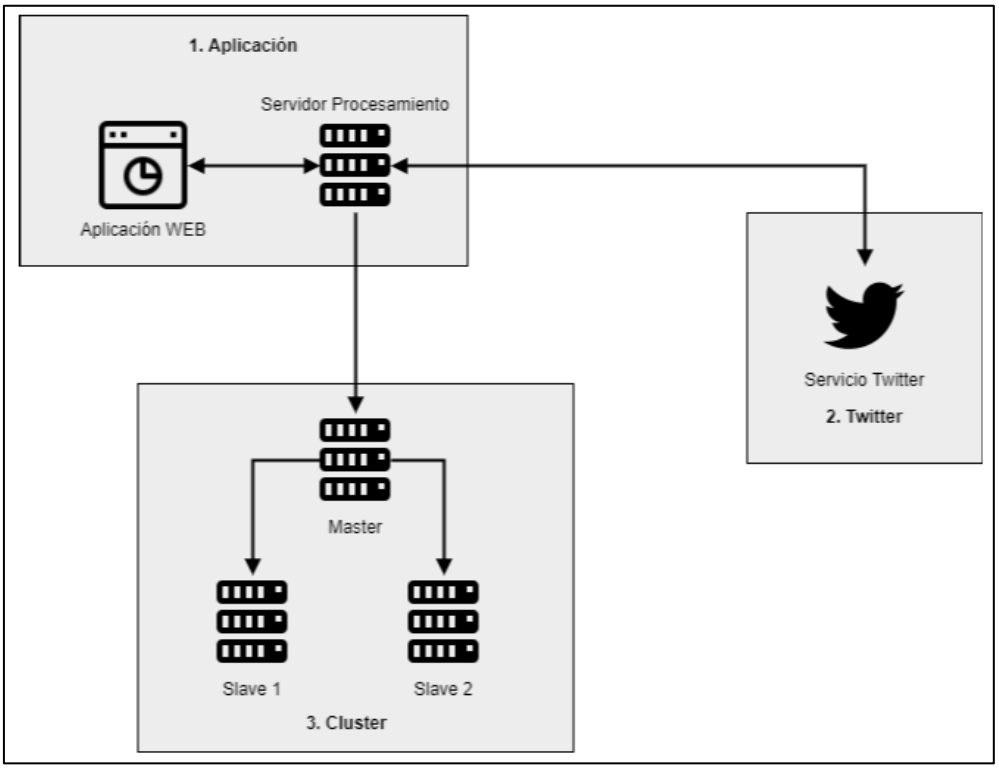


Figura 6. Diseño de arquitectura de aplicación para generación de análisis de datos obtenidos de Twitter

7.1.1. Aplicación

El desarrollo de la aplicación se realizó en lenguaje Python que es un lenguaje de programación flexible que cuenta con una amplia comunidad (sin ánimo de lucro) que se encarga de enriquecerlo y debido a esto tiene librerías integradas para funcionalidades o uso de otras herramientas relacionadas con análisis y procesamiento de datos (Rochina, 2019).

Se realizó una aplicación Web en lenguaje Python 3.0 haciendo uso del framework Django 2.2.4 que permite combinar las ventajas ofrecidas por Python con un entorno Web y de fácil acceso. Para el desarrollo de la aplicación se utilizaron las librerías de Python resaltadas en la Tabla 6.

Tabla 6. Librerías instaladas para el desarrollo de la aplicación WEB

Librería	versión	Función
Arrow	0.15.5	Manejo de fechas y tiempo para Python
Asn1crypto	1.3.0	Serialización de llaves privadas, publicas, certificados, CRL, OCSP, CMS, PKCS#3, PKCS#7, PKCS#8, PKCS#12, PKCS#5, X.509 and TSP
Astroid	2.3.3	Impulsa módulos Pylint
Basemap	1.2.0	Herramienta para crear mapas usando Python.
Blinker	1.4	Manejo de eventos o señales
Certifi	2020.4.5.1	Complemento de la librería Requests para validación de certificados SSL y validación de identidad TLS
Cffi	1.14.0	Interfaz para ejecutar código C
Chardet	3.0.4	Codificación de caracteres
Classifier	2	Herramienta para organizar archivos en directorios
Click	7.1.1	Herramientas de interfaz de línea de comandos
Colorama	0.4.3	Secuencia de caracteres bajo el estándar ANSI para producir una terminal de texto y posicionamiento de cursor en windows

Contextlib2	0.5.5	Utilidades para tareas comunes
Cryptography	2.8	Herramienta para encriptar y desencriptar mensajes
Cycler	0.10.0	Complemento de la librería Matplotlib para la creación de gráficos.
Django	2.2.5	Web framework de Python
Docutils	0.16	Procesamiento de documentación en diferentes formatos
Flask	1.1.1	Web Server Gateway Interface para un framework de aplicación web
Geographiclib	1.5	Conversiones entre coordenadas cartesianas geográficas, UTM, UPS, MGRS, geocéntricas y locales
Geopy	1.21.0	Localiza las coordenadas de direcciones, ciudades, países y puntos de referencia en un mapa del mundo
Idna	2.9	Soporte para los nombres de dominio internacionalizados en aplicaciones
Isort	4.3.21	Ordena las librerías importadas de Python
Itsdangerous	1.1.0	Herramienta para pasar datos en entornos no confiables
Jinja2	2.11.1	Proporciona una sintaxis similar a Django y compila plantillas en código Python
Joblib	0.14.1	Conjunto de herramientas para proporcionar canalización ligera en Python
Kiwisolver	1.2.0	Implementación de C++ en código Python
Lazy-Object-Proxy	1.4.3	Un proxy de objetos
Load	2019.4.13	Carga e inicializa un módulo implementado como un archivo fuente de Python
Lockfile	0.12.2	Provee una API para bloquear archivos
Lxml	4.5.0	Da acceso a las librerías libxslt y libxml2 para manejo de archivos Excel
Markupsafe	1.1.1	Implementa un objeto de texto que valida caracteres por lo que es seguro de usar en HTML y XML
Matplotlib	3.1.3	Crea visualizaciones estáticas, animadas e interactivas
Mccabe	0.6.1	Genera alertas en un archivo
Mkl-Fft	1.0.15	Conjunto de rutinas de matemática vectorizadas que trabajan para acelerar diversas funciones y aplicaciones matemáticas
Mkl-Random	1.1.0	
Mkl-Service	2.3.0	

Nltk	3.5	Librería para procesamiento de lenguaje natural
Numpy	1.18.1	Manejo de arrays multidimensionales para procesamiento de grandes volúmenes de datos y funciones de algebra e integración de C/C++
Oauthlib	3.1.0	Implementa la logia de protocolos de autenticación OAuth1 o OAuth2
Pandas	1.0.3	Proporciona estructuras de datos rápidas, flexibles y expresivas diseñadas para trabajar con series estructuradas (tabulares, multidimensionales, potencialmente heterogéneas) y series de tiempo.
Pillow	7.0.0	Carga y Procesamiento de imágenes
Proj	0.1.0	Administración de carpetas en línea de comandos para archivar y restaurar proyectos
Public	2019.4.13	Reemplaza textos con decoradores
Py4j	0.10.7	Acceso a objetos Java
Pycparser	2.2	Es un analizador completo del lenguaje C
Pylint	2.5.0	Identificación de errores en código Python
Pyparsing	2.4.7	Permite crear y ejecutar gramáticas simples
Pyproj	2.6.0	Interfaz de Python para proyecciones cartográficas
Pyshp	2.1.0	Lee y escribe Shapefiles
Pyspark	2.4.5	API de Python para Spark.
Python-Dateutil	2.8.1	Extensión del módulo datetime de Python
Pytz	2019.3	Cálculos de Zona horaria y multiplataforma
Regex	2020.5.7	Manejo de expresiones regulares.
Requests	2.23.0	Peticiones HTTP (Get, Put, Post, Etc.)
Requests-Oauthlib	1.3.0	Habilita flujos de autorización para sitios web
Scikit-Learn	0.22.2.1	Módulos de Machine Learning y Minería de Datos
Scipy	1.4.1	Funciones matemáticas e ingeniería
Sip	4.19.13	Generador de enlaces de Python para bibliotecas C/C++
Six	1.14.0	Librería de compatibilidad entre Python 2 y 3
Sklearn	0	Módulos de Machine Learning y Minería de Datos
Sqlparse	0.3.0	Analiza, divide y formatea sentencias SQL

Textblob	0.15.3	Procesamiento de texto y análisis de sentimientos
Toml	0.10.0	Analiza y crea TOML
Tornado	6.0.4	Web framework de Python para manejo de redes asíncronas
Tqdm	4.46.0	Implementación de barra de progreso en línea de comandos
Twython	3.8.2	Wrapper de API de Twitter
Typed-Ast	1.4.1	Analizador de librerías para Python 2 y Python 3
Unidecode	1.1.1	Convierte texto Unicode y lo representa en caracteres ASCII
Urllib3	1.25.9	Cliente HTTP/FTP
Werkzeug	1.0.0	Interfaz de puerta de enlace del servidor web
Wincertstore	0.2	Verifica certificados TLS/SSL
Wordcloud	1.7.0	Genera nubes de palabras en Python
Wrapt	1.12.1	Proxy para Python

7.1.2. Clúster

Para la configuración, administración y funcionamiento de los clústers existen diferentes herramientas. Para el objetivo de este proyecto se configuro un clúster con Hadoop versión 2.9.2 que está conformada por dos máquinas virtuales (master y slave) con las características de la Tabla 7.

Tabla 7. Características Maquinas en el Clúster

Master	Memoria: RAM 2GB Disco Duro: 50GB Sistema Operativo: Linux – Ubuntu 18.04 Java 11
Slave	Memoria: RAM 2GB

	Disco Duro: 50GB Sistema Operativo: Linux – Ubuntu 18.04 Java 11

7.1.3. Twitter

Para obtener los *tweets* se creó una cuenta con permisos de desarrollador en la aplicación Twitter. Los permisos de desarrollador sobre la cuenta permiten generar unas llaves con las que haciendo uso de la librería Twython de Python se pueden hacer consultas de los *tweets* asociados a los seguidores que tenga la cuenta creada.

Nota: *Aunque la configuración realizada está asociada a una cuenta creada para este proyecto dicha configuración es fácilmente modificable para que se pueda asociar a cualquier otra cuenta.*

7.2. Configuración Clúster

En este punto se presentarán todas las actividades necesarias para la configuración y puesta en marcha del clúster con Hadoop y Spark.

7.2.1. Configuración Inicial Maquinas Master – Slave

En este capítulo se utilizan las dos máquinas virtuales creadas previamente con las características mencionadas en el capítulo de arquitectura y se realiza la preparación previa que consiste en la configuración de usuarios y red, así como también la instalación de actualizaciones del sistema operativo y software requerido para las configuraciones de seguridad y clúster.

Nota: *Todas las configuraciones a partir de este capítulo se realizan para las dos máquinas master y slave y en los puntos diferenciadores se separarán para identificar las diferencias de configuración.*

- Asignación Password Usuario Root.

```
sudo passwd root
```

- Actualización Ubuntu

```
sudo apt-get update
```

- Instalación de y configuración Java

```
sudo apt-get install default-jdk
ls -l /etc/alternatives/java
sudo nano /etc/environment
. /etc/enviroment
```

- Instalación Hadoop

```
Hadoop -h
Hadoop version
```

- Instalación Python

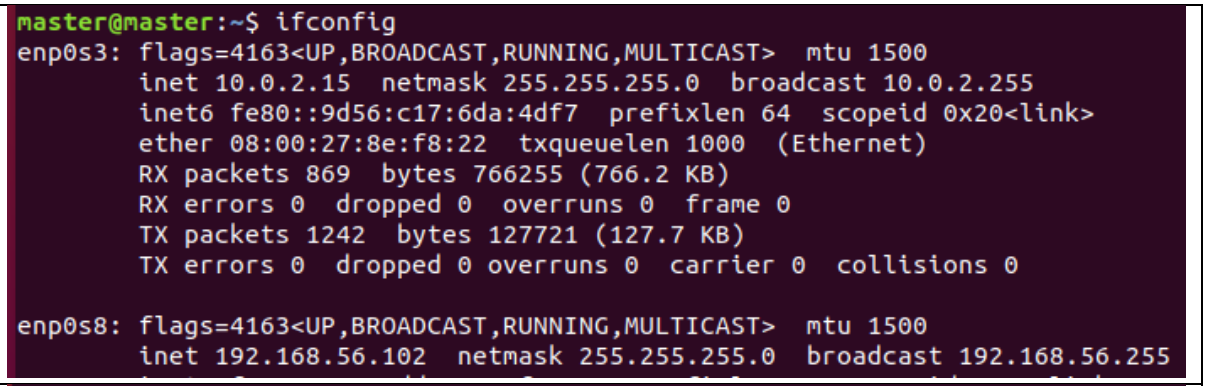
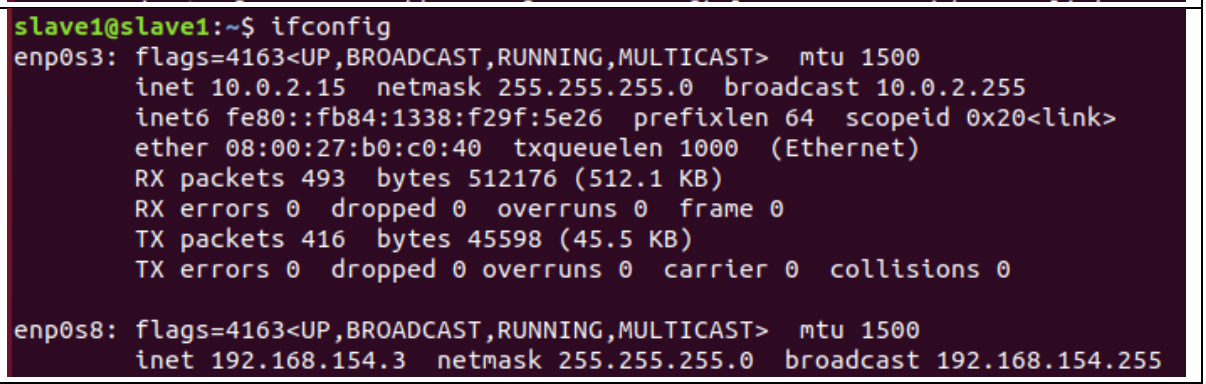
```
sudo apt-get install python
```

- Configuración IP maquinas master y slave1

```
sudo apt-get install python
ifconfig
ping master
ping slave1
```

Con estos comandos se puede validar la configuración IP de cada una de las máquinas y modificando el archivo host y agregando las direcciones IP de cada una de las maquinas se puede validar la conectividad entre ellas.

En la Figura 7 y Figura 8 podemos ver la configuración IP que se llevó acabo para cada una de las maquinas.

MASTER	
SLAVE1	

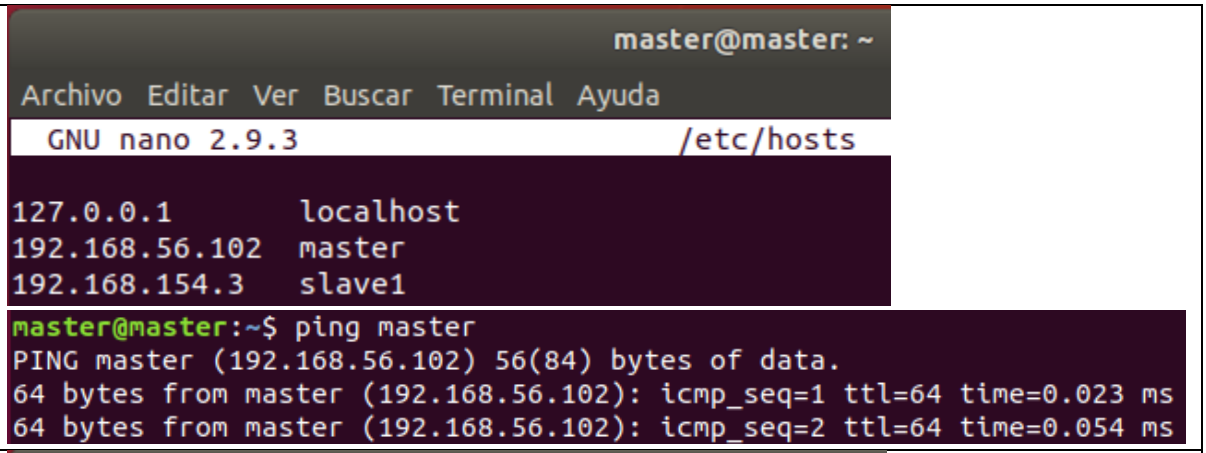
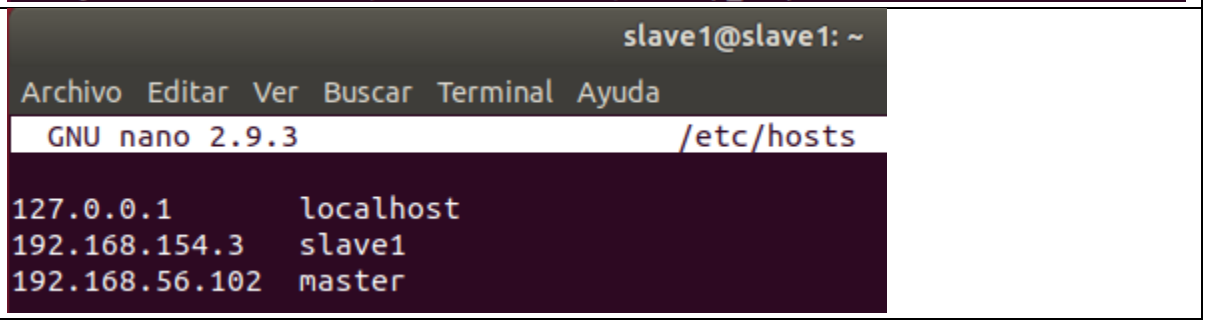
```
master@master:~$ ifconfig
enp0s3: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 10.0.2.15  netmask 255.255.255.0  broadcast 10.0.2.255
        inet6 fe80::9d56:c17:6da:4df7  prefixlen 64  scopeid 0x20<link>
        ether 08:00:27:8e:f8:22  txqueuelen 1000  (Ethernet)
        RX packets 869  bytes 766255 (766.2 KB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 1242  bytes 127721 (127.7 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

enp0s8: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.56.102  netmask 255.255.255.0  broadcast 192.168.56.255
```

```
slave1@slave1:~$ ifconfig
enp0s3: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 10.0.2.15  netmask 255.255.255.0  broadcast 10.0.2.255
        inet6 fe80::fb84:1338:f29f:5e26  prefixlen 64  scopeid 0x20<link>
        ether 08:00:27:b0:c0:40  txqueuelen 1000  (Ethernet)
        RX packets 493  bytes 512176 (512.1 KB)
        RX errors 0  dropped 0  overruns 0  frame 0
        TX packets 416  bytes 45598 (45.5 KB)
        TX errors 0  dropped 0 overruns 0  carrier 0  collisions 0

enp0s8: flags=4163<UP,BROADCAST,RUNNING,MULTICAST>  mtu 1500
        inet 192.168.154.3  netmask 255.255.255.0  broadcast 192.168.154.255
```

Figura 7. Tabla de configuración IP master y slave1

MASTER	
SLAVE1	

```
master@master: ~
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
  GNU nano 2.9.3                    /etc/hosts

127.0.0.1       localhost
192.168.56.102  master
192.168.154.3   slave1
```

```
master@master:~$ ping master
PING master (192.168.56.102) 56(84) bytes of data.
64 bytes from master (192.168.56.102): icmp_seq=1 ttl=64 time=0.023 ms
64 bytes from master (192.168.56.102): icmp_seq=2 ttl=64 time=0.054 ms
```

```
slave1@slave1: ~
Archivo  Editar  Ver  Buscar  Terminal  Ayuda
  GNU nano 2.9.3                    /etc/hosts

127.0.0.1       localhost
192.168.154.3   slave1
192.168.56.102  master
```

```
slave1@slave1:~$ ping slave1
PING slave1 (192.168.154.3) 56(84) bytes of data.
64 bytes from slave1 (192.168.154.3): icmp_seq=1 ttl=64 time=0.028 ms
64 bytes from slave1 (192.168.154.3): icmp_seq=2 ttl=64 time=0.051 ms
```

Figura 8. Configuración hosts y conectividad entre master y slave1

7.2.2. Configuración de SSH

- Generación de llave SSH Master y Slave1

```
ssh-keygen
cd /home/master/.ssh
chmod 0600 authorized_keys
```

- Configuración SSH en master

```
cp id_rsa.pub authorized_keys
scp authorized_keys slave1@slave1:/home/slave1/.ssh
```

- Configuración SSH en Slave1

```
cat id_rsa.pub >> authorized_keys
cat authorized_keys
scp authorized_keys master@master:/home/master/.ssh
```

Una vez realizados estos pasos con el siguiente comando se puede validar la conectividad por medio de SSH de los dos servidores.

```
ssh slave1@slave1
ssh master@master
```

En la Figura 9 se puede identificar la comunicación a través del protocolo SSH una vez realizados los pasos anteriormente mencionados.

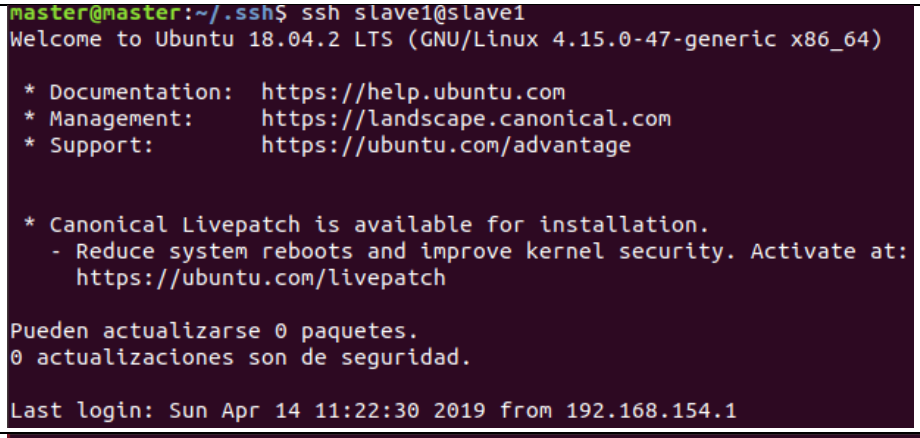
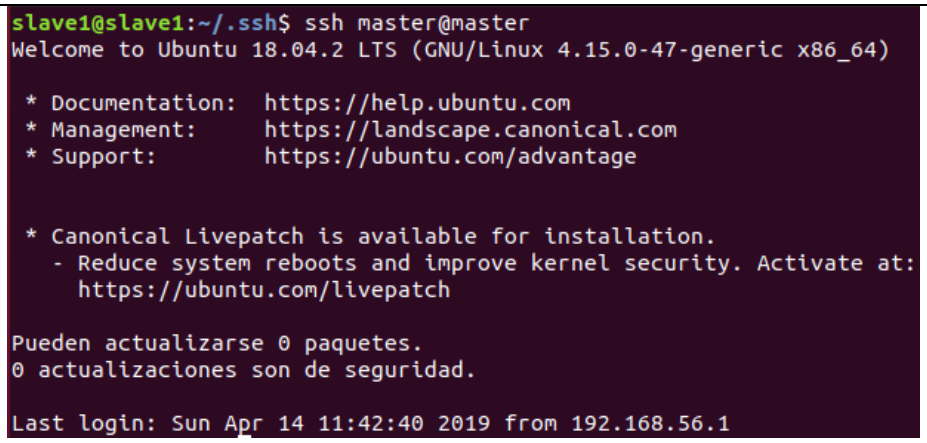
MASTER	
SLAVE1	

Figura 9. Validación de conectividad SSH master y slave1

7.2.3. Instalación Hadoop

El clúster este compuesto por una máquina master y un slave. La máquina Maestra o master es la encargada de administrar la configuración, los recursos y las tareas que son distribuidas a cada una de las maquinas Esclavas o slave.

Las maquinas slave son las encargadas de almacenar y procesar las tareas enviadas.

- Descarga e Instalación de Hadoop

  La descarga para la instalación de Hadoop se llevó a cabo desde el portal principal de la aplicación: https://Hadoop.apache.org/releases.html.

- Descomprimir y ubicar los directorios de Hadoop

```
sudo mkdir Hadoop
sudo tar xvf /home/master/Descargas/Hadoop-2.9.2.tar.gz
sudo mv Hadoop-2.9.2/* Hadoop
su -
```

- Configuración Hadoop

```
gedit .bashrc
```

En el archivo .bashrc se configuran la ubicación de los directorios de java y Hadoop.

En este archivo es donde se configuran las bariables de entorno sober las cuales se va a ejecutar el servidor. Esta configuracion de varibles permite que en cualquier ejecucion de cualquier sistema se puedan utilizar las valores configurados para estas variables.

```
export JAVA_HOME=/usr/lib/jvm/java-11-openjdk-amd64
export HADDOP_HOME=/opt/Hadoop
export PATH=$PATH:/$HADOOP_HOME/bin:/$HADOOP_HOME/sbin
```

Una vez modificados los valores se actualizan las variables con el siguiente comando.

```
. ./.bashrc
```

En el archivo slaves se deben configurar las maquinas que estarán conectadas en el Clúster. Para esto se deben agregar nos nombres de la maquinas que harán parte del

```
sudo nano /opt/Hadoop/etc/Hadoop/slaves
```

Clúster en en srchivo slaves del directorio Hadoop.

- Configuración de la ubicación de la maquina master.

```
sudo nano /opt/Hadoop/etc/Hadoop/core-site.xml
```

```
<configuration>
     <property>
          <name>fs.defaultFS</name>
          <value>hdfs://master:9000</value>
     </property>
</configuration>
```

Nota: *Este archivo se debe configurar para master y slave1*

- Configuración HDFS Replicación, Directorios de NameNode y DataNode

Se crean y otorgan los permisos de lectura y escritura sobre los directorios que se van a configurar en el archivo hdfs-site.xml como directorio namenode y datanode.

```
su - root
cd /
mkdir datos
cd /datos/
mkdir datanode
mkdir namenode
chown -R master:master datos
```

En las siguientes imágenes se puede ver como se realizan estas configuraciones.

```
sudo nano /opt/Hadoop/etc/Hadoop/hdfs-site.xml
```

El archivo se configura de la siguiente manera:

Por último, se da formato a los directorios configurados anteriormente creando los archivos

```
<configuration>
      <property>
            <name>dfs.replication</name>
            <value>1</value>
      </property>
      <property>
            <name>dfs.namenode.name.dir</name>
            <value>datos/namenode</value>
      </property>
      <property>
            <name>dfs.datanode.name.dir</name>
            <value>datos/datanode</value>
      </property>
      <property>
      <name>dfs.namenode.datanode.registration.ip-hostname.check</name>
            <value>false</value>
      </property>
</configuration>
```

necesarios para la ejecución del clúster.

```
Hadoop namenode -format
```

- Una vez terminadas las configuraciones se realiza una prueba de ejecución del clúster Hadoop. Dentro de las validaciones para tener en cuenta se valida la cantidad de nodos habilitados en el clúster. Se pueden ver en la Figura 10 y Figura 11.

```
Hadoop start-dfs.sh
Jps
```

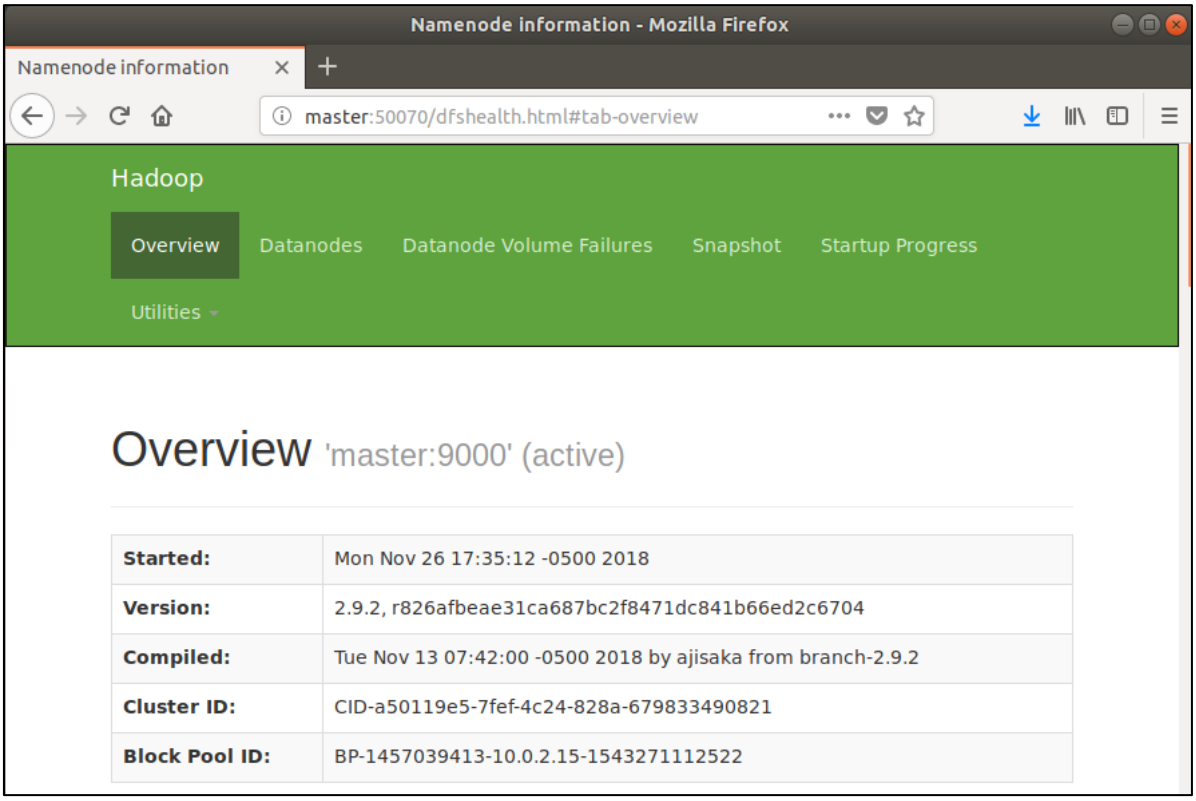


Figura 10. Prueba de ejecución de clúster.

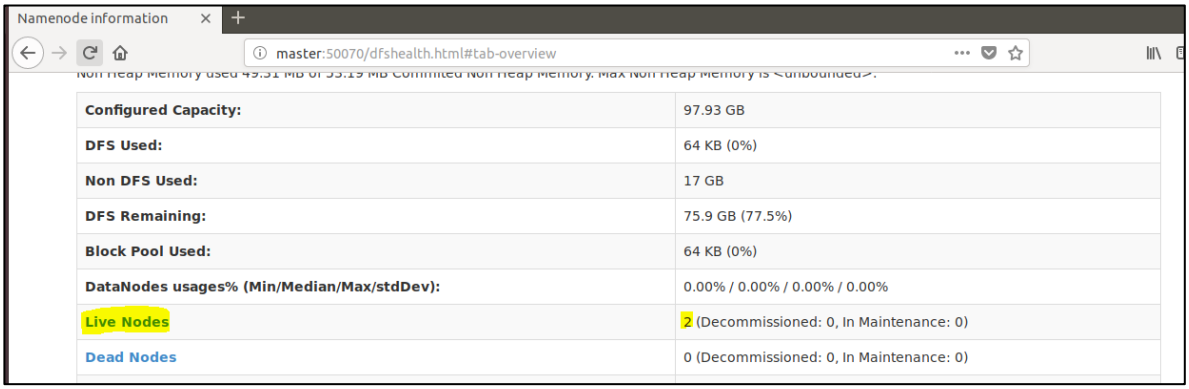


Figura 11. Nodos habilitados en el clúster

7.2.4. Configuración YARN

El proceso YARN es el encargado de la distribución de los trabajos y la gestión de los recursos del Clúster. Este cuenta con dos componentes que son el ResourceManager que arbitra los recursos

entre todos los nodos y los NodeManager que administran los recursos de cada nodo independientemente.

En la Figura 12 se puede observar el proceso que realizan los componentes:

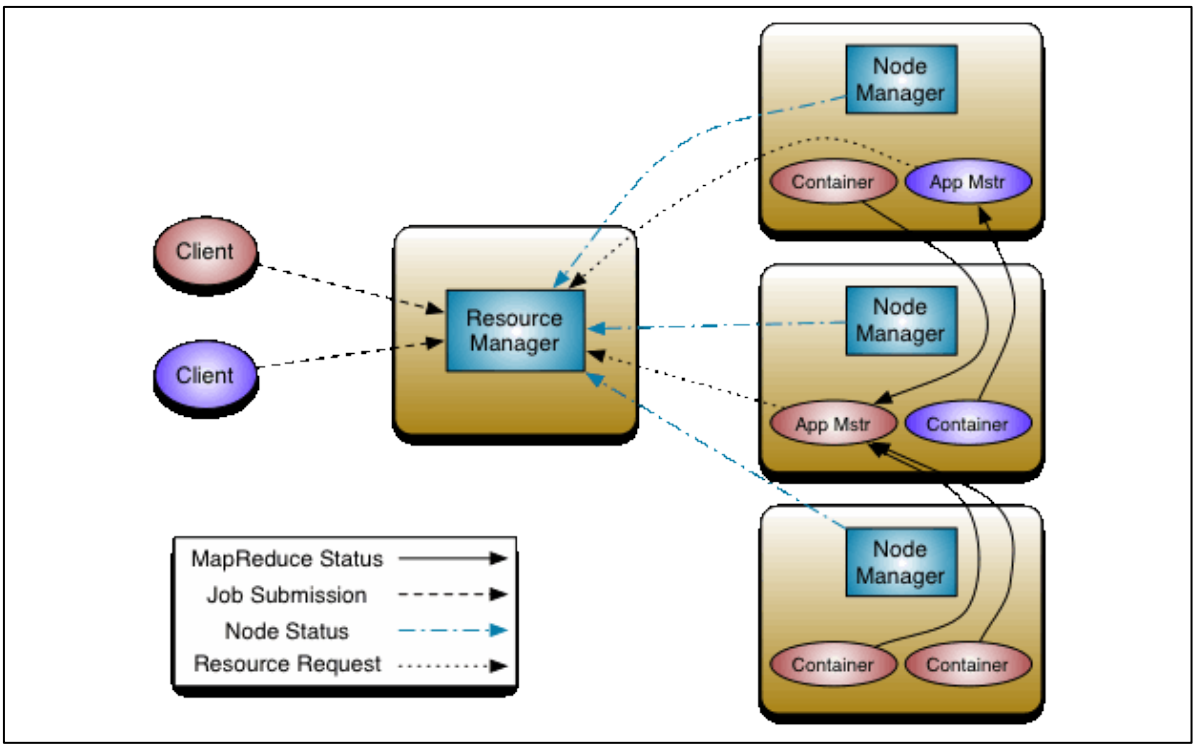


Figura 12. Proceso YARN Clúster Hadoop *(Apache Hadoop, 2019)*

A continuación, en las maquinas master y slave se debe configurar el archivo mapred-site.xml en el cual se realiza la configuración del proceso YARN.

```
<configuration>
      <property>
            <name>mapreduce.framework.name</name>
            <value>yarn</value>
      </property>
</configuration>
```

Se debe realizar la configuración del archivo yarn-site.xml que contiene la configuración de las propiedades sobre las cuales se va a ejecutar el proceso YARN.

```
<configuration>
        <property>
                <name>yarn.resourcemanager.hostname</name>
                <value>yarn</value>
        </property>
        <property>
                <name>yarn.nodemanager.aux.services</name>
                <value>mapreduce_shuffle</value>
        </property>
        <property>
                <name>yarn.nodemanager.auxservices.mapreduce_shuffle.class</name>
                <value>org.apache.Hadoop.mapred.ShuffleHandler</value>
        </property>
</configuration>
```

Realizada la configuración se ejecuta el clúster e ingresando al dominio en la Figura 13 se identifican que los nodos configurados están activos para la ejecución de los procesos.

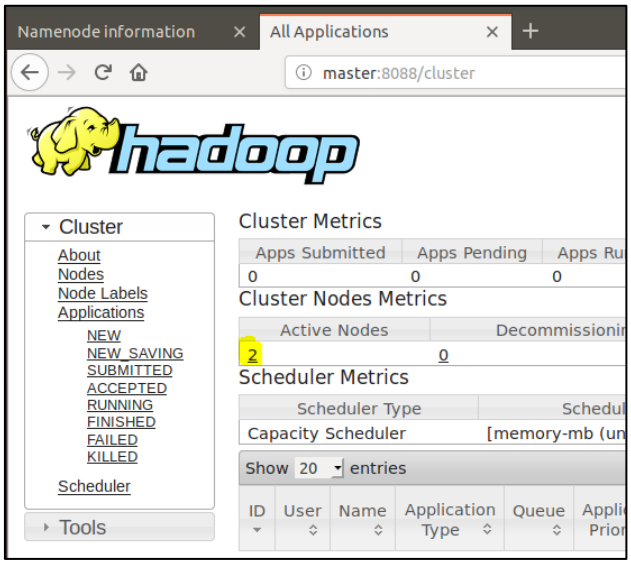


Figura 13. Identificación de nodos activos en clúster ejecutado

7.2.5. Instalación Spark

Spark es un framework de computación de clúster que habilita la ejecución de procesos por medio de diferentes fuentes y diferentes lenguajes y que a su vez permite dividir procesos para ejecutarlos en paralelo y de esta manera mejorar los tiempos de ejecución.

A continuación, se especifica el proceso de configuración que se llevó acabo para el desarrollo de este proyecto.

- Descarga y Configuración Inicial Spark

Se realiza la descarga de los instaladores de Spark desde la página oficial https://Spark.apache.org/downloads.html. Para este proyecto se utiliza la versión 2.4.1.

Cuando se descarga el componente se debe descomprimir y se ubica en la carpeta donde está instalado Hadoop.

```
tar xvf Spark-2.4.1-bin-without-Hadoop.tgz
mv /home/master/Descargas/Spark-2.4.1-bin-without-Hadoop /Spark
gedit .bashrc
```

Adicionalmente se debe configurar nuevamente el archivo .bashrc para incluir la ubicación de los archivos de Spark.

```
export
PATH=$PATH:/$HADOOP_HOME/bin:/$HADOOP_HOME/sbin:/opt/Hadoop/Spark
/bin:/opt/Hadoop/Spark/sbin
export SPARK_DIST_CLASSPATH=${Hadoop classpath}
export SPARK_HOME=/opt/Hadoop/Spark
export HADOOP_CONF_DIR=$HADOOP_HOME/etc/Hadoop
```

Realizada esta configuración se puede probar la ejecución de Python y Scala sobre el clúster como se puede ver en la siguiente tabla:

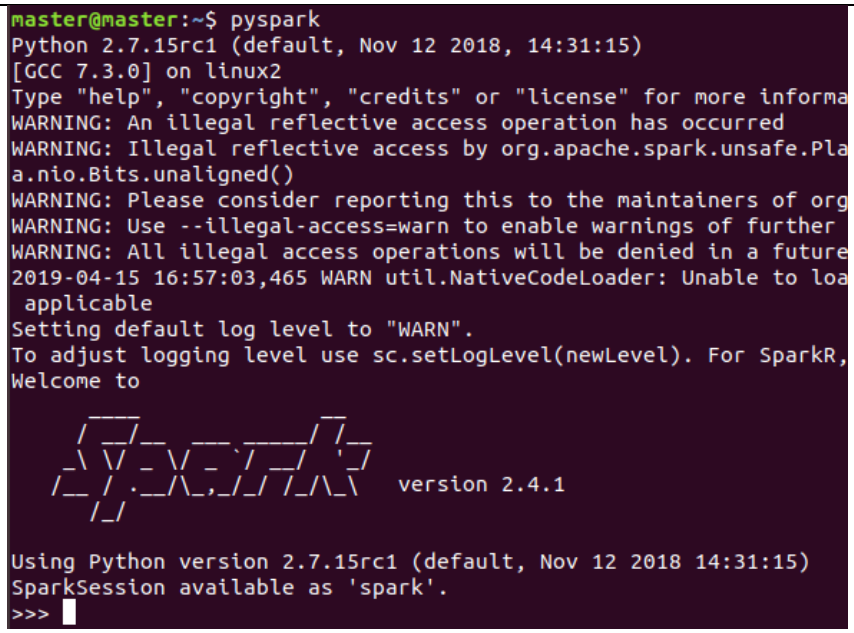
SCALA	
PYTHON	

Figura 14. Prueba de Scala y Python en clúster

7.3. Desarrollo de Aplicación

La aplicación permite al usuario generar diferentes herramientas de análisis consultando cualquier palabra o frase de interés. Las herramientas de análisis implementadas son las siguientes:

- Frecuencia de Palabras
- Nube de Palabras
- Análisis de Sentimientos
- Clusterización

- Geolocalización de *tweets*

Adicional a las funcionalidades de análisis el sistema debe contempla la obtención de *tweets* haciendo uso de los servicios de Twitter y la limpieza y estandarización de dichos *tweets*.

En los numerales a continuación se establecen los requerimientos y el proceso de desarrollo de la aplicación.

7.3.1. Requerimientos

Por medio de la Figura 15 y Figura 16 correspondiente a diagramas de casos de uso y secuencia respectivamente, se definen las funcionalidades que contempla la aplicación y de qué manera interactuara con cada uno de los componentes.

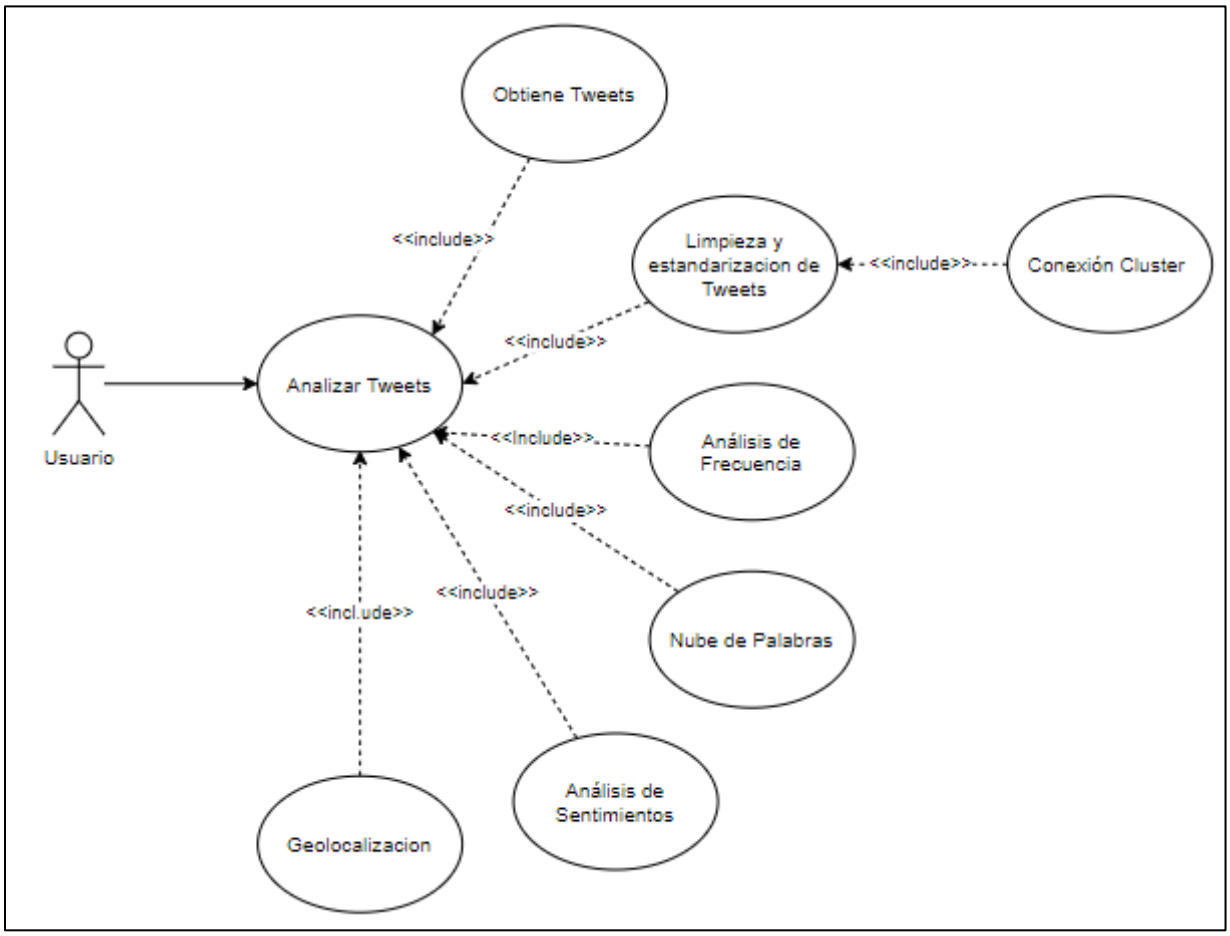


Figura 15. Diagrama de casos de uso aplicación

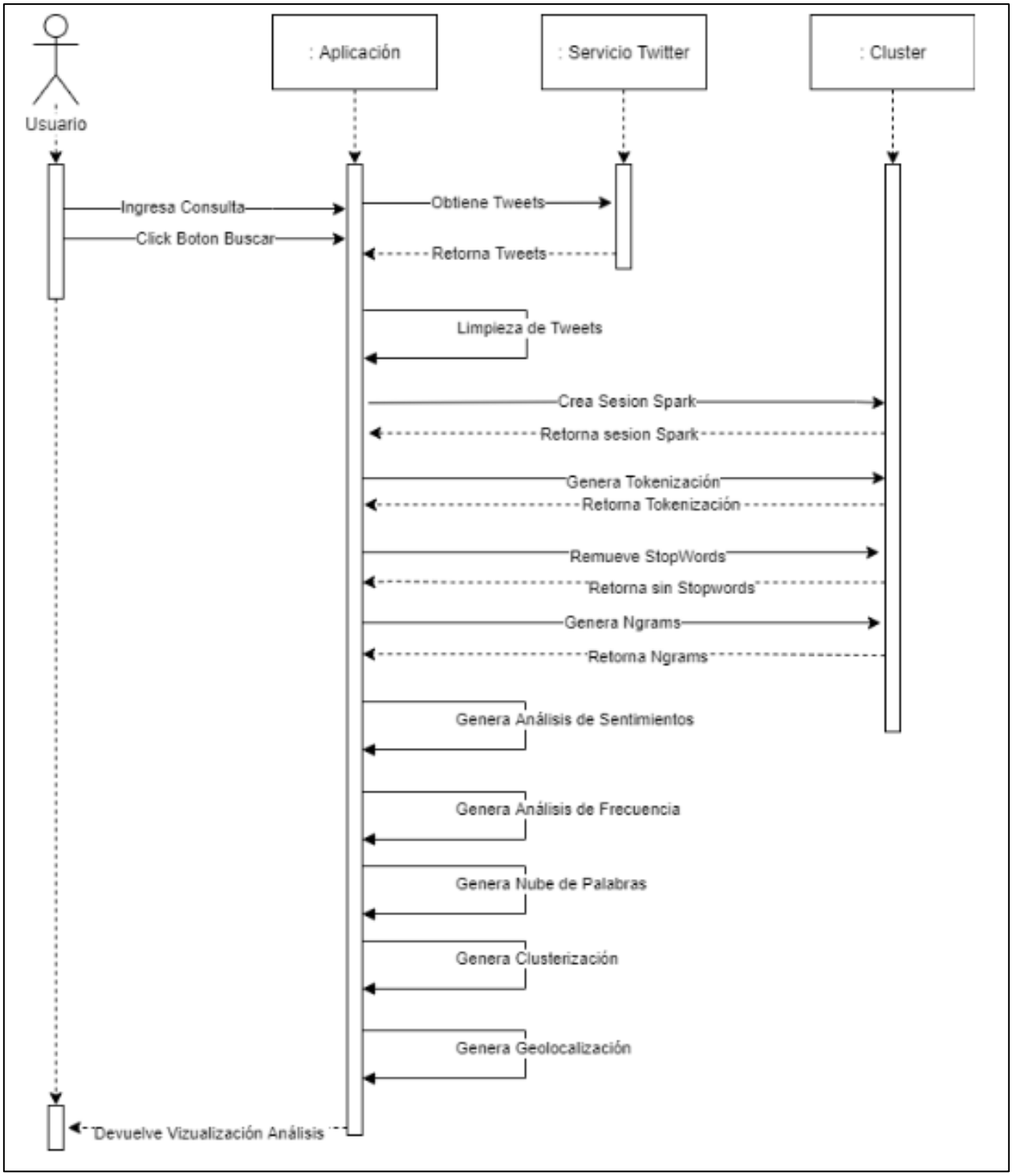


Figura 16. Diagrama de Secuencia de la aplicación

7.3.2. Creación de Aplicación

Se realiza la creación de una aplicación en lenguaje Python haciendo uso del framework Django.

Antes de la creación de la aplicación se requiere la creación de un ambiente de Python para garantizar un ambiente exclusivo y propio para la creación e instalación de las librerías necesarias para este proyecto. Para la creación de dicho ambiente se utiliza el siguiente código:

```
conda create -n Env_Proyecto_Web_Conda
```

Una vez creado el proyecto se realiza la instalación de las librerías necesarias para el proyecto, para lo cual se genera un archivo de texto con las librerías y la versión de cada una de estas y se procede a la instalación con ayuda de la librería freeze que nos permite realizar la instalación masiva por medio del archivo de texto.

```
pip install freeze
pip freeze > requirements.txt
```

Luego de esto se realiza la creación del proyecto y la aplicación en framework Django.

```
django-admin startproject Proyecto_Web_V3_Env2
python manage.py startapp core
```

Este código generara la siguiente estructura de archivos y carpetas:

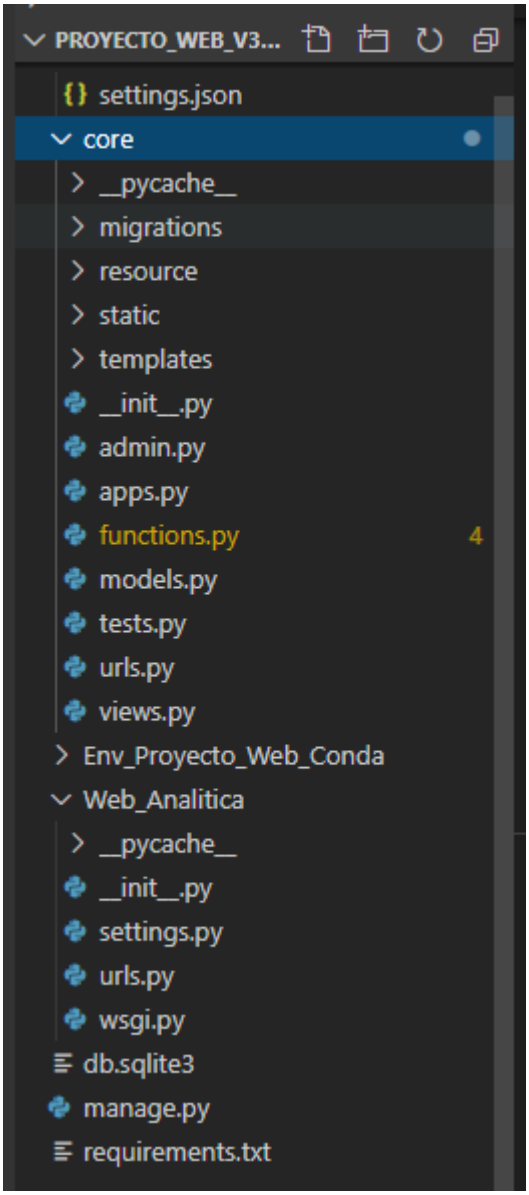


Figura 17. Estructura de Archivos y Carpetas aplicación Django

Una vez realizada la creación se puede ejecutar la aplicación haciendo uso del siguiente comando:

```
python manage.py runserver
```

7.3.3. Consulta y Limpieza de *Tweets*

Para la consulta de los *tweets* se requiere inicialmente de la configuración de una cuenta de Twitter con permisos de desarrollador. Para este proyecto se creó la cuenta @Analitic897 y a esta cuenta se solicitó a Twitter permisos de desarrollador y las credenciales necesarias para poder realizar consultas por medio de sus servicios. Sin embargo, el desarrollo realizado se puede utilizar con cualquier otra cuenta que tenga los permisos de consulta asociados cambiando las credenciales configuradas en el desarrollo

```
credentials = {}
credentials['CONSUMER_KEY'] = 'xxxxxxxxxxxxxxxxxx'
credentials['CONSUMER_SECRET'] = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
credentials['ACCESS_TOKEN'] = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx’
credentials['ACCESS_SECRET'] = 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx'
```

La librería de Python Twython nos permite haciendo uso de las credenciales consultar *tweets* relacionados con una palabra o frase de interés para el usuario. Como respuesta del servicio se obtiene una estructura de datos en lenguaje JSON de la cual se obtiene la información requerida para el desarrollo de las siguientes funcionalidades en un objeto de tipo Dataframe.

Para realizar los análisis de los *tweets* se requiere que la información estandarizada y a la cual se le deben realizar procesos de limpieza para obtener los datos que le den más valor a las herramientas de análisis generadas. Para este fin se hace uso de la conexión con el clúster y se utilizan las funciones de la librería Spark para los procesos de limpieza resaltados en la Figura 18.

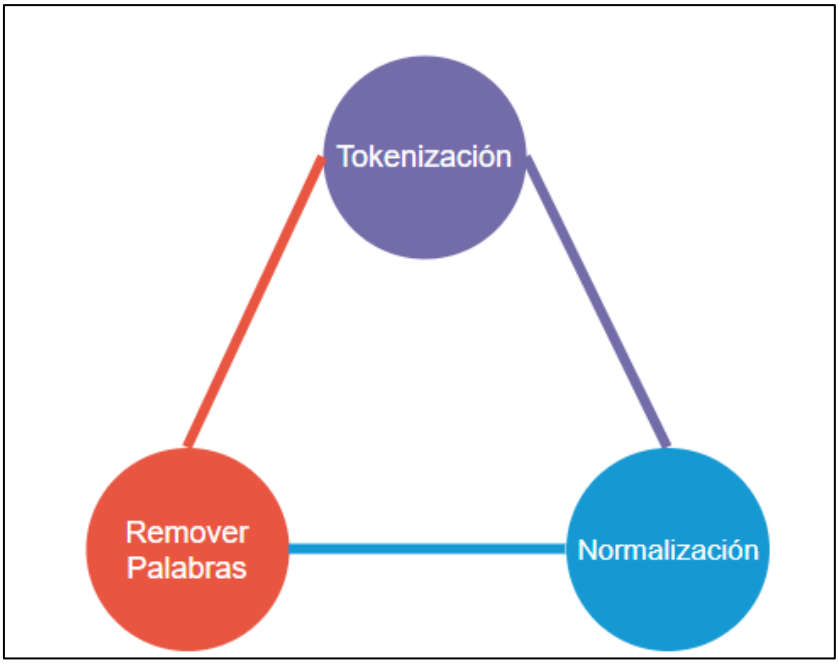


Figura 18. Funciones de limpieza de tweets

a. Tokenización de Palabras

Este proceso consiste en dividir las cadenas de texto en piezas pequeñas o tokens. Para este proceso se utiliza la función RegexTokenizer de la librería pySpark.

b. Normalización

Consiste en colocar todas las palabras en las mismas condiciones eliminando puntuaciones, símbolos, números y colocando todas las palabras en Mayúsculas o Minúsculas.

c. Remover Palabras

Consiste en remover palabras que no agregan valor o que generen ruido a los resultados de los análisis. Para este proceso y el anterior se utilizó la función stopwords de la librería nltk.corpus con el fin de obtener el listado de palabras a remover y para ejecutar el proceso de transformación se StopWordsRemover de la librería pySpark.

7.3.4. Análisis de sentimientos

Haciendo uso de la librería Textblob se realiza en análisis de cada uno se los *tweets* obtenidos y genera un valor entre -1 y 1 y se clasifican de acuerdo a la Tabla 8.

Tabla 8.Tabla de definición de análisis de sentimientos

RANGO	CLASIFICACIÓN
< 0	Negativo
0	Neutral
> 0	Positivo

Esta librería funciona para realizar análisis en textos de idioma inglés. Debido a esto la misma librería incluye una funcionalidad de traducción que realizar la traducción del texto obtenido al idioma que se requiere.

Para este proyecto se realiza el análisis de cada uno de los *tweets* obtenidos y con los resultados se realiza un gráfico *pie chart* o grafico de torta en donde se puede evidenciar la tendencia del tono emocional que tienen los *tweets* obtenidos. A continuación, un ejemplo de la visualización en la Figura 19.

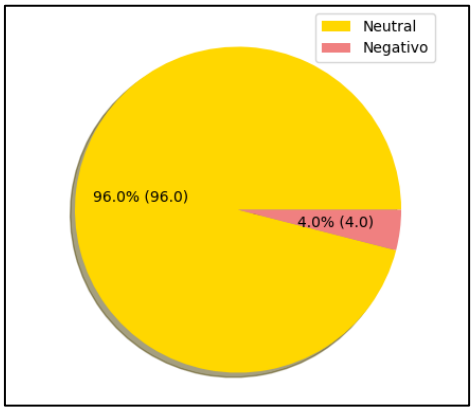


Figura 19. Ejemplo de visualización pie chart análisis de sentimientos

7.3.5. Frecuencia de Palabras

Este proceso consiste en identificar cuáles son las palabras que son más frecuentes y que tiene relación con la palabra o frase consultada en el resultado de *tweets* obtenidos.

Una vez realizado el proceso de limpieza de los *tweets* y con las palabras ya transformadas en tokens, se utiliza la librería de Python Counter cuyo resultado es el número de apariciones de cada una de las palabras en los *tweets* obtenidos. Con esta información se genera un diagrama de barras que muestra cuales son las 30 palabras que tienen más frecuencia en la información obtenida como se puede ver en la Figura 20.

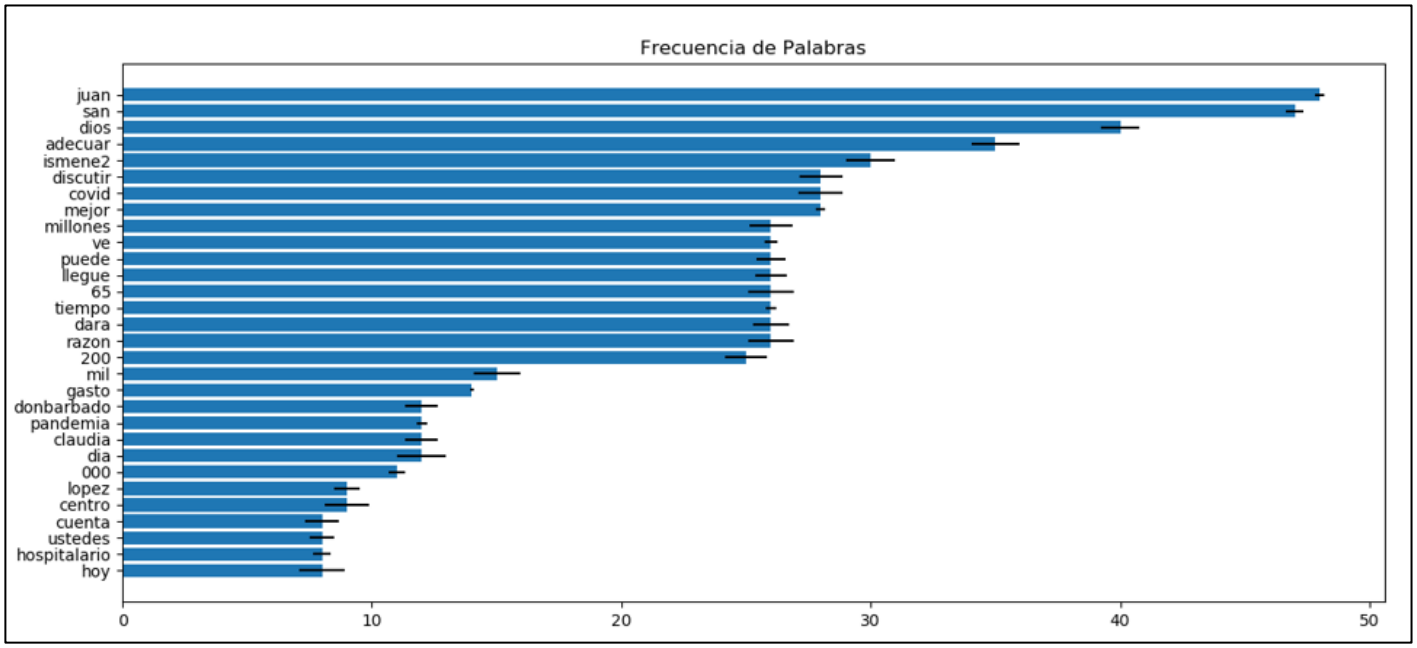


Figura 20. Diagrama de barras frecuencia de palabras

El conteo de apariciones de las palabras también nos permite realizar una nube de palabras que es una representación gráfica de las palabras clave más recurrentes. En este tipo de grafico se puede

observar como las palabras más recurrentes tienen un tamaño de letras más grande y colores más vivos que permiten identificarlos más rápidamente como se puede ver en la Figura 21.

Figura 21, Diagrama nube de palabras

7.3.6. Clusterización de Palabras

Con el fin de complementar el proceso de análisis de frecuencias de palabras y poder realizar una asociación entre estas que generen un mayor significado, se puede hacer uso del proceso de clusterización. Este proceso consiste en hacer agrupaciones de palabras de manera que los objetos en cada uno de los conjuntos o clústers sean similares o tengan cierta asociación. Este proceso nos permite identificar las agrupaciones o categorías en las cuales esta segmentada la información de los *tweets* y poder hacer análisis relacionados con estas.

El método k-means consiste en asignar una etiqueta de uno de los grupos K a cada fila de nuestro conjunto de datos. Cada valor de K tiene como punto de referencia un centroide que contiene las coordenadas de cada uno de los clústers. Inicialmente el algoritmo asigna valores aleatorios para los centroides y por medio de un proceso iterativo, se van ajustando los grupos hasta que se cumplan algunos de estos criterios:

a. Si no hay cambios en los centroides de los grupos

b. Si la suma de las distancias se minimiza

c. Se alcanza un número máximo de iteraciones

A continuación, en la Figura 22 se puede observar cómo cambian los Clústers en relación de cada iteración.

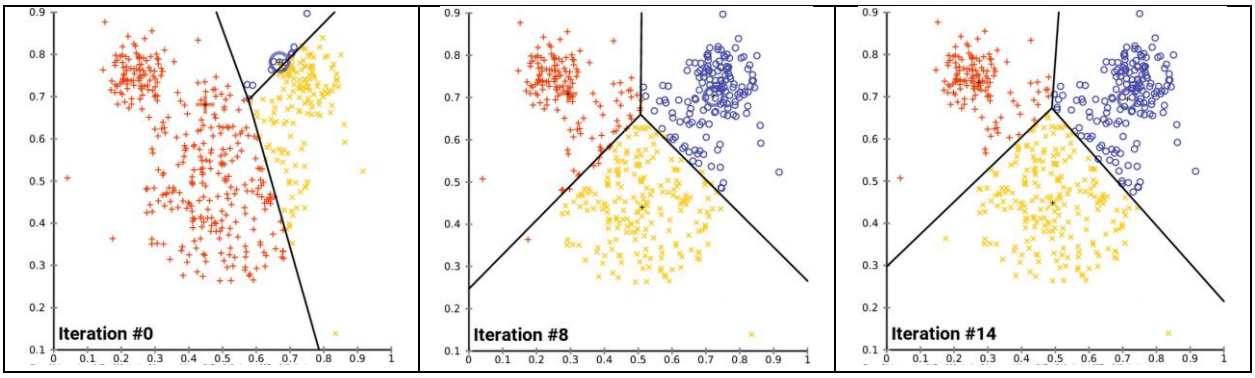


Figura 22. Iteraciones proceso k-means

Para hacer uso del método k-means primero se deben identificar cual es el numero óptimo de clústers para los datos obtenidos. Para esto se va a utilizar el método llamado la Curva de Elbow, que consiste en ejecutar el proceso k-means un número determinado de veces generando el valor de la suma de cuadrados internos de cada iteración y con este resultado identificar en qué punto se registra el mayor cambio en la variación o en otras palabras el mayor punto de inflexión de la curva Como lo podemos ver en la Figura 23.

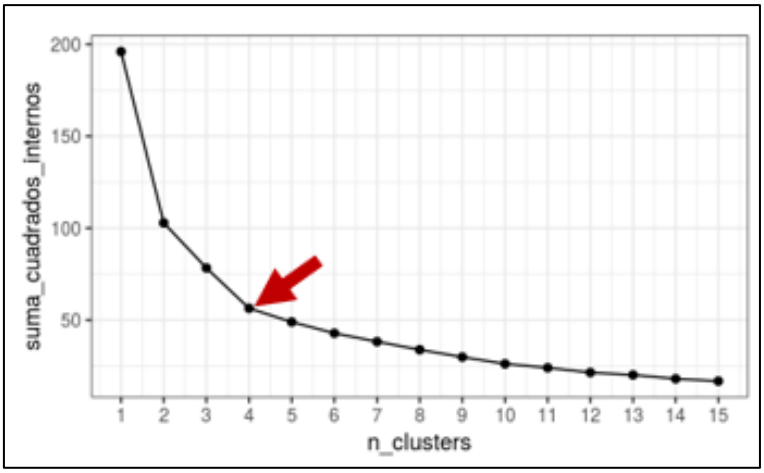


Figura 23. Identificación número de clústers, curva de Elbow

Para realizar este proceso se utiliza la función linkage de la librería scipy pasándole como parámetro el resultado de los k-means y el método *Euclidean*.

Una vez se tienen identificados el número de clústers óptimos se realiza el proceso de ejecución de la función kmeans pasándole como parámetro el número de clústers optimo y seguido de esto se realiza una nube de palabras de cada una de las categorías generadas por el proceso. En la figura 24 podemos ver de que forma se realiza la visualización.

Figura 24. Nube de palabras de categorías generadas.

7.3.7. Geolocalización

En la información obtenida por Twitter existen parámetros de geolocalización que permiten ubicar los mensajes generados en el lugar desde el cual fueron generados cuando el usuario desea compartir su ubicación en el mensaje. Esto permite realizar una ubicación espacial en un mapa del mundo para identificar los lugares de mayor frecuencia desde donde se está generando la información de la palabra consultada.

Debido a que para este prototipo se está utilizando una cuenda de desarrollador, dicha cuenta no nos genera la información de las coordenadas de la información obtenida. Para efectos de poder mostrar cómo se vería la visualización de los *tweets* en la aplicación se cargó una lista de coordenadas de diferentes lugares del mundo y a cada texto se le asigno al azar uno de los lugares de la lista.

Una vez realizado el proceso anterior se utiliza la librería Basemap para ubicar en un mapa cada una de coordenadas asignadas como podemos ver en la Figura 25.

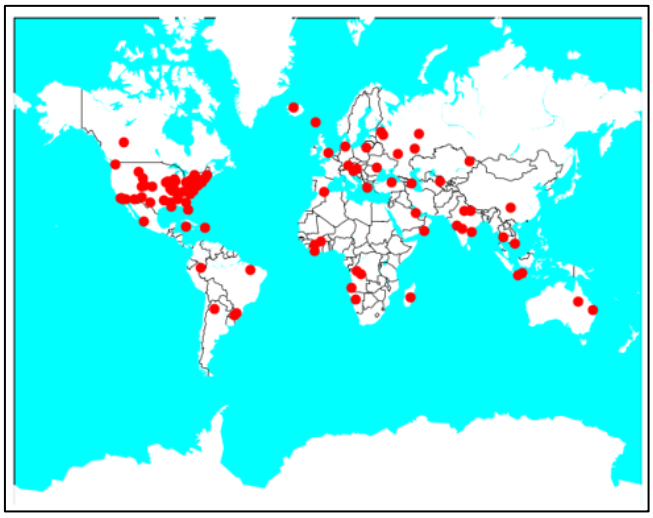


Figura 25. Visualización de geolocalización

7.3.8. Código fuente de la aplicación

En el siguiente link se encuentra el código fuente de la aplicación junto con las indicaciones para la instalación de las librerías y su configuración.

https://github.com/tatiananovoat/Proyecto_Grado

8. Casos de Estudio

Se realizan dos casos de estudio para analizar los resultados obtenidos por medio de prototipo desarrollado en cada una de las funcionalidades elaboradas en el proyecto.

8.1. Caso de Estudio 1: Día de la Madre

Se realiza la consulta de la palabra “Madre” el día 10 de mayo que corresponde al día en el cual se celebra el día de las madres en varios países del mundo como se puede ver en la Figura 26. Se obtuvieron 1000 registros para analizar este día.

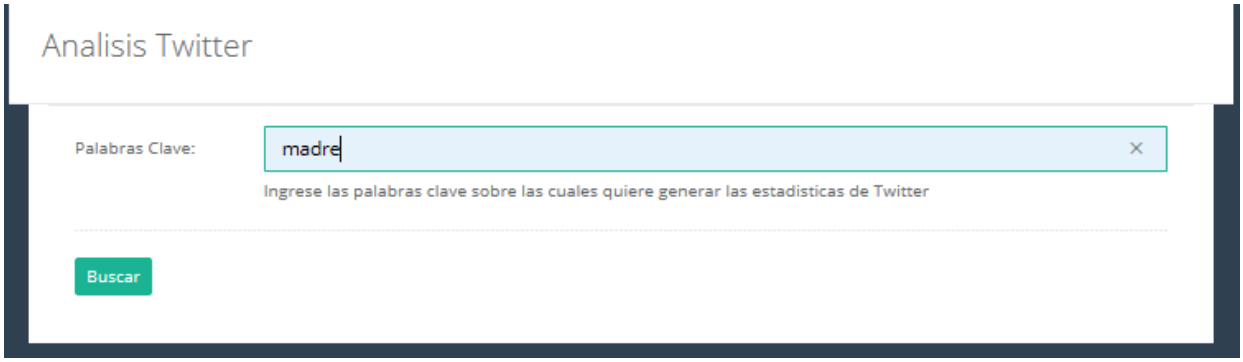
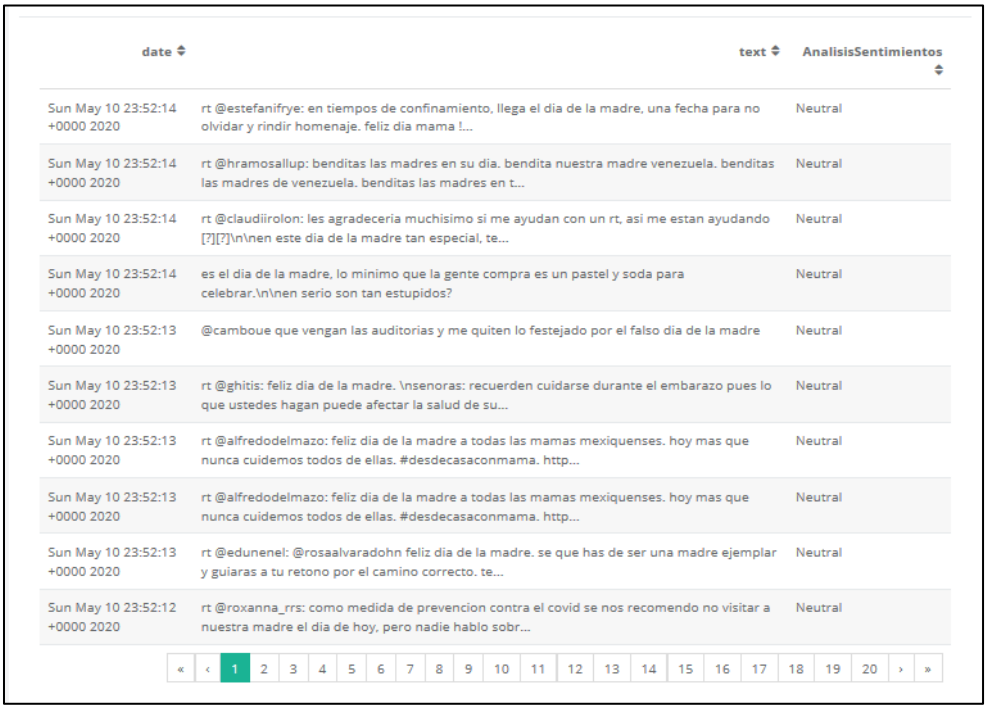


Figura 26. Inicio consulta palabra madre

Los resultados se pueden identificar en la Figura 27 y Figura 28.

Figura 27. Introducción Inicial Pagina de Respuesta



date	text	AnalisisSentimientos
Sun May 10 23:52:14 +0000 2020	rt @estefanifrye: en tiempos de confinamiento, llega el dia de la madre, una fecha para no olvidar y rindir homenaje. feliz dia mama !...	Neutral
Sun May 10 23:52:14 +0000 2020	rt @hramosallup: benditas las madres en su dia. bendita nuestra madre venezuela. benditas las madres de venezuela. benditas las madres en t...	Neutral
Sun May 10 23:52:14 +0000 2020	rt @claudiirolon: les agradeceria muchisimo si me ayudan con un rt, asi me estan ayudando [?][?]\n\nen este dia de la madre tan especial, te...	Neutral
Sun May 10 23:52:14 +0000 2020	es el dia de la madre, lo minimo que la gente compra es un pastel y soda para celebrar.\n\nen serio son tan estupidos?	Neutral
Sun May 10 23:52:13 +0000 2020	@camboue que vengan las auditorias y me quiten lo festejado por el falso dia de la madre	Neutral
Sun May 10 23:52:13 +0000 2020	rt @ghitis: feliz dia de la madre. \nsenoras: recuerden cuidarse durante el embarazo pues lo que ustedes hagan puede afectar la salud de su...	Neutral
Sun May 10 23:52:13 +0000 2020	rt @alfredodelmazo: feliz dia de la madre a todas las mamas mexiquenses. hoy mas que nunca cuidemos todos de ellas. #desdecasaconmama. http...	Neutral
Sun May 10 23:52:13 +0000 2020	rt @alfredodelmazo: feliz dia de la madre a todas las mamas mexiquenses. hoy mas que nunca cuidemos todos de ellas. #desdecasaconmama. http...	Neutral
Sun May 10 23:52:13 +0000 2020	rt @edunenel: @rosaalvaradohn feliz dia de la madre. se que has de ser una madre ejemplar y guiaras a tu retono por el camino correcto. te...	Neutral
Sun May 10 23:52:12 +0000 2020	rt @roxanna_rrs: como medida de prevencion contra el covid se nos recomendo no visitar a nuestra madre el dia de hoy, pero nadie hablo sobr...	Neutral

« ‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 › »

Figura 28. Visualización resultados obtenidos palabra madre

En la Figura 28 se pueden observar los resultados obtenidos de la consulta de la palabra madre, de igual manera se puede ver el resultado del análisis de sentimientos de cada uno de los mensajes generados.

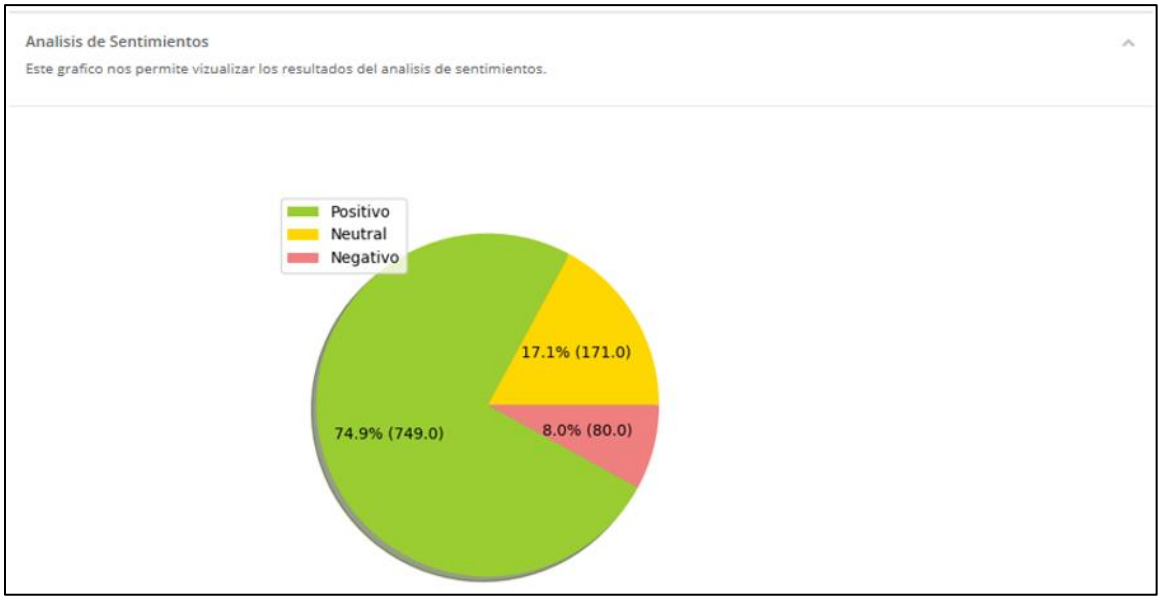


Figura 29. Análisis de sentimientos resultados palabra madre

En la Figura 29 del análisis de sentimientos se puede identificar que la mayoría de los mensajes analizados tienen una tendencia de ser positivos, estos equivalen al 74.9% de la información analizada. El restante se divide entre mensajes neutros con un 17.1% y mensajes negativos con un 8%.

Esto refleja la tendencia que generalmente se da en la fecha en cual se generó la consulta ya que corresponde al día en el cual se realiza la celebración del día de la madre en todo el mundo y es una tendencia que generalmente genera un sentimiento positivo en las personas.

La realizar el proceso de análisis dos veces con el mismo conjunto de datos se observa que los resultados que muestra son totalmente diferentes a los obtenidos en un principio lo que se puede visualizar en la figura 30.

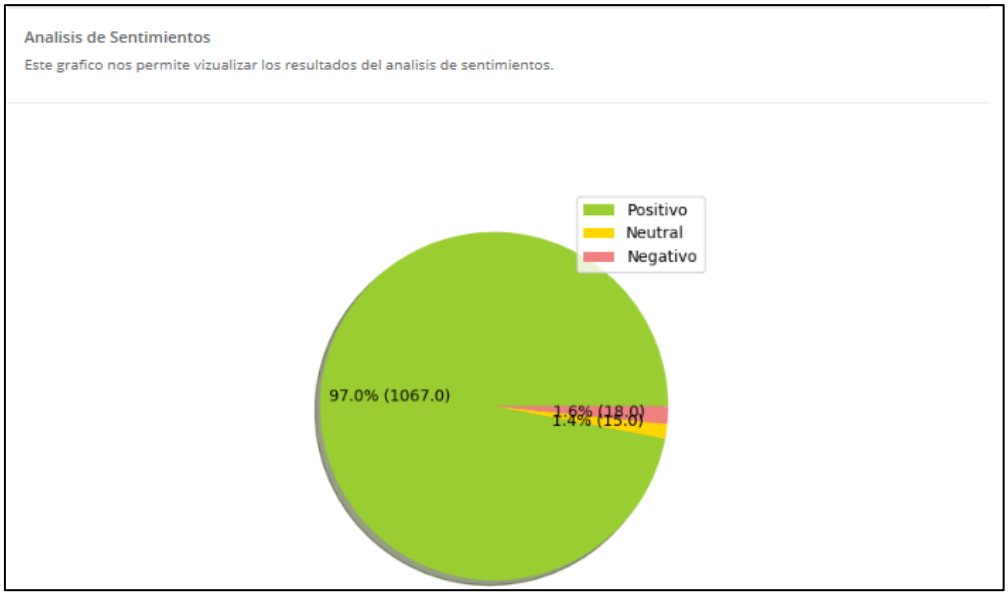


Figura 30. Ejecución de análisis 2 resultados palabra madre

Esto nos lleva a identificar que hay una inestabilidad en los resultados arrojados por la librería que puede afectar la visualización de los resultados obtenidos.

Figura 31. Nube de palabras resultados palabra madre

La nube de palabras de la Figura 31 nos permite visualizar que en las palabras más frecuentes de los resultados obtenidos esta “feliz”, “día”, “mamas”, entre otras, que corresponde también a palabras de celebración por la festividad que se llevó a cabo en ese día.

Adicionalmente se puede identificar el alias de un usuario en a la red social “alfredodelmazo” y la palabra “mexiquenses” que se repite varias veces debido a que la palabra está relacionada varias veces con “hoy” y “mamas” y esto genera bigramas. Al revisar los registros obtenidos se puede evidenciar que es una persona publica de México que realizo una publicación con el mensaje “Feliz Dia de la Madre a todas las mamas mexiquenses. Hoy más que nunca cuidemos todos de ellas”. En el momento de realizar la consulta este mensaje estaba siendo publicado varias veces por otros usuarios y por eso se presenta con mucha frecuencia en la información obtenida.

Los resultados de la nube de palabras se pueden también reflejar en el diagrama de frecuencia de palabras de la Figura 32 que nos brinda información de las treinta palabras más frecuentes y la cantidad de veces que se repiten en el conjunto de textos analizados.

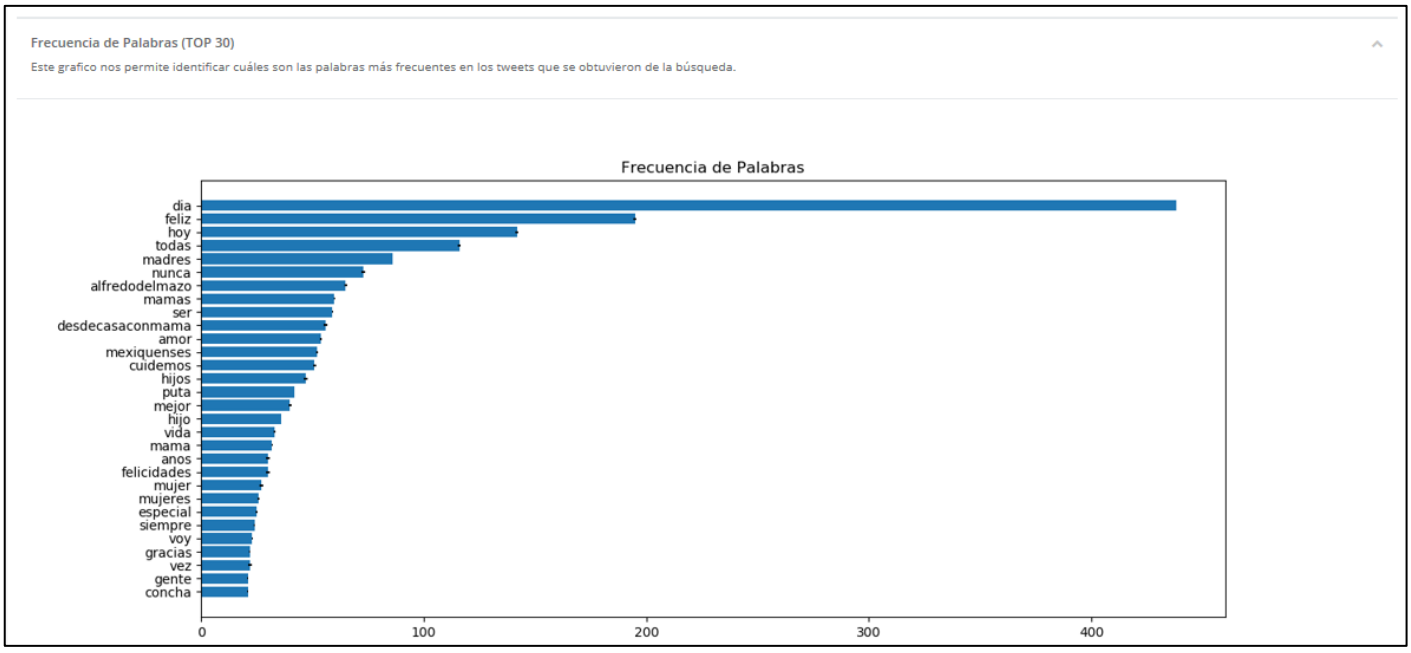


Figura 32. Frecuencia de palabras resultados palabra madre

El proceso de clusterización el conjunto de textos obtenido genero dos categorías que se pueden visualizar en las nubes de palabras de la Figura 33 y Figura 34.

Figura 33. Categoría 0 de resultado de Clusterización de palabras resultados palabra madre

En la primera categoría generada de la Figura 33 se pueden observar las palabras “feliz” y “dia”, palabras que tienen una alta frecuencia y están fuertemente relacionadas, estas corresponden a expresiones que son comúnmente utilizadas en diferentes comunidades para felicitar a las madres en la festividad.

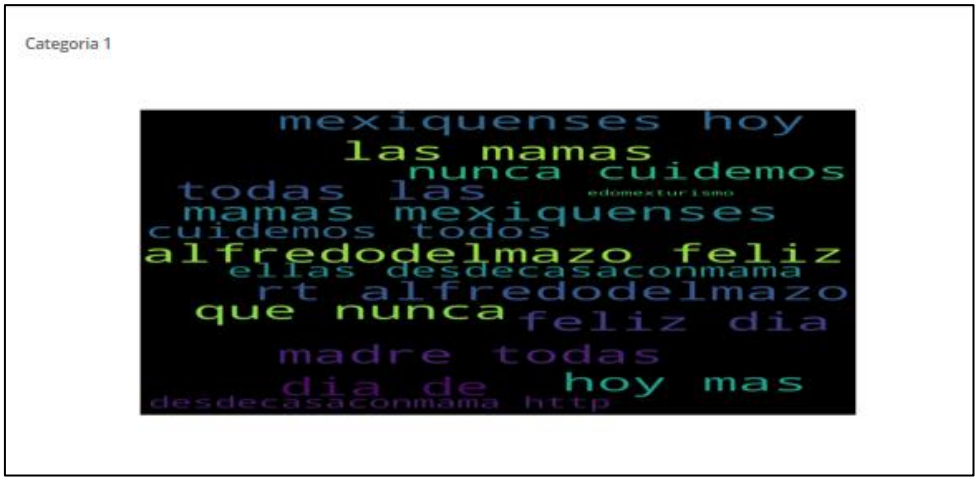


Figura 34. Categoría 1 de resultado de Clusterización de palabras para resultados palabra madre

La segunda categoría de la Figura 34 hace referencia a la publicación mencionada previamente por el usuario “alfredodelmazo” y que fue republicada por varios usuarios lo que hace que tenga una gran relevancia en los análisis efectuados.

Como se indicó antes, el grafico de geolocalización e la Figura 35 tiene información de prueba y debido a esto en este momento no nos genera más información al análisis del caso de estudio.

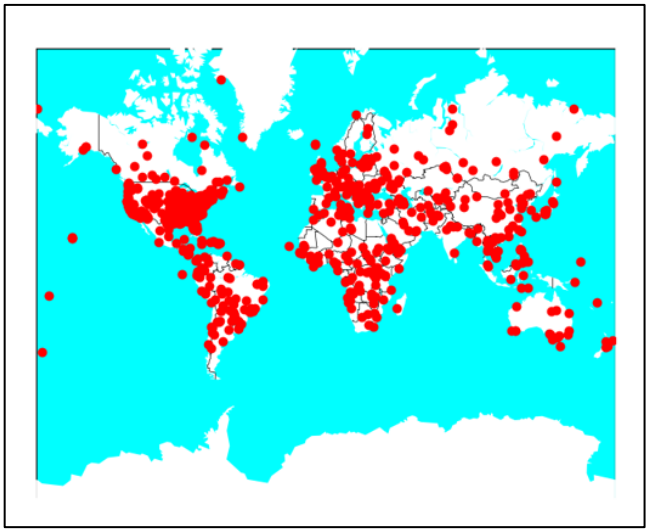


Figura 35. Geolocalización resultados palabra madre

8.2. Caso de Estudio 2: Rappi

El segundo caso de estudio se llevó a cabo el mismo día del caso de estudio 1. Para este análisis como se puede ver en la Figura 36, se consultó la palabra Rappi que hacía referencia a una tendencia que se estaba presentando en Twitter el día de la consulta. Al realizar la consulta se obtuvieron 1000 registros para analizar este día con los cuales se realizaron los análisis que se presentan a continuación.

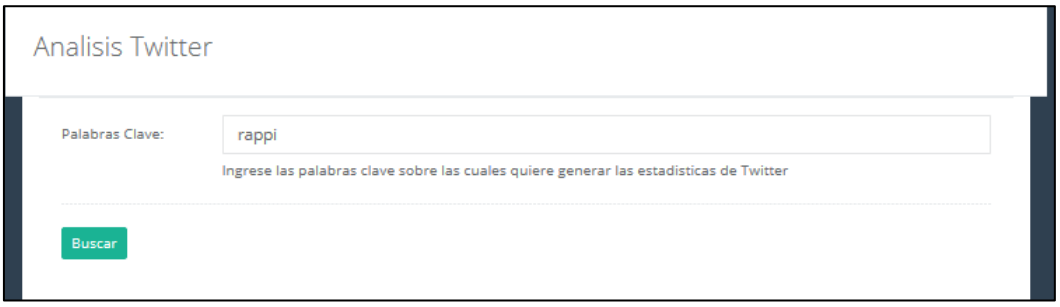


Figura 36.Inicio consulta palabra Rappi

date	text	AnalisisSentimientos
Mon May 11 00:41:55 +0000 2020	negreros,el gobierno permite la esclavitud, por la necesidad de un colombiano#rappi	Neutral
Mon May 11 00:41:51 +0000 2020	rt @jpserna: quedo claro que ningun restaurante estaba preparado para atender los domicilios. colapso total.\n#rappi \n#felizdiadelamadre	Neutral
Mon May 11 00:41:48 +0000 2020	rt @luchovoltios: muchas quejas por los domicilios hoy. la demanda supero la oferta y de alli los problemas en el #diadelamadre los unicos...	Neutral
Mon May 11 00:41:48 +0000 2020	como que rappi ya no es trending? es que ya todos recibieron sus domicilios y se les olvido el mierdero? buuu	Neutral
Mon May 11 00:41:41 +0000 2020	@jotagiralo @juanpatorresv entiendo. habra que ver, porque increible que dadas las circunstancias y la oportunidad... https://t.co/rd6liuifhj	Neutral
Mon May 11 00:41:36 +0000 2020	rt @catarnico: mas maricas ustedes que dejan en manos de rappi el almuerzo de sus mamas.	Neutral
Mon May 11 00:41:33 +0000 2020	rt @rodrigoddhh: el pais de la doble moral ayer se indignaba por la senora que esclavizaron en un edificio de bogota y hoy se indigna porqu...	Neutral
Mon May 11 00:41:19 +0000 2020	rt @romarioqr: se entiende que tengas rabia porque tu pedido de #rappi no llego a tiempo, pero no puedes tratar como una basura al muchacho...	Neutral
Mon May 11 00:41:19 +0000 2020	@nachogreiffenst @rappicolombia aveces rappi es bueno x las promociones	Neutral
Mon May 11 00:41:08 +0000 2020	rt @oorzoo15: los de #rappi intentando entregar el #almuerzoentubarrio de #diadelamadre https://t.co/whc6kbmzcx	Neutral

« ‹ 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 › »

Figura 37. Visualización resultados obtenidos palabra Rappi

En análisis de sentimientos de la Figura 38 nos muestra que hay una tendencia de neutralidad en el 70% de la información analizada. Esto quiere decir que los *tweets* analizados no reflejan el tono emocional de la intención con la que fue escrito el mensaje. El 30% restante se divide entre un 18% de mensajes identificados como positivos y 11% como negativos.

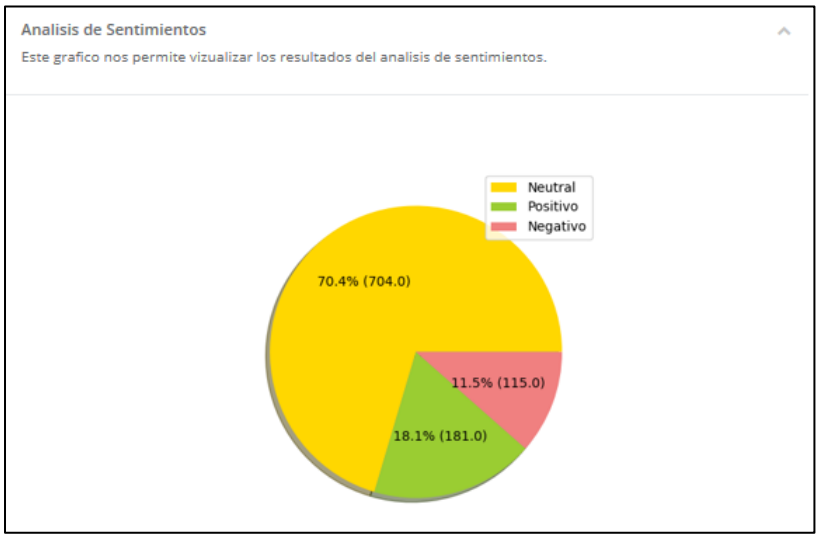


Figura 38. Análisis de sentimientos resultados palabra Rappi

En la nube de palabras de la Figura 39 se identifica que la tendencia de palabras más frecuentes hace referencia a un hecho ocurrido el día hoy y también a clínicas y hospitales. Al contrastar los resultados con la información obtenida se identifica que esto se debe a que se generaron varias repeticiones de un mensaje publicado "*RT @jbagbam74: Lo que paso con Rappi hoy es lo que hubiera ocurrido con clínicas y hospitales si todos nos hubiéramos enfermado al mismo tiempo*". Esto nos indica que hay un alto nivel de afinidad de los usuarios con este mensaje publicado,

El resto de las palabras que se identifican hacen referencia al funcionamiento de la aplicación de domicilios Rappi en el día de la madre.

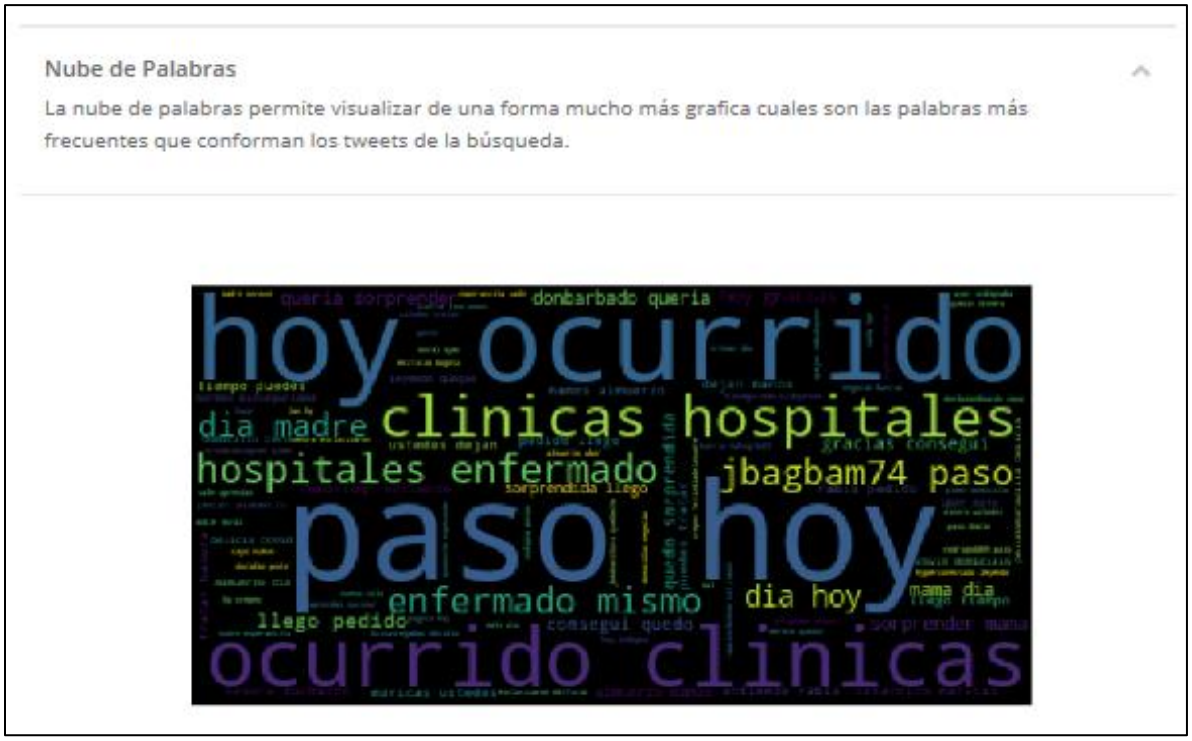


Figura 39. Nube de palabras resultados palabra Rappi

En la frecuencia de palabras de la Figura 40 se pueden identificar las cifras de las 30 palabras más frecuentes en la información obtenida.

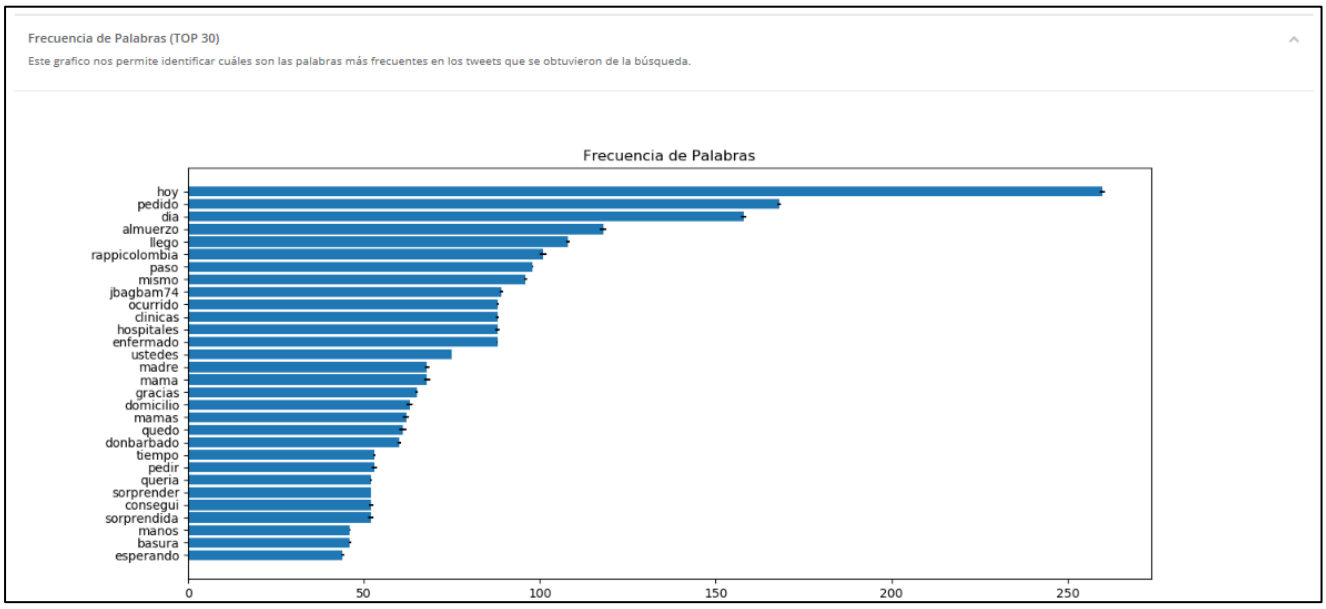


Figura 40. Frecuencia de palabras resultados palabra Rappi

La clusterización de palabras nos genera dos categorías de las Figuras 41 y 42. La primera categoría de la Figura 41 que hace referencia al mensaje que fue replicado varias veces en la red social y de la cual tenemos 174 repeticiones en el conjunto de datos obtenido.

Figura 41. Categoría 0 de resultado de Clusterización de palabras resultados palabra Rappi

La segunda categoría de la Figura 42 hace referencia a los pedidos que se generaron a través de la aplicación. Se identifican palabras como “sorpresa”, “pedido” y “no llego” que nos permiten identificar una insatisfacción en el servicio prestado en este día debido a que no llegaron los pedidos. Y en otro sentido se identifican palabras de agradecimiento a Rappi. Sin embargo, es difícil identificar si estas palabras corresponden a un agradecimiento real o a un sarcasmo generado por medio del mensaje.

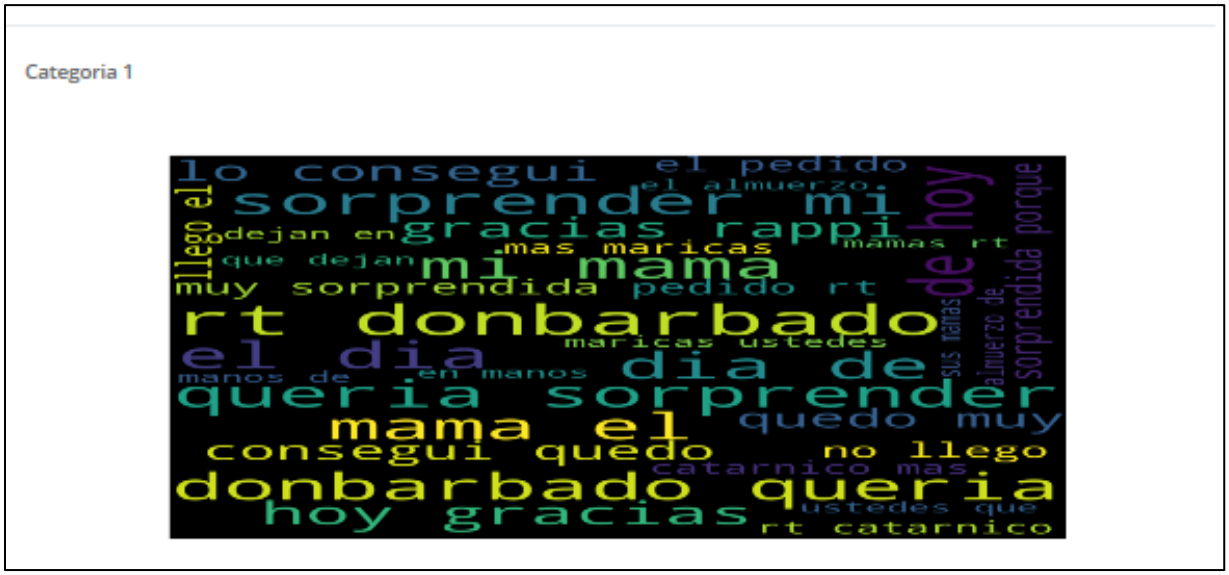


Figura 42. Categoría 1 de resultado de Clusterización de palabras resultados palabra Rappi

Como se indicó antes el grafico de geolocalización de la Figura 43 tiene información de prueba y debido a esto en este momento no nos genera más información al análisis del caso de estudio.

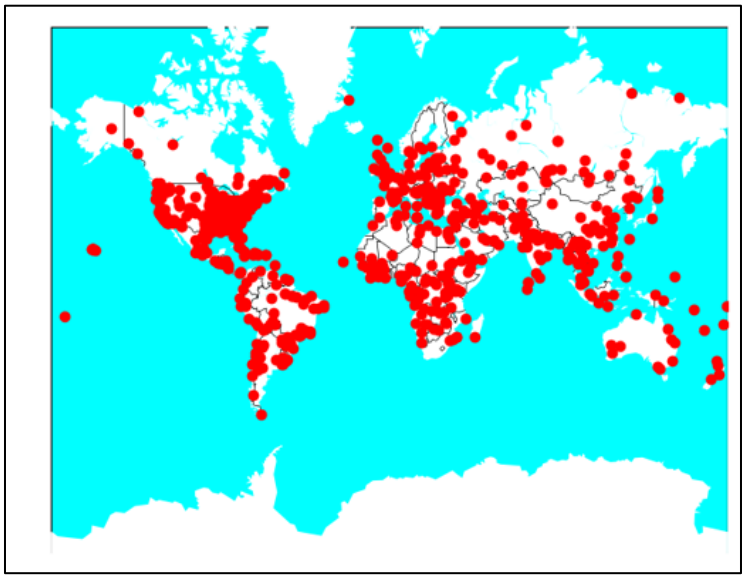


Figura 43. Geolocalización resultados palabra Rappi

9. Conclusiones

- Con el desarrollo del prototipo se identifica una aplicación que es de fácil uso lo que permite que pueda ser utilizado por cualquier persona sin necesidad de tener un conocimiento muy profundo en el desarrollo de las herramientas utilizadas. Adicionalmente las consultas se pueden realizar sobre cualquier palabra o frase de interés lo que nos permite tener un mejor entendimiento de la información que se distribuye a través de la red social.
- El framework Django es una herramienta que permite unificar las ventajas que ofrece Python para los procesos de análisis de datos con la ejecución de tiempo real, Inmediatez de acceso y compatibilidad multiplataforma que ofrecen las páginas web.
- La arquitectura de la aplicación con la utilización del clúster permite que sea escalable de manera que con la configuración de un clúster de mayor capacidad este puede ofrecer un mejor rendimiento para el procesamiento de altos volúmenes de información. Aunque para este proyecto la utilización del clúster fue parcial se podría vincular las más tareas de procesamiento que le brindarían un mayor rendimiento y tiempo de respuesta
- La curva de Elbow permitió optimizar el proceso de clusterización indicando el número de categorías optimas sobre el cual se debía ejecutar el proceso de optimización generando categorías que generaban un mejor entendimiento.

- En los resultados de los análisis de sentimientos se pueden identificar errores y variabilidad en los resultados obtenidos. Esto lleva a que se tenga que hacer un análisis más exhaustivo de la librería utilizada Textblob y analizar otros proyectos colaborativos para el análisis de sentimientos como lo son TASS (Taller de Análisis Semántico de la SEPLN) o VADER (Valence Aware Dictionary and sEntiment Reasoner) entre otras que se pueden evaluar de igual manera para generar un mejor resultado.
- Con los resultados obtenidos se identifican diferentes opciones de mejora en donde se pueda incorporar análisis de otros tipos de datos que se generan a través de Twitter y que pueden realizar una mejor caracterización de las personas que realizan las publicaciones y con esto obtener más información para analizar por parte de los usuarios.

Referencias

Abella García, V. &. (2015). Aprender a usar twitter y usar twitter para aprender. *Revista de curriculúm y formación del profesorado*.

Abellán, M. L. (2012). Twitter como instrumento de comunicación política en campaña: Elecciones Generales de 2011. *Cuadernos de gestión de información*.

Apache Hadoop. (10 de 09 de 2019). Obtenido de Apache Hadoop YARN: https://Hadoop.apache.org/docs/current/Hadoop-yarn/Hadoop-yarn-site/YARN.html

Bhangle, R., & Krishnan, S. (2018). Twitter Sentimental Analysis on Fan Engagement. *Advances in Intelligent Systems and Computing*, 27-39.

Borthakur, D. (22 de 08 de 2019). *Hadoop*. Obtenido de HDFS Architecture Guide: https://Hadoop.apache.org/docs/r1.2.1/hdfs_design.html

Carballar, J. A. (2011). Twitter. Marketing personal y profesional. *RC Libros*.

Congosto, M. L. (2011). Twitter y política: Información, opinión y¿ Predicción?.

Cuell, C. &. (2009). An assessment of climatological synoptic typing by principal component analysis and kmeans Clústering. *Theoretical and Applied Climatology Vol 98*, 3-4.

Dutta, D., Sharma, S., Natani, S., Khare, N., & Singh, B. (2017). Sentimental Analysis for Airline Twitter data. *IOP Conference Series: Materials Science and Engineering*.

Fainholc, B. (2011). Un análisis contemporáneo del Twitter. *Revista de Educación a Distancia*, 3.

Frampton, M. (2015). *Mastering Apache Spark.* Birmingham: Packt Publishing Ltd.

Hostinger, D. (29 de 05 de 2019). *Tutorial Hostinger*. Obtenido de ¿Cómo funciona el SSH?: https://www.hostinger.co/tutoriales/que-es-ssh

Juan. (Febrero de 2020). *https://www.monografias.com*. Obtenido de https://www.monografias.com/trabajos108/impacto-redes-sociales-sociedad/impacto-redes-sociales-sociedad.shtml

Karau, H., Konwinski, A., Wendell, P., & Zaharia, M. (2015). *Learning Spark.* Sebastopol, CA: O'Reilly Media, Inc.

Kumar, S. M. (2014). *Twitter data analytics.* New York: Springer.

López Zapico, M. A. (28 de 02 de 2020). EL USO DE TWITTER COMO HERRAMIENTA PARA LA ENSEÑANZA UNIVERSITARIA EN EL ÁMBITO DE LAS CIENCIAS SOCIALES. *EL USO DE TWITTER COMO HERRAMIENTA PARA LA ENSEÑANZA UNIVERSITARIA EN EL ÁMBITO DE LAS CIENCIAS SOCIALES. UN ESTUDIO DE CASO DESDE LA HISTORIA ECONÓMICA*. Obtenido de https://www.redalyc.org/articulo.oa?id=201028055014

López-García, G. (2015). Nuevos y viejos liderazgos: la campaña de las elecciones generales españolas de 2015 en Twitter.

Nabel, L. C. (2010). Redes sociales y efectos políticos: Reflexiones sobre el impacto de twitter México. *Sociología y tecnociencia: Revista digital de sociología del sistema tecnocientífico*.

Olea, I. (22 de 03 de 2004). *ibiblio*. Obtenido de https://www.ibiblio.org

Petrocelli, D. D. (2017). Procesamiento distribuido y paralelo de bajo costo basado en cloud&movil. *Congreso Argentino de Ciencias de la Computación (Vol. 23)*.

Pla, F. &. (2013). ELiRF-UPV en TASS-2013: Análisis de sentimientos en Twitter. *Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural* .

Prajapati, V. (2013). *Big Data Analytics with R and Hadoop.* Birmingham: Packt Publishing.

Rochina, P. (09 de 10 de 2019). *RevistaDigital Inesem*. Obtenido de Python vs R para el análisis de datos: https://revistadigital.inesem.es/informatica-y-tics/Python-r-analisis-datos/

Rodrigues, A. P., Chiplunkar, N. N., & Rao, A. (2016). Sentiment Analysis of Social Media Data using Hadoop Framework: A Survey. *International Journal of Computer Applications Vol 151 - No 6*.

SAURA, J. R.-M.-S. (2018). Un Análisis de Sentimiento en Twitter con Machine Learning: Identificando el sentimiento sobre las ofertas de# BlackFriday. *Revista Espacios*.

Scott, J. A. (2015). *Getting started with Apache Spark.* San Jose, CA: MapR Technologies, Inc.

Vicente, E. R. (2005). Grasp en la resolución del problema de Clústering. *Revista de Investigación en Sistemas e Informática*.

Villagra, A. G. (2009). Análisis de medidas no-supervisadas de calidad en Clústers obtenidos por K-means y Particle Swarm Optimization. *Ciencia y Tecnología*.

Waxman, A. B. (2017). Rogues of Wall Street (How to Manage Risk in the Cognitive Era) || Twitter Risk and Fake News Risk. 2186—2191.

White, T. (2009). *Hadoop: The Definitive Guide.* Sebastopol: O'Reilly Media, Inc.