UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

Colegio de Ciencias e Ingenierías

Minería de datos en redes sociales por medio de un correlacionador de datos

Proyecto de investigación

.

Christian Alexis Álvarez Espín

Ingeniería en Sistemas

Trabajo de titulación presentado como requisito
para la obtención del título de
Ingeniero en Sistemas

Quito, 30 de marzo de 2017

UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ

COLEGIO DE CIENCIAS E INGENIERÍAS

HOJA DE CALIFICACIÓN
DE TRABAJO DE TITULACIÓN

Minería de datos en redes sociales por medio de un correlacionador de datos

Christian Alexis Álvarez Espín

Calificación:

Nombre del profesor, Título académico Mauricio Iturralde, PhD

Firma del profesor _____

Quito, 30 de marzo de 2017

Derechos de Autor

Por medio del presente documento certifico que he leído todas las Políticas y Manuales de la Universidad San Francisco de Quito USFQ, incluyendo la Política de Propiedad Intelectual USFQ, y estoy de acuerdo con su contenido, por lo que los derechos de propiedad intelectual del presente trabajo quedan sujetos a lo dispuesto en esas Políticas.

Asimismo, autorizo a la USFQ para que realice la digitalización y publicación de este trabajo en el repositorio virtual, de conformidad a lo dispuesto en el Art. 144 de la Ley Orgánica de Educación Superior.

Firma del estudiante: _____

Nombres y apellidos: Christian Alexis Álvarez Espín

Código: 00104108

Cédula de Identidad: 1719928507

Lugar y fecha: Quito, 30 de marzo de 2017

RESUMEN

Twitter es una red social que ha ganado una gran importancia como medio de comunicación. Existen millones de usuarios que utilizan esta red social para informarse sobre diferentes temas de su interés. Al ofrecer una herramienta al ámbito periodístico y comercial que permita analizar tendencias sobre cualquier tema publicado en Twitter, se puede ampliar la utilización de la red social como un medio de comunicación. Los usuarios periodistas pueden obtener resultados sobre temas que ocurren en la actualidad; por ende, conocer el impacto que crea cada tema en esta red social. De la misma manera, un usuario puede utilizar esta herramienta como inteligencia de negocio para visualizar el crecimiento de una marca entre los diferentes usuarios que utilizan Twitter. Existe una cantidad infinita de escenarios de análisis que pueden ser realizados mediante esta herramienta.

Palabras clave: Correlacionador de datos, Twitter, temas de búsquedas, tendencias, análisis, tiempo real.

ABSTRACT

Twitter is a social network that has gained great importance as a mean of communication. There are millions of users who use this social network to learn about different topics of their interest. By offering a tool to the journalistic and commercial scope that allows analyzing trends on any topic published on Twitter, it is possible to extend the use of the social network as a means of communication. Journalist users can get results on any topic that occur every day to visualize the impact that each topic creates in this social network. In the same way, a user can use this tool as business intelligence to visualize the growth of a brand among the different users who use Twitter. There are an infinite number of analysis scenarios that can be performed using this tool.

Key words: Data Correlator, Twitter, search topics, tendencies, analysis, real time.

TABLA DE CONTENIDO

INTRODUCCIÓN

El presente trabajo es una solución orientada a la búsqueda rápida y obtención de datos estadísticos en tiempo real sobre tema publicados en Twitter.

En la actualidad, Twitter se ha convertido en un medio de comunicación muy importante para todo tipo de usuario. Por medio de esta red social, muchas personas se mantienen informadas y comparten sus comentarios y opiniones sobre temas y noticias que ocurren día a día. Poder analizar tendencias sobre estos temas puede ser de gran utilidad en varios ámbitos como en el periodismo, o en la inteligencia de negocios.

Mediante un correlacionador de datos se ofrece una aplicación web que permite realizar búsquedas sobre un tema específico y obtener datos estadísticos sobre el mismo. Esta aplicación es un conjunto de herramientas que trabajan de forma paralela para obtener y analizar resultados de una búsqueda de manera eficiente y confiable.

La aplicación está orientada al periodismo, y a la inteligencia de negocios. En el ámbito periodístico, la aplicación permite al usuario realizar búsquedas sobre un tema específico en Twitter, ofreciendo como resultados diferentes gráficos que muestran la concurrencia de tweets sobre aquel tema, e información sobre los usuarios que enviaron los tweets. En el ámbito de inteligencia de negocios, la aplicación permite realizar búsquedas sobre las tendencias que ha creado un producto, dando la funcionalidad de crear indicadores en tiempo real que muestran el crecimiento del producto en un rango de tiempo, los cuales podrán ser utilizados para toma de decisiones en un negocio.

MARCO TEÓRICO

La aplicación web fue desarrollada utilizando varias tecnologías y en tres diferentes fases. Para la primera fase se utilizó JAVA EE, en la cual se desarrolló la aplicación web encargada de realizar las búsquedas periódicas en Twitter. Esta aplicación mantiene una conexión con un motor de base de datos MYSQL, la cual se utiliza para almacenar la información necesaria para que la aplicación web funcione de manera correcta. En la segunda fase se utilizó un correlacionador de datos conocido como Elastic Stack. Esta tecnología es un conjunto de herramientas que son utilizadas para manejar una gran cantidad de datos, que en este caso son los tweets obtenidos por medio de la aplicación web. Como último paso se realizó la integración de ambas tecnologías utilizando archivos de registro. Estos archivos de texto contienen los resultados de las búsquedas realizadas en la aplicación web, convirtiéndose en la fuente de información para el correlacionador de datos.

Es muy importante poder tener control sobre la aplicación web en temas de mantenimiento y escalabilidad. Por esta razón se escogió utilizar JAVA EE (Enterprise Edition) para el desarrollo de la misma. “Java EE proporciona un amplio conjunto de herramientas para crear, desarrollar, probar, depurar y desplegar aplicaciones empresariales.” (IBM, S.F.) Esta tecnología permite, además, conectarse con cualquier motor de base de datos. Poder escoger el motor de base de datos es una gran ventaja ya que, como desarrollador, uno no se encuentra ligado a una tecnología especifica que puede representar costos, una curva de aprendizaje mayor, o mayor tiempo de desarrollo. Para este proyecto se escogió el motor de base de datos MYSQL ya que, gracias a su rendimiento, confiabilidad, y facilidad de uso es la base de datos de código libre más utilizada para aplicaciones web (Oracle, S.F.). En la base de datos se almacena información como: nombres y claves de los usuarios, búsquedas realizadas, parámetros globales de la aplicación, e información obtenida desde twitter. Por último, para que esta aplicación pueda ser ejecutada, se utilizó Wildfly 8, el

cual es un servidor web que permite desplegar aplicaciones web desarrolladas en JAVA EE. Wildfly soporta los últimos estándares para el desarrollo web, y mejora la productividad de desarrollo ya que provee herramientas fáciles de usar que eliminan la carga técnica. (What is Wildfy, S.F)

Se conoce como un correlacionador de datos a una plataforma que permite almacenar información para poder crear relaciones entre los datos, y encontrar patrones de comportamiento. En este proyecto se utilizó Elastic Stack, el cual es un correlacionador de datos creado para indexar, buscar, y analizar información en tiempo real. Esta plataforma almacena los tweets obtenidos como resultados por la aplicación web, la cual será organizada entre diferentes campos. Entre estos campos se incluye fechas, contenido de los tweets, y datos de los usuarios que realizaron las publicaciones. Una vez creados estos campos, se crea relaciones entre ellos para poder obtener datos estadísticos de manera visual. Es una herramienta que facilita el análisis de datos masivos de manera rápida y confiable.

La información obtenida por la aplicación web es almacenada en un archivo de registro (archivo de log) delimitado por tabulaciones de extensión log.

> “La extensión de archivo de registro se utiliza para un archivo de registro que es utilizado por múltiples programas y sistemas operativos. Actúa como registros mantenidos por los programas de ordenador para cada actividad que se realice. Eventos registro de la actividad de una computadora es relevante para su día a día de procesamiento. Proporciona una manera para que el sistema a ser diagnosticados en los casos en que puede haber problemas encontrados que necesitan soluciones inmediatas. Esto es evidente con los sistemas grandes y complejos en los eventos y actividades que son esenciales en la comprensión de las aplicaciones involucradas. Todos estos eventos se guardan en un archivo de registro junto con los

nombres de archivos y listas instalados de su ubicación.” (.LOG extensión del archivo, S.F.)

En este proyecto, los registros almacenados en este archivo son los resultados del buscador de Twitter. Este archivo tiene un formato pre definido, por lo que cada registro nuevo será ingresado de la misma manera en este archivo. Cada vez que un nuevo registro es ingresado, el correlacionador de datos accede al archivo para obtener información del nuevo registro, y lo almacena en su propia fuente de datos. La lectura de datos que ofrece Elastic Stack es muy amplia, por lo que hay muy pocas limitaciones en el formato del archivo. Sin embargo, en este proyecto se optó por un archivo delimitado por tabulaciones por la velocidad de escritura y facilidad en la lectura del mismo.

ARQUITECTURA DE APLICACIÓN WEB Y ELASTIC STACK

Existen dos partes claves en este proyecto: el buscador de twitter, y el correlacionador de datos. Cada uno tiene su propia arquitectura de desarrollo, la cual será explicada a continuación.

El buscador de Twitter es una aplicación web distribuida en tres diferentes capas: persistencia, servicios, y web. En la capa de persistencia se encuentra la base de datos mapeada en objetos, mejor conocido como ORM. “Object-Relational mapping, o lo que es lo mismo, mapeo de objeto-relacional, es un modelo de programación que consiste en la transformación de las tablas de una base de datos, en una serie de entidades que simplifiquen las tareas básicas de acceso a los datos para el programador” (Que es un ORM, S.F) Además, esta capa se conecta con la base de datos para poder almacenar, actualizar, borrar, y consultar. En este proyecto se utilizó capa de persistencia para consultar las credenciales de usuarios de la aplicación web, almacenar los resultados obtenidos por el buscador, consultar los usuarios de twitter que realizaron publicaciones sobre los temas buscados, y desactivar temas de búsqueda no deseados.

La capa de servicios es aquella que contiene la lógica de negocio de la aplicación. En este caso, esta capa contiene algoritmos utilizados para la generación inteligente de hashtags para la búsqueda, la conexión con los servicios de Twitter para la obtención de datos, y la escritura de registros en el archivo de log. Cuando el usuario de la aplicación web ingresa un nuevo tema de búsqueda, esta capa se encarga de crear varios hashtags, conectarse a los servidores de twitter para obtener los datos, llamar a la capa de persistencia para almacenar los resultados, y por ultimo registrarlos en el archivo de log. Es importante mencionar que esta capa ofrece un manejo de transacciones, por lo que realiza todos los pasos mencionados previamente de manera segura.

Otra funcionalidad que esta capa contiene es la búsqueda de tweets por usuario. Además de buscar los tweets sobre un tema, la aplicación web permite obtener los tweets que un usuario de la red social ha compartido. La implementación de esta funcionalidad es la siguiente: el usuario de la aplicación web ingresa el nombre de un usuario de twitter, y esta capa se conecta con servicios de Twitter para obtener los últimos 200 tweets que ha realizado este usuario. Estos resultados permiten analizar a mayor profundidad el tema de búsqueda ya que se puede observar la participación que tiene cada uno de los usuarios en la tendencia.

Por último, esta capa contiene un temporizador que ejecuta búsquedas sobre los temas deseados cada 30 segundos. Esta funcionalidad permite que la aplicación continúe obteniendo resultados sobre los temas ingresados por el usuario hasta que él decida eliminar el tema de búsqueda.

La capa web es aquella que contiene los controladores, y las paginas JSF (Java Server Faces), en otras palabras, es el front end de la aplicación. Esta capa es muy importante ya que el usuario accede a la aplicación web y todas sus funcionalidades por medio de la misma. Es una combinación de varias tecnologías como html, css, y JSF. “JSF es un framework MVC (Modelo-Vista-Controlador) basado en el API de Servlets que proporciona un conjunto de componentes en forma de etiquetas definidas en páginas XHTML mediante el framework Facelets. Facelets se define en la especificación 2 de JSF como un elemento fundamental de JSF que proporciona características de plantillas y de creación de componentes compuestos” (Dpt. De Ciencias de la Computación e Inteligencia Artificial, 2014).

Por otro lado, el correlacionador de datos está compuesto por cuatro herramientas diferentes. La primera se llama Filebeat, la cual se encarga de leer los registros ingresados por el buscador. Cuando Filebeat está corriendo, ejecuta uno o más trabajadores que buscan archivos de registro en una de las rutas configuradas. Para cada archivo de registro que el trabajador localiza, Filebeat inicia un recolector. Cada recolector lee un solo archivo de

registro para el nuevo contenido y envía los nuevos datos a la salida que se ha configurado para Filebeat. (How Filebeat Works, 2017) Esto quiere decir que, los resultados que se escriben en el archivo de registro son leído por esta herramienta. Una vez que Filebeat lee los nuevos registros, los envía a la segunda herramienta del correlacionador de datos conocida como Logstash.

Logstash es una herramienta de código libre que recibe información de varias fuentes, la transforma y la almacena. Es un motor de recolección de datos con capacidades de canalización en tiempo real. Puede unificar dinámicamente datos de fuentes dispares y normalizar los datos en destinos de su elección. Además, limpia y democratiza todos sus datos para diversos casos de análisis y visualización. (Logtstash Introduction, 2017) En este proyecto, Logstash recibe la fuente enviada por Filebeat en el mismo formato creado por el buscador. Diferentes archivos de registro son estructurados de diferente manera, por lo que puede ser difícil que Logstash interprete diferentes formatos de manera transparente. Para poder realizar la lectura de los registros enviados por Filebeat, Logstash contiene una tecnología llamada filtros, los cuales son creados dependiendo la estructura de los archivos. Además de facilitar la lectura de los archivos, permite organizar la información entre diferentes campos para ser almacenada de manera organizada en Elasticsearch.

Elasticsearch es el corazón del correlacionador de datos ya que almacena toda la información enviada por Logstash. “Es una potente herramienta que nos permite indexar un gran volumen de datos y posteriormente hacer consultas sobre ellos. Al estar los datos indexados los resultados se obtienen de forma muy rápida.” (pico.dev, 2015) En este proyecto, Elasticsearch mantiene la información sobre los tweets organizada entre diferentes campos, los cuales pueden ser relacionados entre ellos para crear datos estadísticos. Entre estos campos se encuentran datos como el usuario de twitter que realizo el tweet, la fecha en

la cual se realizó, el hashtag y todo el contenido del tweet. Elasticsearch crea índices sobre estos campos para ofrecer búsquedas rápidas y cruces de información en tiempo real.

La cuarta herramienta que ofrece Elastic stack es Kibana. Esta es una herramienta que permite visualizar, buscar, e interactuar con los datos almacenados en Elasticsearch de manera gráfica (histogramas, gráficos de barras, gráficos pie, etc). Kibana es la herramienta que permite al usuario final utilizar la información para crear datos estadísticos e inteligencia de negocio en tiempo real. Además de crear varios tipos de visualizaciones, permite crear tableros en donde se agregan los gráficos creados, y realizar búsquedas sobre los mismos. Al igual que las visualizaciones, estos tableros pueden ser almacenados de manera permanente, y compartidos con otros aplicativos.

De lo anterior, se puede concluir el siguiente funcionamiento entre la aplicación web y el correlacionador de datos. El usuario final ingresa una búsqueda en la aplicación web, la cual obtiene resultados directamente de Twitter y los registra en un archivo de log. Cada vez que se crea un nuevo registro en este archivo, Filebeat se encarga de enviarlo a la herramienta Logstash. Por medio de filtros, Logstash convierte la información recibida en diferentes campos, y la almacena en la fuente de datos Elasticsearch. Toda la información almacenada en Elasticsearch es finalmente accedida por el mismo usuario utilizando Kibana, herramienta que permite al usuario visualizar la información de manera gráfica para obtener datos estadísticos sobre la búsqueda ingresada desde la aplicación web.

ALGORÍTMO DE BUSQUEDA

La aplicación web que funciona como un buscador de twitter realiza las búsquedas en base a un tema determinado. Sin embargo, es necesario que la aplicación obtenga resultados que tenga relevancia con el tema que desee el usuario. Por esta razón, esta aplicación ofrece un algoritmo para crear varios hashtags en base a palabras claves ingresadas por el usuario.

El funcionamiento de este algoritmo fue optimizado para que la creación de los hashtags sea rápida y certera. Para que este algoritmo pueda funcionar correctamente, el usuario debe ingresar entre 2 o 3 palabras claves. Una vez que estas han sido ingresadas, el algoritmo empieza a crear combinaciones posibles entre estas palabras claves. Estas combinaciones se convierten en posibles hashtags actualmente siendo compartidos por la red social. Pero, ¿cómo se puede asegurar que estas combinaciones en efecto sean hashtags?

Los posibles hashtags creados previamente son enviados a Twitter para verificar si en la red social existen tweets que contengan los posibles hashtags en las últimas 24 horas desde que se realizó la búsqueda. Entre los hashtags que en efecto han sido mencionados en la red social, se escogen los 3 con mayor mención en los tweets. Estos son los hashtags finales que serán utilizados para que la aplicación web realice las búsquedas respectivas, y los resultados obtenidos puedan ser enviados al correlacionador de datos.

Para mayor información sobre la funcionalidad del algoritmo, revisar Anexo C. Diagrama de Flujo de Algoritmo de Búsqueda.

RESULTADOS

La aplicación está orientada a la obtención de datos estadísticos, e inteligencia de negocio. Es una solución que permite al usuario observar tendencias sobre algún tema de interés en la red social Twitter. Mediante este trabajo se ha podido obtener los siguientes resultados sobre el buscador de twitter y el correlacionador de datos.

La aplicación web contiene el buscador de Twitter, el cual se ejecuta cada 30 segundos para obtener resultados sobre los temas ingresados por el usuario. Al ser un servicio que accede directamente a los servidores de Twitter, ha mostrado obtener resultados confiables. Es un servicio que Twitter ofrece a la comunidad, por ende, ha sido probado por muchos otros usuarios y desarrolladores. Este servicio permite el acceso a los servidores de Twitter cada 15 segundos, lo cual quiere decir que el buscador podría traer información en un periodo de tiempo muy corto. Sin embargo, como fue mencionado anteriormente, el buscador de este proyecto accede al servicio cada 30 segundos por la siguiente razón.

El objetivo principal de esta aplicación es analizar tendencias en un tiempo determinado. “Twitter es una red de contenido que se mueve rápidamente y cambia minuto a minuto. En muchas ocasiones ha sido la plataforma que ha dado a conocer noticias trascendentes, como cuando un avión se estrelló en Nueva York.“ (Smith, 2016) El buscador de Twitter ofrece resultados cada 30 segundos ya que existen casos en los cuales se desea analizar cómo cambia una tendencia minuto a minuto. Al ser un análisis de tendencias, en menos de un minuto no se puede obtener mayores resultados, por lo que sería innecesario cargarle a la aplicación con búsquedas con intervalos menores a 30 segundos. Además, siempre se debe tomar en cuenta el rendimiento de la aplicación. En consecuencia, un temporizador que se ejecuta cada 30 segundos es óptimo para el consumo recursos del servidor en el cual el buscador de Twitter está corriendo. Si se ejecutaría en intervalos

menores, se sobre utilizaría los recursos del servidor. De la misma manera, si los intervalos de tiempo serían mayores, es estaría sub utilizando los recursos.

Una vez obtenidos los resultados por el buscador de Twitter, se utiliza el correlacionador de datos para poder hacer los análisis respectivos. Para esto se utiliza la herramienta Kibana, que permite observar los resultados de manera gráfica. Los resultados que ha dado esta herramienta han sido positivos ya el manejo de información es interactivo para el usuario. La interfaz gráfica de esta herramienta permite al usuario crear relaciones entre los diferentes campos almacenados y visualizar los resultados en tiempo real. Por ejemplo, permite al usuario crear una gráfica en la cual ingresa los campos en los diferentes ejes (x, y, z) para crear las relaciones entre ellos. Estos campos pueden ser ajustados como el usuario desee, es decir, las relaciones entre los campos son infinitas.

Kibana es una herramienta utilizada para la inteligencia de negocios.

> "La Inteligencia de Negocios o Business Intelligence (BI) permite a las compañías contar con la información adecuada para una mejor toma de decisiones. Las compañías que implementan el BI logran sacar mayor provecho de las situaciones de crisis gracias a la posibilidad de contar con un análisis de mercado más acertado debido a que los datos pesados son transformados en importantes estrategias corporativas." (Universidad Esan, 2015)

Al igual que la mayoría de herramientas de inteligencia de negocio, Kibana requiere un conocimiento avanzado para su uso. La herramienta ofrece varios tipos de gráficos (área, barras, tablas, mapas, etc.), y cada uno tiene su propia implementación. Esto quiere decir que el usuario necesita un entendimiento completo para poder relacionar los campos. Caso contrario, el usuario se verá expuesto a crear visualizaciones difíciles de comprender y de

analizar. Sin embargo, una vez que el usuario aprende a manejar Kibana, la flexibilidad y oferta de funcionalidades es infinita.

PRUEBAS REALIZADAS

Control de calidad es una parte clave para el lanzamiento de cualquier producto. El aseguramiento de la calidad es una prueba y revisión de un producto. Se espera que este proceso de control de calidad descubra problemas de diseño y errores de desarrollo mientras se prueba la interfaz de usuario del producto y se mide la experiencia de usuario. (Thompson, 2015) Ya que este proyecto está orientado al análisis de tendencias sobre una noticia, y la inteligencia de negocios, se realizaron dos tipos de pruebas: Análisis periodístico sobre elecciones en el Ecuador 2017, e inteligencia de negocio para la marca BMFIT Gear.

En el año 2017 hubo elecciones presidenciales en el Ecuador. Como muchos otros eventos políticos que suceden en diferentes partes del mundo, fue un tema que creo mucha controversia, y por ende fue una gran tendencia en la red social Twitter. Fue el tema perfecto para ser analizado por el buscador de Twitter y el correlacionador de datos. Para este tema se utilizó el algoritmo de búsqueda y el correlacionador de datos. Se ingresaron algunas palabras claves, y con la ayuda del algoritmo se obtuvo un conjunto de hashtags. Estos hashtags fueron utilizados para obtener resultados de Twitter y generar datos estadísticos con la ayuda del correlacionador de datos.

El conjunto de hashtags obtenido por el algoritmo de búsqueda fue: #fraude, #Lenin, #Guillermo, #lassoguillermo, #leninpresidente, #fraudeecuador, #leninmoreno, #presidentelenin, #guillermolasso, y #fraudeelecciones. Para cada hashtag se realizaron los siguientes análisis. En un periodo de 3 días, ¿cuál es la cantidad de tweets que hablan sobre las elecciones presidenciales del Ecuador?, y ¿cuáles son los 10 usuarios que más hablan sobre este tema?

Para obtener los resultados sobre el primer análisis se utilizaron los gráficos incluidos en Anexo A. Gráficos sobre Elecciones Ecuador 2017. En la Ilustración 1 se puede observar la concurrencia de cada Tweet en la red social. Como eje Y se encuentra el conteo de Tweet, mientras que como eje x está la fecha de publicación. Además, el eje x está dividido entre los diferentes temas de búsqueda (hashtags), cada uno presentado en una línea de diferente color. Esta grafica permite visualizar que los hashtags con la mayor cantidad de Tweets son: #fraude, #lenin, y #guillermo. Además, como se puede observar en la Ilustración 2 del Anexo A, la herramienta ofrece mayores detalles sobre cada registro, lo cual permite realizar un análisis completo. Por ejemplo, mediante el detalle demostrado en la Ilustración 2, se puede concluir que el cambio en la cantidad de Tweets se debe a que la hora de publicación es las 2:30:00 am. Para poder verificar los gráficos con datos numéricos, Kibana ofrece la creación de tablas como se puede apreciar en la Ilustración 3.

Con el objetivo de analizar cuáles son los 10 usuarios que más hablan sobre las elecciones de Ecuador, se creó un gráfico de pie presentado en la Ilustración 4. Este grafico ofrece detalles sobre el número de Tweets que ha realizado cada usuario, como se puede apreciar en la Ilustración 5. Esto quiere decir que Kibana, no solo permite visualizar los datos de manera gráfica, pero también permite al usuario ver detalles sobre los datos presentados en las mismas. Al igual que en análisis previo, se realizó una tabla que permita visualizar los datos numéricos organizados por usuario, y por hashtag. De esta manera se puede interpretar la Ilustración 6, que el usuario de Twitter “DevsBotsECU” ha realizado 271 Tweets sobre los hashtags obtenidos por el algoritmo de búsqueda, mientras que ha publicado 166 con el hashtag “lenin”. Como se puede apreciar con los resultados presentados en esta gráfica, la herramienta permite al ámbito periodístico analizar una infinita cantidad de escenarios, incluso la orientación política de cada usuario de Twitter.

Por otro lado, la inteligencia de negocios permite a un negocio contar con la información adecuada para una mejor toma de decisiones. Para este proyecto, se ha decidido analizar la marca BMFIT Gear. Esta es una marca de equipo de fitness creada en los Estados Unidos por Bradley Martyn. Para este análisis se utilizó como temas de búsqueda los siguientes hashtags: #bmfit, #bradleymartyn. Mediante la creación de varios gráficos, se analizaron dos escenarios para ver el impacto que ha creado esta marca en la red social. Primeramente, se analizó el promedio de tweets que se hace en un lapso de 3 días sobre la marca. El segundo escenario fue determinar cuántos usuarios diferentes hablan de la marca en los 3 mismos días.

Los resultados obtenidos en este análisis son presentados en el Anexo B. Inteligencia de negocio para marca de fitness BMFIT. La tabla que se encuentra en la Ilustración 7 presenta un histograma organizado por los temas de búsquedas utilizados (bmfit, bradleymartyn). En la misma se puede apreciar que durante los tres días analizados, se publicó un promedio de 526 Tweets sobre la marca. Sin embargo, se puede apreciar un fenómeno ya que se realizan más publicaciones sobre el creador de la marca. Este dato puede ser muy útil para tomar una decisión comercial con el objetivo de que cada usuario que publica sobre Bradley Martyn, también publique sobre la marca BMFIT. Esto podría incrementar las ventas de manera significativa.

Por medio de la tabla presentada en la Ilustración 8 en el Anexo B, se puede determinar el número de usuarios nuevos que han realizado una publicación sobre la marca. En esta grafica se puede apreciar que ha habido 21 usuarios nuevos que han publicado sobre BMFIT, sin embargo, sigue habiendo una gran diferencia con el número de usuarios que hablan sobre el creador de la marca. Estos datos permiten apreciar que el creador de la marca Bradley Martyn es una mayor tendencia en las redes sociales que la marca.

CONCLUSIONES Y RECOMENDACIONES

Mediante la realización de este trabajo se puede concluir que un correlacionador de datos es una excelente herramienta para el análisis de tendencias en la red social Twitter. En el ámbito periodístico, un correlacionador de datos permite crear reportes estadísticos sobre noticias que ocurren todos los días. Ya que en este trabajo se utiliza una red social tan conocida como lo es Twitter, se puede analizar noticias locales, al igual que noticias de cualquier parte del mundo en tiempo real. En el ámbito empresarial, un correlacionador de datos sirve como una herramienta de Inteligencia de negocio, lo cual ayuda significativamente para la toma de decisiones en cualquier empresa. Permite visualizar el impacto que ha creado una empresa y cada uno de sus productos en el mercado. Por último, la forma que tiene esta herramienta de presentar los datos permite a los negocios mantener su información de manera distribuida y organizada.

Para el uso efectivo de esta herramienta se dan las siguientes recomendaciones:

Capacitar debidamente al personal que tendrá acceso a la herramienta Kibana. La fácil comprensión de los gráficos depende del análisis previo y experiencia del usuario como analista de datos.

Antes de la creación de cualquier gráfico, realizar un análisis profundo sobre los campos y las correlaciones creadas entre ellos.

No sobrecargar a la herramienta de temas innecesarios. Una vez que se considere al tema obsoleto, borrarlo en conjunto con sus hashtags relacionados.

LIMITACIONES Y TRABAJOS FUTUROS

El correlacionador de datos está compuesto de varias herramientas, lo cual resulta en un gran consumo de recursos de memoria. Mientras más temas de búsqueda sean ingresados por el usuario, mayor será el consumo. Este factor fue una limitación para este proyecto ya que, en un escenario ideal, toda la infraestructura del correlacionador de datos y la aplicación web debe ser ejecutado en un servidor. Sin embargo, este proyecto fue ejecutado en una laptop. Esta es una gran desventaja ya que las herramientas del correlacionador, como Logstash y Filebeat, necesitan acceso a memoria constante, lo cual quiere decir que mucha concurrencia de datos puede consumir muchos recursos y por ende, no permitir a las herramientas trabajar correctamente. En consecuencia, se creó un temporizador que hace una depuración de datos en los archivos log. Esto quiere decir que una vez que los datos son enviados al correlacionador, los mismos son borrados del archivo de registro. Esto permite que Filebeat solo trabaje sobre los nuevos resultados, disminuyendo el tiempo de ejecución.

La mayor limitación para el buscador de Twitter presentada como consecuencia de falta de memoria, es la concurrencia de búsquedas ingresadas en la aplicación. La cantidad óptima de temas siendo buscados al mismo tiempo es de 15. Por cada ingresada, se realizan varios pasos como el acceso a los servidores de Twitter para obtener los resultados, la escritura en el archivo de registros, la transferencia de los datos entre Filebeat, Logstash, y Elasticsearch, y, por último, la visualización de datos en Kibana. Todos estos pasos consumen recursos, y si existe una cantidad mayor a 15 temas concurrentes, las tareas mencionadas pueden fallar. Aunque la falta de recursos es una limitación importante de mencionar, puede ser fácilmente corregida al incrementar la cantidad de memoria y al mejorar el procesador del equipo donde la aplicación está instalada.

Como futuro trabajo se puede implementar la búsqueda de temas sobre otras redes sociales como Facebook, Instagram, y YouTube. Esta agregación mejoraría el análisis de

datos en el ámbito periodístico como en la inteligencia de negocios ya que, al estar vinculada sólo con una red social, se podría considerar que los datos están siendo sesgados.

REFERENCIAS BIBLIOGRÁFICAS (EJEMPLO ESTILO APA)

Dpt. De Ciencias de la Computación e Inteligencia Artificial. (2014). Introducción a Java Server Faces. *Título de Experto Universitario en Desarrollo de Aplicaciones y Servicios con Java EE.* Obtenido el 31 de marzo 2017 de http://www.jtech.ua.es/j2ee/publico/jsf-2012-13/sesion01-apuntes.html#Caracter%C3%ADsticas+de+JSF

Gutzba (2009). Que es Correlacionador de Eventos de Seguridad TI. *Gutzba's Weblog.* Obtenido el 30 de marzo 2017 de https://gutzba.wordpress.com/2009/12/30/que-es-correlacionador-de-eventos-de-seguridad-ti/

Herramientas de desarrollo de Java EE. (S.F.) *IBM Knowledge Center.* Obtenido el 5 de abril 2017 de https://www.ibm.com/support/knowledgecenter/es/SSRTLW_7.5.5/com.ibm.jee5.doc/topics/ctools.html

How Filebeat Works. (2017) *elastic.* Obtenido el 30 de marzo 2017 de https://www.elastic.co/guide/en/beats/filebeat/current/how-filebeat-works.html

Jboss Community (2017). What is Wildfly. *Wildfly.* Obtenido el 5 de abril 2017 de http://wildfly.org/about/

Universidad Esan. (2015) Las 20 herramientas de inteligencia de negocios que debes conocer. *Conexionesan.* Obtenido el 5 de abril 2017 de http://www.esan.edu.pe/apuntes-empresariales/2015/07/20-herramientas-inteligencia-negocios-debes-conocer/

.LOG Extensión de archivo (S.F.). *ReviverSoft.* Obtenido el 5 de abril 2017 de http://www.reviversoft.com/es/file-extensions/log

Logstash Introduction. (2017) *elastic.* Obtenido el 31 de marzo 2017 de https://www.elastic.co/guide/en/logstash/current/introduction.html

Nourie, D. (2016). Java Technologies for Web Applications. *Oracle Technology Network.* Obtenido el 31 de marzo 2017 de http://www.oracle.com/technetwork/articles/java/webapps-1-138794.html

Pico.dev (2015). Introducción a Elasticsearch. *Blog Bitix.* Obtenido el 21 de marzo 2017 de https://picodotdev.github.io/blog-bitix/2014/04/introduccion-a-elasticsearch/

¿Qué es un ORM? (S.F.). *Tu programación.* Obtenido el 5 de abril 2017 de http://www.tuprogramacion.com/glosario/que-es-un-orm/

Sissel, J. (2017). An Introduction to the ELK Stack (Now the Elastic Stack). *Elastic.* Obtenido el 30 de marzo 2017 de https://www.elastic.co/webinars/introduction-elk-stack

World's most Popular Open Source Database. (S.F.) *Oracle Mysql.* Obtenido el 5 de abril 2017 de https://www.oracle.com/mysql/index.html

ANEXO A: GRÁFICAS SOBRE ELECCIONES ECUADOR 2017

Ilustración 1. Gráfico distribuido por Temas de Búsqueda

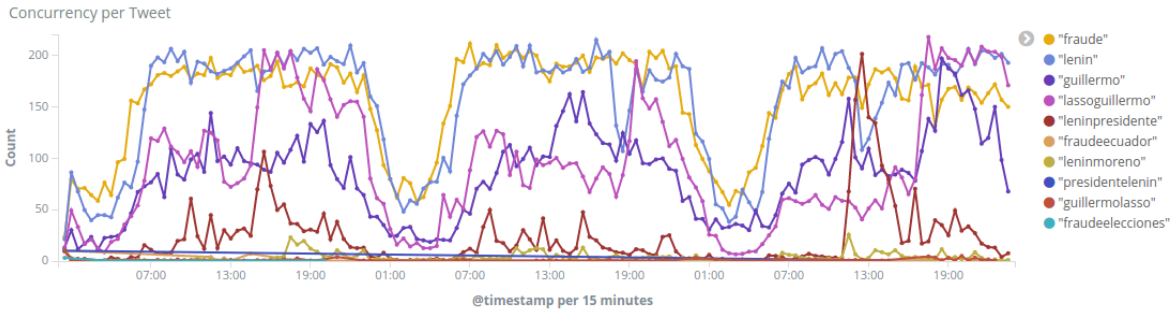

Ilustración 2. Detalle sobre el Tema de Búsqueda

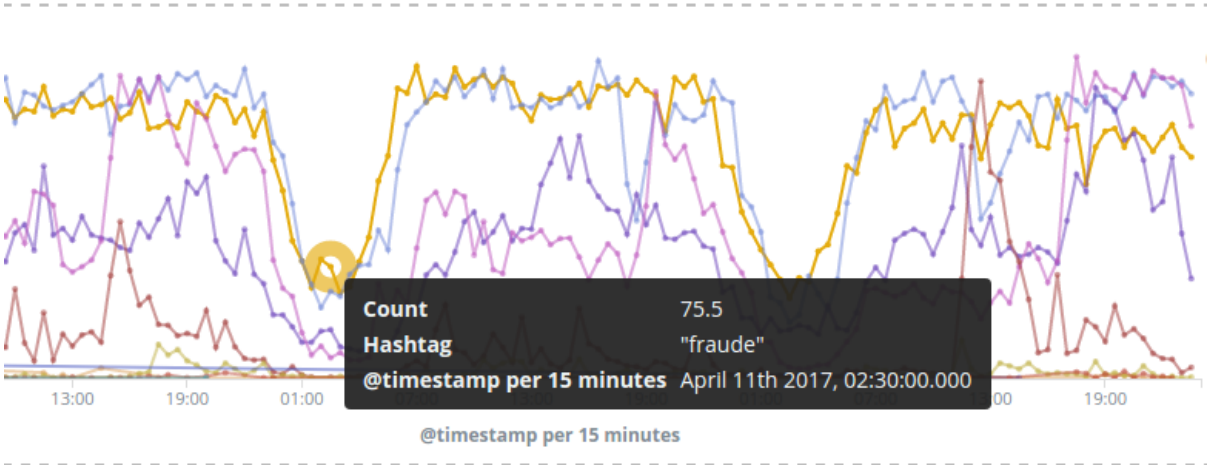

Ilustración 3. Conteo de Tweets dividido por Tema de Búsqueda

t_query: Descending	Count
"lenin"	44,718
"fraude"	44,688
"lassoguillermo"	26,712
"guillermo"	23,826
"leninpresidente"	6,231
"leninmoreno"	1,019
"guillermolasso"	164
"fraudeecuador"	76
"presidentelenin"	21
"fraudeelecciones"	8
	147,463

Ilustración 4. Diez Usuarios con mayor cantidad de Tweets

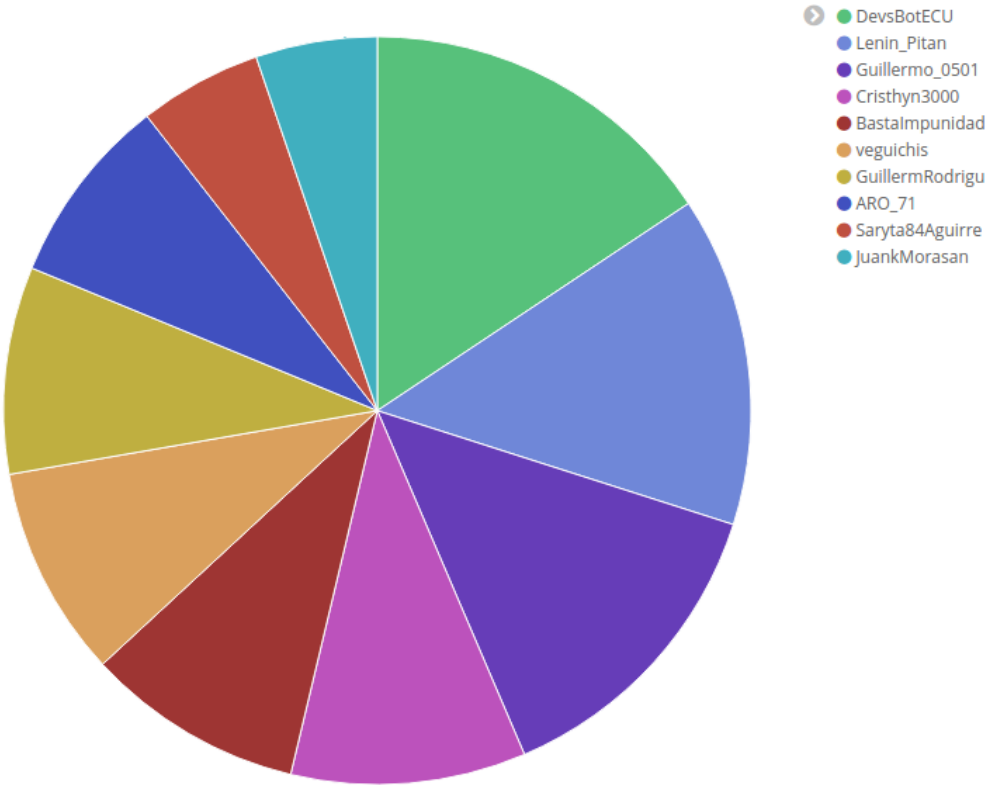

Ilustración 5. Detalle sobre Usuario con mayor cantidad de Tweets

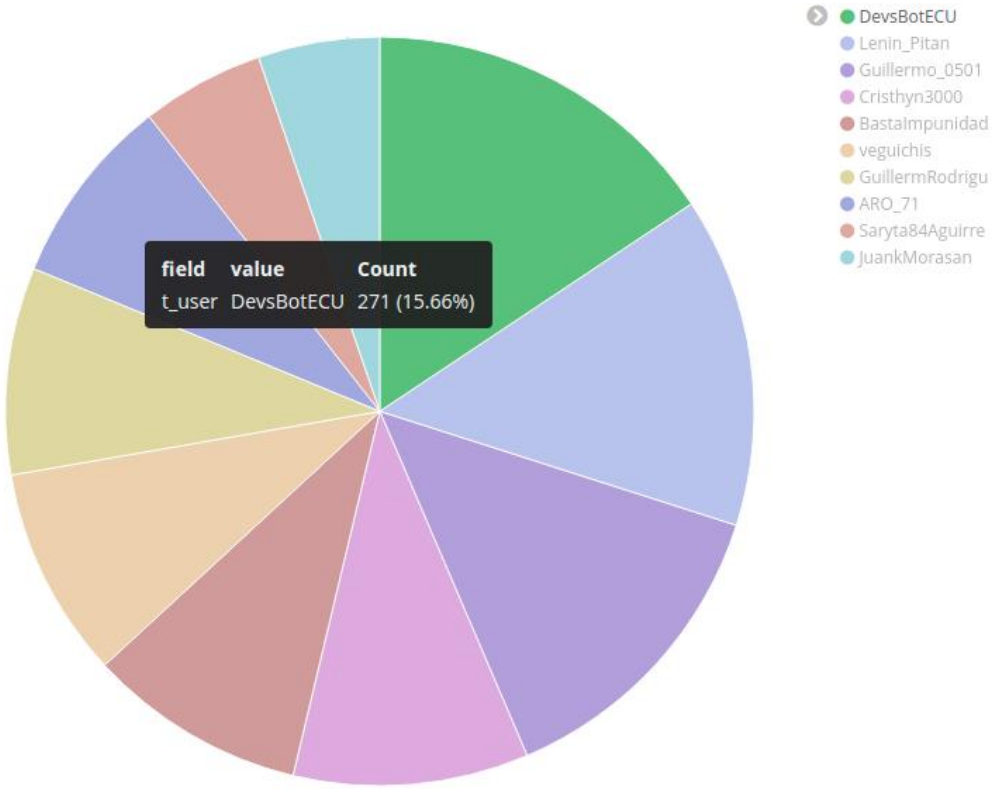

Ilustración 6. Conteo de Tweets de Usuario con mayor concurrencia organizado por Tema de Búsqueda

Twitter user	Count		
DevsBotECU	271	"lenin"	166
DevsBotECU	271	"lassoguillermo"	65
DevsBotECU	271	"leninpresidente"	23
DevsBotECU	271	"fraude"	13
DevsBotECU	271	"guillermo"	4

ANEXO B: INTELIGENCIA DE NEGOCIO PARA MARCA DE FITNESS BMFIT

Ilustración 7. Histograma organizado por Tema de Búsqueda

@timestamp per day	t_query: Descending	Count
April 10th 2017, 00:00:00.000	"bradleymartyn"	1,381
April 10th 2017, 00:00:00.000	"bmfit"	24
April 11th 2017, 00:00:00.000	"bradleymartyn"	715
April 11th 2017, 00:00:00.000	"bmfit"	4
April 12th 2017, 00:00:00.000	"bradleymartyn"	1,031
April 12th 2017, 00:00:00.000	"bmfit"	2
1,491,886,800,000		**526.167**

Ilustración 8. Cantidad de Usuarios que crearon un Tweet sobre cada Tema de Búsqueda

Tema en Twitter	Conteo unico de usuarios
"bradleymartyn"	2,294
"bmfit"	21

ANEXO C: DIAGRAMA DE FLUJO DE ALGORITMO DE BUSQUEDA