HADOOP (PARTE 2)

EXTRACIÓN, ALMACENAMIENTO Y ANÁLISIS DE TWEETS

Juan Casado Ballesteros

GitHub: https://github.com/JuanCasado/ADVANCED_DATABASES/tree/master/P3

ÍNDICE

INTRODUCCIÓN

En esta práctica se continuará trabajando desde la plataforma de explotación de datos de Twitter construida en la práctica anterior. Dicha plataforma implementa una arquitectura lambda con componente del entorno Hadoop y sin realizar la parte de la arquitectura correspondiente con datos en tiempo real.

Para cargar los datos en la arquitectura se utilizan Flume y Pig. Se utilizan un total de cinco configuraciones distintas de los canales de Flume, para extraer datos de Twitter, estas configuraciones se corresponden con:

- Extracción de tweets en Madrid.
- Extracción de tweets en Barcelona.
- Extracción de tweets en la Península Ibérica.
- Extracción de tweets en los que se incluye una lista de palabras
- Extracción de tweets en los que se incluye una lista de palabras y se han publicado dentro de la Península Ibérica.

Una vez recogido aproximadamente un bucket de datos en hdfs para cada uno de los canales estos se procesan con Pig donde se eliminan los Tweets que no contengan ningún match con una expresión regular. Posteriormente se cargan en Hbase y finalmente quedan accesibles desde Hive.

Este proceso puede estar encendido funcionando en paralelo mientras re realiza el análisis almacenando cada vez más y más datos. Según se realicen las consultas estas retornarán cada vez más información, pues cada vez habrá más datos disponibles, no obstante, durante todo el proceso podremos realizar consultas sin en ningún momento notar por el tiempo en que estas se resuelven si se están cargando datos en las bases de datos o no.

El objetivo de todos estos filtros y restricciones en la extracción es obtener datos relacionados con el COVID-19 para su posterior análisis.

El análisis realizado de ha planteado desde un punto de vista explorativo, es decir, para descubrir las posibilidades de los datos recogidos, aprender a trabajar con Hadoop y Hive y poder saber cómo es trabajar con datos provenientes de una fuente con tanto potencial como Twitter. El objetivo del análisis por tanto no ha sido obtener resultados fiables de datos recogidos de forma masiva y de los que se pudiera extraer información de utilidad.

Los datos recogidos son reducidos y se han capturado a lo largo de 5 horas en dos días distintos recogiendo un total de 586 Tweets. En caso de haber utilizado filtros menos restrictivos o haber recolectado datos durante más tiempo esta cantidad sin duda hubiera sido mucho mayor.

TRABAJO REALIZADO

El enfoque original de esta práctica fue muy distinto del que finalmente se ha realizado. Originalmente se planteó utilizar herramientas como Knime, RapidMiner, Pentaho u otras similares. Estas son herramientas profesionales muy utilizadas actualmente pues se integran con Hadoop y permiten utilizar sus componentes de una forma sencilla e intuitiva.

Se exploraron ejemplos en los que se utilizaba Spark como plataforma de cómputo sobre la que poder ejecutar los algoritmos de MLib y ejemplos en las que se utilizaba MapReduce como plataforma de cómputo para utilizar algoritmos de Mahout, todo ello sin salir de la interfaz gráfica de RapidMiner.

No obstante, los problemas encontrados para conectar Hive a cualquiera de estas plataformas por medio de la interfaz jdbc ha impedido utilizarlas.

Como solución se ha implementado una conexión alternativa a Hive que permite enviar consultas por medio de una interfaz HTTP REST y recibir las respuestas en formato JSON. Esta interfaz se ha construido por medio del secuestro del proceso hive-cli de modo tal que se obtiene una puerta al intérprete de consultas de Hive. Posteriormente la habilidad de controlar las consultas realizadas a este proceso se ha expuesto con un servidor REST desde el cual se ofrece un amplio abanico de posibilidades para consultar.

La principal desventaja de realizar esto es que por ser una interfaz no convencional las plataformas mencionadas de análisis de datos no son compatibles con ella.

El análisis realizado utilizando esta interfaz puede clasificarse en tres categorías. Análisis de contenido de los Tweets, en concreto de su texto, análisis de las relaciones entre los Tweets capturados (menciones entre usuarios) y finalmente análisis de la ubicación de los Tweets la cual por desgracia no está presente en todos ellos.

PROBLEMAS ENCONTRADOS

Se mencionan a continuación una serie de problemas generales que surgen a la hora de analizar Tweets, así como problemas concretos surgidos a lo largo de la realización de la práctica.

El principal problema a la hora de analizar los Tweets capturados proviene de los idiomas en los que estos se encuentran. Analizar texto en cualquier idioma que no sea inglés tiene series desventajas. La mayoría de los programas que permiten analizar la información de textos están solo diseñados y probados para trabajar en inglés proporcionando de entre ninguno a muy reducido soporte para otros idiomas. Adicionalmente al extraer datos en la península se obtiene una sorprendente cantidad de Tweets en otros idiomas, principalmente catalán y portugués. Para más inri todos estos idiomas tienen acentos otros caracteres fuera de ASCII por lo que dichas letras aparecen como corruptas y por consiguiente impiden el uso de las palabras que las contienen. En algún punto dentro de Hadoop estos caracteres se convierten en una "?" lo que impide su reparación de forma sencilla.

El segundo problema descubiertos es que gran parte de la información que se transmite en Twitter y otras redes sociales no es en forma de texto. Las imágenes y vídeos juegan un papel cada vez más importante en la comunicación lo cual implica necesitar más tiempo y recursos para realizar análisis correctos.

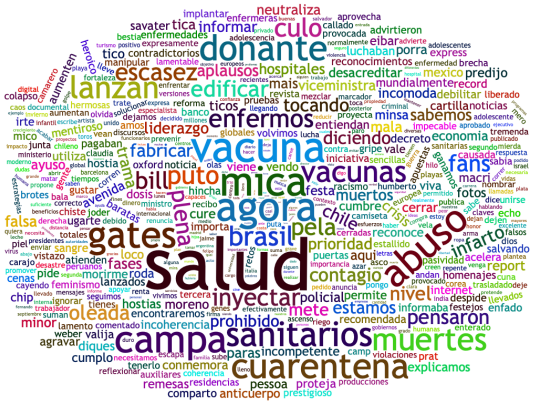
Finalmente se desea reiterar en el problema encontrado al hacer uso de la interfaz jdbc. Estos problemas, se cree, vienen derivados del uso de una instalación propia de Hadoop. Desde el entorno de Hadoop es posible acceder a la interfaz jdbc de Hive, no obstante, desde las plataformas mencionadas de análisis de datos no. En ellas se proponen instalaciones profesionales de Hadoop como Hortonworks o Cloudera cuyas versiones no coinciden con la construida. Adicionalmente En la máquina en la que dichas plataformas fueron instaladas corre JAVA-11 siendo JAVA-8 la versión soportada por Hadoop.

ANÁLISIS DEL CONTENIDO

Para analizar el contenido de los Tweets se ha utilizado la librería Lorca. Esta librería permite el análisis de textos en castellano proporcionando distintas métricas como: tiempo de lectura, sentimiento de las frases, análisis estadístico, importancia de las palabras, concordancia del texto, dificultad de lectura y otros.

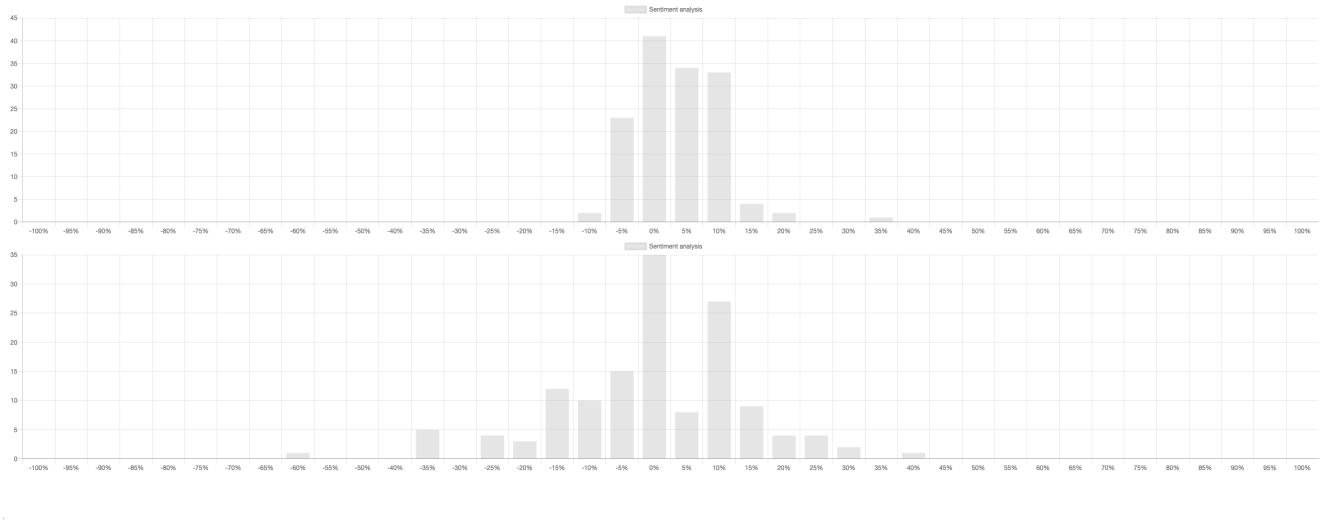
De todos ellos se han explorado el análisis estadístico, el de sentimiento de las frases y el de la importancia de las palabras. En la siguiente gráfica se muestra una captura de una página web que se conecta a la interfaz REST implementada y en la que se dibujan dos nubes de palabras. En la primera nube aparecen las palabras con mayor frecuencia de aparición, estas por supuesto coinciden con las utilizadas en los filtros de Pig.

En la segunda nube aparecen las palabras presentes en los Tweets ordenadas por importancia. La librería Lorca proporciona una base de datos en las que se recogen una gran cantidad de palabras en castellano asignando para cada una un índice de importancia.



Los siguientes histogramas muestran un análisis del sentimiento de los Tweets capturados. Para cada mensaje se ha calculado su sentimiento. Este cálculo se realiza de dos métodos distintos.

En el primer histograma se hace uso de una base de datos que nos relaciona palabras con un porcentaje positivo o negativo del sentimiento al que representa. En el segundo se utiliza esta implementación http://www.sciencedirect.com/science/article/pii/S0957417414001997 para calcular esa misma información.



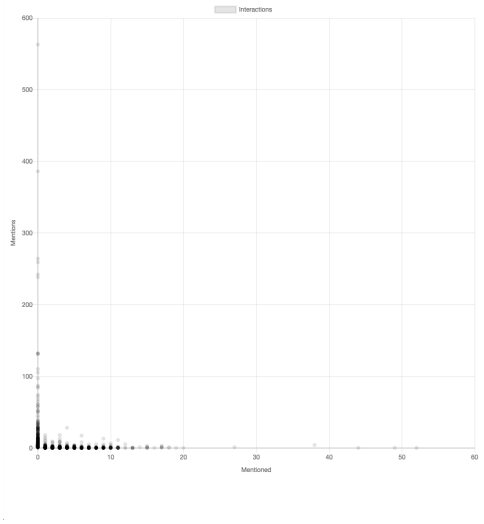
Como vemos los resultados son comparables tendiendo a mostrar un sentimiento neutro. Con la implementación primera se muestra una mayor presencia de sentimiento positivo (hacia los positivos) mientras que con la segunda son más los Tweets categorizados como de sentimiento negativo (hacia los negativos).

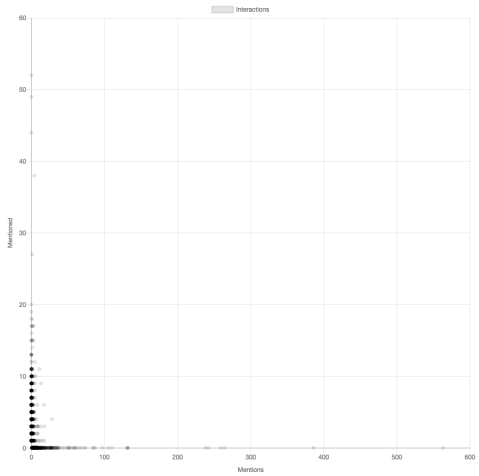
ANÁLISIS DE LAS RELACIONES

Para analizar las relaciones se ha creado de nuevo una página web en la que mostrar una gráfica de puntos 2D. Cada punto se corresponde con las interacciones realizadas por un usuario, es decir, la cantidad de veces que ha sido mencionado con respecto a la cantidad de veces que él ha mencionado.

Se puede ver que la tendencia indica que existe un gran número de usuarios que mencionan y un gran número de usuarios que son mencionados siendo menor la cantidad de usuarios que reciban ambos tipos de interacción.

Ambas gráficas representan los mismos datos, solo cambia que los mencionados aparecen en un eje o en el otro.

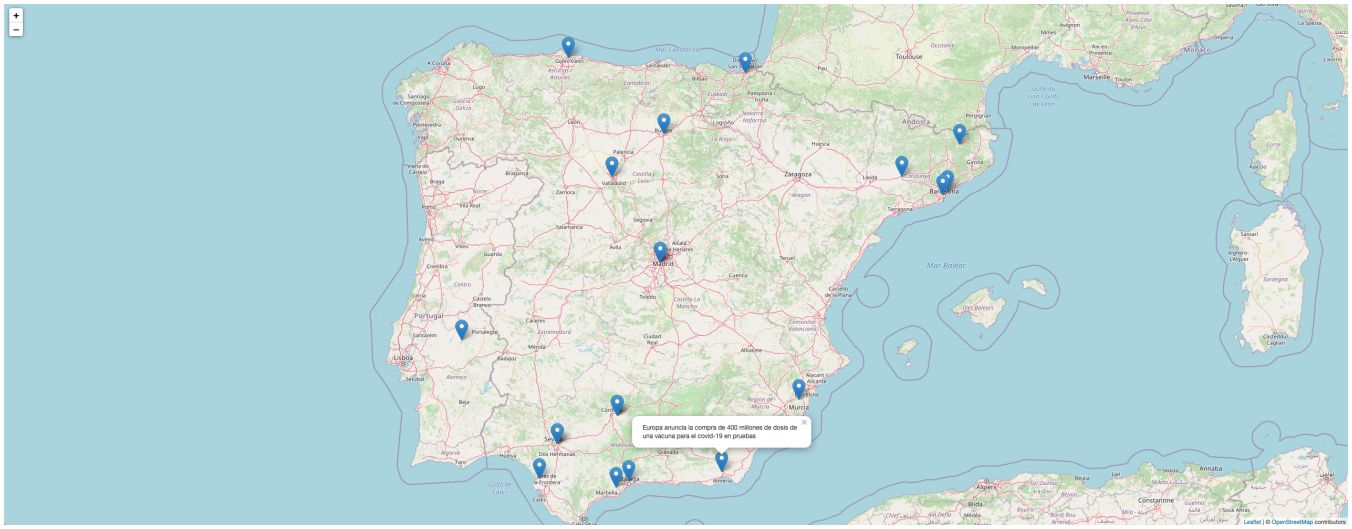




Si la relación entre mencionados y gente que menciona fue más equitativa aparecería una concentración de puntos sobre la recta X=Y o rotaciones desde el origen de esta misma recta si hubiera algún tipo de regresión positiva.

ANÁLISIS DE LA UBICACIÓN

La última página web implementada muestra la ubicación de los Tweets que contienen esta información en un mapa. Como vemos las ubicaciones de estos coinciden con núcleos urbanos dentro de las áreas delimitadas donde extraer datos con Flume.



Al hacer click sobre cada ubicación se desplegará un Pop-Up que nos mostrará el contenido del Tweet relacionado con ella.