Taller 5 – Análisis de Sentimientos en Spark

Prof. Fabio González
Diplomado Big-Data 2016
Universidad Nacional de Colombia

NOTA: *Lea el taller completamente antes de empezar a desarrollarlo, así tendrá una idea clara de lo que toca hacer.*

En el presente taller vamos a entrenar un modelo de análisis de sentimientos sobre tweets. Para ello vamos a usar el corpus descrito en:

http://www.sepln.org/workshops/tass/2015/tass2015.php#corpus

La dirección al conjunto de datos específico es la siguiente:

https://raw.githubusercontent.com/imendibo/SEPLN-TASS15/master/DATA/general-tweets-train-tagged.xml

1. Cargue el archivo en su cuenta de Databricks.

2. Extraiga cada de cada tweet el contenido *content* y la etiqueta del sentimiento *sentiment.polarity.value.*

3. Pre-procese el contenido de cada tweet de la siguiente manera:

    a. Convertir a minúsculas.
    b. Eliminar acentuación y puntuación.

4. Pre-procese la etiqueta del sentimiento de la siguiente manera:

    a. Seleccione únicamente el valor para el cual *sentiment.polarity.entity = "null"*.
    b. Filtre las columnas para las cuales la etiqueta del sentimiento es igual a "*NONE*".
    c. Convierta las etiquetas con el valor "*P+*" a "*P*" y las etiquetas con el valor "*N+*" a "*N*".

5. Utilizando la librería gensim, cargue el modelo entrenado Word2Vec que se encuentra en:

    http://dis.unal.edu.co/~fgonza/courses/eswikinews.bin

6. Una vez cargado el modelo Word2Vec debe construir el vector de características "*features*" para cada tweet de la siguiente manera:

    a. Obtenga el *vector 300* para cada palabra en el tweet utilizando el modelo entrenado.
    b. Construya el vector de características del tweet como el promedio de los *vectores 300* de cada palabra en el tweet.

7. Construya un par de clasificadores (algunos de los utilizados en sesiones anteriores) y evalúe su desempeño utilizando un *random-split* 70-30 (70 entrenamiento, 30 prueba) y utilizando validación cruzada.

8. Recolecte tweets de su interés y clasifíquelos. Evalúe los resultados cualitativamente.

9. Entregables:

    a. Para los puntos donde sea necesaria realizar una descripción, realícela en una celda de tipo markdown o en un documento PDF aparte.
    b. Descargue el notebook completado en formato IPython Notebook y comprímalo junto con los anexos como nombre_apellido_tallerWord2Vec.zip
    c. Cargue el archivo al siguiente enlace Dropbox: https://www.dropbox.com/request/4Tj0YKhyrwiALxVTJ5La

Material de Apoyo:

- Talleres y demos desarrollados en la sesiones anteriores.
- https://radimrehurek.com/gensim/models/word2vec.html
- ¡Google!