Escuela Colombiana de Ingeniería Julio Garavito
Análisis Computacional de Datos
Taller # 2 - Spark
Profesor Cristian C. Garzón Alfonso

ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO

Instrucciones:

- Use el archivo de “trump_tweets.txt” dado por el profesor en clase para el desarrollo de los puntos del taller anterior. Si puede haga uso del archivo completo.
- Deben leer el archivo con los tweets y escribir los resultados del HDFS.
- El taller se debe entregar antes del miércoles 15 de marzo del 2020 vía Moodle.
- Este taller debe desarrollarse en Spark corriendo en el clúster compatible con YARN. El código debe estar hecho en Python.
- El entregable de este taller es un archivo .zip en el que se incluya los 5 archivos .py de cada uno de los puntos incluyendo el código de la gráfica, no se requieren datasets de resultado en el .zip. Además, incluya un pantallazo con los resultados del data set y de la gráfica por cada punto.

1. Cuente cuantas veces aparece la palabra “TRUMP”, “DICTATOR”, “MAGA”, “IMPEACH”, “DRAIN”, “SWAP”, “CHANGE”.
2. Cuente el total de veces que aparece cada palabra eliminando las famosas palabras conocidas como “Stop Words” (a, an, and, are, etc). Deben eliminar TODAS las palabras que se consideran como “Stop Words” no solo las que usted quiera o considere necesarias.
3. Cuente el total de veces que un usuario escribió tweets.
4. Cuente el total de veces que se repitió una etiqueta o hashtag.
5. Cuente por cada lenguaje el total de tweets que se escribieron en este. Haga uso de la etiqueta "lang" que viene en la estructura JSON del tweet.

Usen librerías de Python puede ser D3 o se lo dejo a elección propia de cada grupo que les permita graficar el top de los mejores 10 de cada uno de los aspectos de los 5 puntos anteriores. Este punto es a libre opción y creatividad de cada grupo. Crean que están vendiendo esto a unos inversionistas y les quieren comprar el código que usaron para las gráficas, es decir hagan su mejor esfuerzo por graficar esto.