UNIVERSIDAD
EL BOSQUE

Ingeniería de Sistemas
Big Data Analytics
Taller 3

Enlace de entrega: aquí
Fecha máxima de entrega: **Noviembre 28, 2021 11:59 PM**

A partir del tutorial de análisis de sentimientos en Twitter detallado en este repositorio, plantee algunas estrategias para mejorar las métricas de *precision*, *recall* y *F1* actualmente obtenidas. Considere y experimente sobre los siguientes supuestos:

1. **Para construir un mejor modelo se requieren más datos etiquetados.** Una estrategia para validar este supuesto es mediante el etiquetado incremental de nuevos grupos de datos (por ejemplo, de 100 en 100) y re-entrenando el modelo. Se espera que alguna de las métricas objetivo aumente en la medida en la que incrementa la cantidad de datos etiquetados usados para el entrenamiento.
2. **El entrenamiento se está realizando con datos mal etiquetados.** A partir de las predicciones arrojadas por el modelo sobre los dataset de entrenamiento o de prueba, extraiga los datos que están cayendo en los dos tipos de errores, falsos positivos y falsos negativos. Revise manualmente si cada uno de estos errores puede deberse a que los datos están mal etiquetados, es decir, el modelo los está reconociendo bien pero debido a que tienen una mala etiqueta están siendo considerados como errores. Si se realiza algún ajuste sobre las etiquetas, re-entrene el modelo y vuelva a revisar los resultados.
3. **El modelo de Logistic Regression no es suficiente para lograr una discriminación adecuada de los tweets.** Entrene nuevos modelos utilizando otros algoritmos y evalúe los resultados. Pruebe utilizando otros algoritmos, variando además algunos de sus hiper-parámetros, como:

    a. Decision Trees: *max_depth*: 3, 6, 9
    b. Random Forest: *max_depth*: 3, 6, 9, *n_estimators*: 100, 200, 300
    c. Support Vector Machines: *kernel*: poly, rbf, *degree*: 2, 3, 4, *gamma*: scale, auto
    d. K-Nearest Neighborhoods: *n_neighbors*: 3, 5, 7
    e. Multinomial Naive-Bayes

f. Neural Networks: *hidden_layer_sizes*: (100,), (100, 100), (100, 100, 100), (200,), (200, 200), (200, 200, 200)

(Opcional) Investigue, para cada algoritmo, que otros parámetros son relevantes para el entrenamiento. Para una experimentación más ágil, considere utilizar una búsqueda en grilla.

Forma de entrega

Suba al aula virtual, antes de la fecha estipulada, un repositorio de GitHub en el que adjunte los *scripts* con la implementación de los puntos 1, 2 y 3. Asegúrese de que los scripts contengan los outputs con el resultado de la ejecución de cada bloque.

Grabe y suba a YouTube un video de no más de 5 minutos con la explicación del trabajo realizado. Utilice el archivo Readme para referenciar el video, listar los integrantes del equipo y mencionar cualquier otro aspecto que desee destacar.