MINERÍA DE DATOS WEB - TRABAJO PARA FINAL

Utilizando el dataset del Trabajo Práctico Nro 3 (Review Polarity) desarrollar los siguientes puntos:

1. Realizar el pre-procesamiento de los textos aplicando al menos dos técnicas de las estudiadas. Analice los cambios en el dataset en cada paso (por ejemplo, reducción de dimensionalidad)

2. Utilizando la técnica de Hold-out, divida el dataset en un conjunto para entrenamiento (80%) y otro para prueba (20%)

3. Con el dataset de entrenamiento, evaluar al menos 3 clasificadores de los vistos en clase para la clasificación de reviews en positivos/negativos. Seleccionar los hiperparámetros que considere adecuados en cada clasificador. Utilizando k-fold cross validation, compare y elabore conclusiones a partir de los resultados de los tres clasificadores, considerando no menos de 3 métricas.

4. Investigue y explique en qué consiste la técnica conocida como RandomForest y cuáles son sus principales características.

5. Entrenar un clasificador RandomForest a partir de los datos del dataset de entrenamiento obtenido en el punto 2 y compare sus resultados con los obtenidos por los clasificadores del punto 3 utilizando k-fold cross validation.

6. Evalué los cuatro clasificadores entrenados con el conjunto de datos de prueba reservado en el punto 2. Elabore conclusiones sobre el desempeño de los clasificadores.

ENTREGAR UN INFORME DETALLANDO LOS PASOS ANTERIORES Y EL/LOS NOTEBOOKS CORRESPONDIENTES

TENER EN CUENTA PARA DESARROLLAR EL TRABAJO LOS COMENTARIOS RECIBIDOS EN LA DEVOLUCIÓN DEL TP 3

FECHA DE ENTREGA **15 DE SEPTIEMBRE**