Laboratorio Calificado I

UPC Winter School 2023

Fecha de Entrega: agosto 01 (11pm)

Medio de entrega: https://www.dropbox.com/request/Z5XRRxasFLy6ZBPJiFdn

Condiciones de la entrega:

- No se admiten entregas fuera de tiempo.
- Solo debe entregar el notebook EJECUTADO. Si hay preguntas teóricas, las contesta en el mismo notebook en una celda de texto. No adjunte el dataset en la entrega.
- Formato para el nombre del ipynb nombre_apellido_lab1.ipynb. **Ejemplo**: ruben_marique_lab1.ipynb
- Solo se califica la primera entrega.

Dataset: Text emotion recognition (happy, sad)

- *Solo vamos a usar un archivo de los tres disponibles (train.csv)*: https://www.kaggle.com/datasets/shreejitcheela/text-emotion-recognition?select=train.csv

El objetivo del laboratorio es construir y comparar diferentes clasificadores. La etapa de procesamiento y las relaciones de partición del dataset son comunes para todos.

Procesamiento de texto: No contemple números, pase a minúscula, elimine caracteres individuales, tokenize por palabra, y aplique stemming.

Partición del dataset: Seleccionar 80% para entrenamiento, 20% pruebas.

Clasificador 1

- Representación: Bolsa de palabras con un vocabulario de 500 tokens, eliminando stop words(ingles).
- Modelo: Naive Bayes
- Salida: Reporte de clasificación y matriz de confusión.

Clasificador 2

- Representación: Bolsa de palabras con un vocabulario de 1000 tokens, eliminando stop words(ingles).
- Modelo: Naive Bayes
- Salida: Reporte de clasificación y matriz de confusión.

Clasificador 3

- Representación: Bolsa de palabras con un vocabulario de 5000 tokens, eliminando stop words(ingles).
- Modelo: Naive Bayes
- Salida: Reporte de clasificación y matriz de confusión.

Clasificador 4

- Representación: Bolsa de palabras con un vocabulario de 5000 tokens, tasa de aprendizaje de 0.1, eliminando stop words(ingles).
- Modelo: SGDClassifier
- Salida: Reporte de clasificación y matriz de confusión.

Clasificador 5

- Representación: Bolsa de palabras con un vocabulario de 5000 tokens, tasa de aprendizaje 100, eliminando stop words(ingles).
- Modelo: SGDClassifier
- Salida: Reporte de clasificación y matriz de confusión.

Preguntas en base a los resultados

- ¿Qué efecto sobre el F1-score y el accuracy tiene el incremento del vocabulario?, es bueno o negativo incrementarlo?
- ¿En base a los resultados del clasificador SGDClassifier y experimentación adicional que realice, para este problema que valor de la tasa de aprendizaje es apropiado?, ¿Vale la pena incrementarlo como en el clasificador 5?
- Los coeficientes que se obtienen del SGD son un indicativo de importancia de características. ¿Utilizando el clasificador 4, cuáles son las palabras más relevantes (importantes) para la tarea de clasificación?