Proyecto parte I

Polaridad de opinión usando diccionarios de términos afectivos

Especificaciones

- En equipo de 3 a 4 personas realice lo siguiente:
  - Cargue el corpus Rest_Mex_2022_Sentiment_Analysis_Track_Train
  - Extraiga el texto de las columnas Title y Opinion y obtenga los valores de las columnas Polarity y Attraction
  - Aplique tokenización y lematización al texto de las columnas Title y Opinion
  - Separe el corpus en dos conjuntos, uno de **entrenamiento con el 80%** de los datos y otro de **prueba con 20%** de los datos. Es importante que los datos se revuelvan (shuffle) antes de realizar la separación y que asigne una semilla fija mediante la variable **random_state = 0**
  - Utilizando el **conjunto de entrenamiento (80% de los datos)**, aplique el algoritmo que calcula el valor de la polaridad de opinión (positivo y negativo)

Especificaciones

- Siguiendo el método *rule based approach* planteado en el artículo *A Comparison Between Two Spanish Sentiment Lexicons in the Twitter Sentiment Analysis Task* realice experimentos para determinar el mejor umbral para los valores de polaridad (1-5)
- El mejor umbral es aquel que maximiza el valor de exactitud (accuracy)
- Utilizando el conjunto de prueba (20% de los datos), aplique el algoritmo que calcula el valor de la polaridad de opinión (positivo y negativo) y mediante el umbral determinado en el entrenamiento determine la polaridad de opinión (1-5)
- Calcule la exactitud, precisión, recall y F-measure de la polaridad determinada (predicha) por su método

Evidencias

- Código fuente
- Reporte donde se incluya
  - Una tabla con los rangos de valores de los umbrales probados en el conjunto de entrenamiento y los valores de exactitud (accuracy), precisión, recall y F-measure obtenidos por cada umbral. Puede basarse en la tabla 7 del artículo de referencia para mostrar esta información
  - Una tabla con los valores de exactitud (accuracy), precisión, recall y F-measure obtenidos por el umbral seleccionado en el proceso de entrenamiento aplicado al conjunto de prueba
  - Matriz de confusión de los valores predichos vs los reales en el conjunto de prueba