UNIVERSIDAD EAFIT
MAESTRÍA EN CIENCIA DE DATOS Y ANALÍTICA
ST1800 ALMACENAMIENTO Y RECUPERACIÓN DE INFORMACIÓN, 2023-2
PROFESOR: EDWIN MONTOYA – emontoya@eafit.edu.co

Trabajo 2 – analítica de texto

Fecha de entrega: hasta el 18 de septiembre de 2023

Descripción del trabajo 2

Durante la sesión de sistemas de indexación de documentos para IR, sistemas de recomendación y analítica de texto (NLP y text mining - TM), se pudo ver diferentes técnicas de preparación de texto (tokenización, normalización, remoción de stopwords, stemming y lematización; luego la representación de documentos con diferentes técnicas como BoW, word2vec, doc2vec, principalmente; además de selección de características (features) como bit-vector, tf, tf-idf, word2vec, etc son algunos de los esquemas. Una vez preparado el texto, se tienen diferentes aplicaciones como los motores de búsqueda como Apache Solr, Sistemas de Recomendación y diferentes aplicaciones de la minería de textos. Una vez finalizada la representación de documentos y características, se puede ejecutar una variedad de modelos que nos permitirán extraer información y conocimiento. Se revisaron los sistemas de recomendación en general, modelos de clasificación, principalmente basados en Naive Bayes, si bien hay muchos otros, y un caso particular en la clasificación relacionada con el análisis de sentimientos y detección de tópicos principalmente con LDA. Este trabajo les permitirá aplicar los conocimientos vistos en la clase e investigar, en una de las siguientes alternativas, de acuerdo a su área de interés.

Desarrollo

Parte1: Aplicar las diferentes técnicas y modelos de preparación de datos, los cuales incluyen un proceso de tokenización, optimización del BoW (con reducción de dimensionalidad), representación de características y representación de documentos. El objetivo es obtener el BoW más óptimo (reducido) para pasar a la fase de representación de característicos y de documentos. (se tiene como columna de entrada 'text'). Realizar la preparación de texto tanto en 1) librerías python como nltk, spacy, gensim o una combinación ellas, y 2) en SparkML o SparkNLP utilizando pyspark.

(nota: en el caso de tokenización, ensaye: `from nltk.tokenize import TweetTokenizer`, un tokenizador especial para datos de twitter, compárelo contra el tokenizador estándar/convencional)

Parte2: <será anunciada en la sesión 4>

Parte3: <será anunciada en la sesión 4>

Se va emplear el siguiente conjunto de datos:

https://github.com/st1800eafit/st1800-232/blob/main/datasets/trabajo2/twitterClimateData.csv.zip

Requerimientos:

Cada uno de los grupos, deberá centrarse en realizar diferentes técnicas y métodos de Preparación de texto (tokenización / optimización, representación de características y representación de documentos; <otros requerimientos serán anunciados en la sesión4>

Se tendrán en cuenta 3 aspectos:

1. Preparación de datos, características y representación, así como análisis frecuencial por token o hashtags, nubes de palabras por token o hashtags.
2. <otros requerimientos serán anunciados en la sesión4>
3. <otros requerimientos serán anunciados en la sesión4>

Podrá usar cualquier librería de python de NLP (nltk, spacy, gensim, pyspark, etc)

Criterios y Rúbricas de evaluación

En síntesis, el alcance y criterios/componentes de evaluación que emplearemos en este trabajo son:

Parte1: 30% - preparación de texto, características, representación.

Parte1: 35% - <otros requerimientos serán anunciados en la sesión4>

Parte2: 35% - <otros requerimientos serán anunciados en la sesión4>

Regla de ética y transparencia para todas las alternativas:

Si encuentran soluciones públicas o de reuso de código de alguna de las partes requeridas en este trabajo debe EXPLÍCITAMENTE DECLARAR:

- Declarar explícitamente: De que referencias en kaggle, médium, datacamp, toward data science, o de otro sitio, ud empleo parte del código y la solución para realizar su propio trabajo.
- Declarar explícitamente: cual fue el aporte específico que el grupo realizó en el trabajo.

Entregables:

1. Carpeta con nombre: 'st1800-trabajo2-232' en google drive compartida a: edwin.montoya@gmail.com allí contendrá:

a. Datasets utilizados
b. Notebooks o programas en Python desarrollados
c. Toda la documentación completa debe estar en los mismos notebooks desarrollados

2. Enviar al momento de la entrega POR BUZÓN DE ENTREGA de la plataforma de Interactiva Virtual, especificando claramente los integrantes del trabajo2, links de nuevo a los entregables (no sobra) o adjuntos que considere.