Laboratorio 4.
Clasificación de tweets usando minería de texto.

INSTRUCCIONES:

Utilice el dataset Natural Language Processing with Disaster Tweets de Kaggle. Debe hacer un análisis exploratorio para entender mejor los datos, sabiendo que el objetivo final clasificar si un tweet se refiere a un desastre real no. Genere un informe en pdf con las explicaciones de los pasos que llevó a cabo y los resultados obtenidos. Recuerde que la investigación debe ser reproducible por lo que debe guardar el código que ha utilizado para resolver los ejercicios y/o cada uno de los pasos llevados a cabo si utiliza una herramienta visual. Incluya una nube de palabras que le ayude a detectar las que más se repiten. Este laboratorio debe realizarse en tríos (grupos de 3).

DESCRIPCIÓN DEL DATASET

El conjunto de datos está formado por más de 10 500 filas y 5 columnas:

- id: El identificador del tweet
- keyword: una palabra clave del tweet, puede estar en blanco
- location: la ubicación desde donde fue enviado el tweet
- text: El texto del tweet
- target: La etiqueta de clasificación que especifica si el tweet se trata de un desastre real (1) o no (0).

EJERCICIOS

1. Descargue el archivo train.csv
2. Cargue los archivos de datos a R o a Python, dependiendo de con qué trabaje.
3. Limpie y preprocese los datos. Describa de forma detallada las actividades de preprocesamiento que llevó a cabo.
   3.1. Se pueden hacer tareas como:
   - Convertir el texto a mayúsculas o a minúsculas
   - Quitar los caracteres especiales que aparecen como “#”,”@” o los apóstrofes.
   - Quitar las url
   - Revisar si hay emoticones y quitarlos
   - Quitar los signos de puntuación
   - Quitar los artículos, preposiciones y conjunciones (stopwords)
   - Quitar números si considera que interferirán en la clasificación (quizá debería valorar si quitar o no el 911).
4. Obtenga la frecuencia de las palabras tanto de los tweets de desastres como de los que no. ¿Qué palabras cree que le servirán para hacer un mejor modelo de clasificación?¿vale la pena explorar bigramas o trigramas para analizar contexto?
5. Haga un análisis exploratorio de los datos para entenderlos mejor, documente todos los análisis
   5.1. Puede, para cada archivo:
   - Investigar qué palabra se repite más en cada una de las categorías
   - Hacer una nube de palabras para visualizar las que aparecen con más frecuencia

- Hacer un histograma con las palabras que más se repiten
- Discutir sobre las palabras que tienen presencia en todas las categorías.

6. Elabore una función en la que el usuario ingrese un tweet y el sistema lo clasifique en desastre o no.

EVALUACIÓN

NOTA: La evaluación de cada integrante del grupo será de acuerdo con sus contribuciones al trabajo grupal

(25 puntos) Análisis exploratorio:
- Se elaboró un análisis exploratorio en el que se explican los cruces de variables, hay gráficos explicativos y análisis que permiten comprender el conjunto de datos.

(20 puntos) Limpieza y preprocesamiento de los datos:
- Se documentan las tareas de limpieza, incluyendo los paquetes/módulos que se usaron.

(20 puntos) Generación de los ngramas y cálculo de sus frecuencias y probabilidades:
- Se explica cómo se generaron los n-gramas y se calcularon los valores de frecuencia y probabilidades o cualquier otro análisis que permita clasificar los tweets.

(10 puntos) Algoritmo:
- Se describe el algoritmo que se usó para predecir.

(25 puntos) Función de clasificación.
- Se elaboró una función que permite clasificar nuevos tweets.

MATERIAL A ENTREGAR

- Archivo .pdf con el informe que contenga, los resultados de los análisis y las explicaciones.
- Link de Google drive donde trabajó el grupo.
- Script de R (.r o .rmd) o de Python que utilizó para hacer su análisis exploratorio y predicciones.
- Link del repositorio usado para versionar el código.

FECHAS DE ENTREGA

- **AVANCE:** Descripción de los datos, preprocesamiento y sus explicaciones, unigramas, bigramas, modelo preliminar de predicción: lunes 29 de agosto de 2022 18:55.
- **DOCUMENTO FINAL COMPLETO:** 1 de septiembre de 2022 a las 23:59
  NOTA: Solo se calificará el Documento Final si está entregado el avance con todo lo que se pide.

REFERENCIAS

Daniel Jurafsky, J. H. M. (101AD). Speech and Language Processing (2008), *1*. Retrieved from https://web.stanford.edu/~jurafsky/slp3/

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal Of Statistical Software*, *25*(5), 1–54. https://www.jstatsoft.org/article/view/v025i05

Jurafsky, D., & Martin, J. H. (2014). N-Grams. *Speech and Language Processing*, 2–7. Retrieved from https://lagunita.stanford.edu/c4x/Engineering/CS-224N/asset/slp4.pdf

PAQUETES ÚTILES DE R

- Quanteda
- Wordcloud
- Tm
- Rweka
- Ngram

MÓDULOS ÚTILES DE PYTHON

- Natural Language Toolkit (NLKT)
- Wordcloud