UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FACULTAD DE CIENCIAS FÍSICO-MATEMÁTICAS

Tarea 1
Procesamiento y clasificación de datos

Luis Enrique López Nerio
1487280

PROFESOR

Dra. MAYRA CRISTINA BERRONES REYES

ASIGNATURA

Procesamiento y clasificación de datos

19 de mayo de 2022

Índice

1. Problematica 2

2. Descripción de los datos utilizados 2

3. Preprocesamiento 3

4. Análisis de resultados 3

5. Conclusiones 3

6. Referencias 4

1. Problematica

Dentro de los problemas de procesamiento y clasificación de datos se encuentran los problemas relacionados con datos que contienen información alfanumérica, al ser la forma de comunicación escrita, los datos textuales son abundantes en muchos ámbitos.

Existen una gran cantidad de formas en las que estos textos se pueden utilizar para convertir los datos crudos en información que se traduzca en conocimiento o valor, algunos de los algoritmos utilizados sobre datos textuales son:

- K-vecinos más cercanos
- Análisis de sentimientos
- Sistemas de recomendación
- Clasificador Naive Bayes

Sin embargo, a pesar del amplio abanico de modelos que se pueden implementar con los datos almacenados como texto, para poder alimentarlos a un modelo, se necesita que esta información esté preparada para ser consumida por el modelo, el paso de normalizar los datos alfanumericos de tal manera que puedan ser utilizados por alguno de los algoritmos se conoce como preprocesamiento. Una vez realizado el preprocesamiento se pueden hacer análisis previos a la implementación de los modelos como analizar la frecuencia los términos presentes en los datos o longitud de los mismos.

Algunos de los pasos del preprocesamiento son:

- Transformación de maypusculas.
- Eliminar espacios en blanco.
- Expandir abreviaciones.
- Remoción de signos de puntuación.
- Lematización o derivación de raiz.

Para este reporte se realizara el preprocesamiento de un conjunto de datos particular y se presentaran el análisis descriptivo de este conjunto de datos.

2. Descripción de los datos utilizados

Los datos sobre los que se hará el preprocesamiento son datos provenientes de la plataforma Kaggle, específicamente la base contiene reseñas a libros que se vendieron por la plataforma de amazon y que su temática principal es la ciencia de datos, la base cuenta con 20,647 filas o entradas y 3 columnas que se describen en la tabla 1

Nombre de la columna	Descripción
stars	En esta columna se almacena la cantidad de estrellas que el usuario le dio al libro, este valor puede tomar desde 1 a 5 estrellas
comment	En esta columna se almacena la reseña o comentario realizado sobre el libro comprado.
book url	Esta columna almacena la url que el libro tiene en el sitio de amazon, se cuentan con hasta 836 libros diferentes.

Tabla 1: Descripción de las variables presentes en el dataset

3. Preprocesamiento

Para realizar el preprocesamiento se utilizara el lenguaje de programación python y la libreria `nltk`

1. Remover signos de puntuación.

2. Tokenizar las palabras.
   Este paso consiste en tomar una cadena de texto y separarla en sus palabras, se obtiene como resultado final una lista de las palabras contenidas en la cadena de texto

3. Convertir letras mayúsculas a minúsculas

   En este paso se eliminan signos como acentos, comas, dobles puntos, etc.

4. Remoción de stopwords En este paso se remueven las stopwords, las stop words son palabras que aportan poca información al análisis del texto, esto se puede deber a que se presentan con una alta frecuencia y pueden generar ruido en nuestro análisis o algoritmo. Hay un conjunto de stop words para cada idioma, sin embargo, las stop words también dependen del contexto de la información que se está analizando.

5. Lematización La lematización es el proceso de eliminar los afijos de las palabras y derivar la base lexica de una palabra, también conocida como lema, por ejemplo, transformar riéndose a reír.

A continuación se presenta la función en python que se encarga de realizar el preprocesamiento para una cadena de texto.

```
def preprocesamiento2(comentario):
    comentario =  re.sub('[^a-zA-Z]',' ',comentario)
    palabras = word_tokenize(comentario)
    palabras = [palabra.lower() for palabra in palabras ]
    palabras = [palabra for palabra in palabras if palabra.isalpha()]
    palabras = [palabra for palabra in palabras if not palabra in stop_words]
    palabras = [lemmatizer.lemmatize(palabra) for palabra in palabras]
    return palabras

```

4. Análisis de resultados

Primeramente se calculó una nueva columna que nos calculara la cantidad de palabras en las reseñas antes de eliminar stop words, para analizar la distribución de la cantidad de palabras en las reseñas se realizó un histograma el cual se puede ver en la gráfica 1.

Se puede notar por la escala de logáritmica que la distribución presenta una cola derecha, este significa que la cantidad de palabras por reseña se encuentra del lado izquierdo de la media, sin embargo se presentan reseñas muy extensas, con una gran cantidad de palabras.

Posteriormente se realiza el preprocesamiento, se realizaron dos enfoques, se realizó el preprocesamiento para una reseña en especifico y para todas las reseñas en general, en las gráficas 2a 2b, de manera similar se realiza un gráfico de nube de palabras que se puede observar en las figuras 3a 3b.
Las palabras con mayor frecuencia después de realizar el preprocesamiento están muy relacionadas con el contexto que es reseñas de libros de ciencia de datos.

5. Conclusiones

Con el análisis realizado se puede concluir que el preprocesamiento es un paso clave para el objetivo final, transformar los datos textuales en información que agregue valor. Uno de los siguientes pasos posibles sería alimentar estos datos a un algoritmo en especifico, debido al contexto de los datos, se podría realizar un análisis de sentimientos.

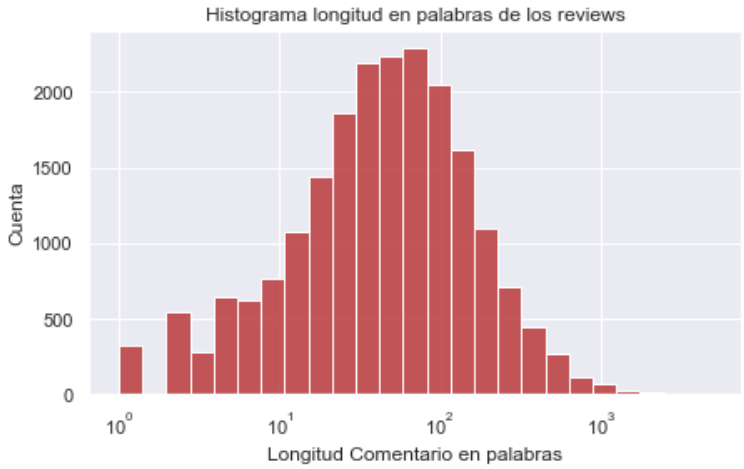
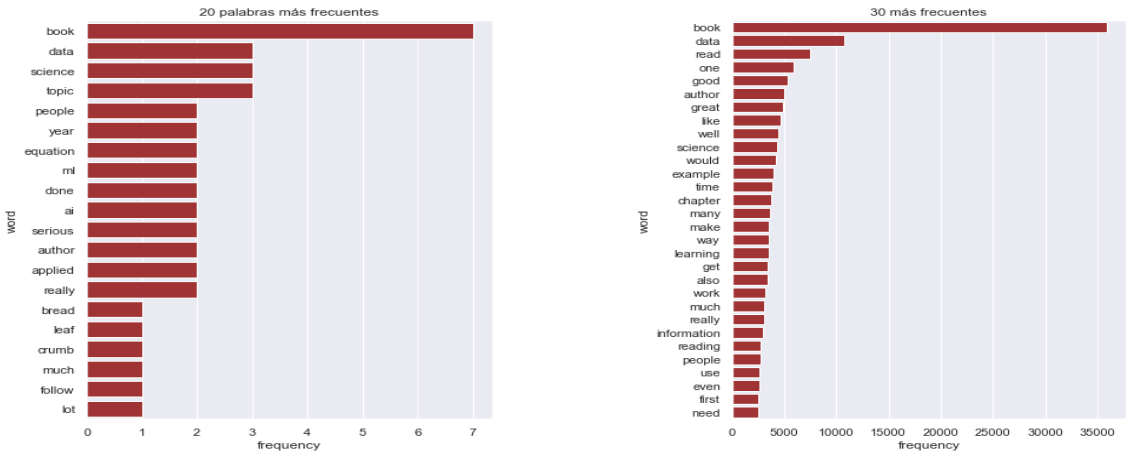


Figura 1: Histograma de la cantidad de palabras por reseña



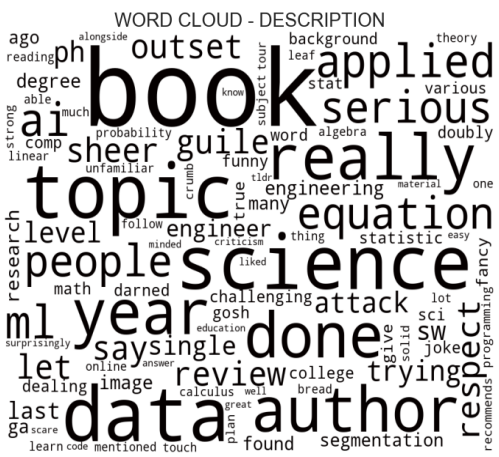
(a) Frecuencia de las 20 palabras más frecuente en un comentario específico

(b) Frecuencia de las 30 palabras más frecuentes en todas las reseñas

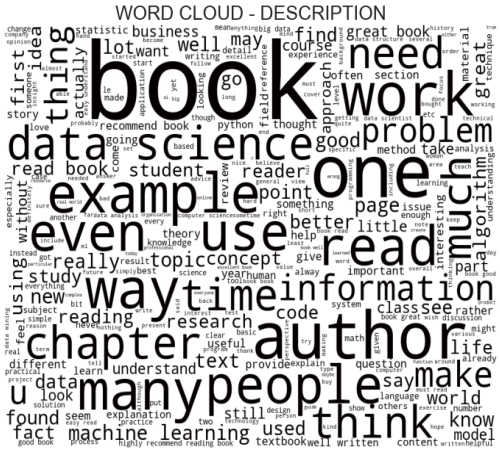
Figura 2: Frecuencias

6. Referencias

- Repositorio de Github
- Origen de los datos en la plataforma de Kaggle



(a) Wordcloud de un comentario específico



(b) Wordcloud de las palabras presenten en todos los comentarios

Figura 3: Wordclouds