Tarea 1: Preproceamiento de Textos

César jair Tamez Juárez

May 2022

1 Introducción

Los datos son la base de la ciencia de datos, es fundamental tener buenos datos para poder realizar un buen análisis y poder entrenar algoritmos de la mejor manera, en la mayoría de las veces los datos no están listos para ser utilizados en un análisis o en un algoritmo, siempre es necesario realizar una adecuamiento de los datos, a esto se le llama preprocesamiento o limpieza de datos. El preprocesamiento de los datos se encarga de dejar listos los datos para ser utilizados, este proceso es diferente dependiendo del tipo de datos con los que estes trabajando, en este caso nos enfocaremos en textos. Una de las cosas mas habituales que se realizan en el preprocesamiento de los textos es la eliminación de ciertos elementos que nos estorban, estos elementos pueden ser signos de puntuación, dígitos, espacios en blanco, texto entre paréntesis, se suele remover todos estos elementos ya que no son relevantes para el texto. Otra de las acciones que se realizan en el preprocesamiento de textos es la normalización, esto es hacer en minúsculas todas las letras del texto, esto debido a que para la computadora una letra en minúsculas es diferente a la misma letra, pero en mayúsculas y esto podría traer confusiones y errores en los análisis. También existen tres proceso muy utilizados en el preprocesamiento de texto, la Tokenizacion, la eliminación de Stop words y el Stemming o Lematización, la Tokenizacion consiste en separar en palabras individuales al texto, esto con el fin de que es más fácil procesar palabra por palabra a procesar el texto completo, por otro lado la eliminación de Stop words se refiere a la eliminación de todas esta palabras de conexión que se suelen utilizar y que realmente no aportan al texto, por ultimo la Lematización se encarga de llevar a su forma de raíz a cada una de las palabras que se encuentran en el texto, esto con el fin de reducir aun más las palabras diferentes con las que se cuenta. Todas estas acciones de preprocesamiento serán abordadas con un ejemplo en el presente trabajo.

2 Datos

Se tomaron como base los textos de la base de datos E-commerse, la cual fue obtenida de Kaggle en formato csv, esta base de datos cuenta con un total de 54,429 textos procedentes de cuatro tipos de sitios web, Household son los textos que vienen de sitios web que venden artículos para el hogar, Electronics que son de los sitios web que venden electrónicos, Books son los textos de sitios web donde se puede adquirir libros y por ultimo Clothing Accesories los cuales son los textos que vienen de sitios en las que se venden ropa y demás accesorios. Por limitaciones de equipo además de que el archivo csv que contenía la base tenía muchas inconsistencias, no estaba bien estructurada, se decidió que para el desarrollo del presente trabajo se utilizara una base más pequeña extrayendo algunos textos de la base original, se obtuvieron 1,004 textos, distribuidos de la siguiente manera:

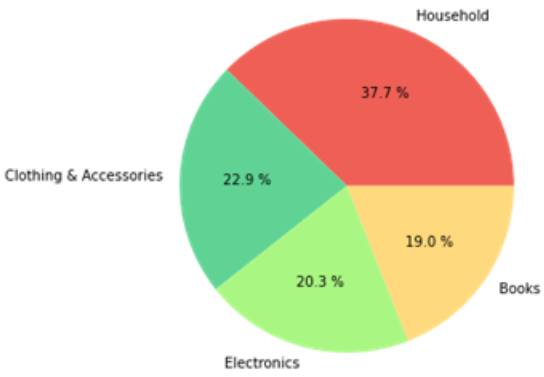

Figure 1: Parte del Dataset

Como podemos ver en la Figura 1 el 37.7% corresponde a la clase Household lo cual representa 379 textos, 230 a la clase Clothing Accesories que equivalen al 22.9%, 204 de la clase Electronics que corresponden al 20.3% y 191 de la clase Books los cuales son el 19% restante. Tenemos mas textos de la clase Household, mientras que de la clase Books es de a que menos textos tenemos, tener más elementos de una clase puede afectar el rendimiento de los algoritmos, pero consideramos que la diferencia entre las cantidades de textos con las que cuenta cada clase no es significativa. A continuación, tenemos un ejemplo de los textos que podemos encontrar en la base de datos:

- Sky Trends Plastic Basket with Artificial Flower and Plant (Violet and Multicolour, 22x10x12cm) A fantastic gift to say you care, for a Birthday present or Anniversary gift, a special gift that will create a talking point.This eye catching decorative miniature rickshaw fabricated Vase with high quality materials. It is replica of the original rickshaw.These

beautiful vase from the house of Sky Trends that will instantly enhance the visual appeal of your room.This is crafted by skilled craftsmen and it is designed for providing elegant and traditional look for home decor.

Este texto pertenece a la clase Household, como podemos ver son textos que promocionan un artículo, están escritos en el idioma inglés y necesitan un preprocesamiento para poder ser utilizados en cualquier algoritmo, el preproceso que se les realizo se vera en la siguiente sección.

3 Preprocesamiento

Como ya sabemos el preprocesamiento de los datos es una parte fundamental para el análisis de datos, en nuestro caso tenemos que aplicar un preprocesamiento o limpieza de los textos con los que contamos. Esta limpieza de los datos se realizó en Python utilizando diversas librerías, tales como pandas y numpy para la manipulación de los datos, la librería re para eliminar diferentes cosas que no son necesarias de los textos, la librería nltk que se utilizó para la limpieza de stopwords, para la lematización, para el steaming y para la Tokenizacion de los textos, también se utilizó las librerías matplotlib y wordcloud para graficar los resultados obtenidos en el preprocesamiento. Utilizando el texto dado de ejemplo en la sección anterior, ilustraremos cada una de las partes del preproceso que le realizamos a los textos.

Lo primero que hicimos fue normalizar los textos, esto se refiere a hacer en minúsculas todas las letras que contiene el texto:

- sky trends plastic basket with artificial flower and plant (violet and multicolour, 22x10x12cm) a fantastic gift to say you care, for a birthday present or anniversary gift, a special gift that will create a talking point.this eye catching decorative miniature rickshaw fabricated vase with high quality materials. it is replica of the original rickshaw.these beautiful vase from the house of sky trends that will instantly enhance the visual appeal of your room.this is crafted by skilled craftsmen and it is designed for providing elegant and traditional look for home decor.

Como podemos ver al normalizar el texto todas las letras se hicieron en minúsculas, ahora el siguiente paso fue eliminar signos de puntuación como paréntesis, comas, puntos, guiones, etc. El resultado fue el siguiente:

- sky trends plastic basket with artificial flower and plant violet and multicolour 22x10x12cm a fantastic gift to say you care for a birthday present or anniversary gift a special gift that will create a talking pointthis eye catching decorative miniature rickshaw fabricated vase with high quality materials it is replica of the original rickshawthese beautiful vase from the house of sky trends that will instantly enhance the visual appeal of your roomthis is crafted by skilled craftsmen and it is designed for providing elegant and traditional look for home decor

Ahora el siguiente paso fue, eliminar todos los dígitos que estén en el texto:

- sky trends plastic basket with artificial flower and plant violet and multicolour xxcm a fantastic gift to say you care for a birthday present or anniversary gift a special gift that will create a talking pointthis eye catching decorative miniature rickshaw fabricated vase with high quality materials it is replica of the original rickshawthese beautiful vase from the house of sky trends that will instantly enhance the visual appeal of your roomthis is crafted by skilled craftsmen and it is designed for providing elegant and traditional look for home decor

Lo siguiente fue eliminar espacios grandes entre palabras que pudieron quedar después de eliminar las cosas anteriores:

- sky trends plastic basket with artificial flower and plant violet and multicolour xxcm a fantastic gift to say you care for a birthday present or anniversary gift a special gift that will create a talking pointthis eye catching decorative miniature rickshaw fabricated vase with high quality materials it is replica of the original rickshawthese beautiful vase from the house of sky trends that will instantly enhance the visual appeal of your roomthis is crafted by skilled craftsmen and it is designed for providing elegant and traditional look for home decor

Después realizamos la Tokenizacion del texto, esto con el fin de separar el texto en palabras y que sea más fácil de procesar para los algoritmos, el resultado fue el siguiente:

- ['sky','trends','plastic','basket','with','artificial','flower',
'and','plant','violet','and','multicolour','xxcm','a','fantastic
','gift','to','say','you','care','for','a','birthday','present',
'or','anniversary','gift','a','special','gift','that','will','cr
eate','a','talking','pointthis','eye','catching','decorative','m
iniature','rickshaw','fabricated','vase','with','high','quality'
,'materials','it','is','replica','of','the','original','rickshaw
these','beautiful','vase','from','the','house','of','sky','trend
s','that','will','instantly','enhance','the','visual','appeal','
of','your','roomthis','is','crafted','by','skilled','craftsmen',
'and','it','is','designed','for','providing','elegant','and','tr
aditional','look','for','home','decor']

Figure 2: Texto Tokenizado

Como podemos ver cada se separo el texto en palabras, ahora es momento de eliminar todas las stopwords del texto:

- ['sky','trends','plastic','basket','artificial','flower','plant'
  ,'violet','multicolour','xxcm','fantastic','gift','say','care','
  birthday','present','anniversary','gift','special','gift','creat
  e','talking','pointthis','eye','catching','decorative','miniatur
  e','rickshaw','fabricated','vase','high','quality','materials','
  replica','original','rickshawthese','beautiful','vase','house','
  sky','trends','instantly','enhance','visual','appeal','roomthis'
  ,'crafted','skilled','craftsmen','designed','providing','elegant
  ','traditional','look','home','decor']

Figure 3: Texto sin stop words

Palabras como and, for, it, is, han sido eliminadas del texto. Y finalmente el último paso del preprocesamiento o limpieza de los textos es la Lematización o el stemming, con el que obtendremos la raíz de cada una de las palabras:

- ['sky','trend','plastic','basket','artifici','flower','plant','v
  iolet','multicolour','xxcm','fantast','gift','say','care','birth
  day','present','anniversari','gift','special','gift','creat','ta
  lk','pointthi','eye','catch','decor','miniatur','rickshaw','fabr
  ic','vase','high','qualiti','materi','replica','origin','ricksha
  wthes','beauti','vase','hous','sky','trend','instantli','enhanc'
  ,'visual','appeal','roomthi','craft','skill','craftsmen','design
  ','provid','eleg','tradit','look','home','decor']

Figure 4: Texto con stemming

Y este es el resultado final del preprocesamiento del texto, como podemos ver el numero de palabras se redujeron bastante, es muy diferente el texto con el que iniciamos, se ve mucho más sencillo el texto final y se pueden ver mucho más fácil las palabras que definen la idea central del texto. A continuación, analizaremos los resultados obtenidos en el preprocesamiento.

4 Resultados

Después de realizar el preprocesamiento es necesario analizar los resultados obtenidos para darnos una idea de lo que podemos hacer con los textos y de que información nos están brindando. Primero analizaremos todos los textos en conjunto, en la siguiente imagen tenemos la nube de palabras formada con todos los textos de la base de datos.

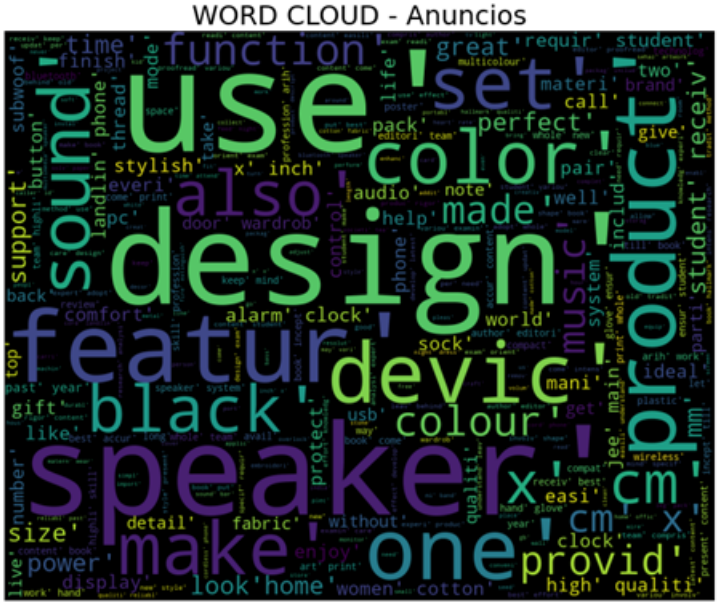

Figure 5: Nube de Palbaras

Podemos ver en la nube de palabras las palabras que más se repitieron en los textos, variando en tamaño según las veces que se hayan repetido, podemos ver que las palabras speaker, design, use, product, son algunas de las que se ven con mayor tamaño, palabras como producto o design tienen sentido que aparezcan entre las más repetidas teniendo en cuenta que los textos son anuncios y estas son palabras que se pueden ver en anuncios. Viendo las palabras que aparecen podemos decir que estas parecieran no pertenecer a una única clase, por lo que podríamos pensar que ninguna clase es mas dominante que las otras. Ahora en la siguiente grafica tenemos las frecuencias de las palabras más repetidas.

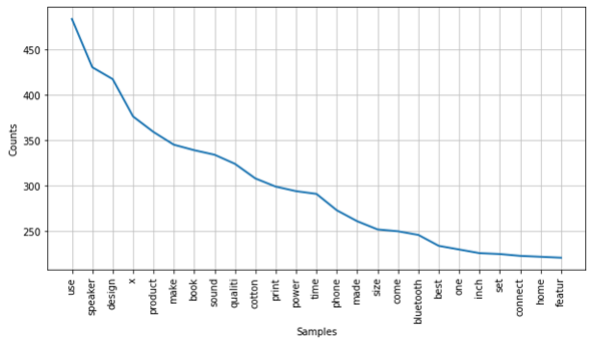

Figure 6: Grafica de Frecuencias

De la Figura 6 podemos ver que la palabra con mayor frecuencia en los textos fue la use con mas de 450 repeticiones, esto coincide con la nube de palabras ya que use es una de las más grandes, algo que se ve extraño es la aparición de la letra x que tiene una frecuencia alta, esto es debido a que la mayoría de los textos pertenecen a la clase Household y muchos de estos anuncios incluían medidas en sus textos del tipo 43x42, por lo que al borrar los dígitos nos quedamos solo con la x, por lo que podemos decir que la x es una palabra representativa de esta clase, también podemos ver palabras como sound, cotton y book que tienen una frecuencia alta y estas parecieran ser representativas de cada una de las clases restantes, por lo que podemos ver efectivamente la diferencia en la cantidad de textos de cada clase pareciera no afectar los resultados dado que son muy diversas las palabras con mayor frecuencia y no parecieran inclinarse hacia una sola clase.
Bien ahora analizaremos los resultados obtenidos por clase. Primero tenemos la nube de palabras obtenida para la clase Books.

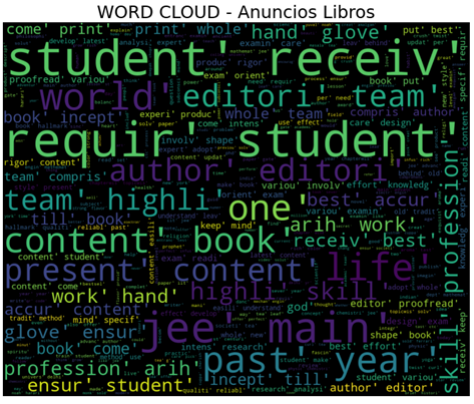

Figure 7: Nube de palabras de la clase Books

En el caso de la clase Books podemos ver que no hay mucha diferencia entre el tamaño de las palabras lo que nos diría que la frecuencia de las palabras no cambia mucho entre sí, podemos notar palabras esperaríamos ver en un anuncio de libros como lo pueden ser author, book, editor, student, word, son palabras que considero describen muy bien la clase que representan. Ahora en la siguiente grafica veremos la frecuencia de las palabras que aparecen en los textos de esta clase. De la Figura 8 podemos ver que la palabra que se despega demasiado

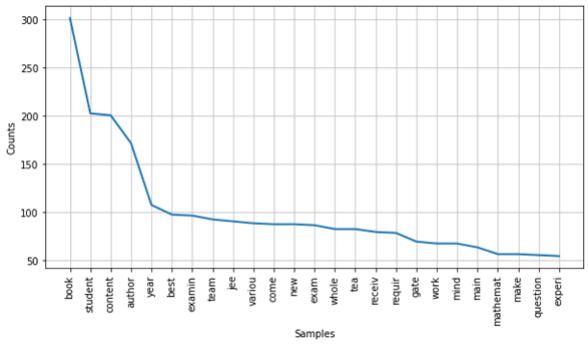

Figure 8: Grafica de frecuencia de palabras de la clase Books

de las demás es book con 300 repeticiones, algo curioso es que en la nube de palabras esta palabra book no es más grande que las demás, el resto de las palabras exceptuando las primeras cinco parecen tener una frecuencia similar lo cual hace sentido con la nube de palabras la cual no tiene mucha diferencia entre el tamaño de sus palabras.
Ahora analizaremos los textos de la clase Household. En la siguiente imagen tenemos la nube de palabras representativas de esta clase:

Figure 9: Nube de palabras de la clase Household

Como ya se explico antes en este tipo de anuncios era muy común ver medidas, por lo que palabras como cm y x es normal verlas de un gran tamaño y considero que estas son las palabras que mejor definen a esta clase, en este caso si están muy marcadas las diferencias de tamaño entre las palabras, podemos ver la grafica de frecuencias para evidenciar aún más esto.

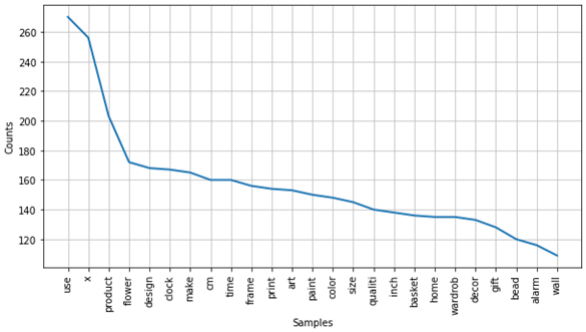

Figure 10: Grafica de frecuencia de palabras de la clase Household

Podemos ver en la Figura 10 como esta pareciera ilustrar una mayor diferencia entre las frecuencias de las palabras, la forma en que desciende la grafica ilustra muy bien esta diferencia. Podemos ver las palabras que tienen mayor frecuencia y podemos decir que son palabras realmente características de este tipo de textos.
Ahora analizaremos las palabras correspondientes a la clase Electronics. A continuación, tenemos la nube de palabras para esta clase.

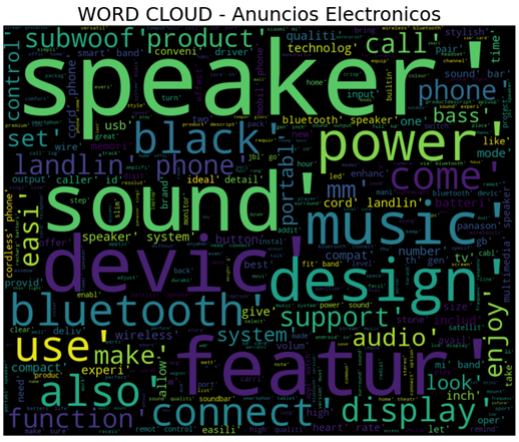

Figure 11: Nube de palabras de la clase Electronics

La palabra más grande y que más resalta de la nube de palabras es speaker, la cual representa una aparto electrónico, podríamos llegar a pensar que la mayoría

de estos textos eran anuncios de bocinas(speakers) dado que también aparecen palabras como music, sound, bluetooth, también podemos ver otras palabras que representa aparatos electrónicos como los es phone, podemos ver palabras muy características de aparatos electrónicos en esta nube de palabras. Ahora analicemos la grafica de frecuencias.

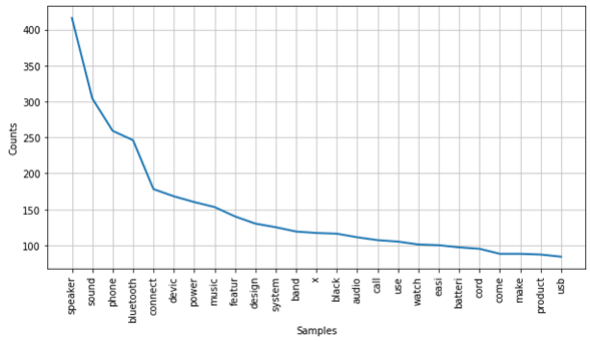

Figure 12: Grafica de frecuencia de palabras de la clase Electronics

Teniendo en cuanta la forma de la nube de palabras, las palabras que se resaltaban por su tamaño son efectivamente las que tienen mayor frecuencia, siendo speaker la de mayor frecuencia, si vemos las primeras ocho palabras con mayor frecuencia podemos ver que son palabras muy características de aparatos electrónicos, y podríamos incluso afirmar que los anuncios promocionaban teléfonos celulares y bocinas.
Por ultimo analizaremos los resultados obtenidos para la clase Clothing Accesories, primero tenemos la nube de palabras de esta clase.

Figure 13: Nube de palabras de la clase Clothing Accesories

Podemos ver que una de las palabras más grandes es sock que precisamente es una prenda, seguido de palabras como comfort y cotton que tambien son

palabras muy usadas para referirse a ropa, por otro lado, podemos ver muy marcadas las palabras woman y girl lo que nos podría indicar que estos anuncios estaban enfocados en vender ropa y accesorios para mujeres en específico. Ahora veamos la grafica de frecuencias de las palabras.

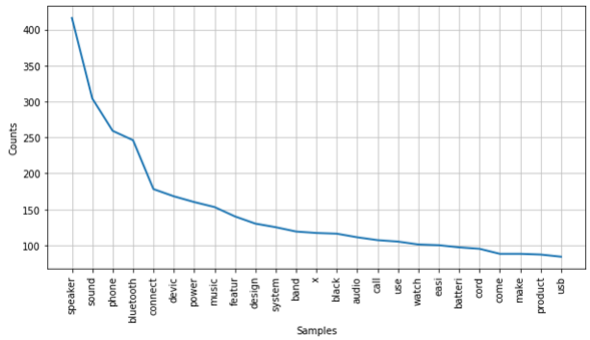

Figure 14: Grafica de frecuencia de palabras de la clase Clothing Accesories

Curiosamente sock es la segunda palabra con mayor frecuencia según la grafica pero esta era la palabra más grande de la nube de palabras, cotton es la palabra de mayor frecuencia, pode,os ver muchas palabras que hacen referencia a prendas como lo son sock, dress, short, al igual que en los casos anteriores parecieran ser palabras que definen muy bien la clase a la que pertenecen.
Con esto concluimos todo el análisis de resultados que obtuvimos, en la siguiente sección se va a discutir y concluir el resultado obtenido en este trabajo.

5 Conclusiones

Como vimos en los resultados anteriores cada una de las clases parece tener muy bien representadas sus palabras características, con esta base de datos se podría realizar un algoritmo de clasificación de textos, al estar tan bien definidas las clases considero que esta base puede ser la adecuada para entrenar un algoritmo para esta tarea. Quizá el único problema que se tendría seria el tamaño de la base que se creó podría llegar a ser algo pequeña para entrenar un algoritmo, pero esto no seria problema ya que se pueden extraer muchos más textos de la base original. Probablemente utilizaría el algoritmo de K-means para realizar esta tarea de clasificación. Definitivamente el preprocesamiento que se le dio a los textos fue el adecuado, se redujeron de una gran manera el tamaño de los textos, se eliminaron todos aquellos elementos que no eran de interés y se logró obtener al final un conjunto de palabras que realmente representaban a cada una de sus clases. Se agregaron un par de pasos más de preprocesamiento a los vistos en clase, y estos fueron la parte de eliminar dígitos y espacios grandes. El hecho de extraer una parte de los textos de la base original para crear una más pequeña puede ser considerado también parte del preprocesamiento teniendo en cuenta que la limpieza parte de la base original. En conclusión, la base que se creo y se le aplico el preprocesamiento es ideal para el desarrollo de algoritmos de clasificación de texto, esto a su vez podría llevar a la creación de sistemas que puedan hacer llegar al público adecuado cada uno de los anuncios según su clasificación.

References

[1] *Link de la base de datos original*
https://www.kaggle.com/datasets/saurabhshahane/ecommerce-text-classification

[2] *Link de Github*
https://github.com/CesarJairTJ/FCFM