UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

UANL

UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

FACULTAD DE CIENCIAS FISÍCO MATEMÁTICAS

Materia

Datos Masivos

Profesora

M.C. Mayra Cristina Berrones Reyes

Tarea de la clase 29-06-2023

Muestreo

Alumna

I.M. María Luisa Argáez Salcido

Matrícula

2173261

Fecha

02-07-2023

Contenido

Clasificación de texto por medio del algoritmo de *K-means*

Instrucción

Se seleccionó la opción 1 como la tarea que se entregará. La opción 1 menciona:

Opción 1: Si sus datos tienen etiqueta, sin utilizar esas etiquetas realizar un algoritmo de agrupamiento para revisar cuales son las agrupaciones que logra su algoritmo, si son las mismas que uds tenían contempladas o son distintas, resuman en sus conclusiones.

Introducción

Con el conjunto de datos que se utilizó la práctica pasada, se realizará un *K-means* como una opción de los algoritmos de agrupamiento. Se tiene la ventaja de que los datos se encuentran etiquetados como emoción positiva o negativa de tal manera que el algoritmo de *K-means* tendrá dos centroides, uno para cada tipo de emoción. Se espera haya alguna distinción significativa entre ambas clases.

Análisis exploratorio de datos

Se utilizó la base de datos Sentiment140 (Kaggle, 2023) que contiene diferentes mensajes de twitter, llamados comúnmente "*tweets".* Dicha base de datos con seis columnas y 1, 600,000 filas, las cuales representan lo siguiente:

- Tipo de emoción, 0 si es negativa o 4 si es positiva, o, en otras palabras, si el texto del *tweet* expresa una emoción negativa o positiva.
- Id del mensaje o tweet, funciona como identificador del mensaje.
- Fecha. Es la fecha en la que se realizó el mensaje.
- Consulta. Si es mediante una consulta o no.
- Usuario, indica el nombre del usuario.
- Texto, es el mensaje que expreso el usuario y se clasificó como positivo o negativo.

Para cada tipo de emoción, negativa o positiva, son 800,000 muestras, lo cual indica una base de datos balanceada.

Preprocesamiento de datos

Se tomo una muestra aleatoria para cada tipo de emoción de 1600 muestras, posteriormente se realizó el preprocesamiento llamado "*stemming*" y "*CountVectorizer*" con el propósito de remover signos de puntuación y eliminar palabras insignificantes para la práctica, así como una transformación de las palabras. El proceso de *Stemming* consiste en reducir la palabra a su forma

raíz de tal forma que ayuda a reducir el grado de escasez de los datos. Y *CountVectorizer* obtiene la frecuencia de una palabra en toda la observación y también ayudó a homogeneizar el texto solo con letras minúsculas.

Modelado

El modelado consistió en un algorimo de *K-means,* con dos centroides, ya que solo hay dos tipos de emociones en este caso, positiva y negativa, se utilizaron 1000 iteraciones ya que en un principio intenté con menos iteraciones y el rendimiento del modelo era menor.

Resultados

Los resultados del algoritmo de *k-means*, no fueron satisfactorios ya que la matriz de confusión que se muestra en la tabla 1, muestra que la mayoría de las observaciones se etiqueto como la clase 1, en lugar de que la mitad de las observaciones positivas fueran agrupadas en un cluster y el restante, es decir las emociones negativas, fueran agrupadas en otro cluster.

Se observa que la exactitud es de 50.49%, el rendimiento del modelo no es el esperado.

Tabla 1: Matriz de confusión de los resultados del algoritmo de K-means y la etiqueta de los mismos.

	0– Emoción negativa	0– Emoción positiva
0 – Emoción negativa	1582	18
1– Emoción positiva	1554	46

Conclusiones

Se concluye que la mayoría de las observaciones se agruparon como emoción negativa y que solo 2.87% de las muestras que corresponden a una emoción positiva fueron agrupadas correctamente. Además, que el porcentaje de los falsos es de 49.12%, que es casi igual a la exactitud, lo cual indica que el rendimiento del modelo es bajo y se concluye que se debe de buscar una alternativa mejor para la segmentación de los mensajes.