PONTIFICIA UNIVERSIDAD CATÓLICA DE VALPARAÍSO
MAGÍSTER EN INGENIERÍA INFORMÁTICA
ESCUELA DE INGENIERÍA INFORMÁTICA

Proyecto 1

Análisis y Clasificación de Tweets Covid-2019

Integrantes

Pablo Oñate
José Ignacio Villar
Catalina Montecinos

Ramo

MII Clasificación Automática de Textos

Dominio y Problema

La enfermedad del coronavirus (COVID-19) es una enfermedad infecciosa causada por el virus SARS-CoV-2.

La mayoría de las personas infectadas por el virus experimentan una enfermedad respiratoria de leve a moderada y se recuperan sin requerir tratamiento especial. Sin embargo, algunos enferman gravemente requiriendo atención médica. Cualquier persona, de cualquier edad puede contraer COVID-19 y enfermar gravemente o morir. Incluso pueden quedar con secuelas varios meses después de haberse sanado o incluso re-infectarse.

El virus puede propagarse desde la boca o nariz de una persona infectada en pequeñas partículas líquidas cuando tose, estornuda, habla, canta o respira. Estas partículas van desde gotículas respiratorias más grandes hasta los aerosoles más pequeños.

Dadas las características de este virus y su alta tasa de propagación y contagio, a fines del 2019 se consideró potencialmente peligroso, convirtiéndose prontamente en un problema mundial generalizado. En Marzo del 2020 se decretaron cuarentenas generalizadas en la gran mayoría de los países, convirtiéndose así en uno de los trending topics más grandes de Twitter (#COVID19 y #StayHome).

Para fines del trabajo propuesto nos pareció interesante y relevante el poder analizar los sets de datos de dicha temática, pues considerábamos que podríamos poder obtener un análisis de sentimientos y nube de palabras bastante satisfactorio, entre otros de los elementos analizados.

Set de Datos

El conjunto de datos consta de dos archivos csv, de los cuales uno se utiliza para el entrenamiento del modelo y el otro para la prueba. Ambos archivos constan de 6 columnas con información, las que se detallan a continuación.

- ***Username:*** Identificador único, el cual solo el usuario puede observar.
- ***ScreenName:*** Identificador único, el cual puede ser observado por todos los demás usuarios de la plataforma.
- ***Location:*** Ubicación en donde se origina el tweet.
- ***TweetAt:*** Es la fecha en la que se originó el tweet.
- ***OriginalTweet:*** Es el contenido o texto del tweet, cuyo idioma está en inglés.
- ***Sentiment:*** Es la clasificación que se le da al tweet dependiendo de su contenido, el cual va desde muy positivo, hasta muy negativo.

El conjunto de datos consta de 44.955 tweets, en donde el archivo de entrenamiento contiene 41.157 de estos, y el de prueba 3.798, los cuales representan el 91,6% y 8,4% del total respectivamente.
Finalmente, con el objetivo de agregar más información para el futuro análisis, se agrega una séptima columna “Length” para ver la cantidad o largo en caracteres del texto escrito en cada tweet.

Notebook Collab

Estos dos archivos, modelo y prueba, los cargamos en el Notebook de Google Collab donde cada uno de nosotros tuvo la oportunidad de explorar el set, revisar las columnas y practicar con los métodos solicitados.

[El archivo se encuentra adjunto]

Análisis

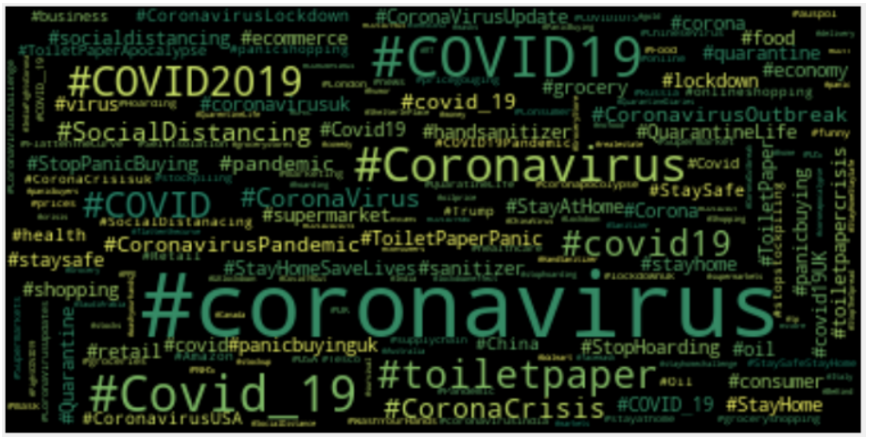
Para realizar un preprocesamiento óptimo, es necesario hacer un pre análisis que permita obtener información exhaustiva de los datos expuestos en los archivos csv.

Con el objetivo de ver las columnas en donde haya pérdida de data, se aplica una visualización de la matriz de valores perdidos, en donde se puede observar que en la columna de “Location” hay 8.590 filas que no contienen información en este apartado.

Luego, se aplica un análisis de las columnas importantes para la extracción de información relevante para las posteriores conclusiones, las cuales se detallan a continuación.

- ***TweetAt:*** En esta columna se realizaron gráficos para identificar en qué mes y fecha es en donde se producen la mayoría de los tweets, lo que da como resultado que es en el mes de abril en donde esto ocurre. Además se analizan los días de la semana que más tweets concurren, en donde no hay resultados que digan o resalten alguno de ellos.
- ***Location:*** Con el fin de ver las localidades que más tweets generan al respecto, se realiza un gráfico de barra, el cual da como resultado que Estados Unidos junto al Reino Unido son los lugares geográficos que más tweets generaron.
- ***Sentiment:*** En esta columna se hicieron dos análisis; El primero consta de ver las cantidades en número y porcentaje de cada tipo a clasificar, los cuales no presentan un gran desbalance ; El segundo, es para ver la distribución del largo de cada comentario según su clasificación, en donde sobresalen los comentarios extremadamente positivos y extremadamente negativos.

Finalmente, con el fin de observar los hashtags y palabras que más se repiten en los tweets, es que se generan nubes de palabras para cada una de estas, como se observan en las figuras adjuntas.



Preprocesamiento

Para el preprocesamiento de los datos se tuvo principal énfasis en los textos, ya que se aplicaron 6 criterios, los cuales se explican a continuación.

1. ***Remover hashtags:*** Debido a que se puede repetir el mismo hashtag en la mayoría de los datos, lo que puede provocar que el algoritmo no aprenda adecuadamente de patrones en los textos. Es por esta razón que se decide eliminar todos los hashtags para nuestra futura bolsa de palabras.
2. ***Remover menciones:*** Es fácil intuir que las menciones no serán un factor clave para la clasificación y además pueden tener consecuencias negativas en el entrenamiento, ya que estas pueden variar mucho o focalizarse en una figura pública.
3. ***Remover URLs:*** Estas se pueden volver repetitivas y simplemente no aportan en detección de patrones en el entrenamiento.
4. ***Considerar solo palabras:*** Este es un preprocesamiento típico, ya que los números, signos de puntuación y caracteres especiales no aportan a la obtención óptima de nuestra bolsa de palabras, ni al análisis y tampoco al entrenamiento del modelo.
5. ***Convertir texto a minúsculas:*** Con el fin de que el algoritmo no haga una distinción entre la misma palabra escrita en mayúscula a otra en minúscula, es que se transforma todo el texto.
6. ***Remover stop words:*** Dada su abundancia en el lenguaje humano, se deben remover estas palabras al ser de poco valor semántico, para que así el algoritmo pueda enfocarse en la información más importante.

Además de esto, se eliminaron las columnas *UserName* y *ScreenName* puesto que son para una identidad y que no afectará nuestro modelo.