Text Mining en Social Media

Master en Big Data Analytics
Universidad Politécnica de Valencia

12 de junio de 2021

SERCOPA
Sergio Campos Pérez
Adrián Cortijo Simarro
Pablo Pons Roger

En primer lugar para realizar el trabajo cargamos todas las librerías que van a ser necesarias para la realización del mismo, instalando previamente los paquetes que no venían en nuestra versión de R.

La carga de las bibliotecas la hacemos desde la línea 1 hasta la línea 9.

Después, se indican los parámetros del número de palabras del vocabulario, el número de pliegues de la validación cruzada y el número de veces que la repite

Los parámetros se indican de las líneas 14 a la 16.

A continuación, se asigna el nombre de los archivos del training y del test y el lenguaje en el que realizamos el intento.

Esta asignación se realiza de las líneas 17 a la 19.

A continuación, implementamos una función para eliminar las palabras que contienen menos de 2 caracteres.

Esta función está definida de la línea 52 a línea 54

Y implementamos otra que servirá para eliminar todas las URLS.

La carga Esta función está definida de las líneas 56 a la 58

Le pasamos los *parámetros long2 y url* a la función Generating Bag of Words (BoW).

Los parámetros están indicados en la línea 94

Modificamos el código para cuando lanzamos el módelo contra el test, no vaya al truth porque la carpeta de test no existe.

La modificación del código se realiza de las líneas 98 a la 102

Metemos las variables creadas anteriormente en el preprocesado del test.

Realizado de las líneas 133 a la 139

Nuestro método ha sido separar el dataset de training en fakers y no fakers para ver cuales son las 5 palabras que usan con más frecuencia los haters.

Este proceso está escrito de la línea 193 hasta la línea 201

Después, incluimos una característica más al dataset, que es cuantas veces ha utilizado

cada autor las palabras hater.

La carga Esta característica está indicada en la línea 204 hasta la línea 207

- **PCA (Principal Component Analysis)**

Se ha decidido hacer un PCA (análisis de componentes principales) para intentar mejorar el modelo.

El PCA está realizado en la línea 211

Finalmente se descarta su utilización ya que hemos visto que no ayuda a reducir la dimensionalidad y por lo tanto no aporta valor en el proyecto.

- **SVM (Support Vector Machine)**

Hemos probado un SVM (Support Vector Machine) con cross-validation y también un Gaussian-Mixture. Como vemos que el Gaussian-Mixture da peores resultados decidimos entrenar el SVM de nuevo con todo el training.

Este modelo está programado de la línea 214 hasta la línea 225

A continuación, generamos el Bag of Words para los datos del test.

Realizado en la línea 228

Añadimos la misma característica que habíamos añadido al training al test, añadiendo la columna de cuantas veces han utilizado las palabras más frecuentes de los haters.

La característica está añadida en las líneas 237 hasta la línea 239

Utilizamos el modelo ya entrenado con el training para hacer una predicción sobre los datos de test y generamos el xml con la predicción para la entrega y evaluación.

Esta última parte la programamos desde la línea 243 hasta la línea 274

Para finalizar el trabajo, y a modo de resumen tras un previo análisis de los resultados podemos comentar que los resultados en inglés son peores porque sus mensajes de odio son más explícitos y en cambio los españoles son mucho más implícitos.