Universidad del Valle de Guatemala
Data Science
Lynette Garcia Perez

José Martínez
Sergio Marchena
Pablo Viana

Laboratorio 5 - Minería de Textos

Análisis Exploratorio

Descripción de los Datos

Los datos fueron extraídos de la página de competencias Kaggle. Fueron compartidos por SWIFTKEY en asociación con el Instituto de John Hopkins de Especialización en Data Science. Los datasets consisten en 3 textos con millones de tweets, post en blogs y noticias en inglés.

El primer texto es un corpus monolingüe en inglés de millones Tweets en inglés. El archivo pesa 154.9MB. Este corpus tiene la característica de tener menos de 140 caracteres por tweet y que hay emoticones.

El segundo texto es un corpus monolingüe en inglés de posts de blogs en inglés. El archivo pesa 200.4MB.

El tercer texto es un corpus monolingüe en inglés de noticias de diferentes fuentes. El archivo pesa 196.3MB.

Preprocesamiento

Las acciones que se realizaron para poder tener datasets más consistentes y mejores para la predicción fueron las siguientes. Cada uno de estos cambios se le realizó a cada corpus por separado.

Se convertir el texto a mayúsculas o a minúsculas para poder evitar problemas a la hora de contar palabras y que no distinga dos palabras iguales solo que la primera letra es mayúscula. Se quitaron los caracteres especiales en cada corpus, como por ejemplo: “#,@, .,%,$,&,*,€, ™,””, etc. Además, se quitaron las url. Para el corpus de Tweets, se quitaron los emoticones. Se quitaron los signos de puntuación. Por último, se quitaron stopwords utilizando la función del paquete de Quanteda, “stopwords()”. Al final se quedó con las palabras más importantes de cada corpus.

n-gramas

Modelo Preliminar