Pasos

Se hace un análisis de frecuencia de palabras y se visualiza

Primer se obtienen y se limpian los datos, después se hace **un análisis exploratorio** de datos (EDA) y la calidad de éstos(que es analizarlo sistemáticamente on an aggregated level). Métodos típicos son resúmenes estadísticos (df.describe) para valores numéricos o recuento de frecuencias para variables categóricas. Histogramas y box plots ilustran la distribución de los valores y las series de tiempo muestran la evolución.

E toma el texto sin procesar, , se limpia, se transforma, y se extraen las características básicas

Un dataset que contiene documentos de texto se llama corpus en NLP.

El hecho de que una palabra aparezca 100 veces no tiene ninguna relevancia pero que la frecuencia relativa de esa palabra se haya duplicado en una semana es relevante

En el dataset se tiene el índice y el primary key. No usar todo el dataset cuando es grande sino una parte (usar df=df.sample(frac=0.1) 10%)

Metadata son los datos de los datos, dan información de los datos. Creo que sería df.info()

Para tener una visión general:

-calcular estadísticas (df.describe)

-checkear valores faltantes (df.isna().sum ) 🡪 df[‘columna’] .fillna(‘desconocido’, inplace=True)

-plotear distribuciones de atributos interesantes

-visualizar evolución en el tiempo. 🡪 si los datos contienen fechas

Los ploteos muestran la dispersión de los valores ( ejemplo las longitudes de los tweets)

A las funciones como map y apply que toman otras funciones como parámetros se las llaman funciones de orden superior en matemática y ciencias de la computación

Palabras y frases frecuentemente utilizadas nos pueden dar un entendimiento básico de los tópicos discutidos. Sin embargo el análisis de las frecuencias de las palabras ignora el orden y el contexto de las palabras. Esta es la idea de l modelo de la bolsa de palabras. Todas las palabras se arroja en una bolsa donde se mezclan. El ordenamiento original del texto se pierde y solo se tiene en cuenta la frecuencia de los términos. El modelo no funciona bien para análisis de sentimientos pero funciona sorprendentemente bien para clasificación y modelado de tópicos. Además es un buen punto de inicio para entender de que trata el texto.

La lista de stop Words en inglés tiene 179 palabras

Hash table = diccionario

Hepaxes = tokens que solo ocurren una vez

Smoothing = técnica de agregar un +1 al denominador para prevenir una división por cero

En procesamiento de texto, usualmente se trabaja con bigramas (oraciones “secuencias”) de longitud 2) y a veces hasta trigramas (longitud 3)

n-gramas de longitud 1 se llaman unigramas. La razón para utilizar n<=3 es que le número de diferentes n-gramas aumenta exponencialmente con respecto a n, mientras que sus frecuencias disminuyen de la misma manera. Por mucho trigramas aparecen solo una vez en el corpus.

Importante: siempre se debe identificar la causa raíz de cualquier token extraño que aparezca. KWIC análisis es una buena herramienta para esto.

La información sobre los tipos de datos se pierde cuando se graba en CSV o solo se graba rudimentariamente con JSON.