Universidad Francisco Marroquín

Data Wrangling

Catedrático: Juan Carlos Girón

Auxiliar: José Josué

Laboratorio #9

Utilizando la librería de twitteR de R,el API de developer de twitter, y las librerías de análisis de sentimiento de Python:

1. Generar una consulta de algún #hashtag, tema o usuario que sea de su interés. (15 pts)
    a. Transformar la respuesta del API de twitter a un dataframe e importarlo a un csv.
    b. Presentar el código de consulta al API de twitter en un Rmarkdown.
        i. **Nota: no desplegar los Tokens de su cuenta**
2. Importar el dataframe generado en R a Python.
3. Utilizando expresiones regulares: (15 pts)
    a. Crear una función normalizadora de texto que limpie los caracteres que pueden generar ruido en el string del tweet y aplicarla a los mismos.
    b. Generar una nueva columna llamada "Handler" que muestre el handle del usuario. ej: @Tepi
    c. Generar una nueva columna llamada "Source" que muestre de donde se generó el tweet. ej: Iphone, Android, etc.
4. Generar un corpus del dataset original que contenga (10 pts):
    a. Handler
    b. screenName
    c. retweetCount
    d. Source.

5. Generar una función que lematize cada tweet y guardar el resultado en una nueva columna llamada "lem_text" (10 pts).
6. Generar una función calcule la polaridad y subjetividad de las columnas "text" y "lem_text" (20 pts).
7. Utilizando pandas, calcular la media de la polaridad y subjetividad de todos las columnas del inciso 6 del corpus. Comparar si la lematización afecta en el resultado (10 pts).
8. ¿Qué puede decir del sentimiento del corpus con base a la agregación del inciso 7? (10 pts)
9. Responder (10 pts):
    a. ¿De dónde se originan la mayoría de tweets de su corpus?
    b. ¿Cuál es el tweet más popular de su corpus?

10. Extra: Generar un wordcloud utilizando Python en el cual se despliegue las palabras más frecuentes de la columna "lem_text" (10 pts)