UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO
CENTRO UNIVERSITARIO UAEM ZUMPANGO
INGENIERÍA EN COMPUTACIÓN

Ciencia de los datos

Proyecto: Minería de textos

Profesores: Dr. Asdrúbal López Chau
Dr. Rafael Rojas Hernández

Fecha: octubre-noviembre 2024

 Este símbolo significa que deberás de incluir lo que se solicita en el reporte final de la práctica.

 Este símbolo significa que deberás de escribir código fuente comentado y que realice exactamente lo que se solicita.

 Este símbolo significa que deberás de realizar una actividad por tu cuenta, pero puedes solicitar ayuda al profesor

 Este símbolo significa un consejo, guía o ayuda que se proporciona

Nota: El plagio se penaliza con la anulación de la calificación a todos los involucrados

Objetivos

Objetivo general: Aplicar webscrapping, visualización y narrativa de datos, vectorización de documentos y otras técnicas de minería de textos en un problema con datos reales.

Objetivos específicos:

1. Recuperar y procesar datos utilizando Python.
2. Procesar los textos y normalizarlos
3. Escribir una narrativa de datos interesante con el apoyo de gráficos adecuados.
4. Aplicar técnicas de minería de textos para extraer conocimiento de grandes cantidades de datos.

Introducción

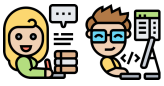
Requerimientos Previos

- Conocimiento las bibliotecas pandas, matplotlib, seaborn, wordcloud, nltk, gensim, etc.

Instrucciones:

Elige uno de los siguientes temas (cada equipo, formado por tres personas como máximo elegirá un tema o propondrá uno, que será aprobado previamente por docente):

- Bitcoin, Ethereum, Binance coin y otras (criptomonedas)
- Donald Trump, Kamala Harris (votaciones en USA)
- Chatgpt, Gemini Copilot y otros (Inteligencia artificial generativa, deep learning y machine learning)
- Cambio climático y sostenibilidad
- Redes sociales en la salud mental y física
- Claudia Sheinbaum, AMLO (4T)
- Metaverso y realidad extendida

 Recolección de datos

- Busca al menos **700 documentos** de distintas fuentes (periódicos, revistas, blogs, redes sociales, etc.) que traten sobre el tema elegido. Aplica webscrapping para descargar los documentos.
- Características de los documentos
    - Total de documentos 700
    - Los documentos deben ser en Español
    - La longitud mínima de cada documento debe de ser de al menos 400 palabras.
    - Documentos de al menos 200 palabras y menores a 400 palabras, se cuentan la mitad, es decir, dos documentos de entre estas longitudes se consideran como un documento. Por lo que la cantidad de archivos a guardar será mayor para estos casos.
    - En el caso de redes sociales, cada 40 comentarios se considerarán como un documento.

Cada documento será almacenado como un documento de texto plano, con extensión .txt en una carpeta o directorio.

-  En el reporte, se agregará una tabla que contenga el nombre del archivo, su longitud, fecha y sitio de donde se descargó.

 EDA

- Aplica las técnicas de procesamiento y normalización a los documentos, elimina lo que no sea relevante. Asegúrate de que no se eliminen tokens importantes.
- Identifica los autores más frecuentes que escriben sobre el tema elegido.
- Identifica las fuentes más frecuentes que publican sobre el tema elegido.
- Aplica análisis de frecuencia de palabras y muestra los resultados en nubes de palabras
    - Por fuente
    - Por autor
-  En el reporte, realiza una descripción clara de los resultados de esta exploración.

 Análisis de sentimientos

- Usa bibliotecas necesarias para identificar la polaridad global de los documentos, pudiendo ser positiva, negativa o neutral.
- Aplica análisis de sentimientos de manera separada
    - Por fuente
    - Por autor

 Modelado de tópicos

- Usa bibliotecas necesarias para los temas principales que se encuentran ocultos o latententes en los documentos. Para esta parte, deberás de aplicar por lo menos dos técnicas diferentes, de *topic modeling*.
- Aplica análisis de topic modeling
    - Por fuente
    - Por autor

 Discutir sobre cuáles fueron los hallazagos, qué temas se tratan, polaridad de los documentos, palabras frecuentemente usadas.

 Publicación de resultados

- Busca y elige **tres revistas** científicas en las cuales podrían publicarse tus resultados, busca en la base de datos siguientes (como ejemplo solamente)
    - https://www.redalyc.org/
    - https://www.revistascytconacyt.mx/
    - https://accensum.org/revistas-indexadas-latinoamerica/revistas-indexadas-mexico/
- Redacta el artículo correspondiente usando el formato de la revista elegida. Observa la estructura otros artículos en la revista como guía.