Tarea 1 - Introducción a la Ciencia de Datos

2025

Para esta tarea, se utilizará una base de datos abierta con discursos de políticos en el marco de las elecciones de Estados Unidos en 2020.

En el repositorio intro-cd se encuentra la base de datos y el código necesario para cargar algunas tablas en un DataFrame de pandas (lenguaje Python). El código está en un Jupyter notebook (us2020_propuesta.ipynb), junto con las instrucciones para correrlo.

La entrega se debe dejar disponible en un repositorio público (por ejemplo, en GitHub o GitLab), y los archivos a evaluar deben estar en la branch principal (main). En dicha rama no debe haber *commits* posteriores a la fecha de entrega estipulada. Los archivos que deben estar presentes en el repositorio son:

- Un informe en formato PDF incluyendo todos los resultados relevantes, y este será en general el trabajo a evaluar.
- Todo el código que haya sido implementado (al menos un notebook y posibles scripts adicionales), pero estos sólo serán revisados en caso de que existan dudas referentes a la implementación.

Agregar un archivo README.md al repositorio, con indicaciones básicas, por ejemplo indicando cuál es el informe, cuál el notebook o script utilizado para responder las preguntas y en caso de haber más de uno, indicar para qué se usó cada uno.

Parte 1: Cargado y Limpieza de Datos

A. Compruebe que puede correr las primeras dos celdas del notebook, observe el contenido de los dataframes cargados. Reporte si existen datos faltantes en algún campo, o cualquier otro problema de calidad de datos que encuentre. En particular, analice la cantidad de discursos por candidato/a, y a partir de este punto trabaje con los cinco candidatos/as con mayor cantidad de discursos.

B. Genere una gráfica que permita visualizar los discursos de los candidatos/as a lo largo del tiempo, con alguna escala temporal adecuada. Comentar si se identifican momentos clave de la campaña. No realizar análisis estadísticos, solamente generar visualizaciones exploratorias.

C. Una de las funciones básicas que se desea realizar, es el conteo de palabras: cuántas veces aparece cada palabra agrupando por distintos criterios. Para ello, primero es

necesario normalizar el texto (i.e: pasarlo todo a minúsculas) y eliminar los signos de puntuación. De no hacerlo, las secuencias "You", "you." y “you,” se contarían como palabras distintas. La función clean_text(...) realiza parte de esta tarea, pero se debe completar agregando algunos signos de puntuación y cualquier otra normalización que considere oportuna. Comprobar el resultado observando el contenido de df_speeches_top_5, algunas celdas más abajo. Comente todas las transformaciones de texto que haya agregado y justifique.

Parte 2: Conteo de Palabras y Visualizaciones

A. Realice una visualización que permita comparar las palabras más frecuentes de cada uno de los cinco candidatos/as.
   Sin necesidad de implementarlo, proponga ideas para modificar esta visualización con el fin de encontrar diferencias entre partidos políticos, fechas, o lugares.
B. Corra el código que permite encontrar los candidatos/as con mayor cantidad de palabras. En caso de encontrar algún problema luego de realizar la visualización, comente a qué se debe y proponga formas de resolverlo.
C. Construya una matriz de 5x5, donde cada fila y columna corresponden a un candiato/a, y la entrada (i,j) contiene la cantidad de veces que el candiato/a “i” menciona al candiato/a “j”. Opcional: genere un grafo dirigido con esa matriz de adyacencia para visualizar las menciones.
D. Proponga al menos tres preguntas que se podrían intentar responder a partir de estos datos, y mencione posibles caminos para responderlas (sin implementar nada).