Tarea 1 - Introducción a la Ciencia de Datos

2023

Para esta tarea, se utilizará una base de datos relacional abierta con la obra completa de William Shakespeare, disponible en:

- https://relational.fit.cvut.cz/dataset/Shakespeare

En el link se puede ver la estructura de la base de datos, y además se incluye la información necesaria para conectarse a la misma utilizando un cliente como *MySQL Workbench*, en caso de que se desee.

En el **repositorio introCD** se encuentra el código necesario para efectuar la conexión y cargar algunas tablas en un *DataFrame* de pandas (lenguaje Python). El código está en un Jupyter notebook (**shakespeare_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.

Si considera que será de utilidad acceder a este trabajo en un futuro, se recomienda agregar un archivo *README.md* al repositorio, con indicaciones básicas.

Parte 1: Cargado y limpieza de datos

a) Compruebe que puede correr las primeras tres celdas del notebook, observe el contenido de los *dataframes* cargados y luego complete el código para cargar el resto de las tablas disponibles.
   Comente la función de cada tabla y la relación entre ellas.
   Reporte si existen datos faltantes en algún campo, o cualquier otro problema de calidad de datos que encuentre.

b) Genere una gráfica que permita visualizar la obra de Shakespeare a lo largo de los años. Por ejemplo, tomando períodos de algunos años y mostrando la cantidad de obras escritos para esos períodos. Comente si se observan tendencias (o no) a lo largo del tiempo, por ejemplo respecto a su producción, o los géneros sobre los que escribió. 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 "Thou" y "thou," (sic) 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_words`, 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, considerando toda la obra.
Sin necesidad de implementarlo, proponga ideas para modificar esta visualización con el fin de encontrar diferencias entre géneros o personajes.
b) Corra el código que permite encontrar los personajes 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) Proponga preguntas que se podrían intentar responder a partir de estos datos, y mencione posibles caminos para responderlas (sin implementar nada).