Introducción a la Ciencia de Datos

Curso 2023

Tarea 1

Grupo 8

Juan Manuel Varela - CI: 4.802.705-4

Nicolás Farías - CI: 4.143.102-6

Contenido

Introducción

Este documento es un informe sobre el trabajo realizado para la Tarea 1 del curso Introducción a la Ciencia de Datos. Se utilizó un conjunto de datos sobre la obra completa de William Shakespeare, extraídos de una base de datos relacional abierta disponible en:

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

El trabajo consistió en extraer los datos de las tablas, procesarlos, y finalmente explorarlos para intentar contestar algunas preguntas sobre la obra de Shakespeare.

Carga y Exploración Inicial

La base de datos está compuesta por las siguientes tablas:

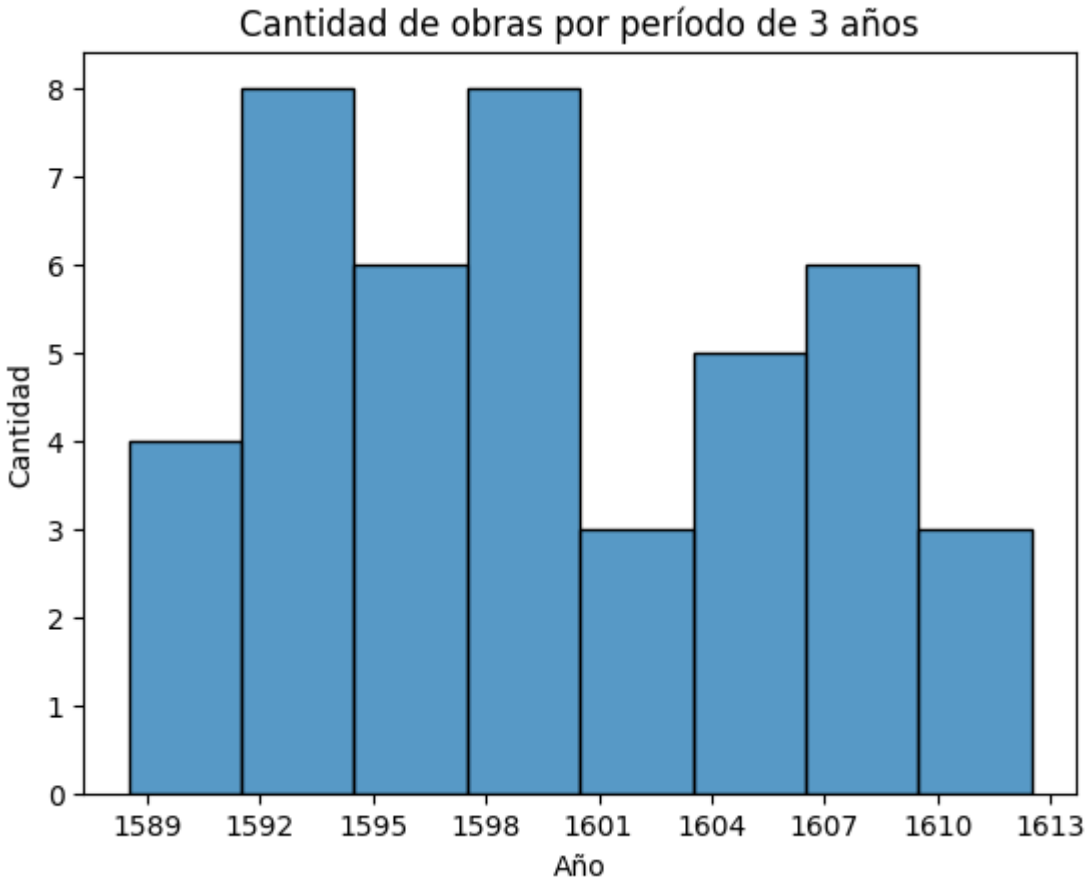
- ***works***: contiene todas las obras de Shakespeare con su año y género.
- ***chapters***: contiene los capítulos o escenas que componen las obras. Cada elemento se relaciona con una obra de la tabla ***works***.
- ***paragraphs***: contiene los párrafos (textos y diálogos) de las obras. Cada elemento se relaciona con un capítulo de la tabla ***chapters*** y con un personaje de la tabla ***characters***.
- ***characters***: contiene los personajes de las obras.

Se realizó una exploración visual y se utilizaron las funciones *info* y *nunique* de la biblioteca Pandas para hacer un diagnóstico inicial sobre la calidad de los datos. A continuación se detallan algunas consideraciones al respecto:

- Los tipos de datos son consistentes con la semántica de cada campo.
- Las columnas de nombre *id* efectivamente tienen valores únicos dentro de cada tabla.
- Muchos de los personajes (tabla *characters*) tienen el campo descripción vacío.
- Hay personajes con distinto *id* que tienen el mismo nombre. Esto puede deberse a que pertenecen a distintas obras.
- Muchos capítulos tienen distinto *id* pero la misma descripción, incluso dentro de una misma obra.
- Hay varios capítulos que no tienen descripción en las obras 28 (“Passionate Pilgrim”) y 35 (“Sonnets”).
- En los diálogos hay muchas contracciones con significados diferentes, algunas que ya no se utilizan en el inglés moderno.
- Hay 1220 personajes distintos en la tabla *paragraphs*, y 1266 personajes en total en la tabla *characters*, por lo tanto hay personajes que no tienen diálogos asignados.
- El personaje que tiene más párrafos asociados, con una distancia importante, tiene como nombre “(stage directions)”. Explorando algunos de estos párrafos se comprobó que efectivamente no son diálogos de un personaje real, sino que son indicaciones para la obra.

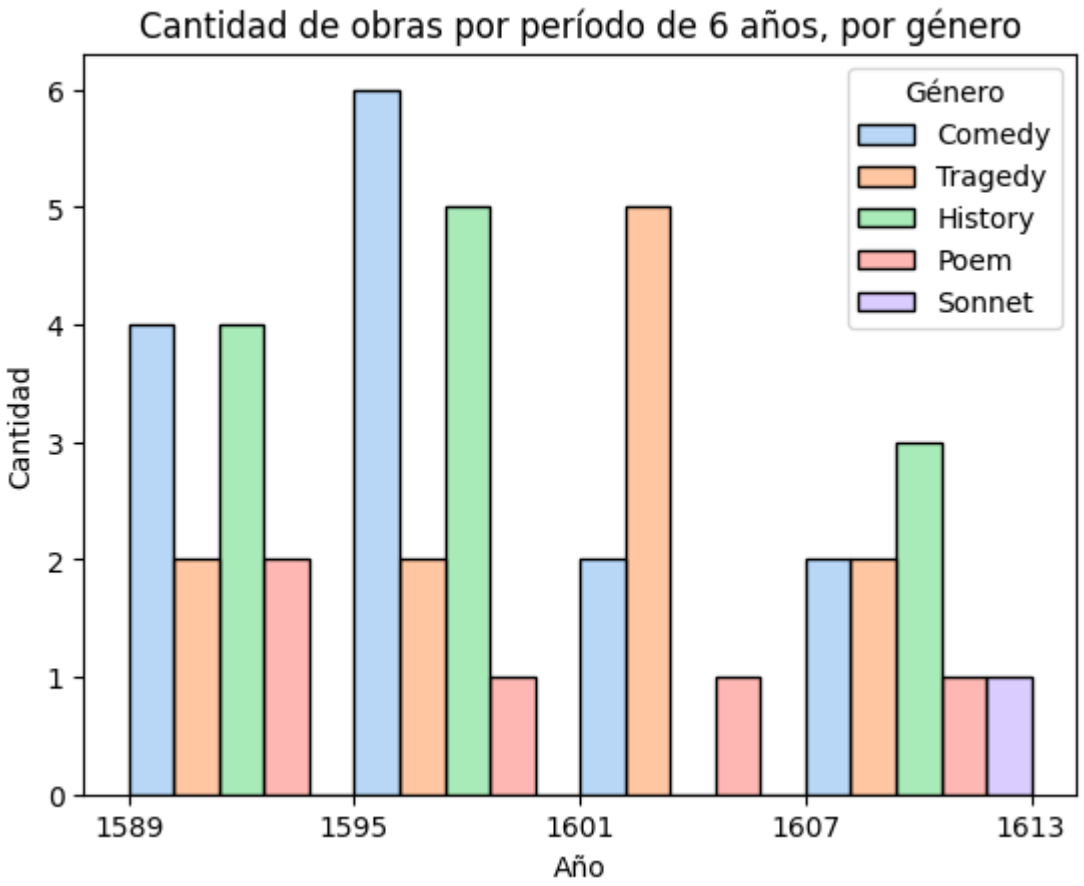
Obra de Shakespeare a los largo de los años

Se presenta a continuación un histograma que muestra la cantidad de obras escritas por Shakespeare agrupadas en períodos de tres años, comenzando en 1589, el año en que escribió su primera obra.



Podemos observar que el período de mayor productividad está comprendido entre los años 1592 y 1601. El resto de sus años de actividad escribió aproximadamente 4 obras cada 3 años.

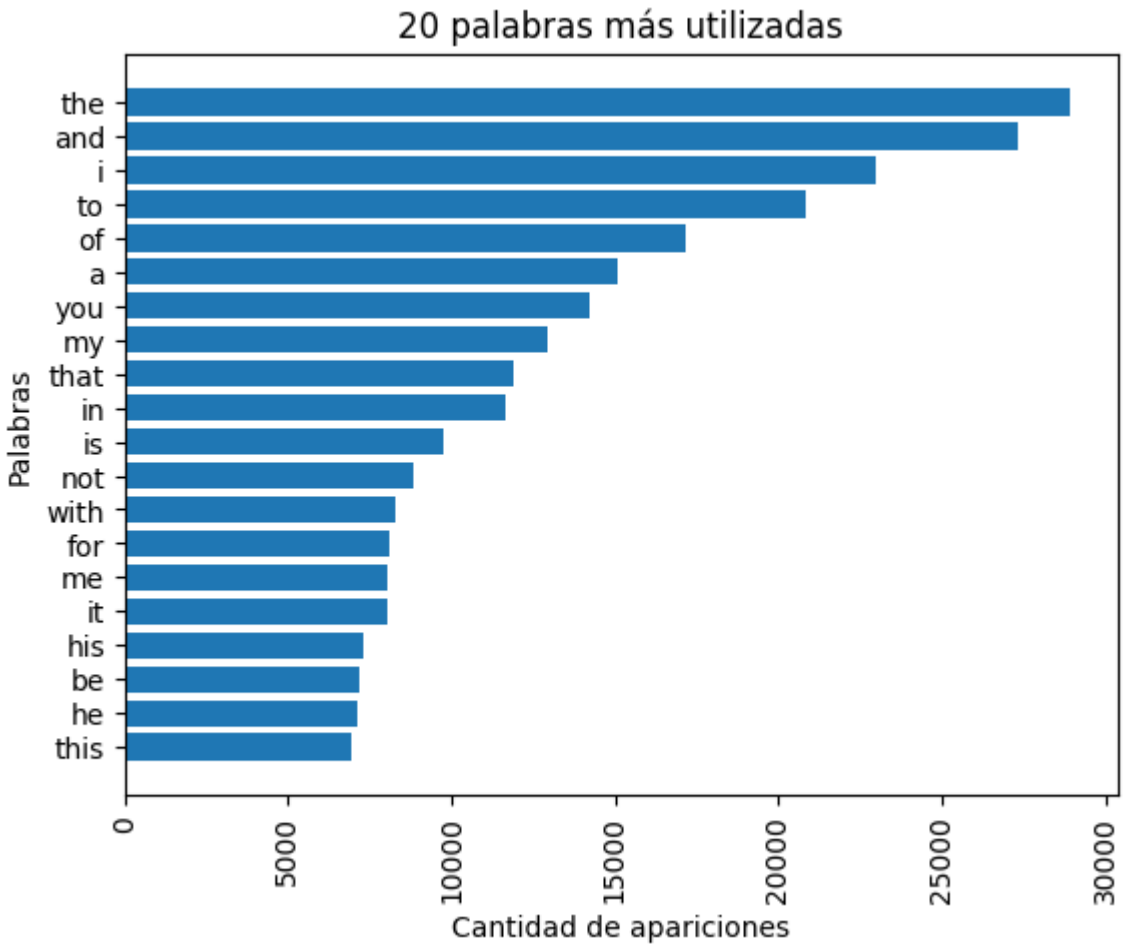
Se presenta a continuación un histograma categorizado por género de las obras. En este caso, para facilitar la identificación de tendencias se agrupó las obras en periodos de 6 años.



Se puede observar como las obras de comedia e historia predominan en los inicios, disminuyendo hacia el final de su carrera, mientras que la tragedia toma un papel preponderante en la mitad de la misma. Los sonetos aparecen recién sobre el período de los últimos 6 años.

Análisis de palabras

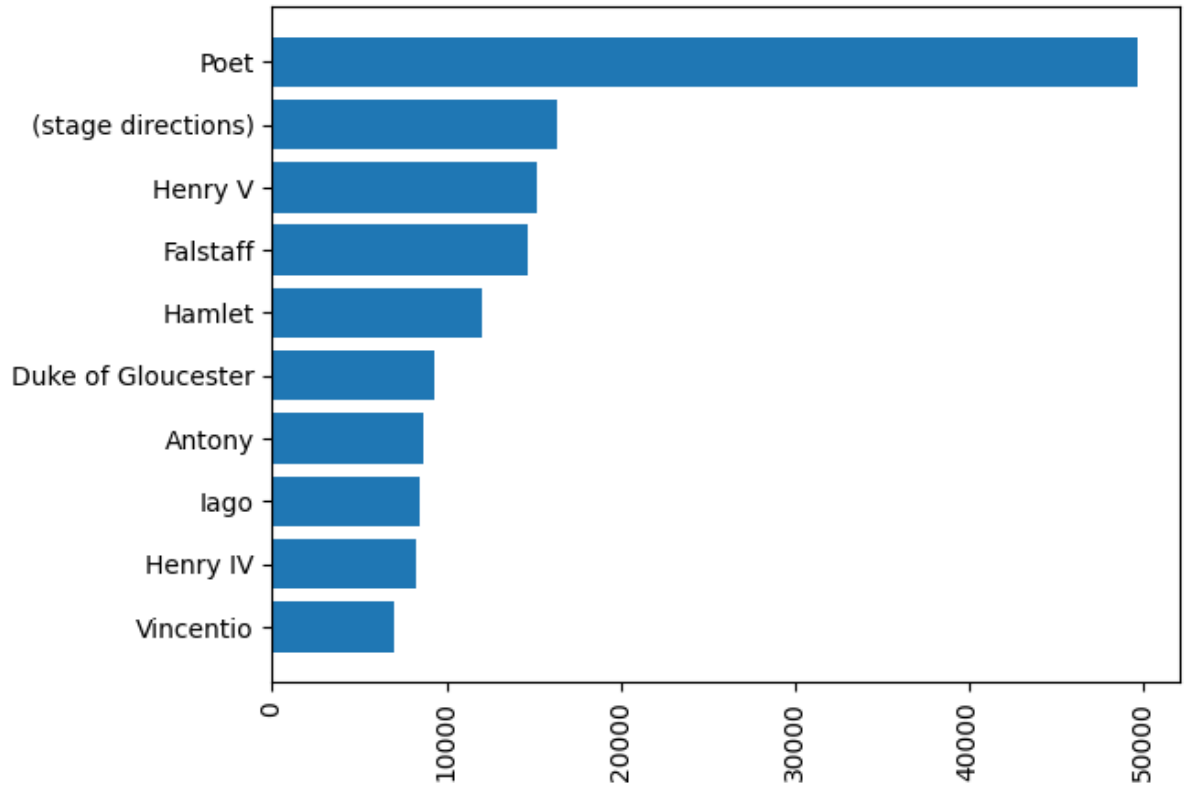
Para poder hacer un análisis de las palabras utilizadas en las distintas obras, es necesario realizar un trabajo de limpieza y normalización sobre el texto. Lo primero que se hizo fue pasar todo el texto a minúsculas. Luego de eso se eliminaron signos de puntuación y contracciones, con el objetivo de quedarnos únicamente con las palabras. Específicamente los caracteres y fragmentos eliminados fueron: "[", "\n", ",", ";", "?", ".", "!", "]", ":", "'d", "'s", "'ll", """, "-". Si bien se podría intentar hacer un procesamiento más avanzado con respecto a las contracciones, consideramos que la cantidad no era representativa en el total de palabras, por lo que optamos simplemente por eliminarlas. Finalmente se generó un dataframe nuevo “df_words”, separando las palabras de cada diálogo y colocándolas en filas diferentes. A partir de este dataframe se cuenta la cantidad de apariciones de cada palabra y se ordena los resultados de mayor a menor.
Se presenta a continuación una gráfica mostrando las veinte palabras más usadas a lo largo de toda la obra, junto a la cantidad de veces que aparecen.



Se puede observar que la mayoría de las palabras corresponden a artículos, pronombres y preposiciones. Podría ser interesante utilizar un diccionario para quitar estas palabras de la comparación.
Por otro lado, dado que en el *DataFrame* de las palabras tenemos el identificador del capítulo, podemos hacer un *join* con las obras para tener disponible el género asociado. Esto nos permitiría realizar un filtrado previo y graficar las palabras más frecuentes para cada género, para comparar el vocabulario utilizado. También podríamos buscar los personajes con más palabras (como haremos más adelante) y graficar para cada uno de ellos las palabras más utilizadas, con el fin de identificar diferencias entre los distintos personajes.

Personajes con más palabras

Se presenta a continuación una gráfica con los personajes que tienen más palabras asociadas a lo largo de toda la obra.



El personaje con más palabras es "Poet", con una amplia diferencia. Esto se debe a que este personaje tiene asignados todos los textos de las obras de poesía. El segundo personaje con más palabras es "(stage directions)", que como se vio anteriormente, tampoco es un personaje real, sino acotaciones. Se podrían descartar previamente todas las filas correspondientes a estos dos personajes, con el objetivo de visualizar la cantidad de palabras utilizadas por los que consideramos relevantes o diferenciar entre “Poet” o “(stage directions)” de distintas obras, para que no tengan tanto peso en el total de palabras.

Preguntas adicionales

A continuación presentamos una lista de preguntas adicionales que podríamos responder con estos datos y una breve idea de como hacerlo:

- ¿Cuál es la obra más larga de Shakespeare?
  Se puede simplemente contar la cantidad de palabras de cada obra.

- ¿Cómo evoluciona el vocabulario de Shakespeare a lo largo de su carrera?
  Se puede graficar las palabras más utilizadas en distintos periodos de tiempo.

- ¿Las obras con más personajes son más largas?
  Se puede estudiar la correlación entre estas dos variables.

- ¿Qué tan amplio es el vocabulario utilizado en las obras?
  Se puede contar la cantidad de palabras diferentes utilizadas en cada obra. También se puede agrupar por género para ver si hay alguna relación.

- ¿Cuál es el personaje más alegre?
  Se puede aplicar técnicas de análisis de sentimientos para evaluar el tono emocional de las palabras utilizadas por los diferentes personajes.