UNIVERSIDAD DE LA REPÚBLICA

FACULTAD DE INGENIERÍA

Curso: Introducción a la Ciencia de Datos

Informe de Tarea 1:

"Análisis de base de datos de Shakespeare"

Grupo: 9

Andres Santos, 4.646.015-1

Pedro Ignacio Lasa, 3.401.437-0

Fecha de entrega: 20/05/2024

CONTENIDO

INTRODUCCIÓN

La tarea propuesta, pretende dar al estudiante un primer acercamiento a la Ciencia de Datos explorando una base de datos sobre las obras de William Shaskpeare. La misma, requirió incorporar una serie de herramientas informáticas, a las que en muchos casos nos eran ajenas como las librerías Pandas, lenguaje de código Python, el uso de Jupyter notebooks, así como Git, Github o Gitlab, herramientas que permiten trabajar de forma colaborativa y ordenada.

PARTE 1. – EXPLORACION DE LA BASE DE DATOS

Sección A

Exploración de la base de datos

Los datos se presentan como una base de datos relacional, se estructura en 4 tablas (o "dataframes") vinculadas entre sí por restricciones de integridad del tipo clave principal-clave foránea. Las tablas son: "paragraphs", "chapters", "characters", y "works".

1. Tabla "works": contiene información sobre las obras de Shakespeare, sus títulos, fechas de publicación y sus géneros.

2. Tabla "paragraphs": Esta tabla contiene párrafos o fragmentos de texto de las obras de Shakespeare. Cada registro podría representar un párrafo específico de una obra. Además, se vincula a la tabla de "characters" y "chapters" a través de campos foráneos con sus respectivos id.

3. Tabla "chapters" (capítulos): contiene información estructurada sobre las obras, como divisiones en actos, escenas o capítulos. La función de esta tabla sería organizar las obras en secciones más grandes, facilita la navegación y el análisis. Esta tabla se vincula con la tabla "Works", a través del campo *works_id*

4. Tabla "characters" (personajes): información sobre los personajes que aparecen en las obras. Cada registro podría representar un personaje con detalles como nombre, descripción, roles en obras específicas. La función de esta tabla es proporcionar información detallada sobre los personajes en las obras.

Datos Faltantes:

En la tabla *works* no hay datos faltantes en ninguna de las columnas. Todos los valores están presentes. Lo mismo sucede con las tablas *paragraphs* y *chapters.*

Por otro lado, la tabla *characters* presenta 5 datos faltantes en la columna "Abbrev", y la columna "Description" tiene 646 datos faltantes.

La mayoría de los dataframes tienen todos los datos completos, lo cual es bueno para el análisis. Depende de los objetivos se podría eliminar o sustituir los datos faltantes, así no se entorpece futuros análisis.

Párrafos por personaje:

Para determinar la cantidad de párrafos por personaje se cruzaron las tablas *paragraphs* y *characters* por medio del campo “character_id”.

En el cuadro a continuación se listan los 10 personajes con más párrafos:

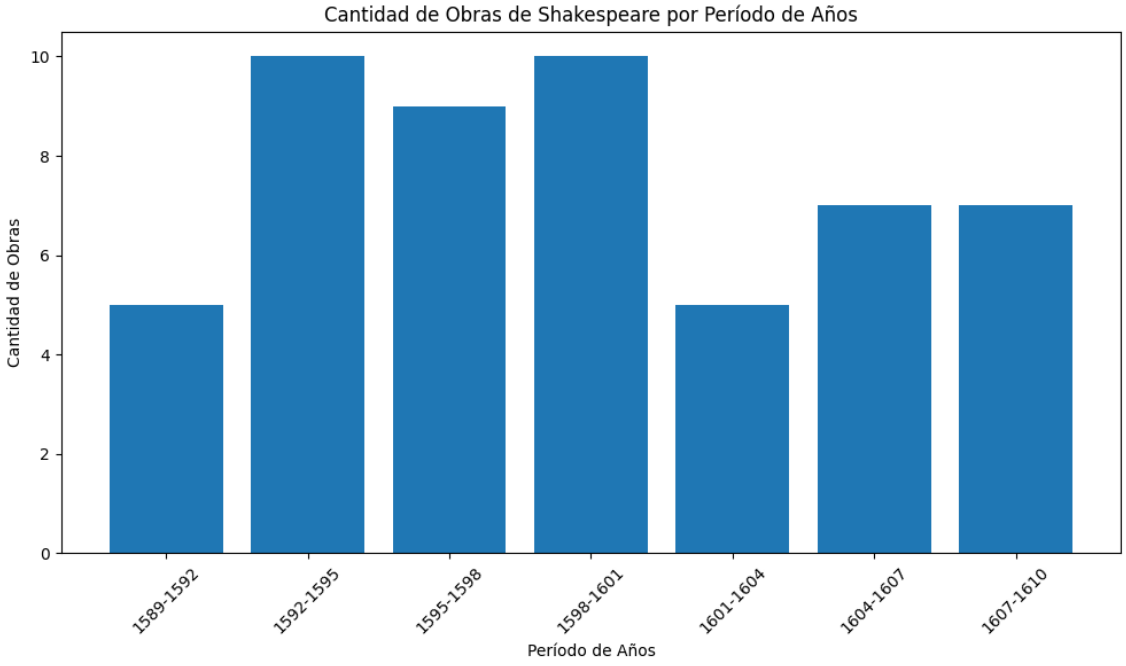
Personaje	Líneas
(stage directions)	3751
Poet	766
Falstaff	**471**
Henry V	377
Hamlet	358
Duke of Gloucester	285
Othello	274
Iago	272
Antony	253
Richard III	246

Desestimamos las dos primeras filas por no ser personajes propiamente dichos. Sin normalizar los párrafos y sin hacer cambios, el personaje con mayor presencia en los párrafos es 'Falstaff', con un total de 471. Veremos en el próximo apartado si cambia al normalizar las palabras.

Sección B - Cantidad de producciones y géneros por año.

El objetivo ahora es utilizar diferentes herramientas de visualización gráficas para estudiar la obra de Shakespeare a lo largo de los años, comparando los niveles de producción y las tendencias en cuanto a los géneros más utilizados en diferentes períodos de su carrera.

A continuación, se presenta la cantidad de obras editadas, tomando períodos de 3 años, partiendo del año 1589 hasta el 1610.

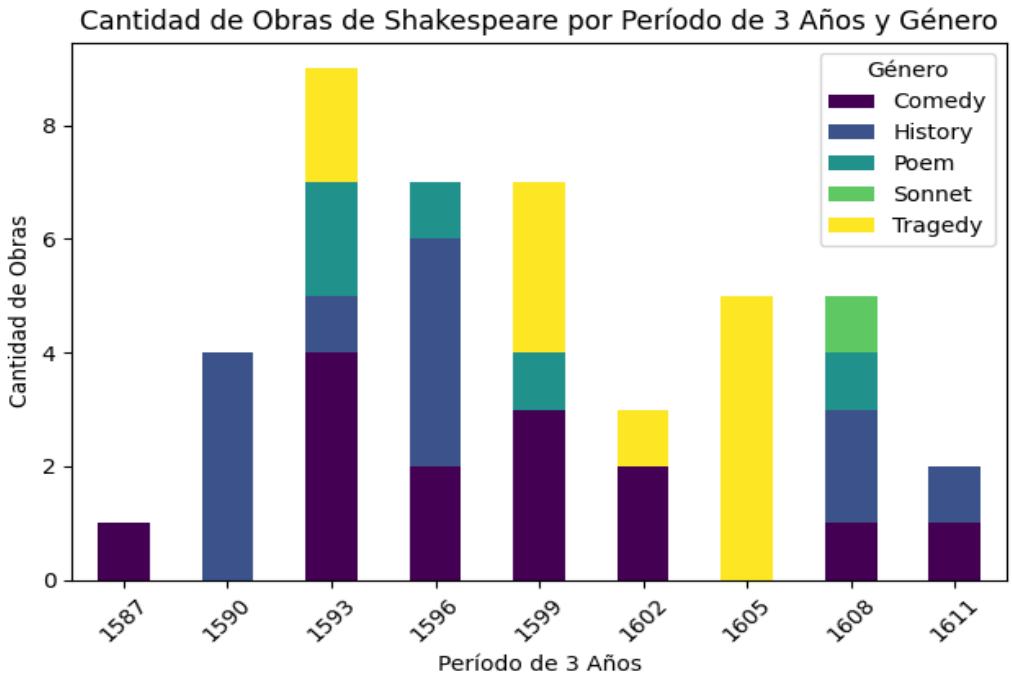


La primera impresión que se desprende del gráfico, es que Shakespeare fue un artista extremadamente prolífico a lo largo de toda su carrera. Su período de mayor producción se ubica aproximadamente entre los años 1592 y 1601, cuando tenía entre 28 y 37 años de edad (cabe señalar que Shakespeare nació en abril de 1564).

Es importante señalar, que medir el nivel de producción de un artista solo por la cantidad de obras editadas puede resultar un tanto arbitrario o al menos impreciso. En primer lugar, porque incluso si nos acotamos a un enfoque meramente cuantitativo, las obras pueden tener extensiones muy diversas. En este sentido, además de la cantidad de obras, podría ser útil considerar la cantidad de palabras o párrafos en cada una de ellas.

En segundo lugar, si nuestro objetivo fuera estudiar cuál fue la etapa de mayor esplendor o auge del artista, deberíamos ponderar otras cualidades asociadas a la calidad de sus obras, ya sea por su trascendencia o popularidad. En este sentido, si tuviéramos los datos, sería interesante comparar para las diferentes obras, la cantidad de ventas de libros, los idiomas a los que fueron traducidos, o las entradas vendidas en el teatro y el cine (esto último parece un dato bastante inaccesible).

En el siguiente gráfico se hace un análisis similar al anterior, pero discriminado los diferentes géneros por colores:



No parece haber una clara preferencia por un género en particular, aunque si se puede ver que en determinados períodos se volcó más a unos que a otros. Digamos que fue alternando principalmente entre el drama, la comedia y el drama histórico. En menor medida que estos tres anteriores aparece la poesía. En el año1609, ya llegando al final de su carrera, pública “Sonnets”, única publicación del género homónimo.

Explorando otras formas de visualización, a continuación, generamos una nube de palabras con los géneros de las obras. Para ello se utilizó la librería “WordCloud”, el tamaño de la fuente está asociado a la cantidad de veces que fueron adoptados cada uno de los géneros en sus obras:

Sección C - Limpieza del Texto

Se plantea realizar el conteo de palabras en todos los textos del autor y estudiar cuáles son las más frecuentes. Para ello, es necesario realizar previamente un tratamiento específico del texto:

1. Normalización: convertimos todas las mayúsculas en minúsculas.
2. Reemplazo de contracciones: suplantamos las diferentes contracciones por las palabras completas correspondientes. Por ejemplo, donde figura el apóstrofe seguido de "re" ('re) la suplantamos por " are".
3. Limpieza: Eliminamos todos los signos de puntuación, reemplazándolos por espacios.

Estas modificaciones nos permitirán identificar de forma consistente palabras en diferentes contextos. Por ejemplo, podremos reconocer la palabra "go" tanto en "go" como en "GO!".

Se debe aclarar que la eliminación de las contracciones en algunos casos nos puede llevar a errores. Cuando la contracción es "´s" puede estar indicando que algo pertenece a alguien o algo, o puede ser la contracción del verbo "is", o incluso, puede ser la contracción del verbo "has". En este caso, hemos decidido que se suplante por el verbo "is" ya que consideramos que así es en la mayoría de los casos.

Ocurre algo similar con el "'d" que puede ser una contracción del vervo modal "would" o del verbo "had".

En el cuadro a continuación se muestra para cada contracción, las palabras por las cuales fueron sustituidas y la cantidad veces:

Contracción	Sustitución	Cantidad
can't	cannot	3
won't	will not	0
n't	not	276
're	are	144
's	is	8812
'd	would	6482
'll	will	2493
't	not	3036
've	have	18
'm	am	110

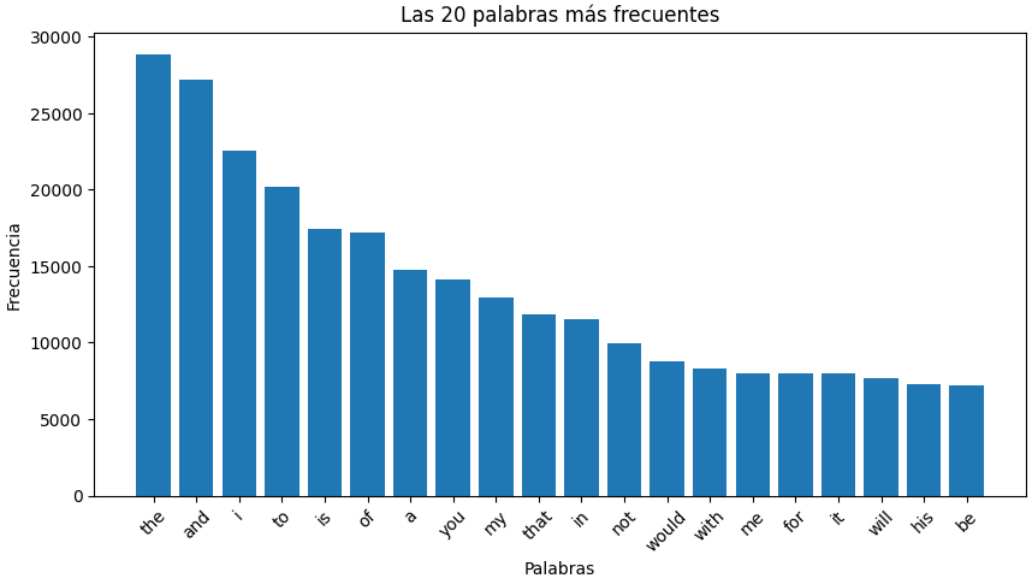
Vemos que justo "´s" y "'d" son las contracciones más veces encontradas. Veremos que esto va a influir apreciablemente en los resultados a los que arribemos más adelante.

PARTE 2: CONTEO DE PALABRAS Y VISUALIZACIONES

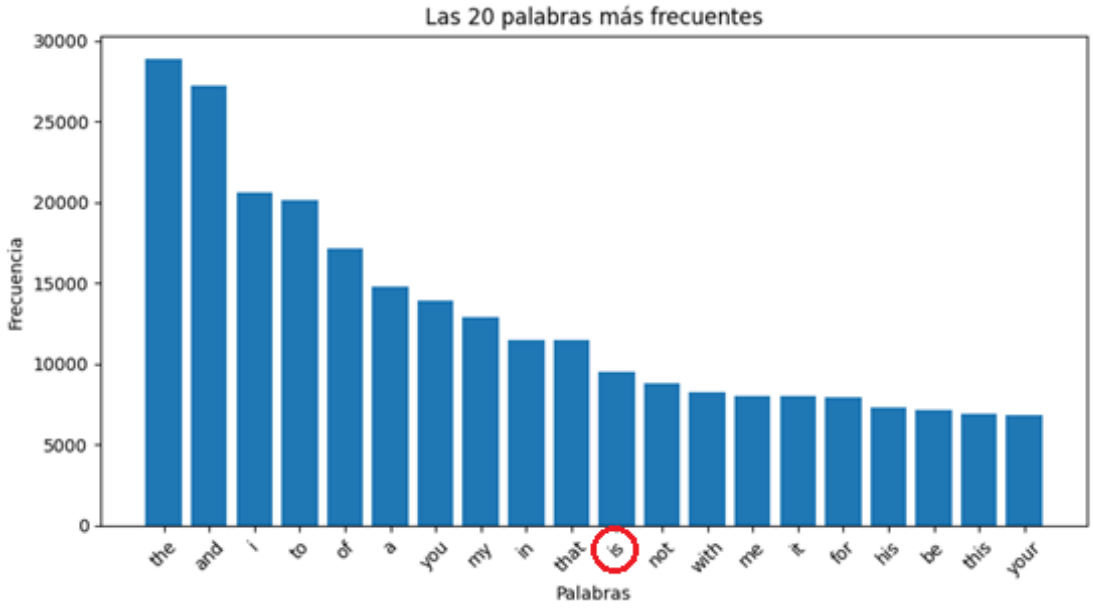
Sección A: Conteo de palabras

Una vez que obtuvimos un texto uniforme y sin signos de puntuación, estamos en condiciones de avanzar con el conteo de palabras propiamente. Para ello debimos generar un nuevo dataframe, al que llamamos "words". Partiendo de la tabla "paragraphs", transformamos el texto limpio en listas de palabras, luego creamos una nueva columna con una sola palabra por campo. Generamos un nuevo dataframe con esta información, el cual tiene 5 columnas: "id" (identificador del párrafo), "ParagraphNum" (número de párrafo), character_id (identificador del personaje), "chapter_id" (identificador del capítulo), "word" (palabra).

A continuación, se presenta un gráfico de barras con las 20 palabras más frecuentes:



El resultado muestra dos de las palabras que sustituyen contracciones en el tratamiento previo. Si no hubiésemos realizado tal modificación, el resultado obtenido hubiese sido el siguiente:

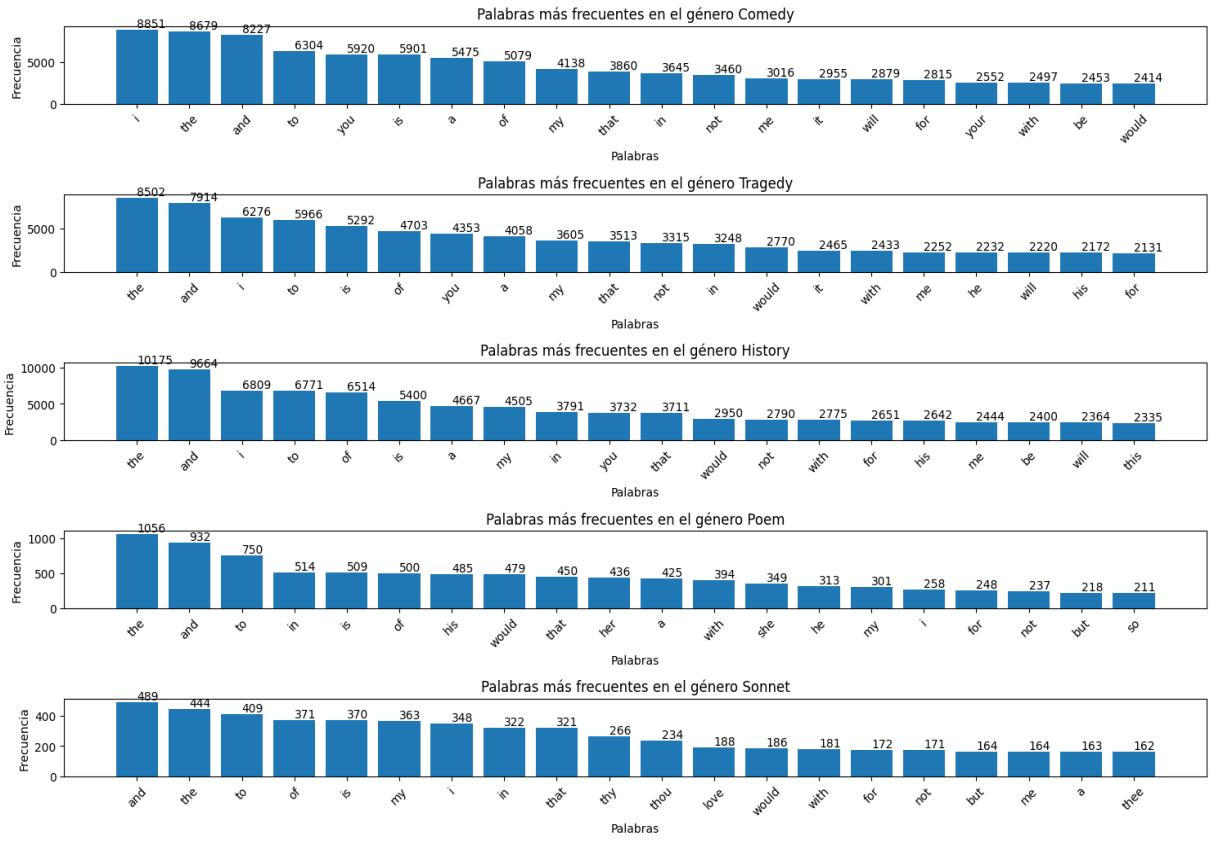


La palabra “is” cae 6 posiciones en la lista de las más frecuentes, la palabra “would” ya no figura entre las 20 más repetidas. Un análisis más exhaustivo, si el objetivo lo ameritara, debería identificar cada vez que una de estas palabras figura en el texto y, en función de su contexto, aplicar la sustitución correspondiente. Consideramos que este tipo de análisis excede los propósitos del curso.

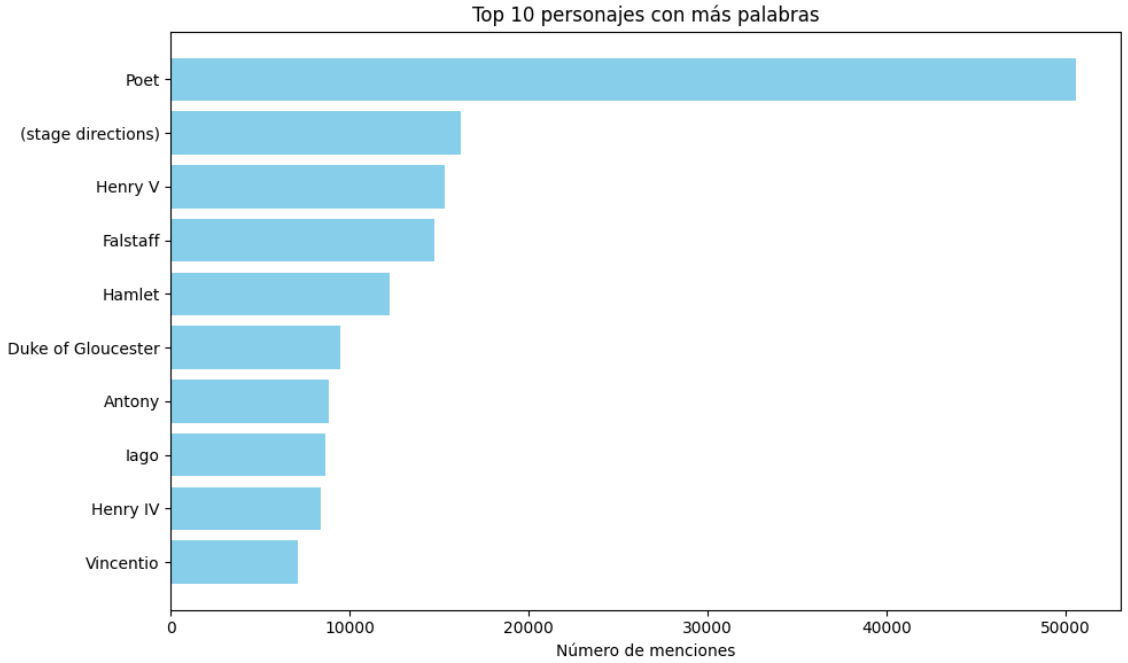
Conteo de palabras por género

Por último, analizamos las palabras más usadas por género y lo representamos en un gráfico por cada uno. Unimos la tabla “df_chapters” con “df_works” usando work_id e id respectivamente y obtenemos así, el género de cada capítulo. Luego, guardamos el resultado en “df_chapters_merged”, unimos “df_paragraphs” con “df_chapters_merged” usando chapter_id y id_chapter, añadiendo el género a cada párrafo. Guardamos el resultado en “df_paragraphs_merged”. Durante estas uniones se eliminan columnas redundantes para mantener los datos organizados.

Para cada género se concatenaron los textos de los párrafos, se contaron las palabras y se seleccionaron las 20 palabras más comunes. Visualizamos los resultados:

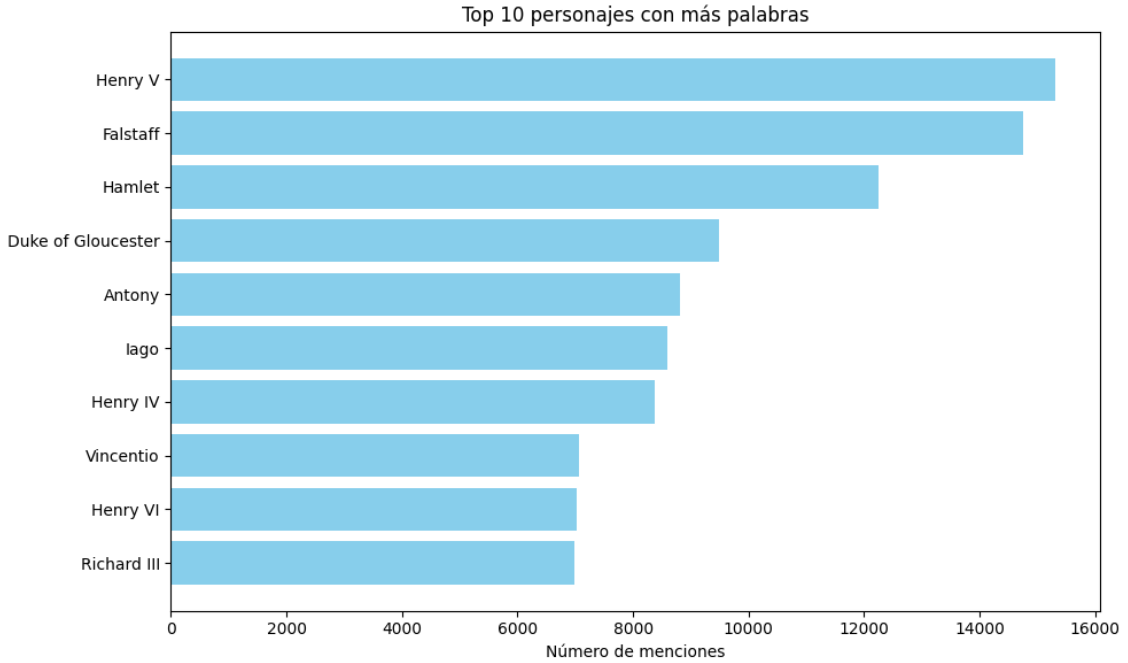


Sección B:
El objetivo ahora es contar la cantidad de palabras dichas por cada personaje. En una primera instancia la búsqueda nos arroja el siguiente resultado:



Nuevamente el resultado se ve distorsionado por la inclusión de "Poet" y "(stage directions)" en el análisis. Estos no son personajes propiamente dichos. Se trata de los framentos de texto en que el autor relata la historia por fuera de los diálogos (Poet), y las indicaciones por fuera de la historia que dan pautas y direcciones para la representación de la obra en escena (stage directions). Para resolver este problema, descartamos los nombres "Poet" y "(stage directions)" del análisis de palabras por personaje.

Siguiendo la sugerencia anterior, el personaje con más párrafos es *'Henry V'*, con un total de 15076.



Sección C:

Aquí dejaremos planteado algunos análisis que podríamos realizar con la misma base de datos, y daremos una orientación de cómo se podrían abordar.

Uno de los análisis que se dejó planteado anteriormente, es el de analizar el nivel de producción de Shakespeare a lo largo de su vida, pero no tomando como medida la analizar si existe alguna relación entre los géneros de las obras y la cantidad de palabras utilizadas por personaje cantidad de obras editadas, sino la cantidad de palabras escritas. Sería una forma mucho más precisa a la hora de determinar cuál fue su época más prolífica. Este implicaría cruzar las tablas "words", "paragraphs" y "works" para determinar el año que fue escrita cada palabra.

Otra posibilidad es preguntarse cuáles son los personajes con mayor protagonismo en las obras de Shakespeare, en términos de la cantidad de párrafos que poseen. A su vez esto se podría discriminar por género (sexual, no literario). Esto implica asignarle un sexo a cada personaje, datos con los cuales no contamos a priori, pero se podrían generar o conseguir fácilmente.

Por último, nos parece importante. Se podría contar la cantidad de palabras utilizadas por cada uno, en diferentes géneros por obras. Luego, se puede realizar un análisis estadístico para determinar si existe una correlación significativa entre los géneros de las obras y la cantidad de palabras utilizadas por personaje.

CONCLUSIONES:

El análisis realizado sobre las obras de William Shakespeare, a través de una base de datos relacional, ha proporcionado un valioso primer acercamiento a la Ciencia de Datos, aplicando herramientas como Python, Pandas, Jupyter notebooks, Git y GitHub; este proceso nos permitió explorar y limpiar los datos, además de generar visualizaciones significativas.

Esta tarea ha demostrado la utilidad de la Ciencia de Datos para extraer conocimientos valiosos de textos literarios, subrayando la importancia de la limpieza de datos, la visualización y la exploración cuidadosa.

Fue un proceso progresivo, desde comprender datos con estructura relacional, hasta la visualización. Nos deja una muy buena aproximación al entendimiento de la librería Pandas, Matplolib y Wordcloud.

La ciencia de datos, no solo se basa en encontrar patrones o tendencias en los datos, sino también en saber comunicar de forma asertiva los resultados.