INTRODUCCIÓN A LA CIENCIA DE DATOS

Tarea - 1

AÑO 2024

GRUPO 14:

Francisco Galletto
Fernando Sellanes
Lucas Sellanes

FECHA: 21 de mayo de 2024

Tabla de contenidos

1. Cargado y limpieza de datos

1.1. A

Para realizar la tarea se cuenta con cuatro tablas utilizadas como entradas para el código realizado en jupyter notebook. A continuación se describe el contenido de cada tabla:

- **works**: Contiene información del título, año y género de cada obra realizada por William Shakespeare, además de un respectivo id para cada obra.
- **characters**: Contiene información de nombres y descripción de los personajes.
- **chapters**: Contiene información de cada capítulo, asociándole una escena (a través de un id), descripción de esa escena, obra asociada (id de la obra), además de un id para cada capítulo.
- **paragraphs**: A cada párrafo se le asocia un id, número de párrafo, texto sin un formato definido, id del personaje e id del capítulo asociado.

De la descripción anterior y la Figura 1 se deducen las relaciones entre tablas.

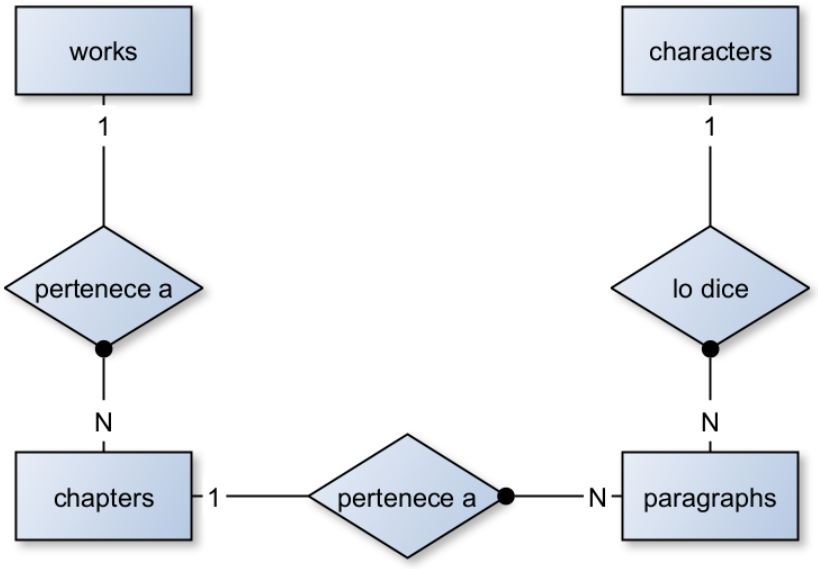

Figura 1: Diagrama representativo de la base de datos

.

Se verifica si existen datos faltantes en alguna columna de las tablas descriptas anteriormente, dando como resultado valores faltantes en las columnas **Abbrev** (5 de 1266) y **Description** (646 de 1266), de la tabla **characters**.

Luego, para analizar la cantidad de párrafos por personaje, se filtra el DataFrame de la tabla **characters** para poder ver los 5 personajes con más párrafos (ver Figura 2). De esto último se tiene que, el personaje con más párrafos es Falstaff (Sir John Falstaff), dado que stage directions y Poet no refieren a un personaje como tal.

	Unnamed: 0	id	CharName	Abbrev	Description
558	558	559	Hamlet	Ham	son of the former king and nephew to the prese...
572	572	573	Henry V	HENRY5	Prince, King of England
392	392	393	Falstaff	FALSTAFF	Sir John Falstaff
893	893	894	Poet	Poet	the voice of Shakespeare's poetry
1260	1260	1261	(stage directions)	xxx	NaN

Figura 2: DataFrame de la tabla **characters** filtrado por los 5 personajes con más párrafos.

.

1.2. B

En la Figura 3 se puede ver un histograma de cantidad de obras realizadas a lo largo de los años, en este último también se realiza una diferenciación por género. Se puede ver que todas las obras se realizaron entre los años 1589 y 1612, habiendo realizado entre una y cuatro obras por año, a excepción del año 1603 donde no realizó ninguna. Si se mira su producción por género, se puede ver que los predominantes son la comedia, tragedia e historia, siendo este último el género (de los tres predominantes) con mayor interrupción en términos de lapso temporal. Por último, el genero con menor cantidad de trabajos es el soneto, contando con un solo trabajo realizado.

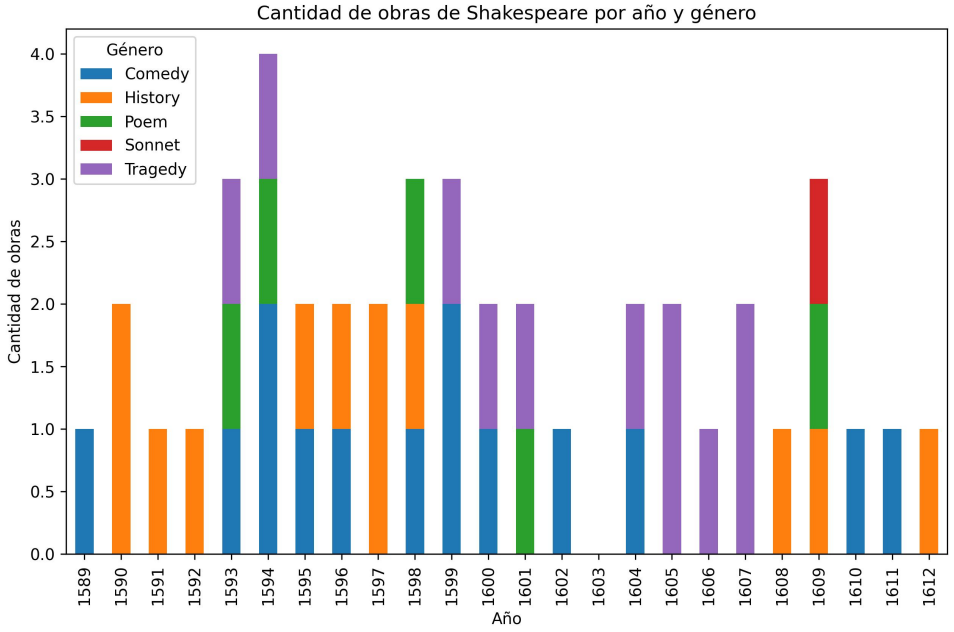

Figura 3: Histograma de obras realizadas por Shakespeare a lo largo de los años diferenciadas por género.

.

Otra forma de visualizar la evolución temporal de las obras diferenciadas por género se presenta en la Figura 4, en esta se muestra el acumulado de obras separadas por género a través de los años. Se puede observar que el género de comedia e historia son los predominantes en los primeros 10 años, luego deja de escribir historias y toma mayor protagonismo la escritura de tragedias. De igual forma, se observan períodos de no escritura de diferentes géneros, como para las historias y poemas.

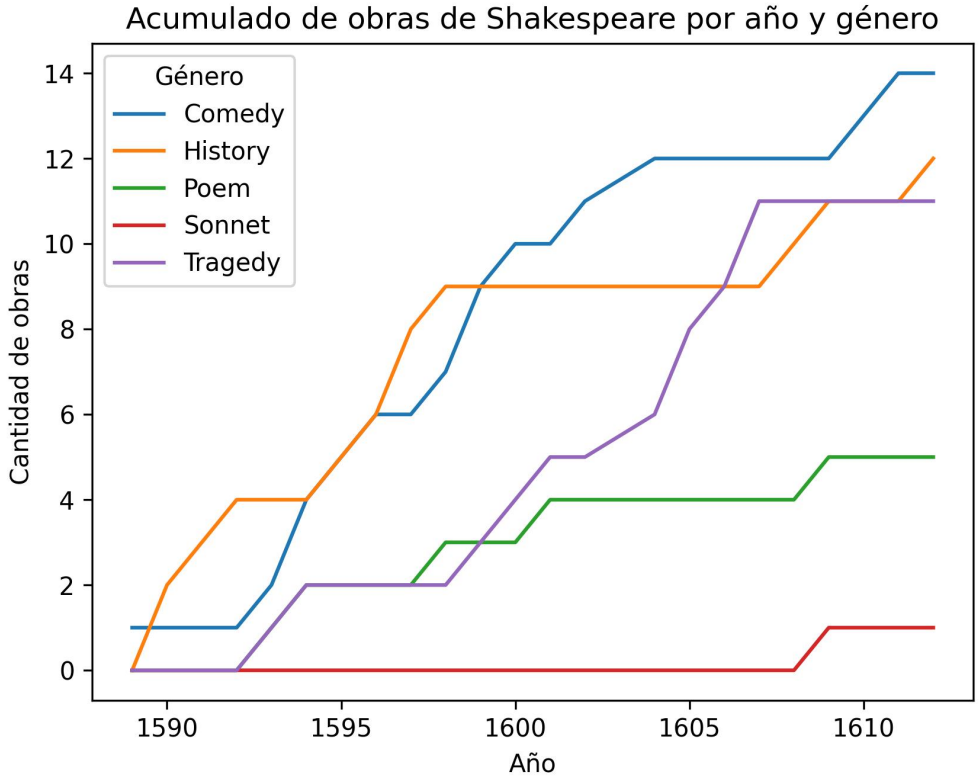

Figura 4: Acumulado de obras realizadas por Shakespeare a lo largo de los años diferenciadas por género.

.

1.3. C

Para realizar el conteo de palabras se normaliza el texto, para esto se utiliza la función `clean_text` que realiza la normalización respecto a las mayúsculas, es decir, aplicando esta función normalizamos el texto para eliminar las mayúsculas. Para completar la normalización, a la función `clean_text`, se le agrega la posibilidad de normalizar respecto a signos de puntuación, números y demás expresiones presentes en el texto.

Para ver los signos (todo lo que no sean letras) presentes en el texto, primero se crea una columna que contiene listas formadas con las palabras y signos de cada párrafo, luego se crea una columna con las entradas de todas las listas, y por último se cuenta la cantidad de ocurrencias de cada letra y signo obtenida. De este conteo se puede visualizar todo lo que no es palabra y está presente en el texto. Finalmente, se normaliza en base a los siguientes signos encontrados: `[`, `]`, `\n`, `\r`, `,`, `;`, `?`, `.`, `!`, `:`, `'`, `-`, `_`, `(`, `)`, `&`, `"`, `0:9`.

2. Conteo de palabras y visualizaciones

2.1. A

En la Figura 5 se puede ver un histograma que muestra el conteo de las 10 palabras más frecuentes en la obra de Shakespeare.

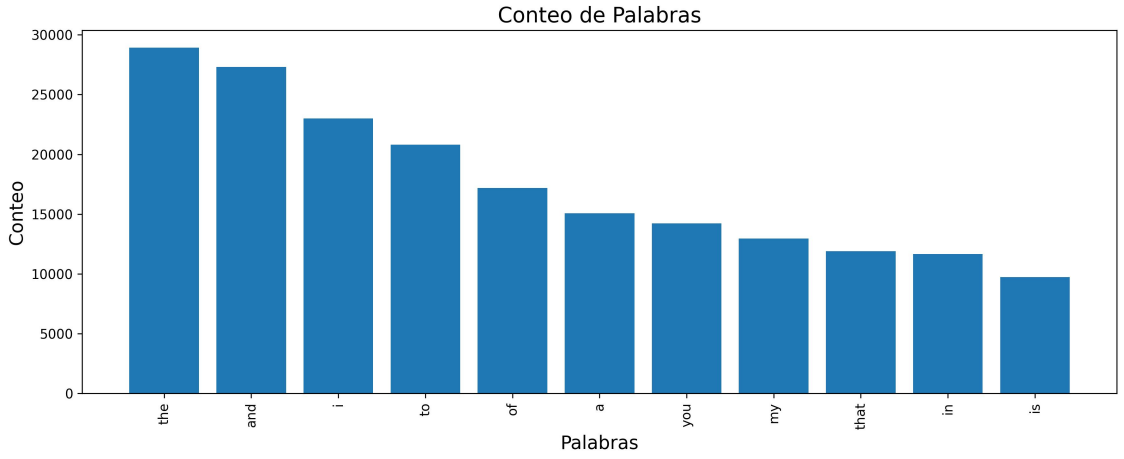

Figura 5: Histograma de las 10 palabras más frecuentes en toda la obra.

.

Con el fin de encontrar diferencias entre géneros, se podría obtener el porcentaje (o conteos) del total de conteos asociado a cada género para las 10 palabras más frecuentes. Luego, se realiza un histograma donde cada barra del histograma, asociada a una palabra, se divide en 5 barras (una por cada género) como en la Figura 3). Algo similar a lo explicado anteriormente se puede aplicar para visualizar a qué personajes están asociadas las palabras más frecuentes, es decir, se podrían ver los 5 personajes que más frecuentemente utilizaron cada palabra y hacer un histograma.

2.2. B

En la Figura 6 se puede ver un histograma que muestra los 10 personajes con mayor cantidad de palabras en las obras. Sucede que, los dos primeros, no refieren a personajes como tal, es por eso que se crea la Figura 7 donde se eliminan los dos primeros nombres que no refieren a personajes.

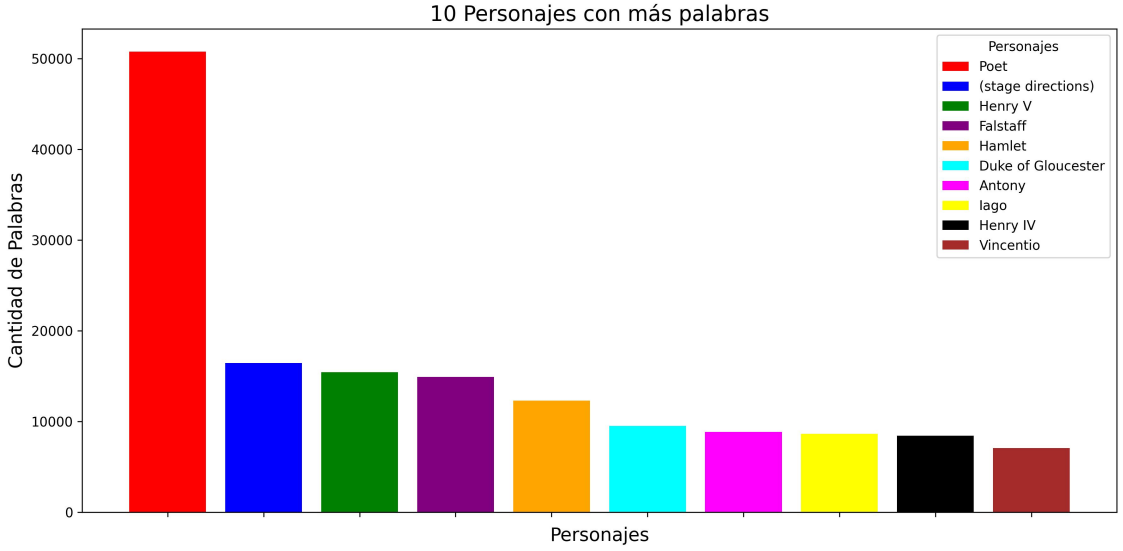

Figura 6: Histograma de los 10 personajes con más palabras en la obra.

.

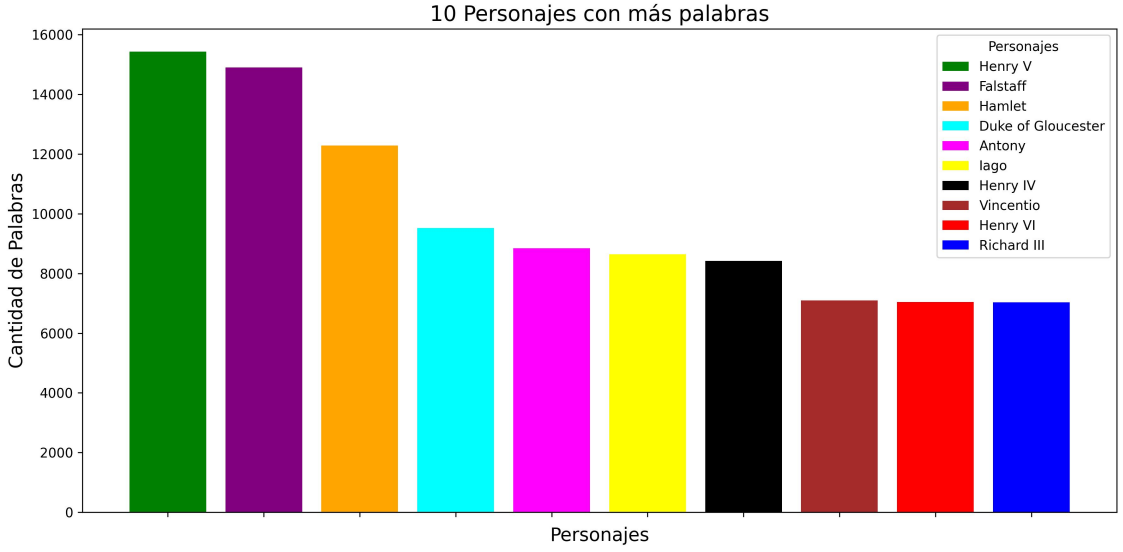

Figura 7: Histograma de los 10 personajes con más palabras en la obra.

.

2.3. C

Finalmente, se identifican posibles problemas a responder a partir de la base de datos procesada.

El conteo de palabras por obras permite tener una medida aproximada del peso de los atributos de las obras. Por ejemplo, se puede responder que influencia tiene cada personaje en las obras realizando un conteo de las palabras al igual que se lo realizó anteriormente para la totalidad de las obras.

También es posible visualizar cuál es la importancia de los diálogos utilizando la variable CharName de la tabla **characters**, contando las palabras por personajes y diferenciando los personajes de la voz referenciada a Shakespeare y las direcciones de escenario. De esta forma, se puede diferenciar que obras o géneros son más hablados o actuados.