Tarea 1

Introducción a la Ciencia de Datos

Materia: Introducción a la Ciencia de Datos

Sofía Zeballos

Estudiante de Maestría en Bioinformática

Pablo Schiavone

Estudiante de Maestría en Ciencia de Datos y Aprendizaje Automátic

Montevideo 23 de mayo de 2023

Introducción

El objetivo de la tarea es realizar un análisis exploratorio de la base de datos que representa las obras realizadas por Shakespeare. El análisis está compuesto por dos partes. Por una lado carga y limpieza de datos y por otro conteo de palabras y visualización. Por más detalles ver Anexo Cargado y limpieza de datos, y Conteo de palabras y visualizaciones.

Cargado y limpieza de datos

Las datos a analizar provienen de una base de datos publicada en el siguiente link https://relational.fit.cvut.cz/dataset/Shakespeare la cual cuenta con documentación que explica sus relaciones.

Antes de comenzar con el análisis de datos se estudia el esquema relacional (ER) para entender las relaciones y verificar si por la carga semántica de sus tablas y atributos se puede entender a qué objeto de la realidad hace referencia. A continuación (Figura 1) se muestra el ER.

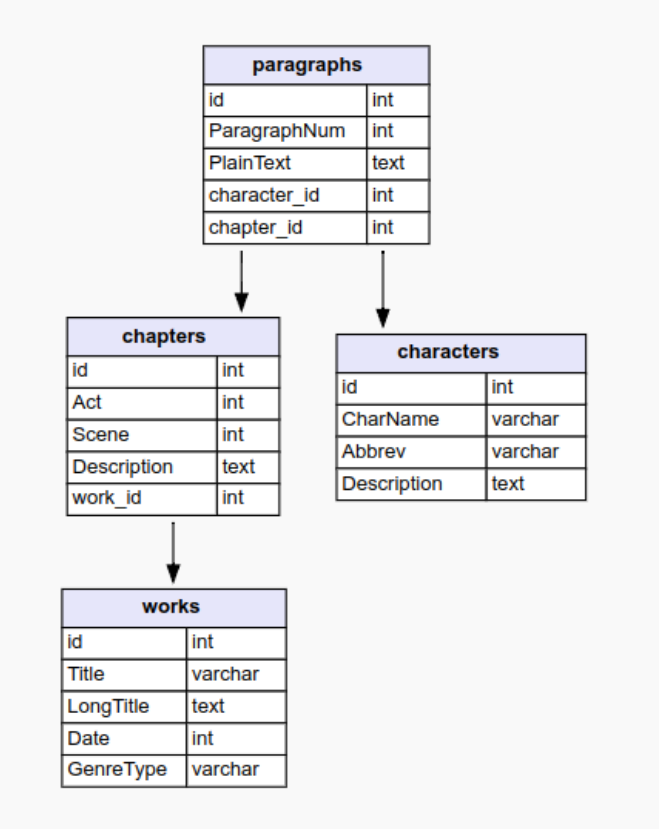

Figura 1: Descripción de ER

De acuerdo al estudio del ER, entendemos que se puede interpretar el objetivo de cada tabla y atributo.

A continuación se muestra cada tabla y lo que se espera de cada atributo.

Paragraphs	
id	Identificador de párrafo
ParagraphNum	Número de párrafo
PlainText	Líneas del párrafo
character_id	Identificador del personaje (Relación con Characters)
chapter_id	Identificador de capítulo (Relación con Chapter)

Chapters	
id	Identificador de capítulo
Act	Acto
Scene	Escena
Description	Descripción de l escena
work_id	Identificador de obra (relación con Works)

Characters	
id	Identificador de personaje
CharName	Nombre del personaje
Abbrev	Abreviatura del personaje
Description	Descripción del personaje

Works	
id	Identificador de obra
Title	Título de la obra
LongTitle	Título largo de la obra
Date	Fecha de publicación
GenreType	Género literario

Una vez entendido el ER se comienza a revisar que la información cargada en cada tabla cumpla con lo esperado.

Durante la primera exploración de datos se observa que en la tabla Paragraphs existen datos en el atributo PlainText que no coinciden con el dominio esperado (líneas de un párrafo). También se observa que en los casos que no coincide PlainText con el dominio se repite el charapter_id. En particular el character_id 1261(ver figura 3).

```
In [11]: #Visualización paragraphs
df_paragraphs
```

Out[11]:

	id	ParagraphNum	PlainText	character_id	chapter_id
0	630863	3	[Enter DUKE ORSINO, CURIO, and other Lords; Mu...	1261	18704
1	630864	4	If music be the food of love, play on;\nGive m...	840	18704
2	630865	19	Will you go hunt, my lord?	297	18704
3	630866	20	What, Curio?	840	18704
4	630867	21	The hart.	297	18704
...	...	...	...	...	...
35460	666323	3460	That she is living,\nWere it but told you, sho...	866	19648
35461	666324	3467	You gods, look down\nAnd from your sacred vial...	584	19648
35462	666325	3475	There's time enough for that;\nLest they desir...	866	19648
35463	666326	3483	O, peace, Paulina!\nThou shouldst a husband ta...	667	19648
35464	666327	3504	[Exeunt]	1261	19648

Figura 2: Exploración de tabla Paragraphs sin filtros

Por otro lado, se detecta que la tabla characters no solo contiene personajes sino que se utiliza para identificar otro tipo de entidades como directrices. A modo de ejemplo se muestra en la figura 3 lo ocurrido con el id 1261 de la tabla characters.

Revisión de df_characters con id 1261

```
df_characters.loc[df_characters['id'] == 1261]
```

	id	CharName	Abbrev	Description
1260	1261	(stage directions)	xxx	NaN

Figura 3: Datos del personaje con id 1261

Debido a que no sabemos si el registro con id 1261 es el único que no pertenece a un personaje, se consulta todos los personajes con descripción nula o vacía observándose nuevos registros que no acompañan la interpretación de ER. Ejemplo el que posee el id 1 (ver Figura 4).

Revisión de df_characters con Description null o vacía

```
df_characters.loc[df_characters['Description'].isnull() | (df_characters['Description'] == "")]
```

	id	CharName	Abbrev	Description
0	1	First Apparition	First Apparition	
1	2	First Citizen	First Citizen	
2	3	First Conspirator	First Conspirator	
3	4	First Gentleman	First Gentleman	
4	5	First Goth	First Goth	
...	...	...	...	...
1252	1253	Simpcox's Wife	Wife	
1257	1258	Cardinal Wolsey	CARDINAL WOLSEY	
1259	1260	Earl of Worcester	EARL OF WORCESTER	
1260	1261	(stage directions)	xxx	
1262	1263	Young Clifford	YOUNG CLIFFORD	

Figura 4: characters con descripción nula o vacía

Durante el análisis exploratorio de datos se intentó encontrar un patrón que nos indique que registro alojado en characters no es un personaje sin encontrar alguno categórico. La mejor aproximación encontrada es con description vacía pudiendo detectar grupos de personajes y acciones como por ejemplo First Apparition, All, messenger , etc

Por otro lado se detectan nombres de roles, grupos y personajes repetidos. Ver figura 4:

ChaName repetidos en df_characters

```
count_char= df_characters['CharName'].value_counts()
count_char.loc[count_char > 1].reset_index(name='Count')
```

	index	Count
0	All	23
1	Messenger	23
2	Servant	21
3	Lord	9
4	Page	8
...	...	...
120	All Lords	2
121	Lucilius	2
122	Lieutenant	2
123	Varrius	2
124	First Musician	2

125 rows × 2 columns

Figura 4: CharName repetidos

Respecto a la tabla de capítulos (df_chapters) se corroboró que tuviera la misma cantidad de obras que la tabla de obras (df_works) y que no hubiera valores nulos en ninguna columna (datos no mostrados, ver notebook adjunto).

Finalmente, la tabla de obras no posee datos nulos ni nombres de obras duplicadas. Adicionalmente se corroboró que la columna ‘GenreType’ no tiene ningún error de escritura (datos no mostrados, ver notebook adjunto).

Personaje con más párrafos

Un requerimiento específico de la tarea es detectar el personaje con más párrafos. Tomando en cuenta el análisis previo se establecen los siguientes criterios:

- Personajes con el mismo nombre y distinto ID se identifican como personajes distintos.
- Los nombres que identifican a grupos o roles se toman como personajes individuales.
- No se adicionan los párrafos si un personaje participa de forma grupal.

Debido a que la tabla characters es utilizada para múltiples propósitos (además de proporcionar el nombre del personaje), no es posible asegurar que el join de paragraphs y characerts agrupados por charName y character_id nos de la cantidad de párrafos por personaje.

Como se observa en la figura 5, el “personaje” con la mayor cantidad de párrafos claramente no corresponde a un personaje.

Exploración de Datos

```
merged_df = pd.merge(df_paragraphs, df_characters, left_on='character_id', right_on='id')
merged_df['CharName'] = merged_df['CharName'].str.upper()
#cantidad de parrafos por personaje
merged_df.groupby(["character_id","CharName"])["ParagraphNum"].count().sort_values(ascending=False).reset_index(name
```

	character_id	CharName	Count
0	1261	(STAGE DIRECTIONS)	3751
1	894	POET	733
2	393	FALSTAFF	471
3	573	HENRY V	377
4	559	HAMLET	358
5	531	DUKE OF GLOUCESTER	285
6	844	OTHELLO	274
7	600	IAGO	272
8	120	ANTONY	253
9	945	RICHARD III	246

Figura 5: Cantidad de párrafo agrupados por nombre de personaje y ID

El siguiente “personaje” con mayor cantidad de párrafos es Poet. Para verificar que efectivamente sea un personaje se filtra la tabla characters por el charactr_id obtenidos en el top 10 de cantidad de párrafos. Dada la descripción del Poet podemos afirmar que tampoco es un personaje(Ver figura 6).

```
merged_df = pd.merge(df_paragraphs, df_characters, left_on='character_id', right_on='id')
merged_df['CharName'] = merged_df['CharName'].str.upper()
#cantidad de parrafos por personaje
character_id=merged_df.groupby(["character_id","CharName"])["ParagraphNum"].count().sort_values(ascending=False).res
character_id=character_id['character_id']
df1_filtrado = df_characters[df_characters['id'].isin(character_id)]
df1_filtrado
```

	id	CharName	Abbrev	Description
119	120	Antony	ANTONY	(Marcus Antonius)
392	393	Falstaff	FALSTAFF	Sir John Falstaff
530	531	Duke of Gloucester	GLOUCESTER	brother to the King
558	559	Hamlet	Ham	son of the former king and nephew to the prese...
572	573	Henry V	HENRY5	Prince, King of England
599	600	Iago	IAGO	Othello's ancient (?)
843	844	Othello	OTHELLO	A noble Moor in the service of the Ventian state
893	894	Poet	Poet	the voice of Shakespeare's poetry
944	945	Richard III	RICHARD3	son of Richard Plantagenet, duke of York; was ...
1260	1261	(stage directions)	xxx	NaN

Figura 6: personajes que aparecen en el top de agrupación por párrafo

Dada la agrupación por párrafo más la descripción de charaters podemos decir que el personaje con más párrafos es Sir John Falstaff.

Gráfica obras de Shakespeare

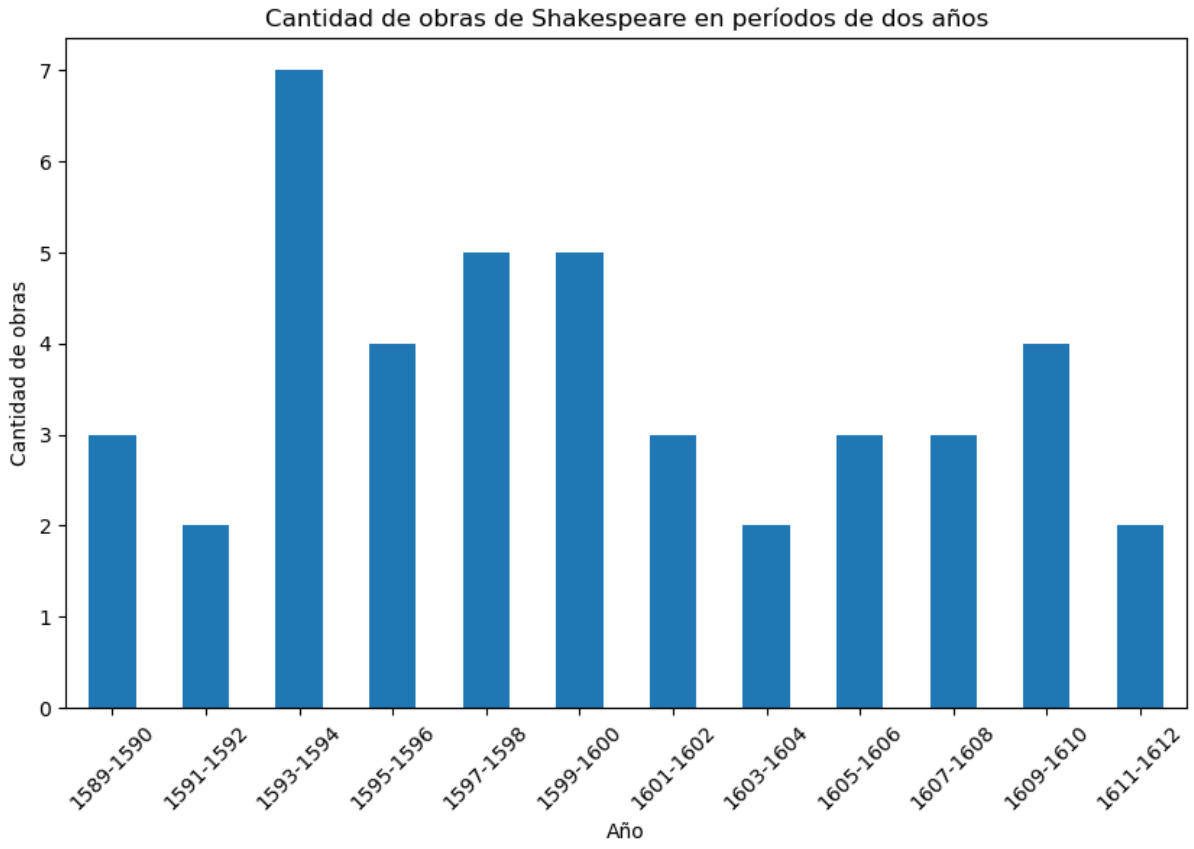

Figura 7: Cantidad de obras de Shakespeare cada 2 años.

Para visualizar la producción de obras de Shakespeare, se decidió hacer una gráfica de barras en donde muestra la cantidad de obras realizadas en períodos de dos años (ver Figura 7). Las 43 obras del autor fueron escritas en un período de 23 años, con un promedio de aproximadamente dos obras por año. Escribió el 60% de su obra en los primeros 12 años de su carrera, con un pico en su producción entre 1593 y 1594.

Si se observa la naturaleza de sus obras (Figura 8) hay una clara tendencia a producir comedias, tragedias y obras históricas, con una menor proporción de poemas y sonetos.

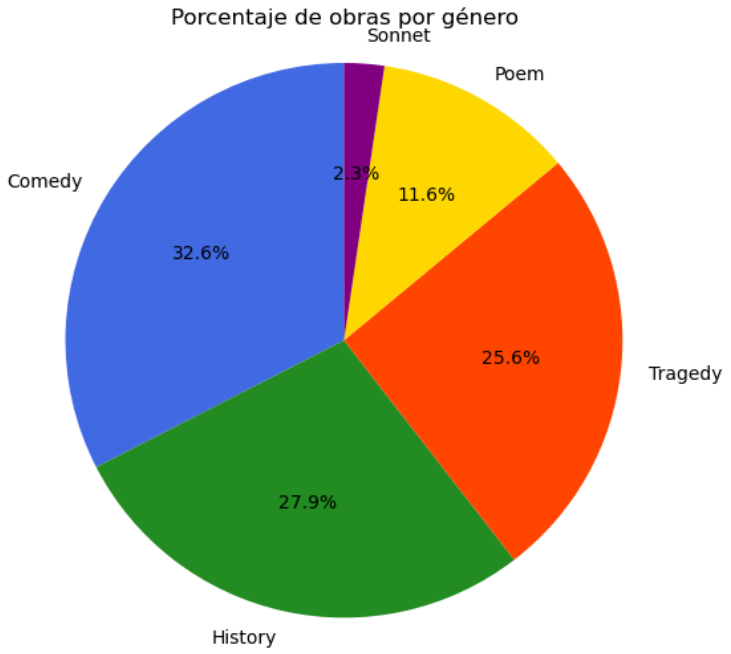

Figura 8: Porcentaje de obras según el género.

Al observar el género de su producción, resulta interesante identificar tendencias en sus obras en el correr de los años. Esto se puede observar en la figura 9, en donde resulta claro que la producción de comedias fue constante a lo largo de su carrera, mientras que las obras históricas se concentran en los primeros y últimos años. En los años que disminuye su publicación de estas obras se ve una concentración en la producción de tragedias (entre 1599 y 1608). Respecto a los poemas, no parece haber una tendencia clara, mientras que sólo escribió un soneto en toda su carrera.

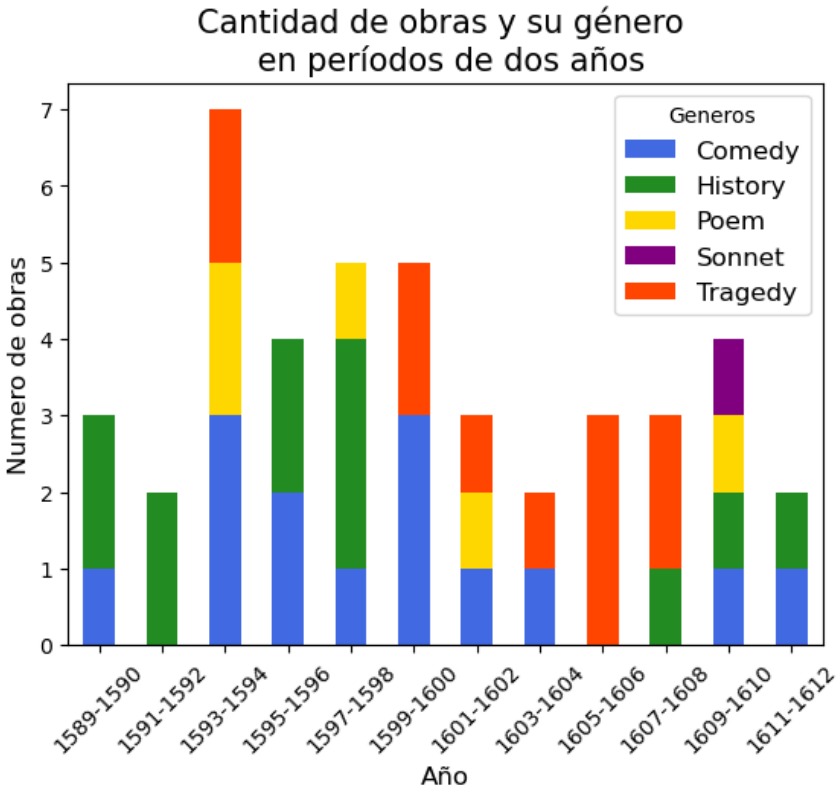

Figura 9: Obras agrupadas cada 2 años y género

Parte 2: Conteo de palabras y visualizaciones

Conteo de palabra por obra

Luego de realizar la revisión de palabras más utilizadas en las obras de Shakespeare no se observa ninguna que pueda aportar información adicional o relevante respecto a su obra (ver figura 10). Por tal motivo se decide quitar de la búsqueda las palabras denominadas stop word (ver figura 11).

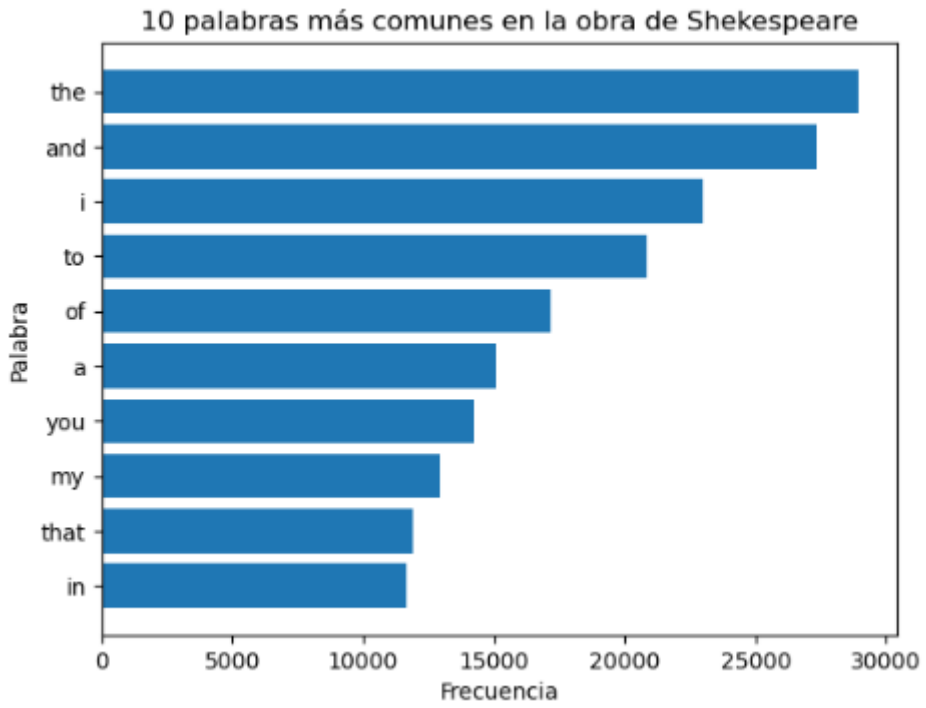

Figura 10: top 10 de palabras más usadas

La aparición de títulos nobiliarios en las palabras más usadas nos hace suponer que se está utilizando un lenguaje formal y un contexto histórico pero si no tuviéramos ningún tipo de conocimiento previo al análisis podríamos suponer que es una ambientación histórica, una narrativa de fantasía, un lenguaje formal, un lenguaje arcaico o simplemente un estilo más elaborado y cortés.

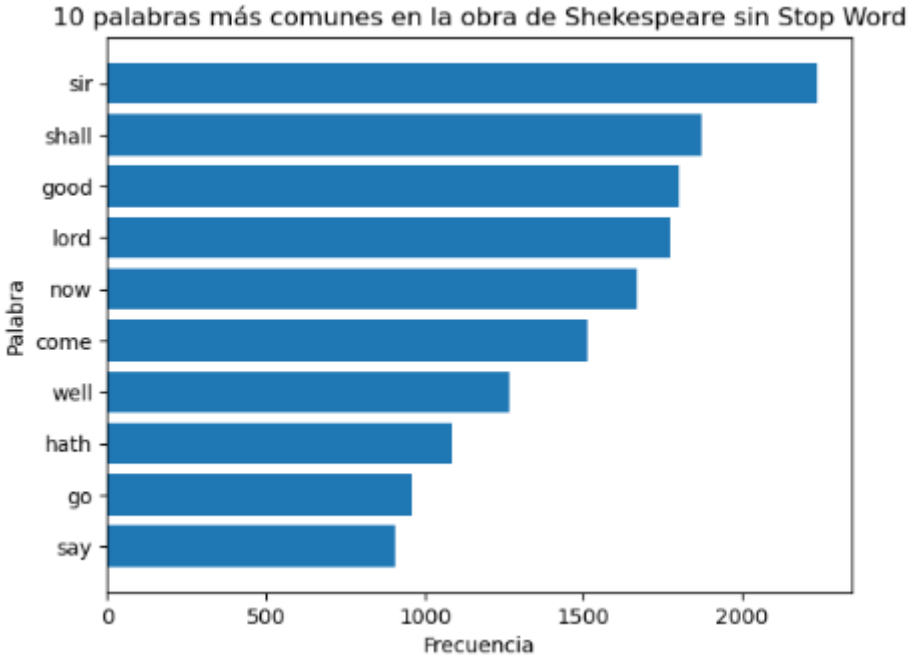

Figura 11: top 10 palabras sin stop word

Un siguiente estudio de palabras podría ser qué combinación de 2 palabras aparecen más seguido. Viéndolas aisladas nos hace suponer que good puede ser seguida por lord o sir.

Para obtener una idea del lenguaje utilizado por género de las obras, se podría realizar una visualización de barras horizontales como la de la figura 11, pero con una barra por cada género en cada palabra. De la misma forma, se podría mostrar la palabra más utilizada en las obras escritas en cada año, de forma de poder evidenciar un cambio en el estilo del lenguaje.

Respecto a la visualización de las palabras más usadas por personaje, entendemos que una gráfica de CloudWord por personaje nos podría llegar a dar una idea del rol que ocupan los 10 personajes con mayor cantidad de palabras.

Personajes con mayor cantidad de palabras

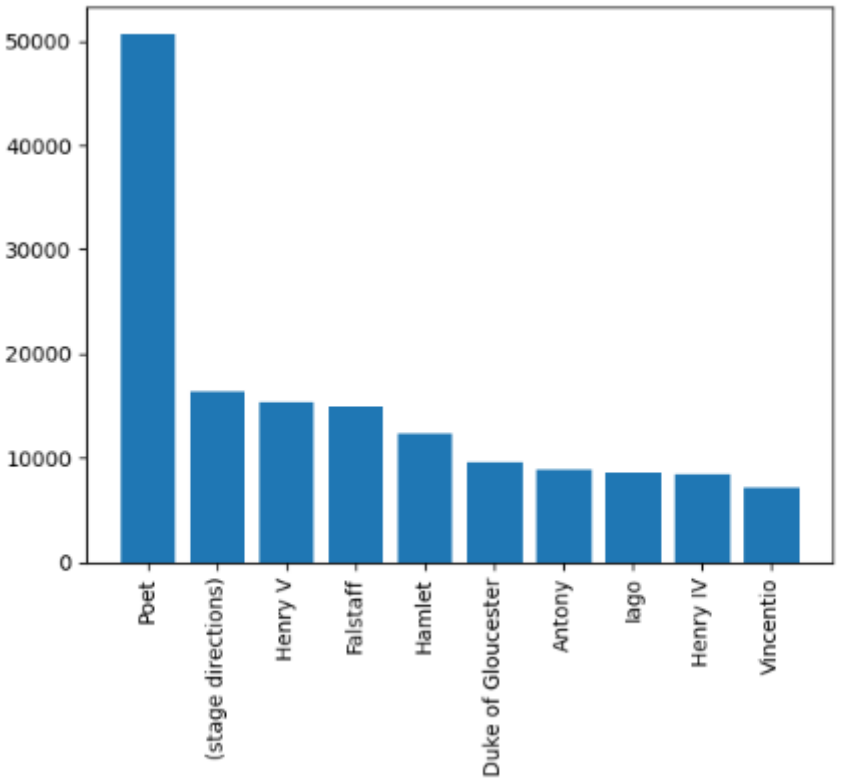

Figura 12: Cantidad de palabras por personaje

Se vuelve a dar el error derivado de la utilización polifuncional de la tabla characters. Se observa en el puesto 1 y 2 Poet y Stage Directions que no son personajes.

Una posible solución es filtrar todos los párrafos donde character_id correspondan al identificar de Poet o Stage direction. Esta solución eliminaría el texto de los poemas. Otra solución sería buscar una gramática en el texto de los poemas que nos permita obtener los personajes (si es que existen), agregarlos a la tabla characters y asignar el párrafo al personaje correspondiente.

Preguntas que se podrían intentar responder

Algunas preguntas que se podrían llegar a responder con los datos que tenemos son:

1) Personajes con mayor cantidad de interacciones entre sí:
   Tendríamos que buscar una gramática que nos permita identificar con qué personaje está interactuando, mapear el personaje de la tabla characters, recorrer la tabla paragraph y sumar 1 en el contador correspondiente al par de personajes.

2) Personajes que aparecen en más de una obra:
   En este caso excluimos a los poemas dado que no existe una relación directa entre párrafo y personaje. Se haría un join entre paragraph con chapter agrupados por work_id y character_id

3) Obras con mayor intervención del director:
   Se podría observar con una gráfica de barra del resultado de la agrupación entre párrafos y capítulos donde character_id=1261.

4) Tendencias en la directrices del director:
   Se haría un análisis exploratorio entre palabras más usadas, frases y/o conjuntos de palabras utilizadas por el personaje con id=1261.
   Para profundizar en este tema se podría agrupar por año o género o ambos.
   La visualización depende va a depender de lo complejo en identificar tendencias o de la tendencia detectada.

Anexos

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. En particular, analice la cantidad de párrafos por personaje. ¿Cuál es el personaje con más párrafos?

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 escritas 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.

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)