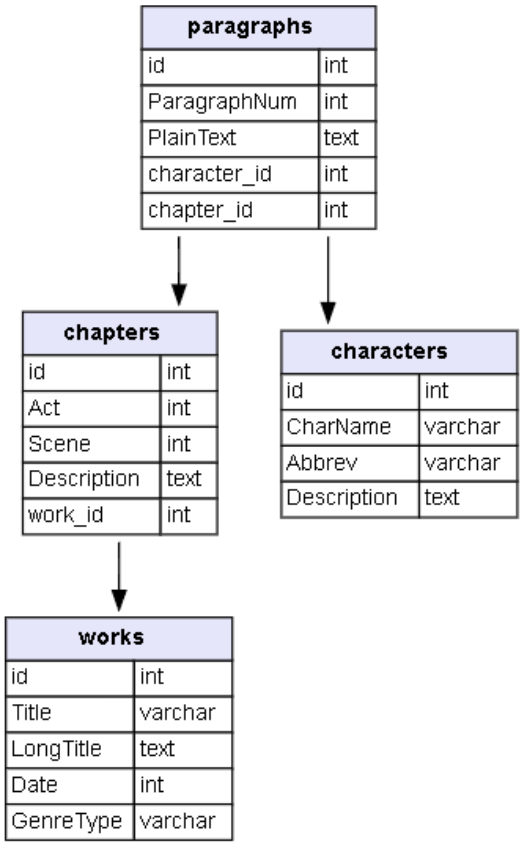
Introducción a la ciencia de datos

Tarea 1

Parte 1: Cargado y limpieza de datos.

a) 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?

La base de datos de Shakespeare contiene 4 tablas que diagraman de la siguiente manera:



Con respecto a las relaciones entre las tablas podemos indicar lo siguiente:

1. Cada obra (work - id) tiene 1 o N capítulos (chapter - id).
2. Cada capítulo (chapter - id) corresponde a 1 obra (work_id).
3. Cada capítulo (chapter - id) tiene 1 o N párrafos (paragraph - id).
4. Cada párrafo (paragraph - id) corresponde a 1 capítulo (chapter - id).
5. Cada párrafo (paragraph - id) corresponde a 1 personaje (character - id).
6. Cada personaje (character - id) tiene 0 o N párrafos (paragraph - id).

A continuación, se analizarán cada una de las entidades del diagrama.

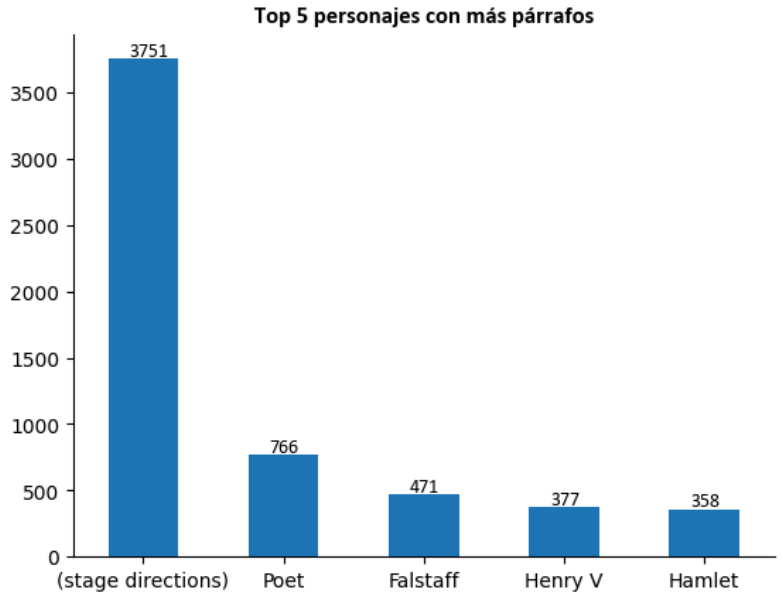
Paragraphs

Contiene los párrafos de todos los trabajos de Shakespeare (*PlainText*) junto con el personaje que lo interpreta (*character_id*) y el capítulo en el que sucede (*chapter_id*). Todos los párrafos están identificados por número (*ParagraphNum*) y por un identificador numérico único (*id*). Esta tabla se relaciona con las tablas **Chapters** y **Characters** mediante las claves foráneas *character_id* y *chapter_id*, claves primarias en estas tablas.

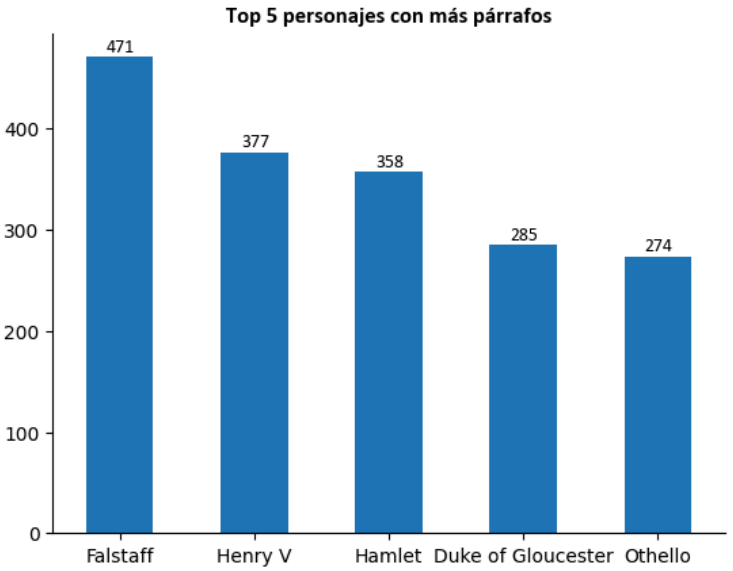
La tabla de párrafos contiene 35.465 registros entre los cuales no se encuentran valores nulos.

```
Index: 35465 entries, 0 to 35464
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   id            35465 non-null  int64 
 1   ParagraphNum  35465 non-null  int64 
 2   PlainText     35465 non-null  object
 3   character_id  35465 non-null  int64 
 4   chapter_id    35465 non-null  int64 
 5   CleanText     35465 non-null  object
dtypes: int64(4), object(2)
```

A partir de esta tabla podemos conocer los personajes con mayor cantidad de párrafos.



Observemos que tanto "*stage directions*" como "*Poet*" con son personajes como tales. Descartando estos sujetos como objeto de análisis, el personaje con mayor cantidad de párrafos es *Falstaff*, con 471 párrafos.



Characters

Contiene los nombres todos los personajes de las obras de Shakespeare (*CharName*) junto con su abreviación (*Abbv*) y una descripción de los mismos (*Description*), así como un identificador numérico único (*id*).

La tabla de personajes tiene 1266 registros. La columna de abreviaciones y descripción contienen 5 y 646 registros nulos.

```
Index: 1266 entries, 0 to 1265
Data columns (total 4 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           1266 non-null   int64
 1   CharName     1266 non-null   object
 2   Abbrev       1261 non-null   object
 3   Description  620 non-null    object
dtypes: int64(1), object(3)
```

Chapters

Contiene todos los capítulos de las obras de Shakespeare (*work_id*) por acto (*Act*) y escena (*Scene*) y una breve descripción donde se desarrolla el capítulo (*Description*). Cada uno identificado por un número único (*id*). Esta tabla se relaciona con la tabla **Works** mediante la clave *work_id*.

La tabla de capítulos contiene 945 registros sin valores nulos.

```
Index: 945 entries, 0 to 944
Data columns (total 5 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   id           945 non-null    int64
 1   Act          945 non-null    int64
 2   Scene        945 non-null    int64
 3   Description  945 non-null    object
 4   work_id      945 non-null    int64
dtypes: int64(4), object(1)
```

Durante el análisis de esta tabla se encontraron algunos datos atípicos en la columna *Description*. En primer lugar, las descripciones "---" y "---\n" no satisfacen las reglas de dominio de esta columna, es decir, no representan ningún lugar geográfico donde se desarrollan el acto y la escena. En particular, estas descripciones corresponden a las siguientes obras:

	id	Title	LongTitle	Date	GenreType
27	28	Passionate Pilgrim	The Passionate Pilgrim	1598	Poem
34	35	Sonnets	Sonnets	1609	Sonnet

Es esperable que estas obras no tengan una ubicación asociada ya que no se desarrollan en ningún lugar geográfico particular, pero quizás, un valor nulo podría ser un valor más adecuado para poder ser capturado con mayor facilidad.

En segundo lugar, existen 12 obras con escenas que se describen como "The same", haciendo referencia a la escena que se desarrolla anteriormente. El problema aquí es que estas escenas pueden no ser consecutivas, por lo que un mismo valor "The same" puede hacer referencia a un lugar completamente diferente. Veamos un ejemplo:

En la escena 2 del acto 3 de la obra 5 (*Comedy of Errors),* el valor "The same" hace referencia al sitio donde ocurre la escena 1 del mismo acto, es decir, "*Before the house of ANTHIPHOLUS of Ephesus*".

	id	Act	Scene	Description	work_id
109	18813	3	1	Before the house of ANTIPHOLUS of Ephesus.	5
110	18814	3	2	The same.	5

Por otro lado, los valores "The same" en el acto 4 del a obra 41 (*Two Gentlemen of Verona*), hacen referencia a Milan.

	id	Act	Scene	Description	work_id
920	19624	4	1	The frontiers of Mantua. A forest.	41
921	19625	4	2	Milan. Outside the DUKE's palace, under SILVIA...	41
922	19626	4	3	The same.	41
923	19627	4	4	The same.	41

Esto podría inducir a valores erróneos, por ejemplo, si quisiéramos saber que ubicación es la más frecuente entre todas las obras. En conclusión, tenemos un mismo valor "The same" que corresponde a valores geográficos diferentes, lo que evidencia un claro problema de calidad de datos.

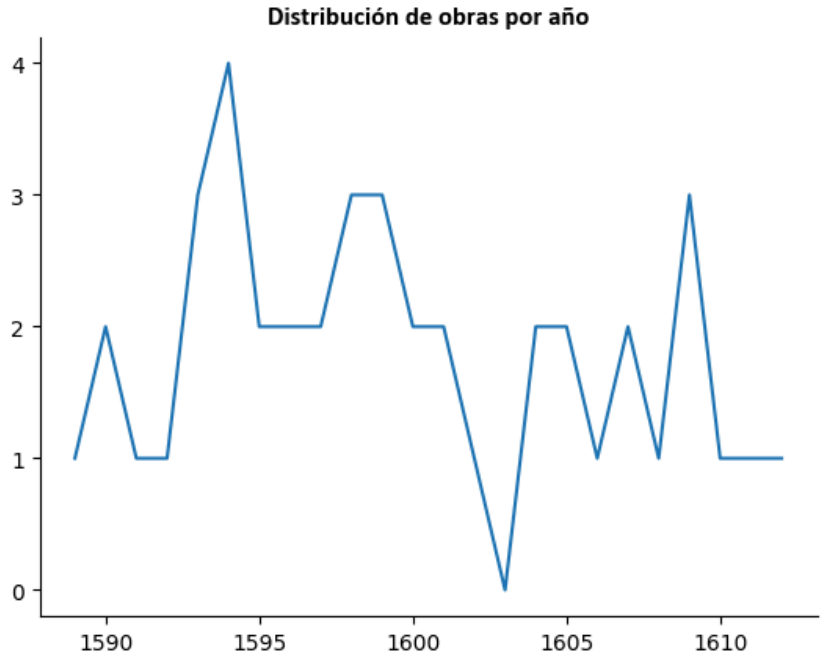
Works

Contiene el nombre de todas las obras de Shakespeare (*Title*), su nombre completo (*LongTitle*), la fecha en que fue escrita (*Date*) y su género (*GenreType*), así como un identificador único (*id*).
La tabla de obras contiene 43 registros de los cuales ninguno posee valores nulos.

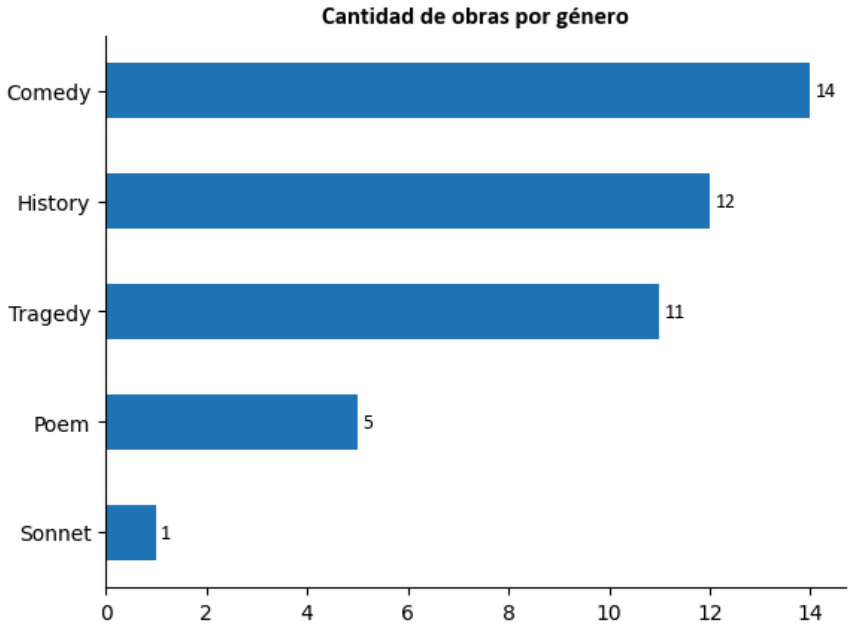
```
Index: 43 entries, 0 to 42
Data columns (total 5 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   id         43 non-null     int64
 1   Title      43 non-null     object
 2   LongTitle  43 non-null     object
 3   Date       43 non-null     int64
 4   GenreType  43 non-null     object
dtypes: int64(2), object(3)
```

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.

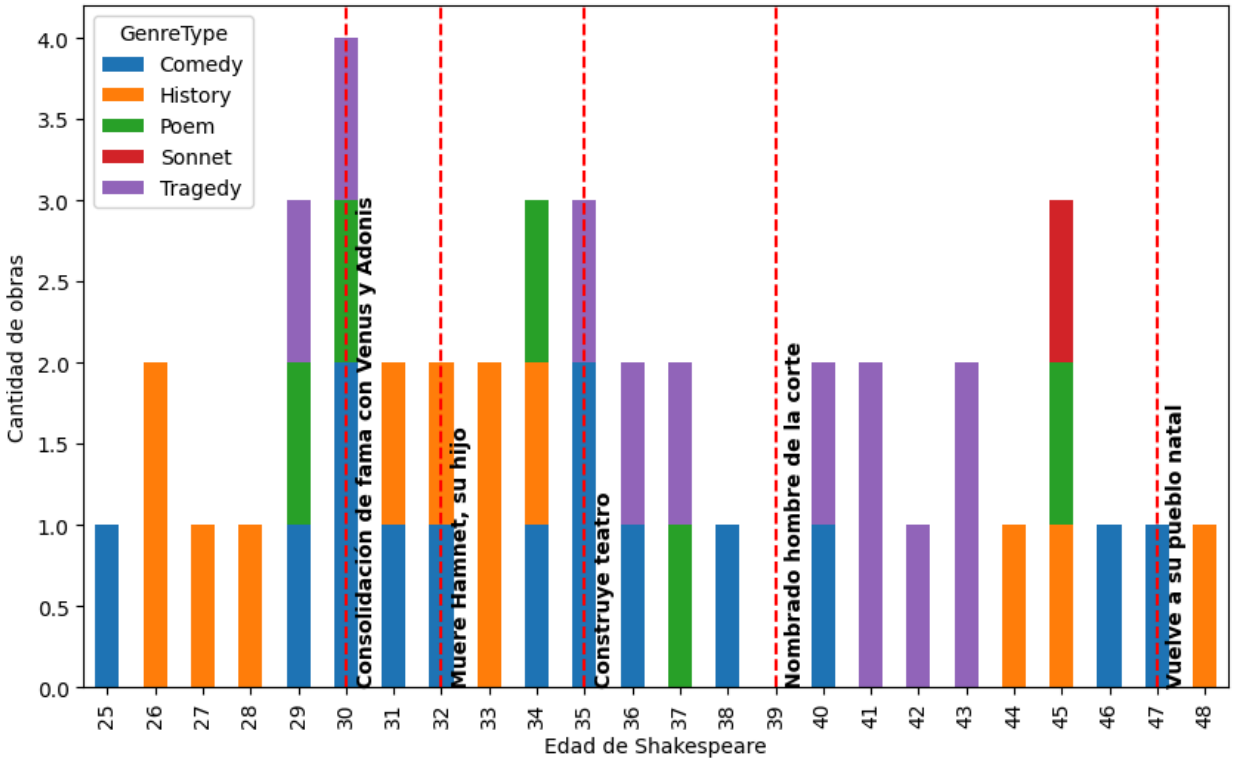
A lo largo de los años, Shakespeare publicó al menos 1 obra por año, a excepción del año 1603 en el que no publicó ninguna (su obra Otelo fue escrita entre 1603 y 1604). En promedio, Shakespeare ha escrito, entre 1589 y 1612, 1.79 obras por año.



Respecto a los géneros escritos por Shakespeare, podemos decir que la mayoría de sus obras se basan en comedias, seguidas por una importante cantidad de historias y tragedias.



Tomando en cuenta que William nació en 1564, podemos apreciar que publicó su primera obra a los 25 años y luego lo hizo de forma consistente todos los años siguientes, publicando al menos 1 obra al año (menos en 1603). Su pico de productividad lo tuvo a sus 30 donde publicó 4 obras en un mismo año calendario.
Con respecto a sus géneros más populares, podemos ver que durante su vida lo que más escribió fue Comedia, seguido de Historia y Tragedia.



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.

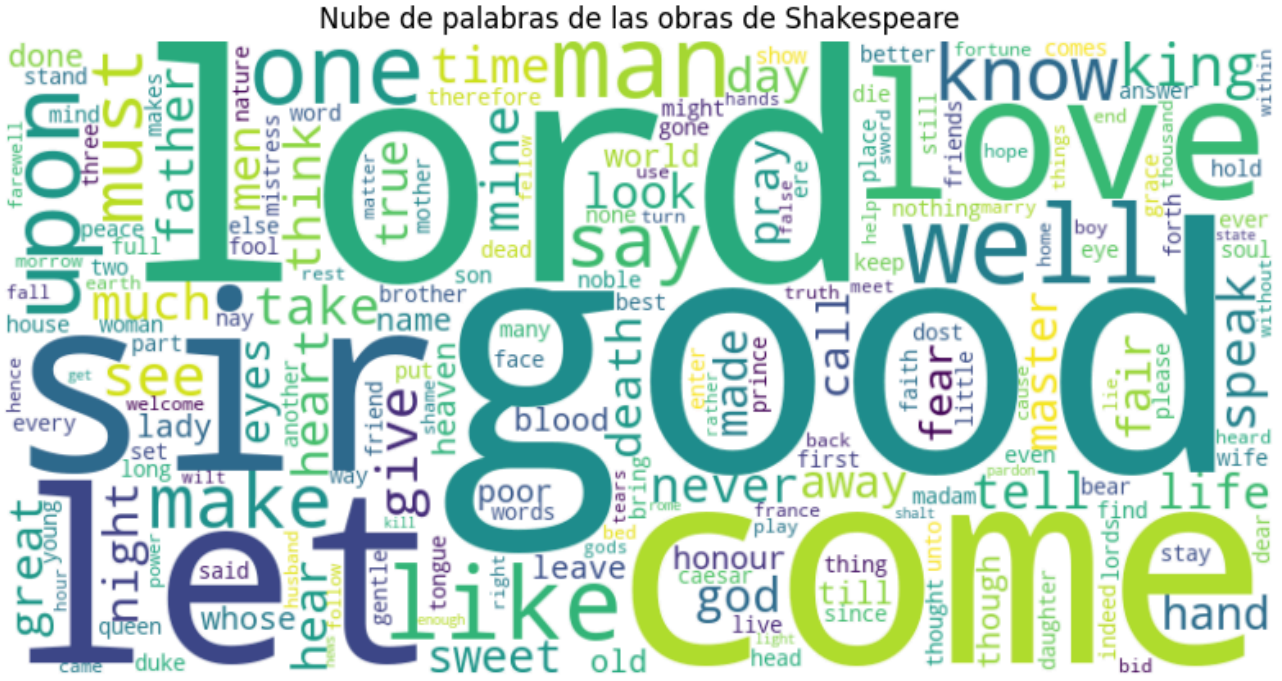
Antes de realizar el conteo de las palabras de las obras de shakespeare se realizaron las siguientes reglas de limpieza:

1. Eliminación de expresiones teatrales que se encontraban en el texto, como por ejemplo [Exeunt], buscando todo texto encerrado entre corchetes y eliminandolo.
2. Eliminación de los caracteres especiales de saltos de línea, tabulaciones, etc que puede contener el texto en la base de datos.
3. Eliminación de los signos de puntuación usando la constante de python string.punctuation.
4. Eliminación de los dígitos que se encuentran en el texto.
5. Eliminación de las stopwords o palabras vacías, que son palabras que se consideran de bajo valor semántico en un texto. Tuvimos que agregar stopwords tanto en inglés moderno como en Middle English ya que los textos tienen palabras que hoy no son consideradas vacías ya que se dejaron de usar. Por ejemplo, thou.
6. Eliminación de todas las palabras de longitud menor o igual a 2.

Con las transformaciones anteriores, lo que buscamos es dejar el texto lo más limpio posible de palabras con valor semántico y significado en las obras de Shakespeare, y así lograr un análisis que se centre principalmente en las palabras más relevantes de su obra.

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.

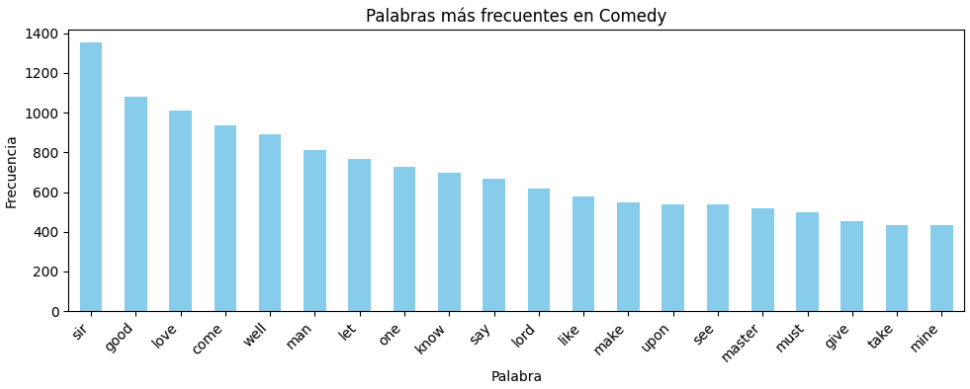


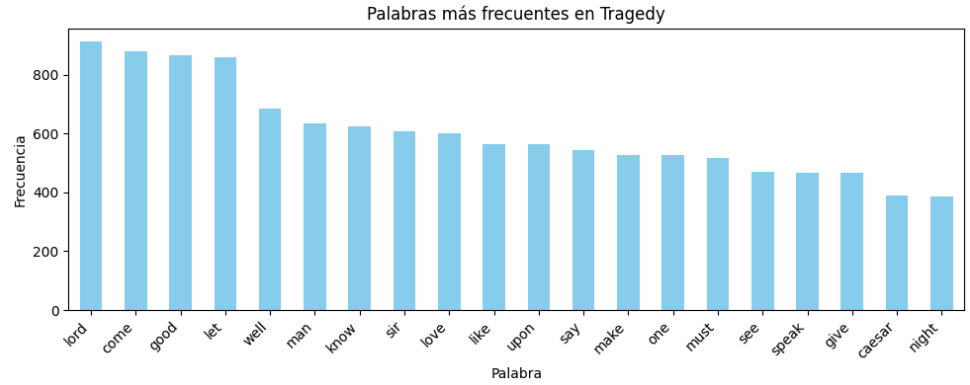
Sin necesidad de implementarlo, proponga ideas para modificar esta visualización con el fin de encontrar diferencias entre géneros y personajes.

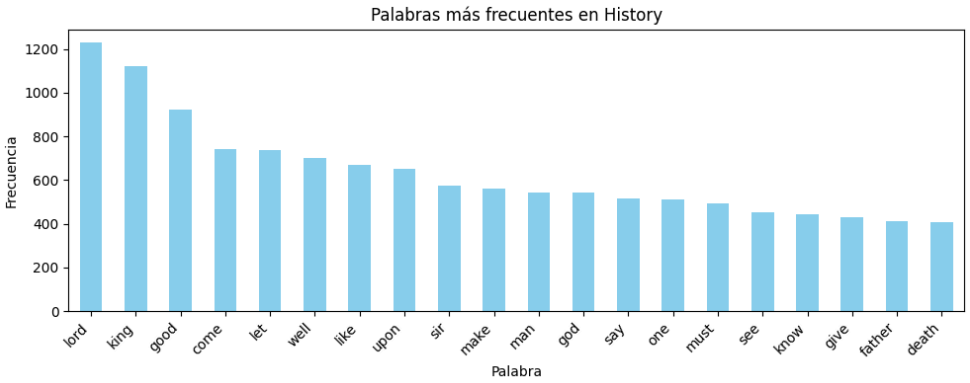
Lo que me parecería interesante para destacar géneros o personajes en la visualización anterior sería:

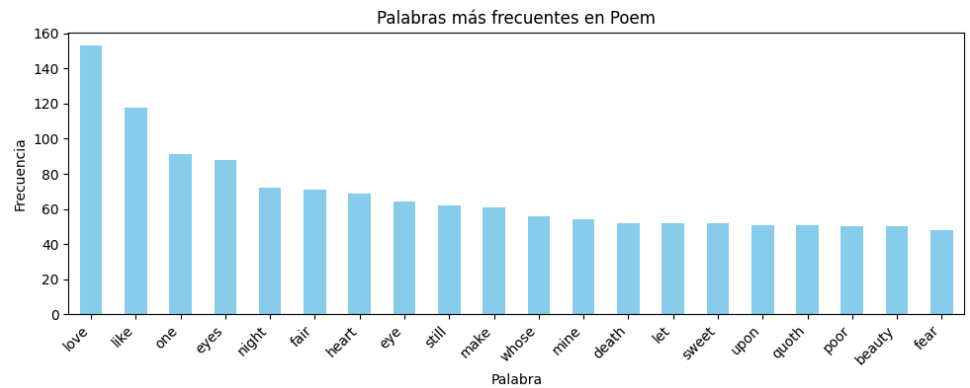
- Para diferenciar géneros, usar diferentes colores por género, entonces si bien la estructura de la gráfica se mantendrá, a partir de los colores fácilmente podría identificar si las palabras más repetidas son dichas por hombres, mujeres o cualquier otro género.
- La misma técnica de usar colores o tipografías , podría aplicarlo a personajes aunque en este caso la cardinalidad de las variantes aumentan y no sería tan claro de analizar a simple vista.

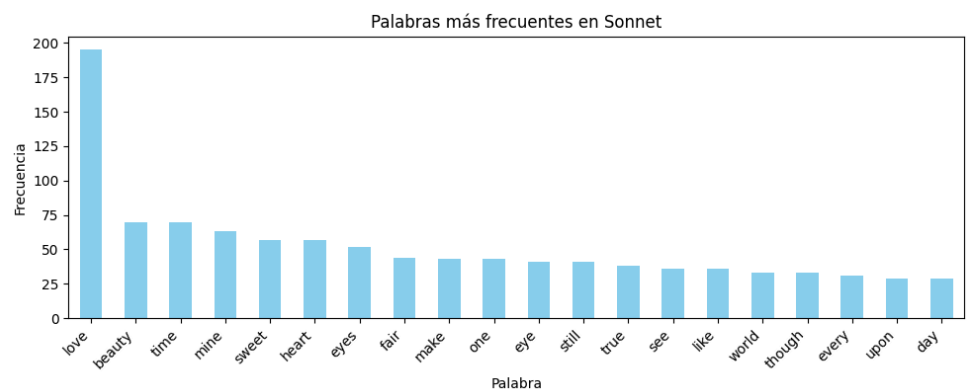
Nos pareció interesante armar un gráfico de frecuencia de palabras discriminado por género ya que nos permite ver la importancia que tiene la palabra LOVE en los géneros Poemas y Sonetos, y como en los otros géneros toma preponderancias palabras más formales como Lord, Sir, etc.


Palabras más frecuentes en Comedy
Frecuencia
Palabra
sir good love come well man let one know say lord like make upon see master must give take mine


Palabras más frecuentes en Tragedy
Frecuencia
Palabra
lord come good let well man know sir love like upon say make one must see speak give caesar night

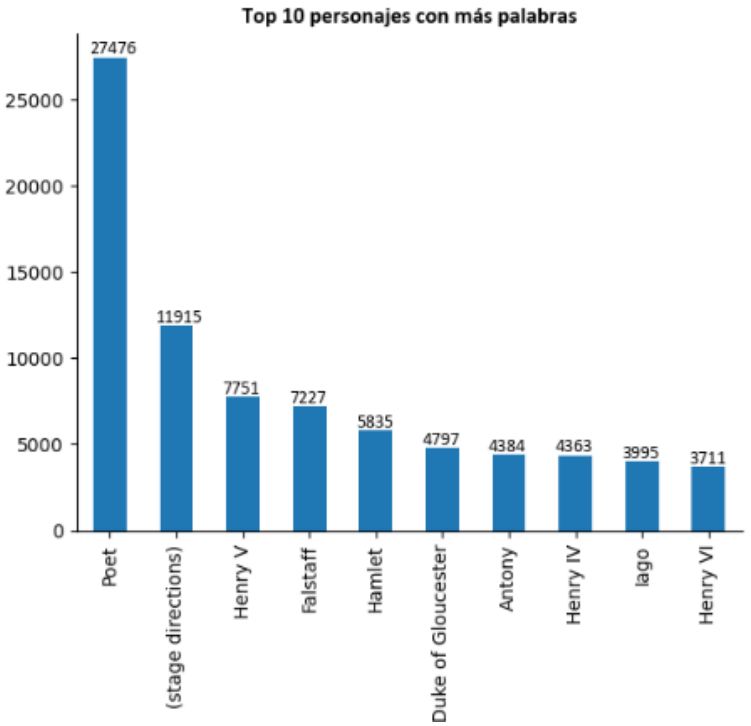

Palabras más frecuentes en History
Frecuencia
Palabra
lord king good come let well like upon sir make man god say one must see know give father death


Palabras más frecuentes en Poem
Frecuencia
Palabra
love like one eyes night fair heart eye still make whose mine death let sweet upon quoth poor beauty fear

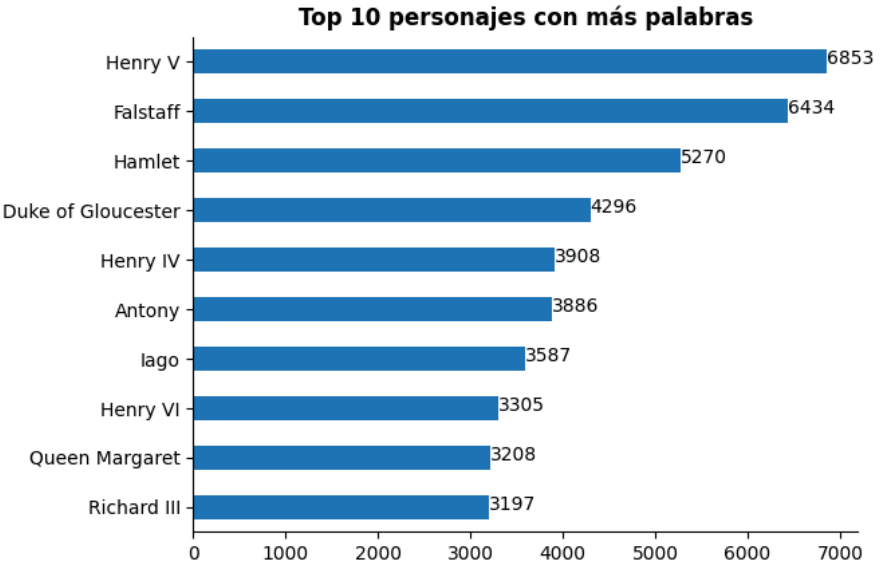

Palabras más frecuentes en Sonnet
Frecuencia
Palabra
love beauty time mine sweet heart eyes fair make one eye still true see like world though every upon day

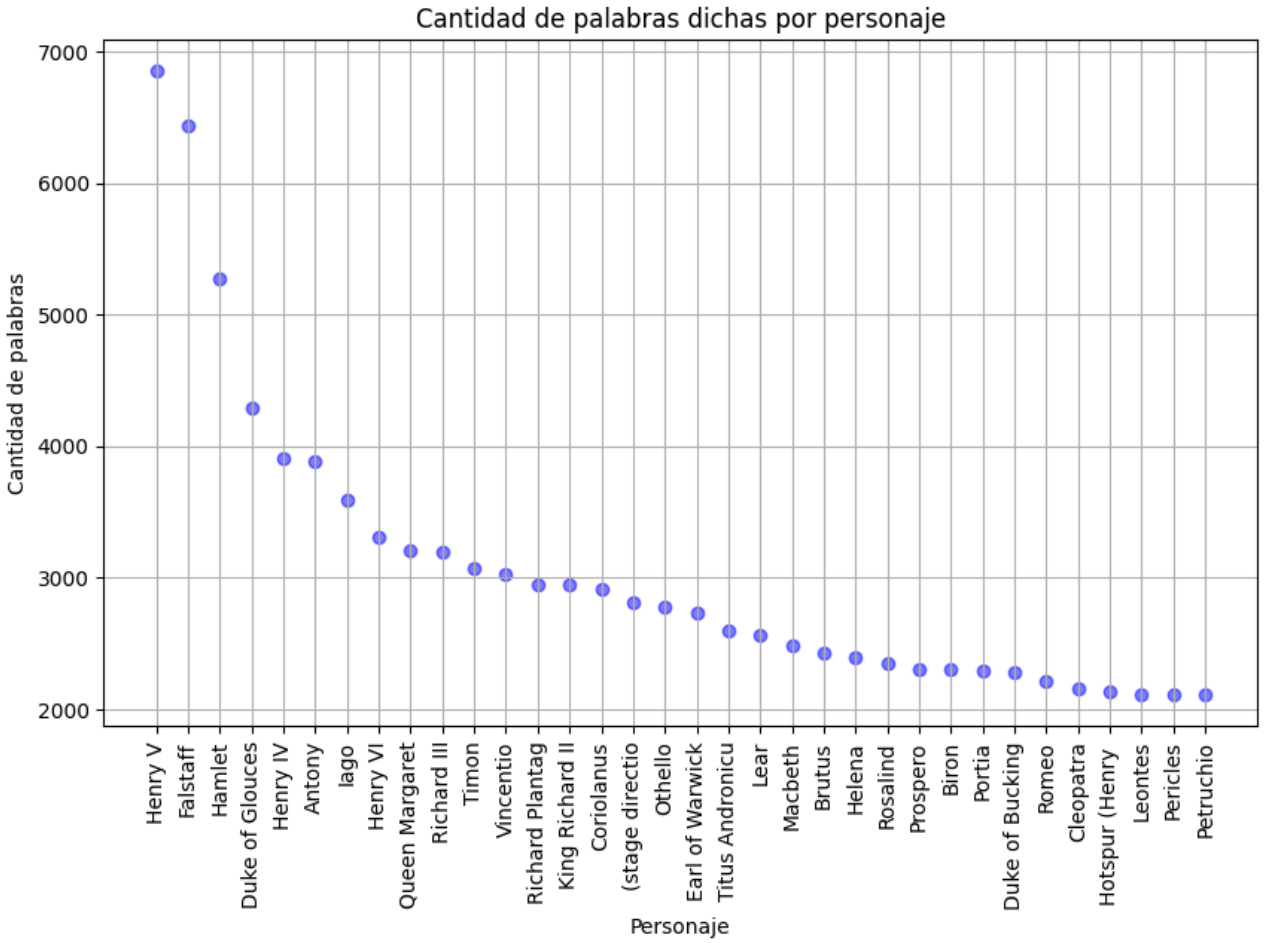
b) Corra el código que permite encontrar los personajes con mayor cantidad de palabras. En caso de encontrar algún problema luego de realizar la visualización, comente a qué se debe y proponga formas de resolverlo.

Al realizar el gráfico de cantidad de palabras por personaje nos encontramos en la misma situación que al momento de realizar el gráfico de personajes con más párrafos: aparece “stage directions” y “poet” como personajes, aún cuando éstos no lo son.



Siguiendo el mismo razonamiento que hicimos con el gráfico de párrafos por personajes, descartamos estos dos “personajes” para obtener la cantidad de palabras por personajes “reales”. Así, obtenemos lo siguiente:





c) Proponga preguntas que se podrían intentar responder a partir de estos datos, y mencione posibles caminos para responderlas (sin implementar nada).

Algunas preguntas que podrían surgir son:

i. ¿En qué género/obra estos personajes tienen mayor participación?
ii. ¿Cuáles son las palabras más usadas por estos personajes?
iii. ¿Se puede predecir la personalidad de los personajes a partir de las palabras que usan?
iv. ¿Cómo se vinculan los personajes los unos con los otros?
v. ¿Qué tipo de relación tienen los personajes los unos con los otros?
vi. ¿Cómo se relaciona la cantidad de palabras por personaje con su rol en las obras?

Sin duda alguna las respuestas a estas preguntas yacen en un análisis más profundo sobre los datos del que ya hemos realizado. El procesamiento del lenguaje natural podría sernos de ayuda a la hora de entender un poco más sobre las palabras de cada uno de estos personajes, así como el contexto en el que se desarrollan, la intención de sus palabras, y quizás también, por qué no, el motivo de las mismas, ¿estás palabras demuestran enojo, alegría, tristeza? ¿aparecen como respuesta alguna pregunta? ¿en qué lugar se están diciendo estas palabras? ¿hacia quién van dirigidas? Un trabajo interesante de realizar sería poder separar los personajes por obra para así entender cuáles se relacionan entre sí y evitar análisis de personajes que nada tienen que ver los unos con los otros. Quizás el género de las obras pueda darnos mayor información sobre estos personajes y la naturaleza de sus palabras.