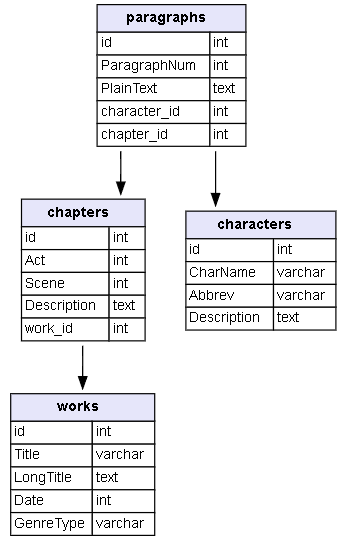
Introducción a la ciencia de datos

*Tarea 1*

**Parte 1: Cargado y limpieza de datos.**

1. 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:

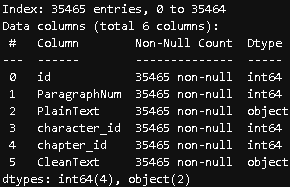


A continuación, se analizarán cada una de las entidades del diagrama.

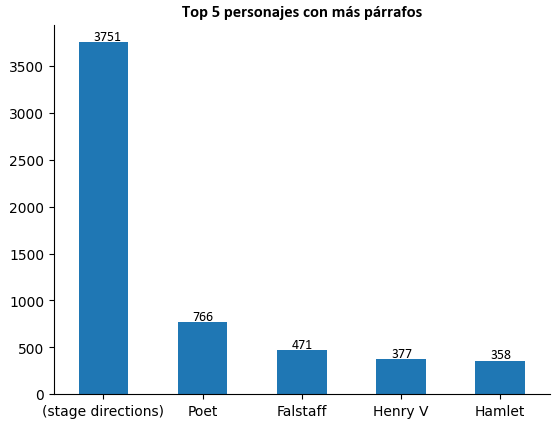
**Paragraphs**

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.

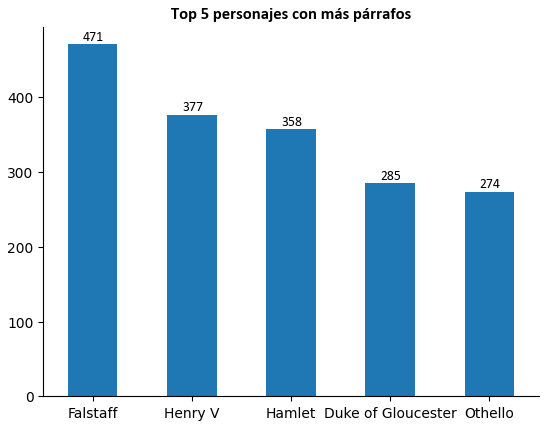
La tabla de párrafos contiene 35.465 registros entre los cuales no se encuentran valores nulos.



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



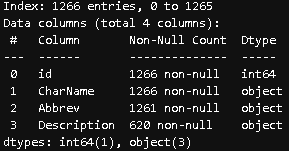
Observemos que tanto “*stage directions*” como “*Poet*” con son personajes como tales. Descartando estos sujetos como objeto de análisis, el personaje con mayor cantidad de párrafos es *Falstaff*, con 471 párrafos.



**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*).

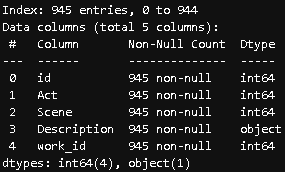
La tabla de personajes tiene 1266 registros. La columna de abreviaciones y descripción contienen 5 y 646 registros nulos.



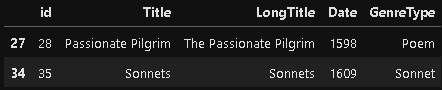
**Chapters**

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.



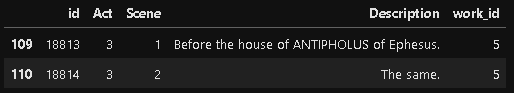
Durante el análisis de esta tabla se encontraron algunos datos atípicos en la columna *Description*. En primer lugar, las descripciones “---” y “---\n” no satisfacen las reglas de dominio de esta columna, es decir, no representan ningún lugar geográfico donde se desarrollan el acto y la escena. En particular,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



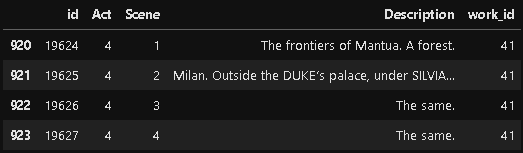
Es esperable que estas obras no tengan una ubicación asociada ya que no se desarrollan en ningún lugar geográfico particular, pero quizás, un valor nulo podría ser un valor más adecuado para poder ser capturado con mayor facilidad.

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*”.



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

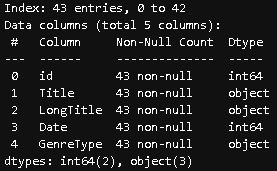


Esto podría inducir a valores erróneos, por ejemplo, si quisiéramos saber que ubicación es la más frecuente entre todas las obras. En conclusión, tenemos un mismo valor “The same” que corresponde a valores geográficos diferentes, lo que evidencia un claro problema de calidad de datos.

**Works**

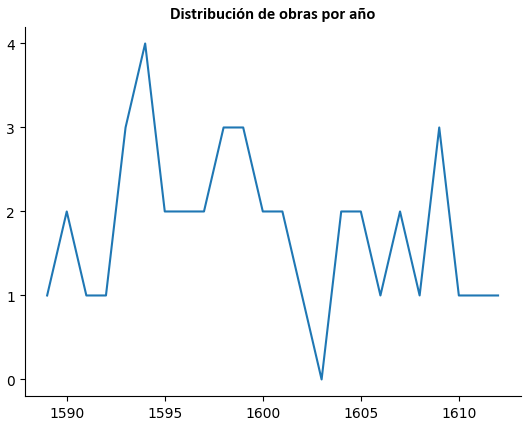
Contiene el nombre de todas las obras de Shakespeare (*Title*), su nombre completo (*LongTitle*), la fecha en que fue escrita (*Date*) y su género (*GenreType*), así como un identificador único (*id*).

La tabla de obras contiene 43 registros de los cuales ninguno posee valores nulos.

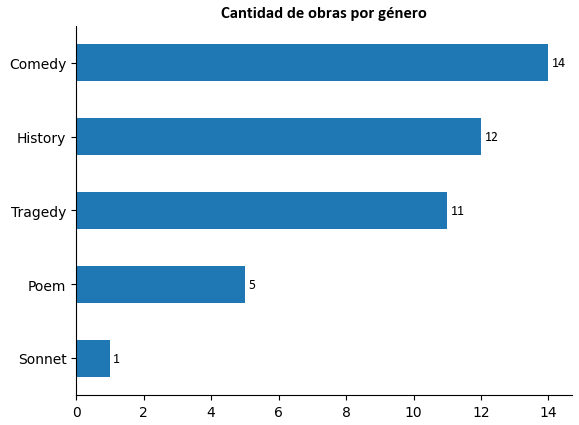


1. Genere una gráfica que permita visualizar la obra de Shakespeare a lo largo de los años. Por ejemplo, tomando períodos de algunos años y mostrando la cantidad de obras escritas para esos períodos. Comente si se observan tendencias (o no) a lo largo del tiempo, por ejemplo, respecto a su producción, o los géneros sobre los que escribió. No realizar análisis estadísticos, solamente generar visualizaciones exploratorias.

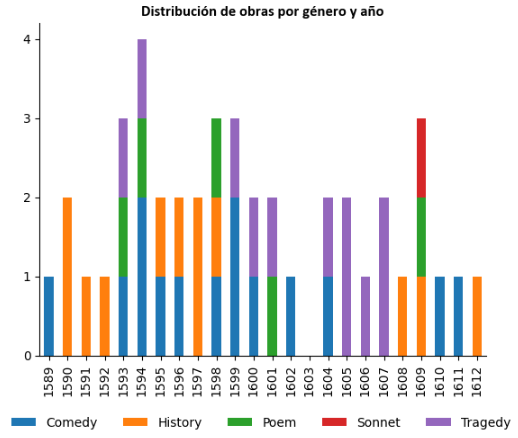
A lo largo de los años, Shakespeare publicó al menos 1 obra por año, a excepción del año 1603 en el que no publicó ninguna (su obra Otelo fue escrita entre 1603 y 1604). En promedio, Shakespeare ha escrito, entre 1589 y 1612, 1.79 obras por año.



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.



Podemos ver la tendencia de géneros escritos por Shakespeare en el siguiente gráfico. En sus comienzos, se dedicó casi exclusivamente a escribir historias, mientras que, hacia finales de su carrera como escritor, la tragedia fue el género tomó mayor relevancia.

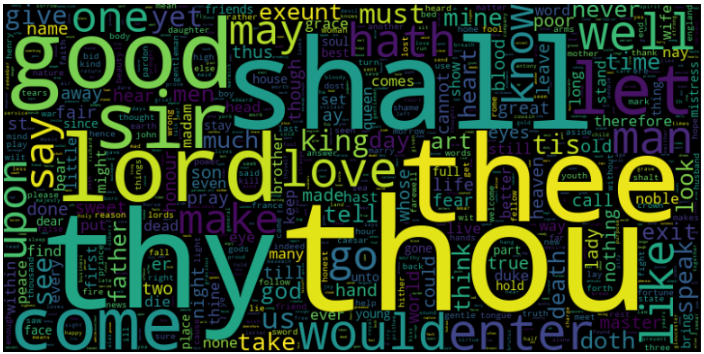


1. Una de las funciones básicas que se desea realizar, es el conteo de palabras: cuántas veces aparece cada palabra agrupando por distintos criterios. Para ello, primero es necesario normalizar el texto (i.e: pasarlo todo a minúsculas) y eliminar los signos de puntuación. De no hacerlo, las secuencias "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.

Para la limpieza de texto se utilizó la función “*string.punctuation*”, junto con los saltos de página (*\n*) y tabulaciones (*\t*). Además, se descartaron los caracteres numéricos y palabras consideradas “*stop words*”, que son muy utilizadas en el idioma inglés y no aportan información relevante, por ejemplo “*the*”, “*and*”, “*is*”, “*are*”, etc. Una vez aplicada la función *clean\_text*, se filtraron aquellas palabras con longitud mayor a 1, descartando así palabras provenientes de abreviaciones del inglés, por ejemplo “*re*”, “*ve*”, “*em*”, que surgen de “*they’re*”, “*should’ve*” o “*catch’em*”. Finalmente, con aquellas palabras que aportan información relevante sobre las obras de Shakespeare, se realizó un gráfico de nube para mostrar cuáles son las 1000 palabras más frecuentes.

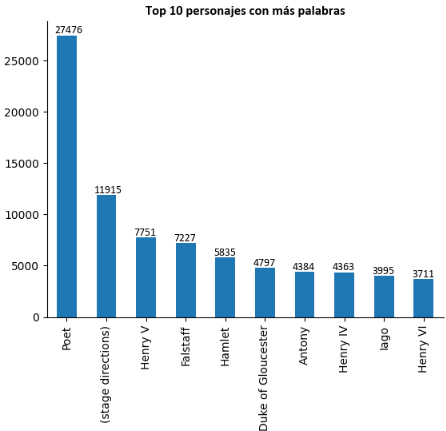
**Parte 2: Conteo de palabras y visualizaciones.**

1. 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.

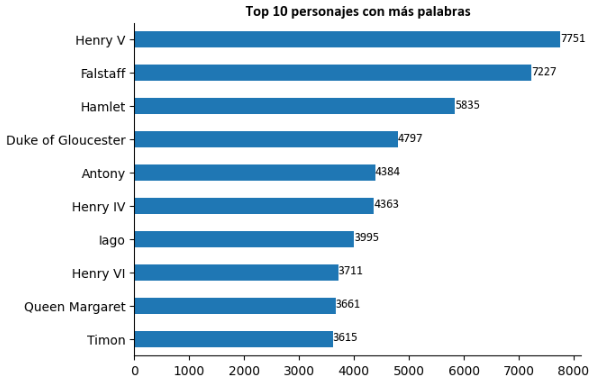


1. Corra el código que permite encontrar los personajes con mayor cantidad de palabras. En caso de encontrar algún problema luego de realizar la visualización, comente a qué se debe y proponga formas de resolverlo.

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:



1. 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:

1. ¿En qué género/obra estos personajes tienen mayor participación?
2. ¿Cuáles son las palabras más usadas por estos personajes?
3. ¿Se puede predecir la personalidad de los personajes a partir de las palabras que usan?
4. ¿Cómo se vinculan los personajes los unos con los otros?
5. ¿Qué tipo de relación tienen los personajes los unos con los otros?
6. ¿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.