21 de mayo de 2024

Introducción a la Ciencia de Datos

Tarea 1 - Grupo 5

Agustín Ghazarian
Sofía Gervaz

La Tarea consiste en el análisis de una base de datos que contiene información sobre la obra completa de William Shakespeare. La misma está conformada por las siguientes 4 tablas:

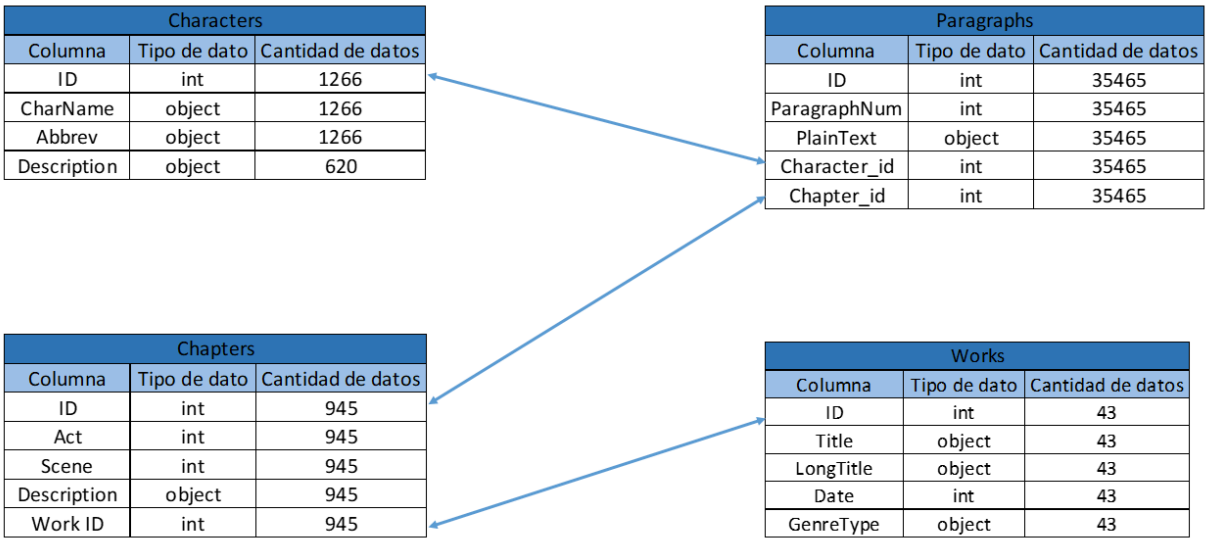
- *Characters*. Tiene los personajes de todas las obras de Shakespeare. Tiene un ID, el nombre y una abreviación para cada personaje. Además posee una columna dedicada a la descripción del mismo. Tiene 1266 entradas (filas), que en general se encuentran casi completas, salvo la columna descripción donde faltan 646 valores.

- *Chapters*. Tiene los actos y escenas de todas las obras de Shakespeare. Tiene un ID para cada acto-escena de cada obra. Otros dos campos numéricos indican por separado el número de acto y de escena. Además posee una descripción de cada escena que menciona la ubicación de la misma y un work_Id, que debe relacionarse con la tabla *Works* indicando de qué obra se trata. Tiene 945 entradas, y no falta ningún dato.

- *Works*. Tiene el nombre de las obras de Shakespeare, en versión corta y larga. Tiene un ID para cada obra que debe relacionarse con el que apareció en la tabla *Chapters*. Además especifica el año y el género para cada obra. Tiene 43 entradas, y no falta ningún dato.

- *Paragraphs*. Tiene información de todos los párrafos de las obras de Shakespeare, incluyendo todo el texto y tanto un ID como un número de párrafo para cada uno. Además, muestra un ID para cada personaje que se debe relacionar con los datos de *Characters* y para cada capítulo que se deben relacionar con los datos de *Chapters*. Tiene 35465 entradas, y no falta ningún dato.

La Figura 1 resume la información relevada:



Characters		
Columna	Tipo de dato	Cantidad de datos
ID	int	1266
CharName	object	1266
Abbrev	object	1266
Description	object	620

Paragraphs		
Columna	Tipo de dato	Cantidad de datos
ID	int	35465
ParagraphNum	int	35465
PlainText	object	35465
Character_id	int	35465
Chapter_id	int	35465

Chapters		
Columna	Tipo de dato	Cantidad de datos
ID	int	945
Act	int	945
Scene	int	945
Description	object	945
Work ID	int	945

Works		
Columna	Tipo de dato	Cantidad de datos
ID	int	43
Title	object	43
LongTitle	object	43
Date	int	43
GenreType	object	43

Figura 1 - Descripción de base de datos de la obra de Shakespeare.

Respecto a la calidad de los datos, se observa que las tablas están en general completas, salvo la columna *Description* en la tabla *Characters* en la que faltan más de la mitad de los datos. Los tipos de datos de cada columna son los adecuados, los ID de las tablas son

únicos y hay consistencia en las referencias entre ellas (por ejemplo: los *Work_ID* en *Chapters* existen en *Works* y así sucesivamente). No hay filas duplicadas en las tablas.

En la tabla *Characters*, aparece *(stage directions)* como un personaje (siendo además el personaje con más párrafos). Además, existen casos de personajes como *First Apparition*, *First citizen*, *First Gentleman*, etc., que pueden referir a personajes distintos dependiendo de la obra.

En cuanto al análisis de las obras, se observa que las mismas se ubican entre el año 1589 y 1612. El máximo de obras escritas en este periodo fue de 4 para el año 1594, existiendo varios años de 3 obras (1593, 1598, 1599 y 1609) y un solo año en el periodo donde no publicó ninguna obra, 1603. En la Figura 2 se presenta un histograma con la distribución de la cantidad de obras publicadas por año.

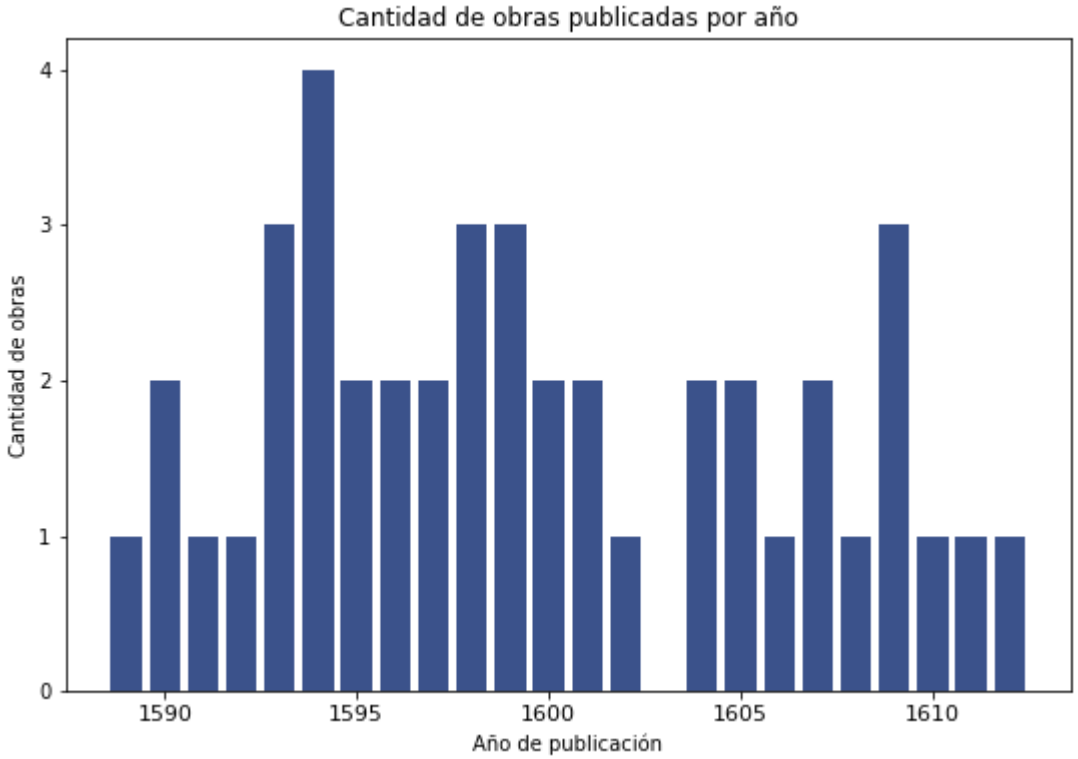


Figura 2 - Histograma de cantidad de obras de Shakespeare publicadas por año.

Al analizar la evolución por género (ver Figuras 3), se observa que las comedias fue lo que escribió de forma más constante a lo largo de toda su carrera (publicando 14 en total), mientras que sólo escribió un soneto. En la primera mitad de su carrera escribió tres veces más de novelas históricas que en la segunda mitad (9 y 3 respectivamente), mientras que las tragedias tienen un comportamiento opuesto.

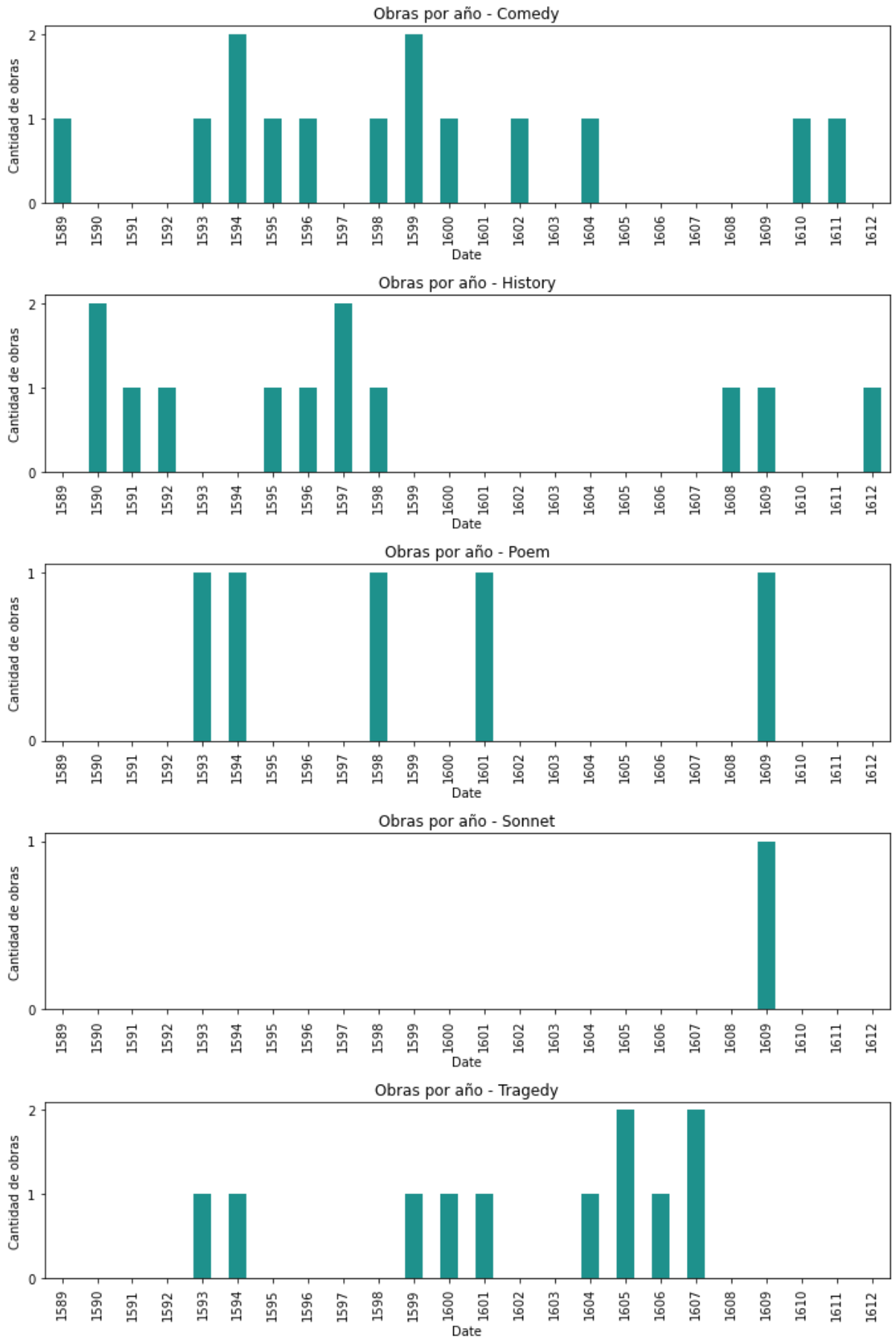


Figura 3 - Histograma de obras de Shakespeare por año para cada género.

Se realizaron además distintas visualizaciones que resumen los gráficos presentados previamente. La gráfica de la Figura 4 permite comparar rápidamente el total de obras en periodos de 4 años y da una idea de cómo se distribuyeron esas obras según el género, pero puede dificultar la comparación de la evolución temporal por género. Por ejemplo, se dificulta obtener de forma rápida la evolución de tragedias a lo largo del tiempo por encontrarse “apilado” encima del resto de los géneros. lo mismo sucede con todos los géneros, salvo la comedia que se encuentra debajo.

A partir de la Figura 8 se puede observar que entre 1593 y 1600, el autor publicó 21 de sus 43 obras. De estas 21, la mayoría son comedias seguidas por obras históricas, tragedias y poemas, en ese orden. Se observa también que al inicio de su carrera los géneros preponderantes eran las comedias y las obras históricas, que fueron dando lugar a las tragedias. En particular, entre 1605 y 1608, de las 6 obras que publicó, 5 fueron tragedias. En el último tramo de su carrera, sin embargo, la producción estuvo repartida entre comedias, género histórico y poemas y, además, publicó su único soneto.

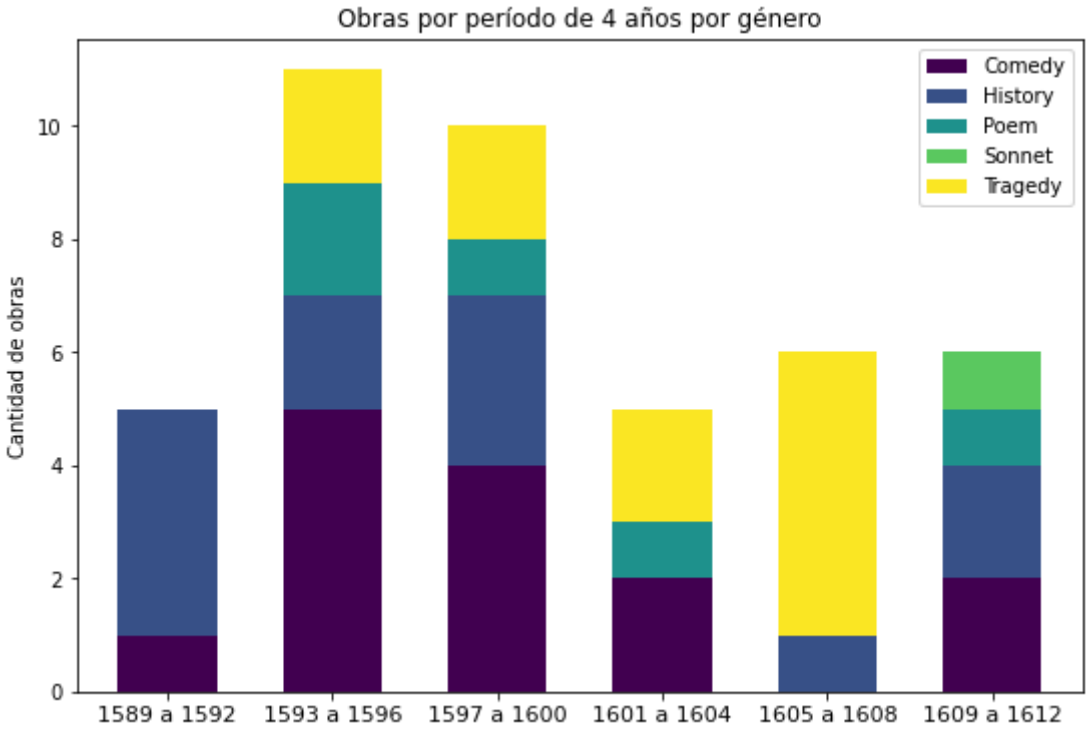


Figura 4 - Histograma obras de Shakespeare por año, columnas apiladas según género.

Se realizó la siguiente visualización en modo de mapa de color (ver Figura 5), se cree que esta alternativa es eficiente para analizar tanto evoluciones temporales por género como distribución por géneros en un periodo de tiempo, no así para analizar la evolución del total de obras presentados por periodo.

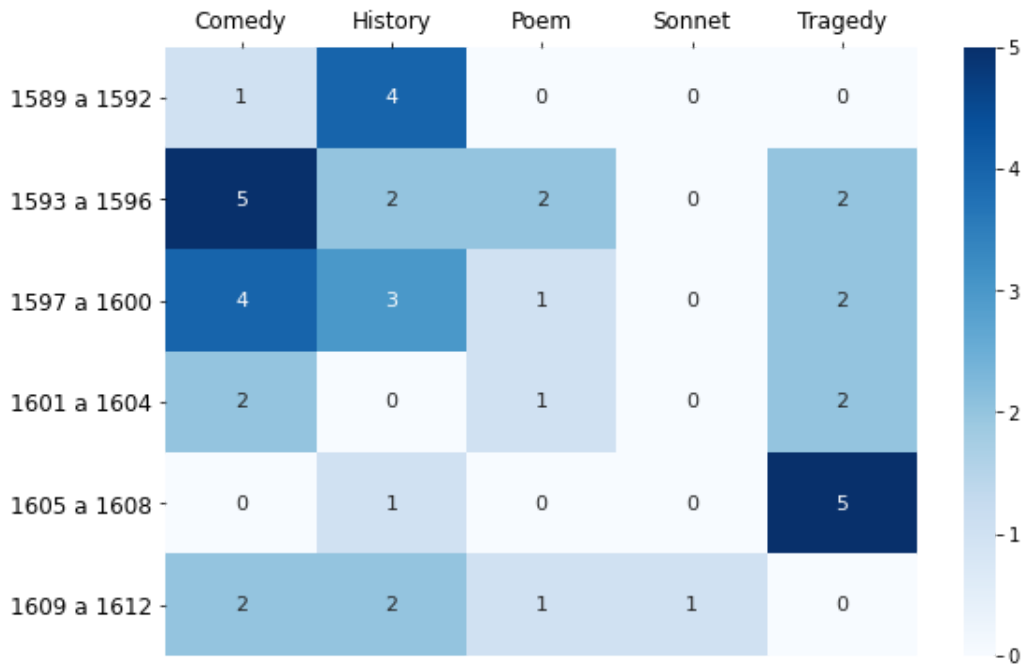


Figura 5 - Cantidad de obras publicadas por Shakespeare por periodos de 4 años y por género.

Para poder analizar el texto de las obras de Shakespeare, se debe procesar el texto de forma de eliminar mayúsculas y signos de puntuación. Para esto se utilizó la función *CleanText*, que pone todo el texto en minúscula y elimina los signos de puntuación ",",";",":","!","?","-","""," ". Tras realizar esta transformación, analizamos las palabras más populares en la obra de Shakespeare, que se muestran a continuación.

Para analizar las palabras más populares por género o personaje se podría agregar una columna género y otra columna personaje al dataframe. La visualización de mayor cantidad de palabras por género y personaje se podría observar en un mapa de calor como la imagen previa, donde en un eje estuvieran los distintos géneros, en el otro eje las palabras más populares y se muestre la cantidad de apariciones de cada una de esas palabras para cada género.

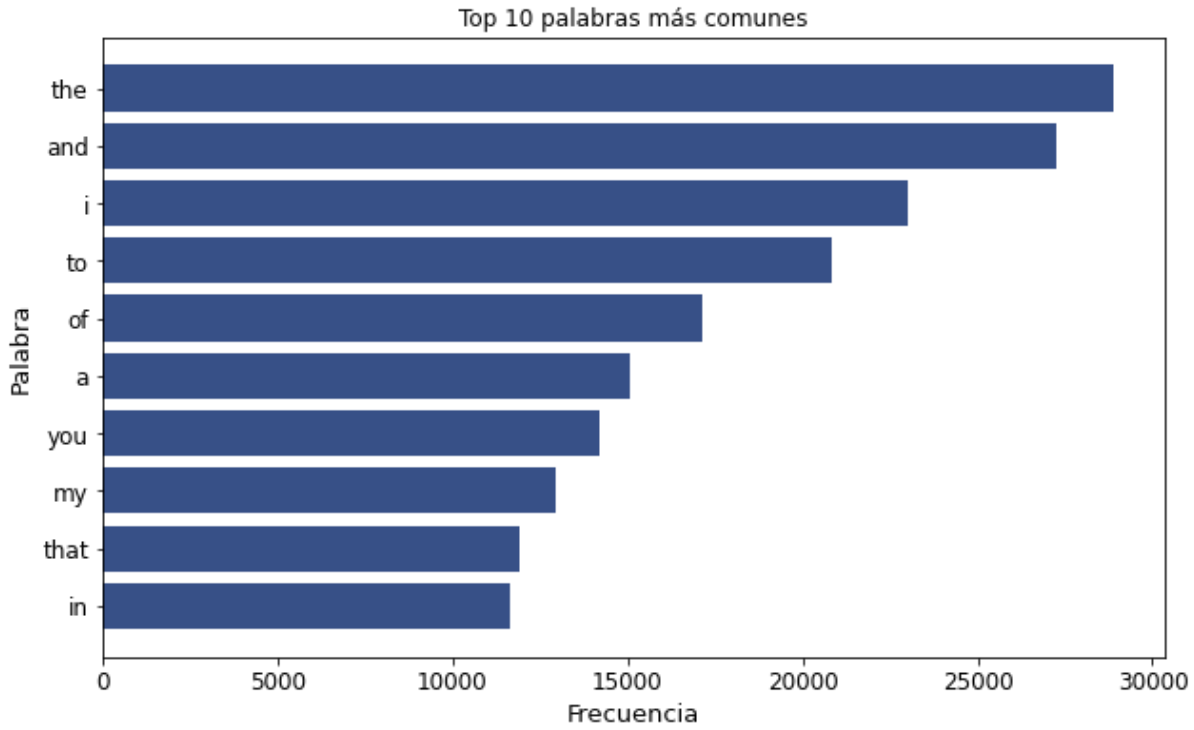


Figura 6 - Cantidad de menciones a las palabras más frecuentes en la obra de Shakespeare.

En caso de sólo querer hacer un análisis cualitativo de las palabras usadas con más frecuencia se puede optar por presentarlas en forma de *wordmap* (ver Figura 7).

Figura 6 - Palabras más frecuentes en la obra de Shakespeare.

Si bien el personaje que aparece con más palabras es “Poet”, se entiende que el mismo puede referirse a poetas distintos lo que además explicaría la amplia diferencia que presenta frente a “ Henry V” y “Falstaff”. Por lo tanto se considera que el personaje con mayor cantidad de palabras es Henry V. Para ver efectivamente los 10 personajes con más palabras se podría eliminar la primera entrada del dataframe.

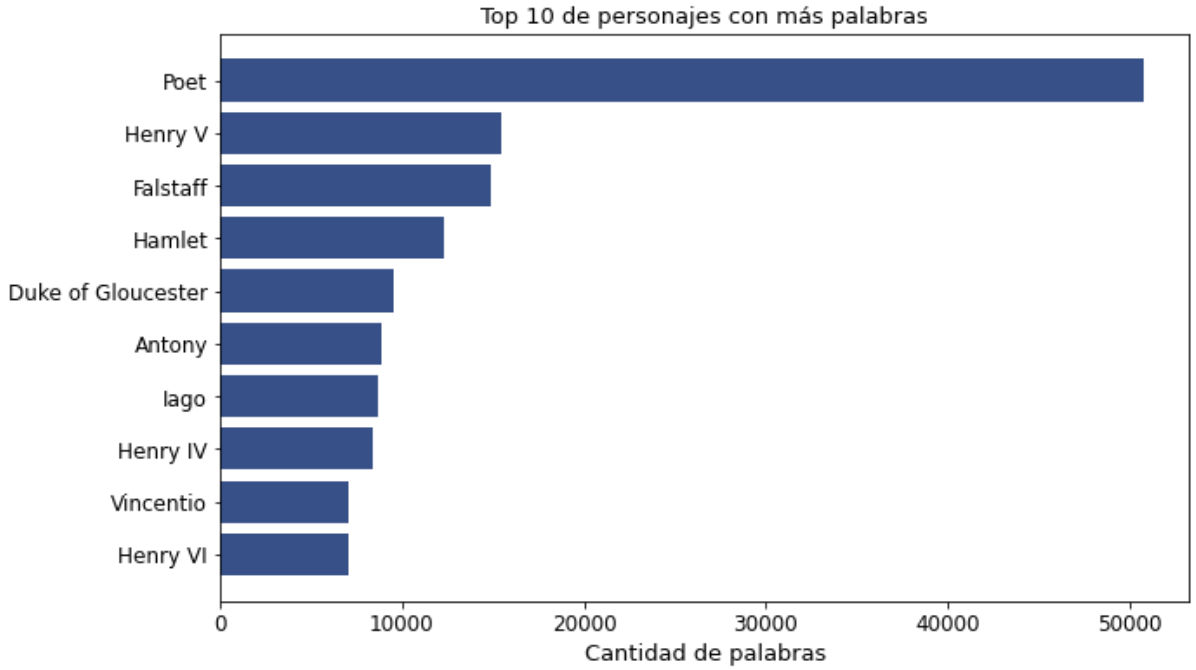


Figura 11 - Cantidad de palabras por personaje en las obras de Shakespeare.

Más allá del análisis realizado, el set de datos permite extraer otras conclusiones como:

- ¿Cuál es la obra más larga de Shakespeare? ¿Y la más corta?
- ¿Qué género presenta el mayor largo por obra promedio? ¿Y el menor?
- Análisis de la estructura de las obras ¿existe alguna tendencia temporal en cuanto al largo de sus obras? ¿Y en cuanto a la cantidad de escenas y actos?¿Y según el género?
- ¿Qué personaje aparece en más párrafos? ¿Y en más capítulos?
- Para alguna obra en particular ¿cuáles son los personajes principales?¿Cómo es la evolución de la participación de los personajes?
- Comparación de palabras más recurrentes diferenciando por género y/o por año de publicación.