Universidad de la República

Facultad de Ingeniería

Maestría en Ciencia de Datos y Aprendizaje Automático

Introducción a la Ciencia de Datos 2023

Análisis y Visualización de la Obra de William Shakespeare utilizando Ciencia de Datos

Ana Cortazzo
Luciana Olazábal

23 de mayo de 2023

Contenido

1. Introducción

Este informe presenta métodos y resultados utilizados para el análisis de algunos aspectos relacionados a la obra de William Shakespeare[1]. Se utilizó una base de datos relacional abierta con la obra completa de William Shakespeare, disponible en este link. En la Figura 1.1 se presenta la estructura relacional de la base de datos utilizada.

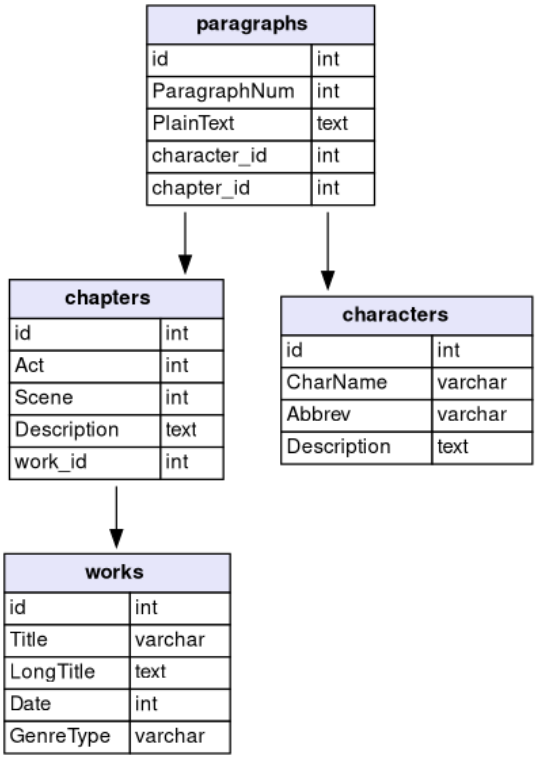

Figura 1.1: Estructura de la base de datos utilizada

El trabajo se realizó en dos etapas, iniciando con el **procesamiento de datos** que se verá en la Sección 2, dónde se cargan los mismos y se realiza una exploración y limpieza de los datos proporcionados, **análisis de datos**, a verse en la Sección 3, donde se realizó una visualización de la obra de Shakespeare a lo largo de los años y un conteo de palabras y párrafos (considerando la obra completa y luego discriminando por trabajo). Al final del informe (Sección 4) se presentan los resultados más relevantes y una discusión de los mismos. Se puede acceder al código completo en el repositorio de GitHub.

[1] Dramaturgo inglés (1564 - 1616). https://es.wikipedia.org/wiki/William_Shakespeare

2. Procesamiento de datos

2.1. Cargado y exploración de datos

El entorno utilizado para el desarrollo de este proyecto fue Jupyter Notebook, una aplicación de código abierto que proporciona un entorno de programación interactivo basado en lenguaje Python, que facilita la exploración de datos, el análisis y la visualización. Para realizar las tareas de manipulación, análisis y visualización, el ambiente se configuró con las siguientes bibliotecas:

[]:
```
from time import time
from pathlib import Path
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
from sqlalchemy import create_engine
from wordcloud import WordCloud, ImageColorGenerator
```

La conexión con la base de datos se realizó utilizando el motor `SQLAlchemy` y la cadena de conexión adecuada, las tablas se importaron desde la base de datos y se almacenaron en el formato DataFrame de `Pandas` y se guardaron en archivos `.csv` en un directorio específico. Para facilitar la tarea se define la función `load_table` con todos los parámetros necesarios.

[]:
```
df_works = load_table("works", engine)
df_paragraphs = load_table("paragraphs", engine)
df_chapters = load_table("chapters", engine)
df_characters = load_table("characters", engine)
```

Una vez cargadas las tablas se analiza la cantidad de registros y las propiedades que contiene cada una de ellas:

- `df_works`: información detallada de las obras de Shakespeare. Tiene 43 filas y 5 columnas:
  - `id` identificador de la obra
  - `Title` título de la obra (versión corta)
  - `LongTitle` titulo completo de la obra
  - `Date` año de publicación de la obra
  - `GenreType` género literario de la obra
- `df_chapters`: identifica los diferentes actos y escenas dentro de cada obra. Tiene 945 filas y 5 columnas:
  - `id` identificador del capítulo
  - `Act`número del Acto
  - `Scene` número de la escena
  - `Description` breve descrićión o título de la escena
  - `work_id` relaciona el `id` de la obra (relación con tabla `df_works`)
- `df_characters`: identifica los diferentes personajes que intervienen en las obras. Tiene 1266 filas y 4 columnas:
  - `id` identificador del personaje
  - `CharName` Nombre del o los personajes
  - `Abbrev` abreviación del nombre
  - `Description` información que describe al personaje
- `df_paragraphs`: Contiene el texto principal de las obras. Tiene 35465 y 5 columnas:
  - `id` identificador del párrafo
  - `ParagraphNum` número de párrafo dentro de la obra
  - `PlainText` texto real del párrafo
  - `character_id`indica qué personaje está hablando en el párrafo (si corresponde) (relación con tabla `df_character`)
  - `chapter_id` indica en qué capítulo (acto y escena) se encuentra el párrafo (relación con tabla `df_chapters`)

2.2. Calidad de datos y limpieza

En primer lugar se hizo un relevamiento de datos faltantes (null, NA o NaN). Para esto se utilizó la función `isna()` de `Pandas`. De las cuatro DataFrame analizadas, se detectó que solo `df_characters` presenta 646 valores faltantes en la columna `Description` y 5 valores faltantes en la columna `Abbrev`. Dado que estas dos columnas no cumplen un papel relevante en el análisis que se realizó, se dejan los valores faltantes en las tablas.

En segundo lugar se definió la función `puntuacion` que dada una columna de un DataFrame, realiza un búsqueda de los caracteres de puntuación (utilizando la cadena predefinida `string.punctuaction`) y devuelve una lista con todos los signos encontrados.

Dado que el texto a ser analizado se encuentra en idioma inglés, existen contracciones típicas del idioma, por ejemplo I'm (I am), you're (you are) que al remplazar el caracter ' por un espacio, se separa la contracción, pero no se obtienen las palabras exactas que ella representa. Esto podría representar un problema a futuro en análisis más detallados sobre el contenido de la obra, o bien análisis gramaticales de la misma, pero dentro del alcance de este trabajo, no se pierde infromación relevante al realizar la sustitución del caracter por espacio, por lo que se decide realizar la sustitución.

Para la limpieza de datos se definió la función `clean_text`, la cuál remplaza todos los signos de puntuación por espacios, y convierte las mayúsculas a minúsculas, de esta forma nos quedamos con un texto limpio que facilita las tareas de análisis y conteo. Se aplicó la función `clean_text` a las siguientes columnas, que serán utilizadas posteriormente:

[]:
```
df_paragraphs["CleanText"] = clean_text(df_paragraphs, "PlainText")
df_characters["CleanName"] = clean_text(df_characters, "CharName")
df_works["CleanTitle"] = clean_text(df_works, "Title")
```

En la tabla 2.1 se presenta, a modo de ejemplo, el antes y el después de la columna `PlainText` luego del proceso de limpieza.

Tabla 2.1: Ejemplo de texto antes y después de la limpieza de datos

PlainText	CleanText
[Enter DUKE ORSINO, CURIO, and other Lords; Mu... If music be the food of love, play on;\nGive m... Will you go hunt, my lord? What, Curio?	enter duke orsino curio and other lords mu... if music be the food of love play on give me... will you go hunt my lord what curio

3. Análisis de datos

3.1. La obra de Shakespeare a través de los años

¿Es posible analizar aspectos creativos de la obra de Shakespeare a partir del análisis de los datos? ¿Es posible identificar algunos períodos de su obra? Estas son algunas preguntas disparadoras para la discusión.

Si se empieza por analizar el histograma de la Figura 3.1, donde se muestrala cantidad de obras de Shakespeare a lo largo del tiempo en intervalos de 4 años, se pueden ver algunas tendencias. Por ejemplo, que su productividad en términos de producción individual de obras tiene un máximo entre los años 1593 y 1601, donde el autor llegó a escribir 20 obras en un intervalo de solamente 8 años. Esto es seguido por el período de 4 años de menor concentración de obras (5 obras en 4 años).

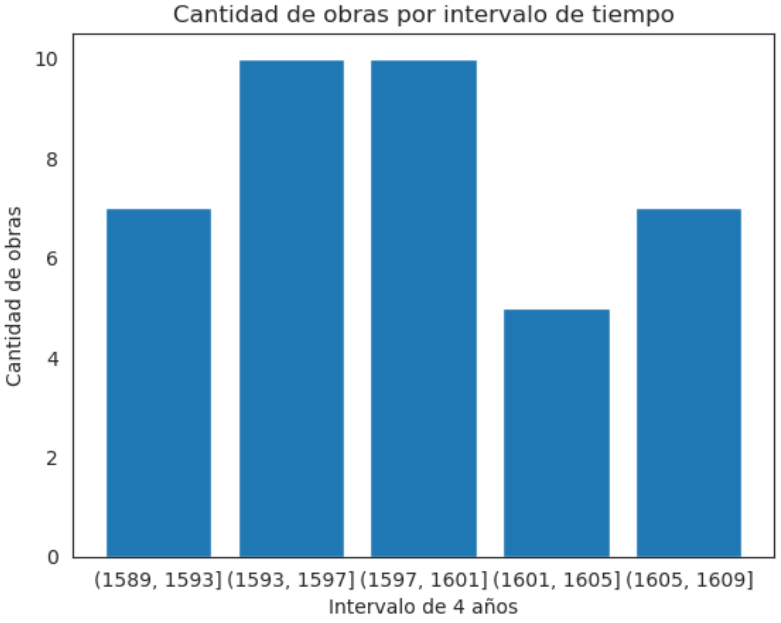

Figura 3.1: Obras publicadas en periodos de 4 años

Al analizar el histograma, es importante tener en cuenta que las obras de Shakespeare abarcan una amplia gama de géneros, que van desde tragedias y comedias hasta historias históricas y sonetos. Cada género puede tener una representación desigual en el histograma, lo que puede influir en la distribución de las obras a lo largo del tiempo. Para realizar un análisis más detallado en la Figura 3.2 se dividen las obras por género:

- Comedia: Algunas de las comedias más conocidas de Shakespeare son “Sueño de una noche de verano”, “Como gusten”, “Mucho ruido y pocas nueces” y “Noche de reyes”.
- Tragedia: Ejemplos de tragedias de Shakespeare incluyen “Romeo y Julieta”, “Macbeth”, “Hamlet”, “Otelo” y “El Rey Lear”

- Historia: donde narra eventos de la historia de Inglaterra, como los reinados de los monarcas ingleses. Ejemplos de esto incluyen “Ricardo III”, “Enrique IV”, “Enrique V” y “Julio César”.
- Poemas y Sonetos: Los sonetos son poemas de 14 líneas escritos en verso

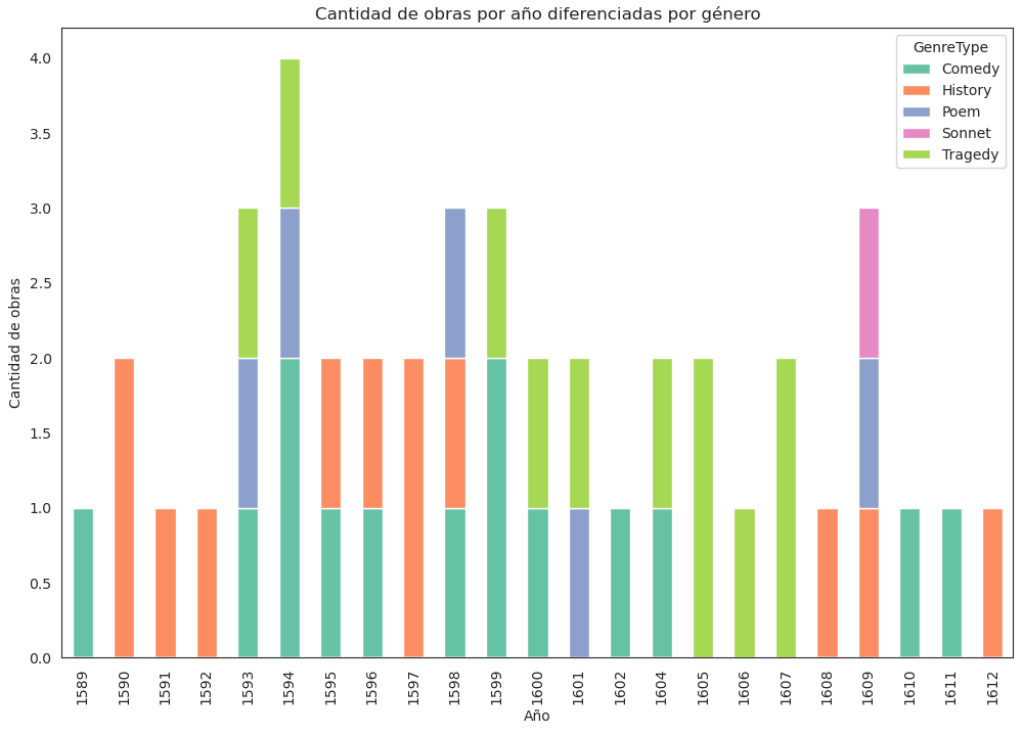

Figura 3.2: Obras publicadas por año diferenciadas por género

Analizando la gráfica de la Figura 3.2 se puede concluir que el autor era muy versátil, ya que logro abarcar a lo largo de su vida una variedad de géneros sumamente distintos entre sí. En sus primeros años como escritor predominan las comedias, y a medida que su carrera avanza, vemos un aumento en las tragedias, que exploran temas más profundos y oscuros, identificando un perído en el sólo escribió obras de este tipo (1605-1607)[2]. Tanto al comienzo como al final de su carrera literaria se ve una gran cantidad de obras históricas, probablemente influenciadas por el contexto histórico y cultural en el que vivió el autor.

Entre estos géneros mencionados, que representan la mayor parte de la producción literaria de Shakespeare, se pueden encontrar algunos poemas y sonetos, pero no se ve una clara tendencia temporal en los mismos. Probablemente tanto las obras identificadas como poemas (`Poem`) y sonetos (`Sonnet`) refieran a un mismo género literario y en la base de datos se registó de forma desagregada, llamando soneto al libro debido a su título[3].

Un dato interesante para determinar la evolución de la carrera literaria de este autor es determinar el número de palabras escrita por año. Esto nos puede dar una idea del volumen de trabajo que

[2] La etapa 1604-1608 es conocida como la de las grandes tragedias, en las que Shakespeare bucea en los sentimientos más profundos del ser humano Fernández and Tamaro [2004]

[3] Para profundizar en esa discusión ver Fernández and Tamaro [2004]

manejó el escritor a lo largo de su vida. Observando únicamente la cantidad de obras en el tiempo, como se hizo en la Figura 3.1, no se tiene información de la carga de trabajo que significaron para el autor, pues algunas obras puede ser menos extensas que otras.

En la Figura 3.3 se puede ver la cantidad de palabras escritas por año. 1594 fue el año de mayor volumen de trabajo en términos de palabras escritas, superando ampliamente a los demás. En este año Shakespeare escribió varias obras teatrales que se consideran importantes en su carrera como "Romeo y Julieta", "El mercader de Venecia" y "Sueño de una noche de verano".

En el año 1603 no se tiene registro de escrituras del autor, y los períodos de menor concentración de palabras escritas se dan sobre el inicio y el final de su carrera como escritor. Sus años más productivos parecen ser los abarcados desde 1593 hasta 1600, cuando los géneros predominantes en su escritura eran la comedia y la historia. Considerando que Shakespeare nació el 26 de abril de 1564, se podría afirmar que el autor vivió su etapa más productiva en términos literarios desde sus 29 a sus 39 años de edad.

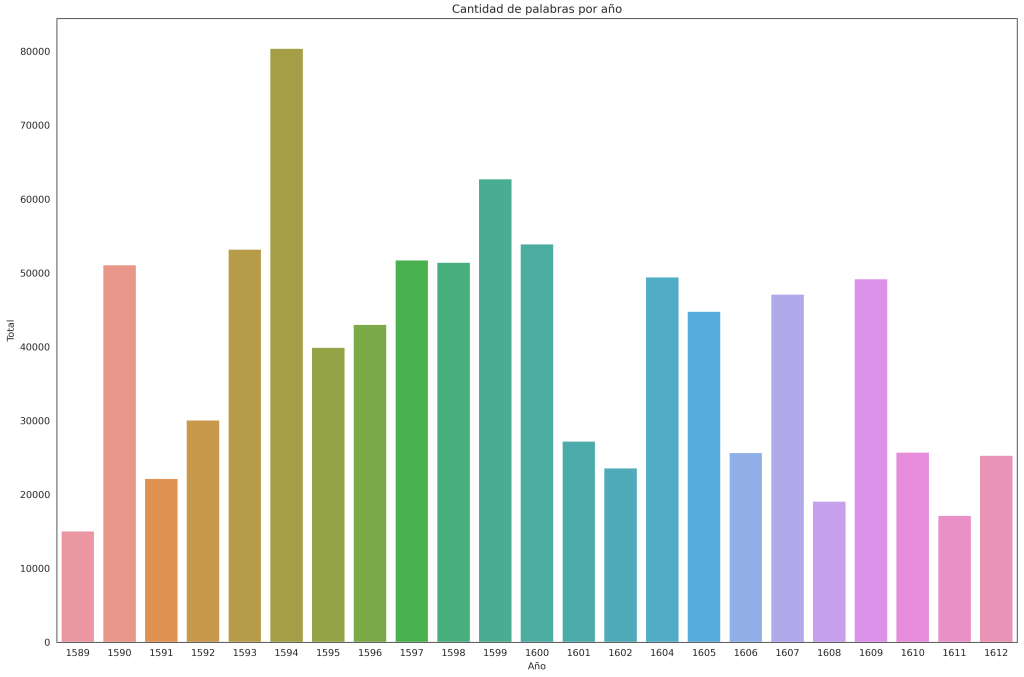

Figura 3.3: Cantidad de palabras escrita por año

3.2. Personaje con mayor número de párrafos

Si realizamos un conteo directo de la cantidad de párrafos que dice cada personaje (sin considerar en que obra aparece), obtenemos el resultado que se muestra en la Tabla 3.1.

Shakespeare no publicó sus obras de teatro en forma de libros durante su vida, sino que sus obras fueron escritas originalmente para ser representadas en el escenario y se publicaron póstumamente en forma de folios y cuartos. Por esto es esperable que la dirección de escenario (*stage directions*) posea un gran número de párrafos, pues se están contando todas las líneas de la dirección de escenario de todas las obras.

Tabla 3.1: 10 personajes con mayor número de párrafos sin discriminar por obra

N.Párr.	Personaje
3751	(stage directions)
733	Poet
471	Falstaff
377	Henry V
358	Hamlet
285	Duke of Gloucester
274	Othello
272	Iago
253	Antony
246	Richard III

Tabla 3.2: 10 personajes con mayor número de párrafos por obra

N.Párr.	Personaje	Obra
358	Hamlet	Hamlet
274	Othello	Othello
272	Iago	Othello
269	Poet	Rape of Lucrece
210	Timon	Timon of athens
204	Cleopatra	Antony and Cleo.
202	Antony	Antony and Cleo.
201	Rosalind	As you like it
201	Poet	Venus and Adonis
194	Brutus	Julius Caesar

Por otra parte, el personaje del poeta (*Poet*), el Duque (*Duke of Gloucester*) y *Falstaff* aparecen en más de una obra, por lo que también es esperable que tengan un gran número de párrafos cuando los datos no se discriminan por obra.

Para resolver esta situación se agrupan los datos por personaje y obra, dando por resultado la Tabla 3.2. Comparando esta tabla con la Tabla 3.1 se ve la diferencia de que algunos personajes tienen mayor cantidad de párrafos en números absolutos, pero no en una obra completa.

Haciendo estas consideraciones y discriminando por obra se puede decir que el personaje con mayor cantidad de párrafos por obra es Hamlet.

3.3. Conteo de palabras

¿Cuáles son las palabras que más aparecen en la obra de Shakespeare? ¿Son representativas y características de su trabajo?

Si realizamos un conteo inicial de las palabras que aparecen en la obra (sin realizar ningún ajuste inicial) obtenemos el resultado que se muestra en la Tabla 3.3.

Tabla 3.3: 10 palabras más frecuentes en la obra (sin limpieza de contenido)

Palabra	Conteo
the	28933
and	27312
i	23006
to	20820
of	17179
a	15084
you	14227
my	12951
that	11910
in	11656

¿Qué conclusión se puede sacar de esta lista? ¿Son representativas de la obra estas palabras? Probablemente, si se analizara cualquier otro texto u obra literaria, el resultado obtenido sería

similar.

Para realizar este análisis debe empezarse por eliminar el set de palabras conocido como *Stop Words* o palabras vacías[4], este es un grupo de palabras usadas comúnmente en los lenguajes. Eliminar las *Stop Words* puede proporcionar una visión más precisa de las palabras clave o términos significativos presentes en el texto, permitiendo realizar un análisis más enfocado al contenido y no puramente cuantitativo.

Para la eliminación de las *Stop Words* se utilizó una lista modificada de *Early Modern English* y *Old English*, ya que la obra se Shakespeare se ubica en esa época[5] La Figura 3.4 muestra las 15 palabras más frecuentes en la obra de Shakespeare luego de eliminar las palabras vacías.

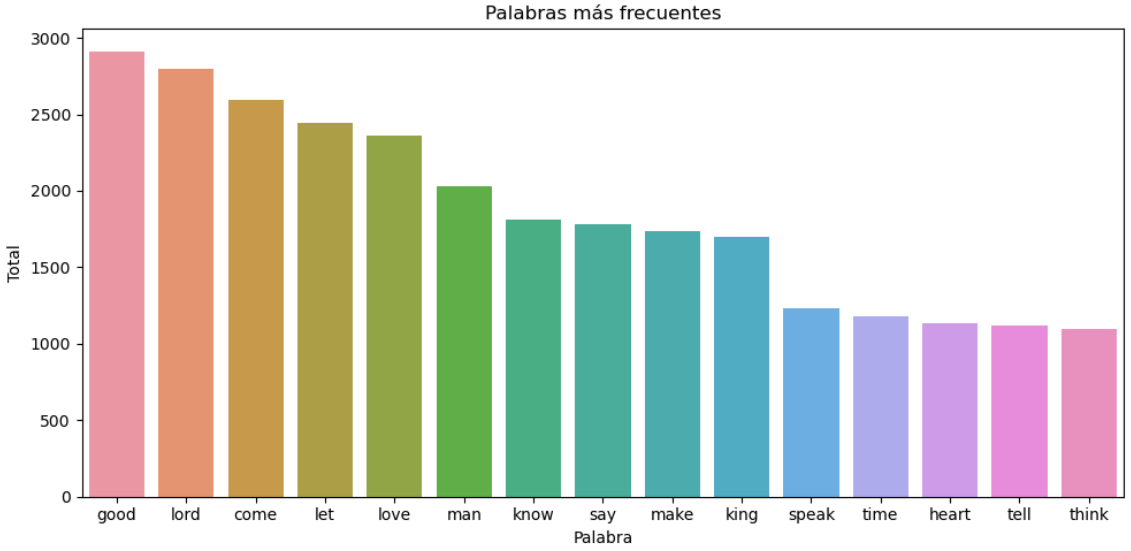

Figura 3.4: Palabras que más se repiten en la obra completa

Las palabras "lord" y "man" pueden estar relacionadas con el contexto social y religioso en el que vivió el autor, ya que "Lord" podría hacer referencia tanto a la nobleza como a figuras divinas, y "man" podría representar la humanidad en general, no solamente referir a un hombre en particular, de la misma forma la palabra "king" se asocia al momento histórico y al contenido de algunas de sus obras.

Las palabras "say", "speak", "tell", "know" y "think" están relacionadas con el lenguaje y la comunicación. Tiene sentido que se repitan muchas veces ya que la palabra y el diálogo son elementos claves de las obras teatrales del autor.

Las palabras "heart", "tell", "think", "speak" y "make" podrían estar relacionadas con aspectos emocionales de los personajes, ya que el autor exploraba en muchas de sus obras las emociones humanas. El concepto del tiempo o "time" como aparece en el gráfico puede estar relacionado con reflexiones sobre el paso del tiempo y la experiencia humana, que también eran temas recurrentes en la obra del autor.

[4]Palabras vacías es el nombre que reciben las palabras sin significado como artículos, pronombres, preposiciones, etc. Wikipedia - Acceso: 18 de mayo 2023

[5]La lista utilizada fue adaptada de Clark [2018]

Si miramos ahora la palabra más frecuente “good” y analizamos el contexto en el que aparece (en este caso, indagando la palabra que le sigue) podemos deducir algunas expresiones comunes en la obra de Shakespeare. La Tabla 3.4 muestra las 5 palabras que más se repiten luego de “good”. Vemos que la expresión “good lord” es la más utilizada, seguida de “good morrow”[6] y “good night”.

Tabla 3.4: Las 5 palabras más frecuentes que siguen a la palabra “good”

Palabra	Conteo
lord	241
morrow	113
night	110
master	53
man	40

En la nube de palabras de la Figura 3.5 se puede visualizar de otra forma la predominancia de ciertas palabras en los textos de Shakespeare.

Figura 3.5: Visualización de las 30 palabras que más se repiten en la obra completa

3.3.1. Palabras por personaje

En el análisis de palabras por personaje sucede algo similar a lo que se vio en el análisis de párrafos por personaje. El poeta, *Falstaff* y la dirección de escenario tienen un gran número de palabras cuando no se discrimina por obra (Figura 3.6).

[6] “good morrow” = “good’ morning”. https://www.dictionary.com/browse/good-morrow

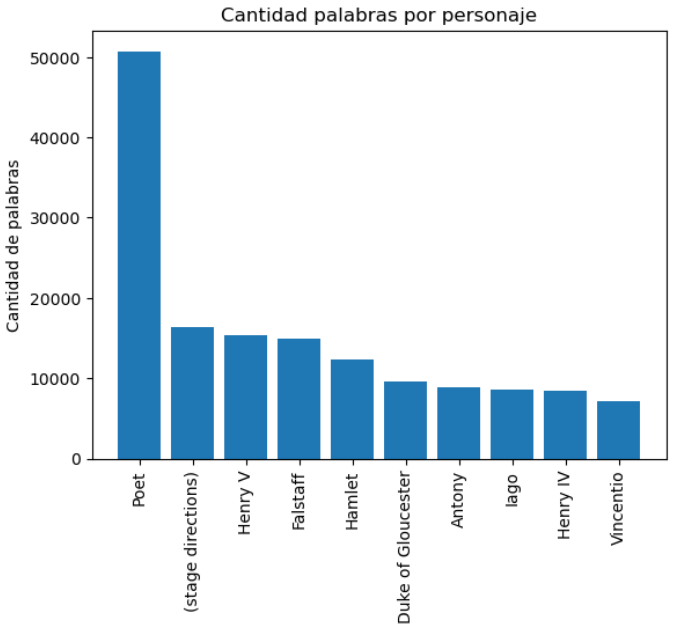

Figura 3.6: Visualización de los 10 personajes con mayor cantidad de palabras considerando la obra completa

Al realizar el análisis discriminando por obra (Figura 3.7) se ve como el poeta se mantiene como el personaje con mayor cantidad de palabras (en números absolutos y por obra). Hamlet al tercer, dejando atrás a Henry V, personaje que aparece en más de una obra, por lo que el conteo total es mayor que el conteo por obra. Por otro parte, tanto Falstaff como la dirección de escena salen del ranking de los 10 personajes con mayor cantidad de palabras.

Es interesante notar cómo la cantidad de palabras del personaje Poeta se encuentra dispersa en varias obras. El conteo inicial muestra al poeta con 50762 palabras, sin embargo, en una obra completa, el número máximo de palabras dicha por el personaje es de 18036 (en la obra Sonnets), y el segundo lugar es de 15530 (en la obra Rape of Lucrece).

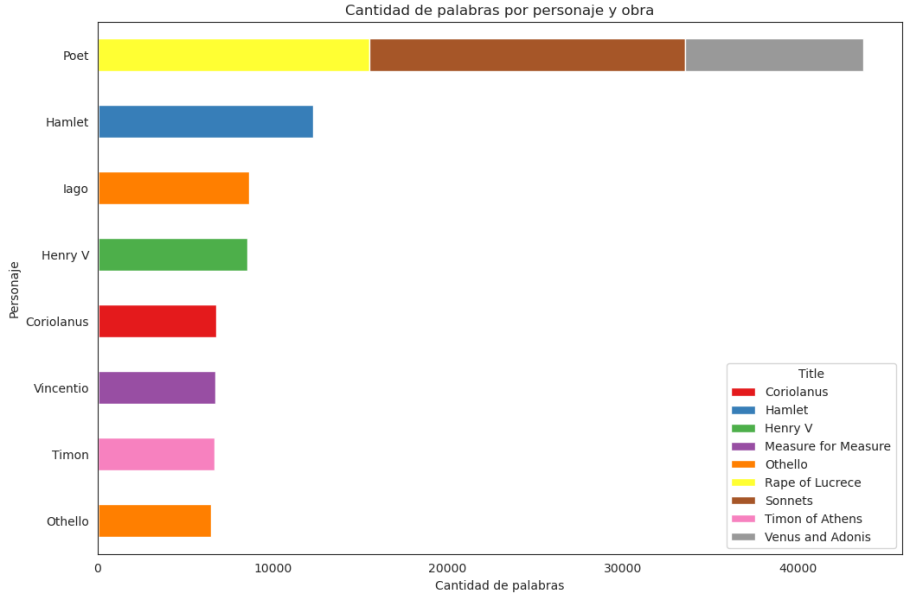

Figura 3.7: Personajes con mayor cantidad de palabras discriminados por obra

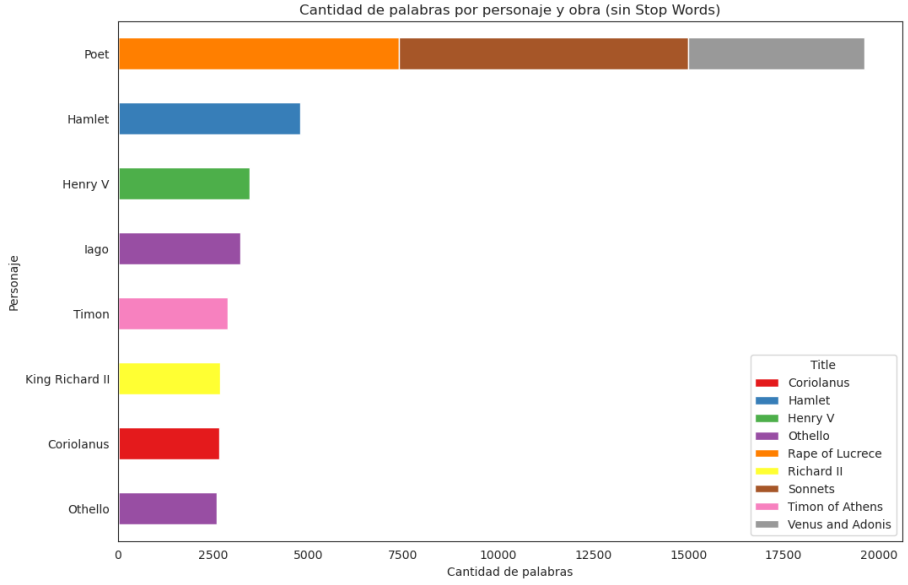

Figura 3.8: Personajes con mayor cantidad de palabras discriminados por obra (sin considerar las Stop Words)

Al quitar las *Stop Words* se puede ver como la concentración global de palabras por personaje disminuye considerablemente y a su vez cambia el ranking de personajes con mayor número de palabras. Por ejemplo, entra en los 10 primeros puestos King Richard II y desaparece Vincentio.

Algunos personajes modifican su lugar en el ranking, pero el poeta y Hamlet mantienen los primeros lugares del conteo.

4. Discusión y conclusiones

Las visualizaciones obtenidas respecto a la obra se Shakespare a lo largo de los años permitió identificar algunos períodos claves en su trabajo, etapas dónde el género predominante es la tragedia, y otras dónde son las comedias o historias. Estos resultados están en concordancia con algunos estudios realizados dentro del campo literario. Para ampliar el análisis con base en estos datos, podrían generarse visualizaciones de palabras escritas por cada género literario, esto permitiría identificar si hay géneros se destaquen en la cantidad de palabras o no.

Realizar el conteo de párrafos por personaje, discriminado por obra, ayuda a tener una idea sobre la relevancia de determinados personajes o no en las obras, por ejemplo Hamlet es el personaje más relevante en la obra homónima, así como Othelo. Además, son dos de los personajes con mayor número de párrafos en la misma obra.

Sin duda, los resultados con mayor potencial para el análisis de la obra son aquellos que realizan conteo de palabras. Al eliminar las *Stop Words* es posible observar elementos representativos de la escrita de Shakespeare, que se relacionan con la época histórica en la que el escritor vivió. Se podrían generar visualizaciones que muestren en qué géneros literarios los personajes tienen mayor cantidad de palabras, por ejemplo.

El análisis de expresiones que se realizó brevemente para las expresiones con "good" puede extenderse a otras expresiones y con técnicas más avanzadas, por ejemplo Procesamiento de Lenguaje Natural (PLN). También se podría utilizar PLN para realizar, por ejemplo, un análisis de sentimientos para determinar las emociones de los textos e intentar identificar los momentos trágicos o cómicos. Un análisis detallado a partir de las palabras más utilizadas para obtener información sobre los temas principales, los personajes más relevantes y las características distintivas de las obras de Shakespeare. Se podría realizar un análisis de las relaciones entre los personajes.

Este trabajo es un primer acercamiento al tema, y quedan abiertas las preguntas para futuras investigaciones.

Referencias

Michael Clark. *An Introduction to Text Processing and Analysis with R*, 2018. URL https://m-clark.github.io/text-analysis-with-R/. Accedido el 14 de mayo de 2023.

Tomás Fernández and Elena Tamaro. *Biografia de William Shakespeare. En Biografías y Vidas. La enciclopedia biográfica en línea [Internet]*, 2004. URL https://www.biografiasyvidas.com/biografia/s/shakespeare.htm. Accedido el 15 de mayo de 2023.