Tarea 1

Introducción a la Ciencia de Datos
Grupo 16

María Luciana Martínez
4.421798-2
luchymd89@gmail.com

Lucía Lemes
5.127.425-6
lucia.lemes@fing.edu.uy

Introducción

En el presente trabajo se utilizó la colección completa de trabajos de William Shakespeare, recopilada sobre la base de datos relacional “Open Source Shakespeare” (OSS) [1]. La misma recopila información de cada título escrito por Shakespeare, su género y año de publicación. También, se tiene el detalle de los personajes, capítulos y escenas que componen la obra, así como los diálogos (cada párrafo y sus contenidos) y direcciones de escena.
El objetivo del trabajo fue explorar los datos dados mediante funciones y visualizaciones que permitan entender mejor el estado de los mismos (si existen campos vacíos, sus contenidos y particularidades) así como extraer información sobre la obra de WS, previo a la realización de ningún tipo de análisis estadístico (por ejemplo, cantidad y características de las palabras, los personajes y sus diálogos, protagonistas de cada historia, así como la distribución temporal de los géneros publicados).

Datos utilizados

El esquema de la base de datos utilizada se puede ver en la Figura 1. La misma cuenta con un total de 43 obras, divididas en 945 capítulos, donde existen 1266 personajes y 35465 párrafos.
Las particularidades de cada tabla se detallan a continuación.

- **works:** Contiene información de las obras publicadas, con un valor de id, año de publicación, género y título.
- **chapters:** Refiere a los capítulos de cada obra, se relaciona con la tabla anterior por medio del atributo *work_id*. Cada capítulo tiene asignado, además, un número de identificación (*id*), un número de acto, número de escena y una descripción del mismo.
- **characters:** Guarda los personajes que aparecen en cada uno de los trabajos escritos, cada personaje se encuentra vinculado a sus párrafos asignados a través de su identificador (*id* del personaje) en la tabla paragraphs. Como datos guardados en la tabla, se tiene un identificador único del personaje, su nombre, la abreviación del mismo y una breve descripción.
- **paragraphs:** Refiere a los párrafos de cada capítulo, vinculados con el capítulo y personaje correspondiente a través de los identificadores *chapter_id* y *character_id*. Cada párrafo tiene un identificador (*id*), un número de párrafo, el texto que contiene, y los identificadores al capítulo (*chapter_id*) y personaje (*character_id*) al que hace referencia.

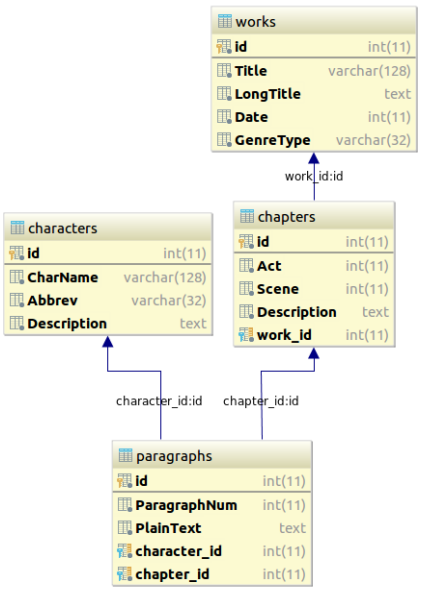


Figura 1: Modelo relacional de la base de datos “Open Source Shakespeare”. En el mismo se detallan los nombres de cada tabla con sus atributos (nombre y tipo).

Métodos

Se trabajó sobre un cuaderno de Jupyter Notebook [2] para descargar las tablas de la base de datos, aplicar transformaciones sobre las mismas y extraer las visualizaciones. El mismo se encuentra disponible como parte de un repositorio de GitLab [3].
Se utilizó inicialmente el paquete de funciones *sqlalchemy* (en su versión anterior a 2.0) para descargar los datos de OSS en archivos .csv. Posteriormente, se utilizó el paquete *pandas* para la carga y manipulación de cada tabla mediante el uso de la clase *DataFrame.* En particular, esta clase permite identificar campos faltantes (espacios vacíos, null, etc) y extraer un informe inicial de cuántas ocurrencias hay por columna.
Para la limpieza de datos se utilizó la función *clean_text()*, usada para pasar el texto de los párrafos a minúscula, y eliminar la mayoría de los símbolos de puntuación. Cabe mencionar que se probó con el paquete *contractions* para eliminar las contracciones permitidas en el idioma inglés (por ejemplo, pasar de *It’s* →*It is*) y así poder contar palabras individuales. Sin embargo, el uso de inglés antiguo en los textos impidió eliminar la mayoría de ellas (de 27819 se resolvieron menos de 3000), ya que muchas veces se utilizaba el apóstrofe para la conjugación de verbos en pasado (kill’d, dissever’d, perform’d, etc.). Considerando el desempeño observado, se decidió omitir este paso y conservar los apóstrofes (contando en cada caso una única palabra). Una vez realizado el procesamiento del texto se realizaron las visualizaciones a partir de los paquetes de *matplotlib* y *seaborn*.

Resultados y discusiones

Exploración de datos

La exploración de datos permitió identificar campos faltantes en algunas filas de la tabla *characters*, donde se vió que no todos los personajes contaban con una abreviación de su

nombre (5 faltantes identificados) y más de la mitad tampoco contaba con una descripción (hay 646 sin ninguna descripción). Además, los campos de texto no estaban normalizados, habiendo varios que mezclaban mayúsculas con minúsculas.
También, en el texto correspondiente a los párrafos existían abreviaciones, símbolos, signos de puntuación y contracciones, además de mezclar mayúsculas y minúsculas. Para trabajar con el contenido de los párrafos (por ejemplo, agrupar por palabra y contar cuántas veces se usó cada una) fue necesario normalizar y corregir los valores de estos campos.

Limpieza de texto

Previo a aplicar transformaciones que permitiera modificar o quitar el contenido no deseado de la columna *PlainText*, fue necesario evaluar qué signos o caracteres se utilizaban en conjunto con el texto de interés. Para ello, se llevó el texto a minúsculas y se retiraron todos los caracteres del abecedario (de la “a” a la “z”), junto con aquellos que representaban números (del “0” al “9”). Luego, a partir del texto resultante, se extrajeron los símbolos que permanecían y que se deseaban eliminar en la etapa de limpieza.
En total, se obtuvieron 16 símbolos diferentes, incluyendo el apóstrofe (utilizado en las contracciones). De éstos se añadieron 15 en la función *clean_text*, donde se incluyen: la coma, el salto de línea (\n), punto, apóstrofe, punto y coma, dos puntos, signo de interrogación, signo de exclamación, guión, paréntesis rectos, paréntesis curvos, et (&), comillas doble y el salto de tabulación (\t).
Todas las ocurrencias de estos símbolos fueron sustituidas por espacios en blanco. Respecto del apóstrofe, se exploró una librería que permitiera resolver las contracciones, sin embargo, de las 27819 veces en que se utilizó este símbolo, un total de 19882 quedó sin resolver. Esto ocurrió debido a la aparición del apóstrofe utilizado en verbos con conjugación en pasado, además de otros casos particulares que la librería no logró resolver. Considerando que la misma podría arrojar resultados erróneos (por ejemplo, confundir usos de “is” con “has”) se prefirió omitir este paso en la limpieza del texto, y conservar todas las contracciones.
Finalmente, se separaron las palabras individuales en listas (una por párrafo) y se creó otro DataFrame que guardara todas las ocurrencias de cada palabra existente en la obra de Shakespeare.

Conteo de Palabras

Con el texto ya sin símbolos se procedió a contar las palabras en la totalidad de las obras, obteniéndose un total de 27083 palabras distintas. En la Figura 2 se muestran las 50 palabras con mayor cantidad de apariciones sobre el total de las obras.
Entre las palabras más frecuentes se puede distinguir una alta prevalencia de monosílabos (principalmente conectores), pronombres y verbos (como el to-be). Es interesante la alta prevalencia de dos palabras: lord y sir, que dan pauta del alto protagonismo que la aristocracia tiene en la obra de Shakespeare. Un factor que no se tuvo en cuenta para el conteo de palabras, y que sería importante considerar, fue separar las indicaciones para la puesta en escena de los diálogos de los personajes. En la Figura 2 no es posible distinguir si la prevalencia de algunas palabras es debido a su uso en los diálogos o en las indicaciones.
En caso de contar con una categorización de palabras por temática, se podría llegar a tener una noción de qué trata cada obra dependiendo de las temáticas que prevalecen en sus textos.

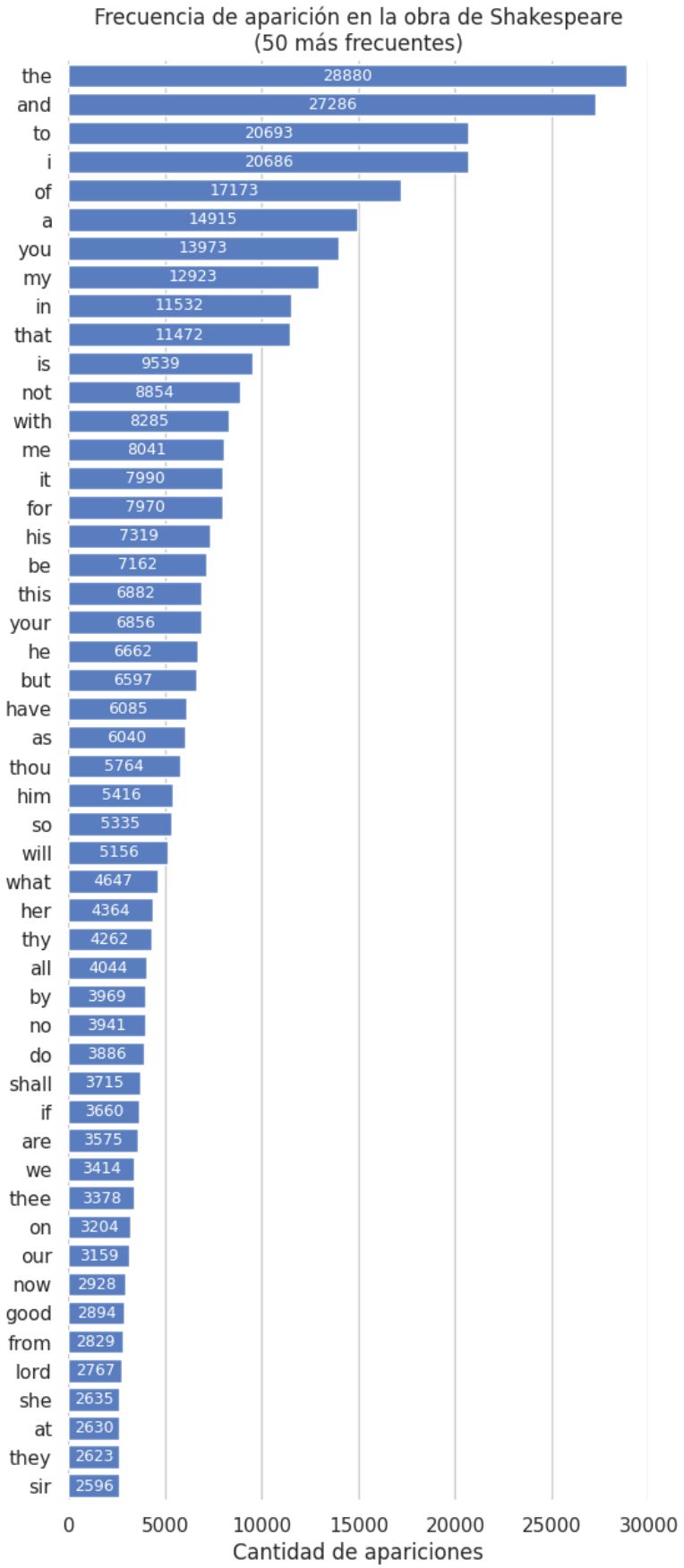


Figura 2: En la imagen puede observarse un gráfico de barras comparando la cantidad de veces que aparecen las 50 palabras más frecuentes (el largo de la barra se indica en el centro) en la obra de William Shakespeare.

Puede resultar de interés explorar las palabras más usadas de acuerdo al género literario del trabajo, y evaluar si la frecuencia de uso de ciertas palabras arrojan información sobre dicha clasificación (por ejemplo, ¿se puede predecir el género de la obra a partir de explorar las palabras más o menos usadas?). Previo a ello habría que realizar un agrupamiento (merge) de los datos de palabras con las tablas de capítulos (por *chapter_id*) y con la tabla de obras (*por work_id*). Con los datos obtenidos se podría estudiar las palabras por obra (*work_id*) y por género (*genreType*) mediante técnicas de clasificación supervisadas, como k-vecinos más cercanos, o SVM. Se podría evaluar también si existen clusters identificables al graficar las frecuencias de palabras en cada obra.
De la misma manera se puede obtener una idea de las palabras más frecuentes por personaje, haciendo un agrupamiento con la tabla de personajes (por *character_id*). Esto permitiría, por ejemplo, estudiar si las palabras utilizadas permiten identificar ciertos personajes en una obra (¿cuál es el personaje que está hablando?) o distinguir si dicho personaje pertenece a la nobleza (muchos de los nombres indican el título nobiliario). Como este caso también se trataría de un problema de clasificación, podría estudiarse con estrategias similares a las ya planteadas antes.
Por último, sería interesante explorar si, para los personajes que hablan con mayor frecuencia (como el Poeta), es posible estimar la respuesta que darían ante una situación de diálogo con otro personaje. Este problema parece ser, a simple vista, mucho más complejo que el anterior, ya que requiere que no solo se infieran las palabras que el personaje usaría a partir de conocer la identidad del personaje y qué líneas de diálogo fueron dichas antes, sino que debe ordenarse las palabras de modo que se forme una frase coherente de acuerdo a las reglas del idioma en el que responde. Quizá podría abordarse mediante técnicas de aprendizaje profundo.

Personajes principales

Por cantidad de palabras

Fue de interés para el trabajo encontrar qué personajes tenían asignadas la mayor cantidad de palabras.
Como un primer acercamiento se agregó al conjunto de palabras los datos de los personajes: nombre e identificador, luego se los agrupó por nombre y se contó la cantidad de palabras que correspondían a cada nombre.
Los primeros 15 resultados obtenidos se pueden observar en la tabla de la Figura 3a. Sin embargo, agrupar las palabras por el nombre del personaje que las dijo presenta problemas cuando éste no es único. Haciendo un breve análisis, se constató que los nombres se repetían en las diferentes obras, por lo que la agrupación por nombres no era correcta, sino que se debía agrupar por el identificador de personaje. En la Figura 3b se muestra una tabla con los primeros 15 nombres de personajes repetidos en las obras, sobre un total de 125 nombres repetidos.
Al agrupar por identificador de personaje se obtuvo una lista similar a la de la Figura 3a, pero era posible apreciar que, por ejemplo, “Poet” (el nombre de personaje con más palabras en toda las obras) tiene un total de 49730 palabras, pero sólo 48950 son las correspondientes al personaje con identificador 894 y nombre “Poet”. Los valores separados por identificador para los 15 personajes con más palabras pueden observarse en la Figura 3c.

CharName	count
Poet	49730
(stage directions)	16408
Henry V	15223
Falstaff	14626
Hamlet	11961
Duke of Gloucester	9331
Antony	8632
Iago	8475
Henry IV	8251
Vincentio	6970
Henry VI	6907
Richard III	6872
Queen Margaret	6834
Coriolanus	6613
Timon	6478

(a)

CharName	count
All	23
Messenger	23
Servant	21
Lord	9
Page	8
First Lord	8
First Gentleman	8
Second Gentleman	8
Gentleman	7
First Servant	7
Both	7
Captain	7
First Citizen	6
Second Servant	6
Second Lord	6

(b)

character_id	count_x	id	CharName
894	48950	894	Poet
1261	16408	1261	(stage directions)
573	15223	573	Henry V
393	14626	393	Falstaff
559	11961	559	Hamlet
531	9331	531	Duke of Gloucester
120	8632	120	Antony
600	8475	600	Iago
572	8251	572	Henry IV
574	6907	574	Henry VI
945	6872	945	Richard III
736	6834	736	Queen Margaret
1236	6617	1236	Vincentio
283	6613	283	Coriolanus
1198	6478	1198	Timon

(c)

Figura 3: En (a) puede observarse una tabla con los nombres de los personajes que aparecen en la obra de Shakespeare, y la cantidad de palabras dichas por éste (en toda las obras). En (b) se indican los nombres que se repiten en las obras (el mismo nombre refiere a personajes distintos), junto con la cantidad de personajes que lo comparten. Por último, en (c) se muestra la cantidad de palabras por personaje, pero esta vez agrupado por identificador. Esto garantiza que se agrupan las palabras por un único personaje. Adicionalmente, se añade el nombre del personaje y cuántas palabras dijo.

Algo interesante a señalar, asociado a un problema ya mencionado en la sección de conteo de palabras, es que en esta base de datos se consideran las indicaciones para la puesta en escena como texto correspondiente al personaje “(stage directions)”, con un id distinto para cada obra. Se sabe que éstos no son personajes de las obras de Shakespeare, sino que recopila información necesaria para ejecutarlas; por eso, al momento de realizar una visualización comparativa entre personajes se decidió dejar fuera las indicaciones de escena.

En la Figura 4 se grafica de forma comparativa las cantidades de las Figuras 3a y 3c, para los 25 personajes con más palabras. Como añadido se incluye, para aquellos personajes con nombre repetido, la cantidad total de palabras dichas por personajes con ese nombre, y la cantidad total de palabras dicha por un único personaje con ese nombre (el de mayor cantidad de palabras). El número al final de cada barra indica la cantidad de palabras dicha por ese personaje, de acuerdo a lo mostrado en la Figura 3c.

A simple vista, entre los personajes con mayor cantidad de diálogo, “Poet” y “Vicentio” son los dos que tienen un nombre asignado a más de un personaje en distintas obras.

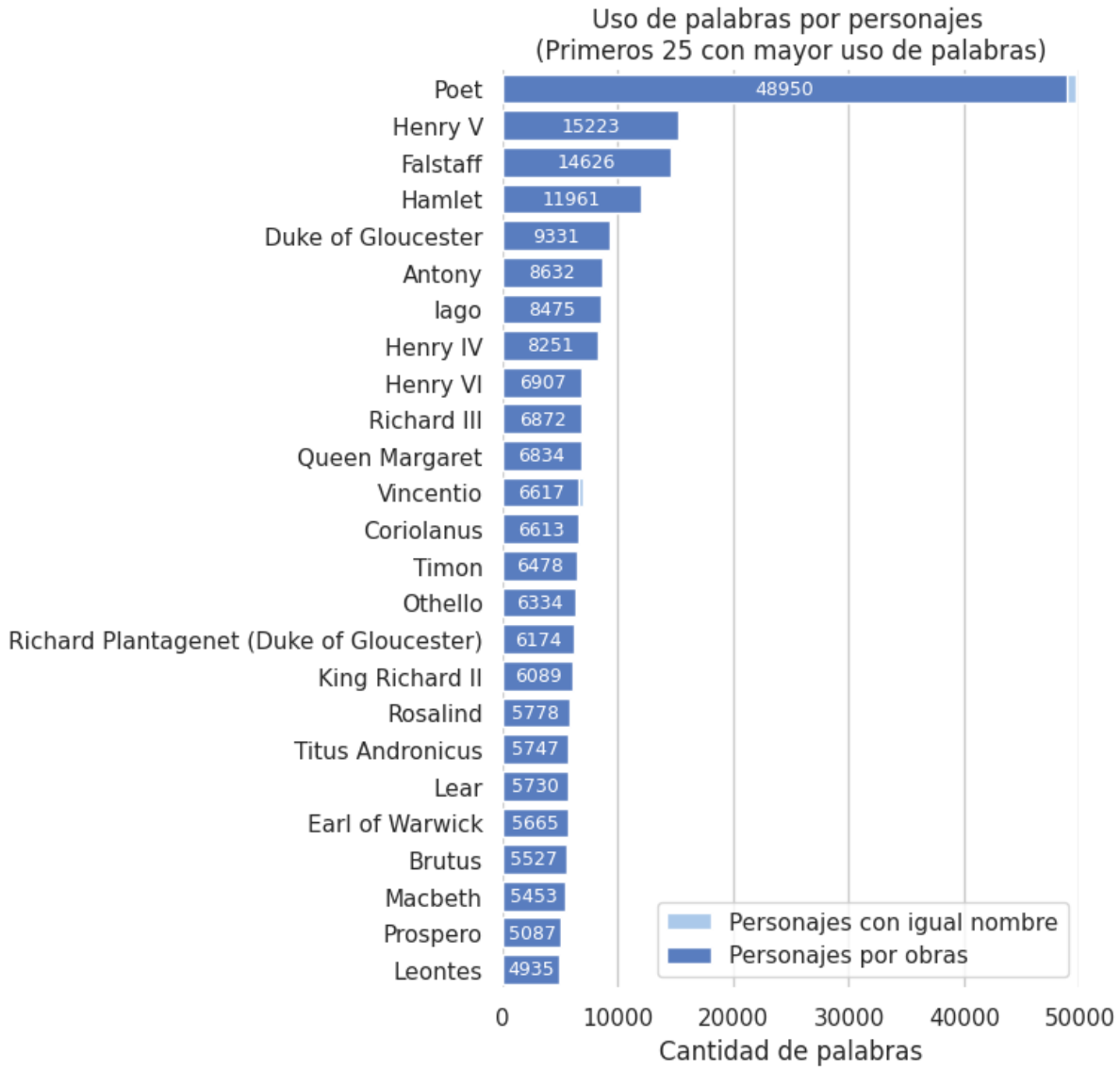


Figura 4: En la imagen puede observarse un gráfico de barras comparativo de la cantidad de palabras dichas por los 25 personajes con más palabras en la obra de William Shakespeare. En azul oscuro se muestra la cantidad de palabras dichas por un único personaje cuyo nombre es igual al que se indica a la izquierda (el largo de la barra se escribe en el centro de la misma), mientras que en celeste claro la cantidad de palabras dicha por todos los personajes que comparten dicho nombre (a modo de comparación).

Por cantidad de párrafos

Se realizó el mismo procedimiento descrito anteriormente, pero esta vez utilizando la cantidad de párrafos de las obras en vez de la cantidad de palabras.

La cantidad de párrafos en una obra de teatro corresponden por lo general a la cantidad de diálogos que realiza el personaje. Al contar la cantidad de párrafos por personaje se tiene una idea un poco más certera de cuáles son los personajes principales. En la Figura 5 se muestran gráficamente los resultados obtenidos para los primeros 25 personajes (se quitó de la lista el primer personaje “stage directions” dado que corresponde a las direcciones de escena).

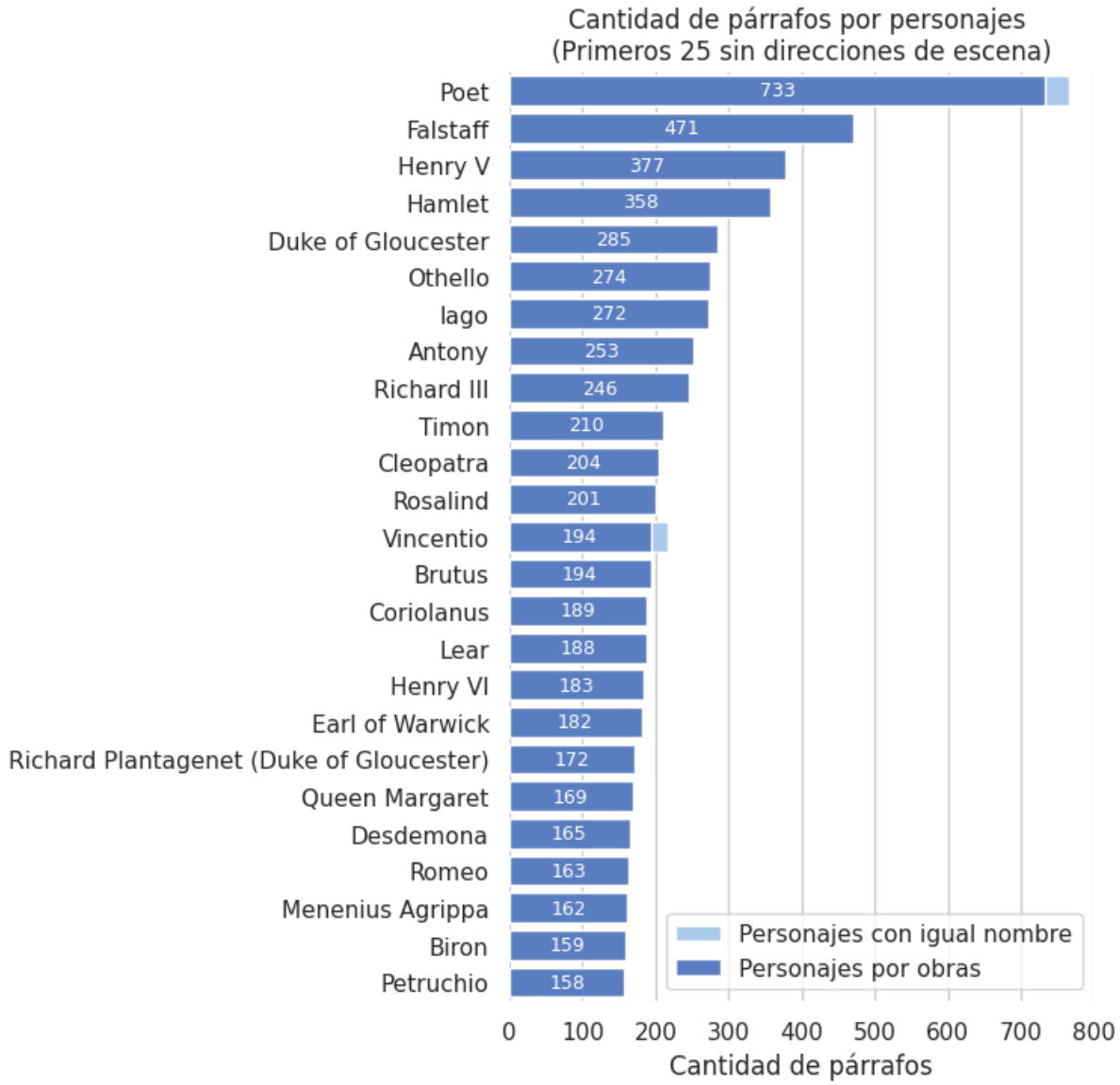


Figura 5: En la imagen puede observarse un gráfico de barras comparativo de la cantidad de diálogos (párrafos) dichos por los 25 personajes con más intervenciones en la obra de William Shakespeare. En azul oscuro se muestra la cantidad de párrafos dichos por un único personaje cuyo nombre es igual al que se indica a la izquierda (el largo de la barra se escribe a la derecha de la misma), mientras que en celeste claro la cantidad de párrafos asignados a todos los personajes que comparten dicho nombre (a modo de comparación).

Por cantidad de párrafos por Obra

Fue de interés explorar una visualización que extrajera el personaje con más líneas de diálogo en cada obra, y mostrarlo asociando dicha cantidad y la obra a su nombre.
Para obtener el personaje principal de cada obra se realizó un procedimiento similar a los comentados antes: a los datos de párrafos por personaje se le agregaron los capítulos correspondientes a cada párrafo y las obras correspondientes a cada capítulo; de esta manera se tiene un conjunto de datos donde se puede agrupar los personajes con mayor cantidad de párrafos dentro de cada obra. Luego, se repitió esto pero contando la cantidad de palabras.
En las figuras 6 y 7 se muestran ambos resultados, y en la Figura 8 se propone una visualización alternativa para la Figura 6 (mediante one-hot encoding), apilando las barras

correspondientes a las obras donde el mismo personaje es protagonista. La visualización alternativa para palabras no se realiza porque es análoga.

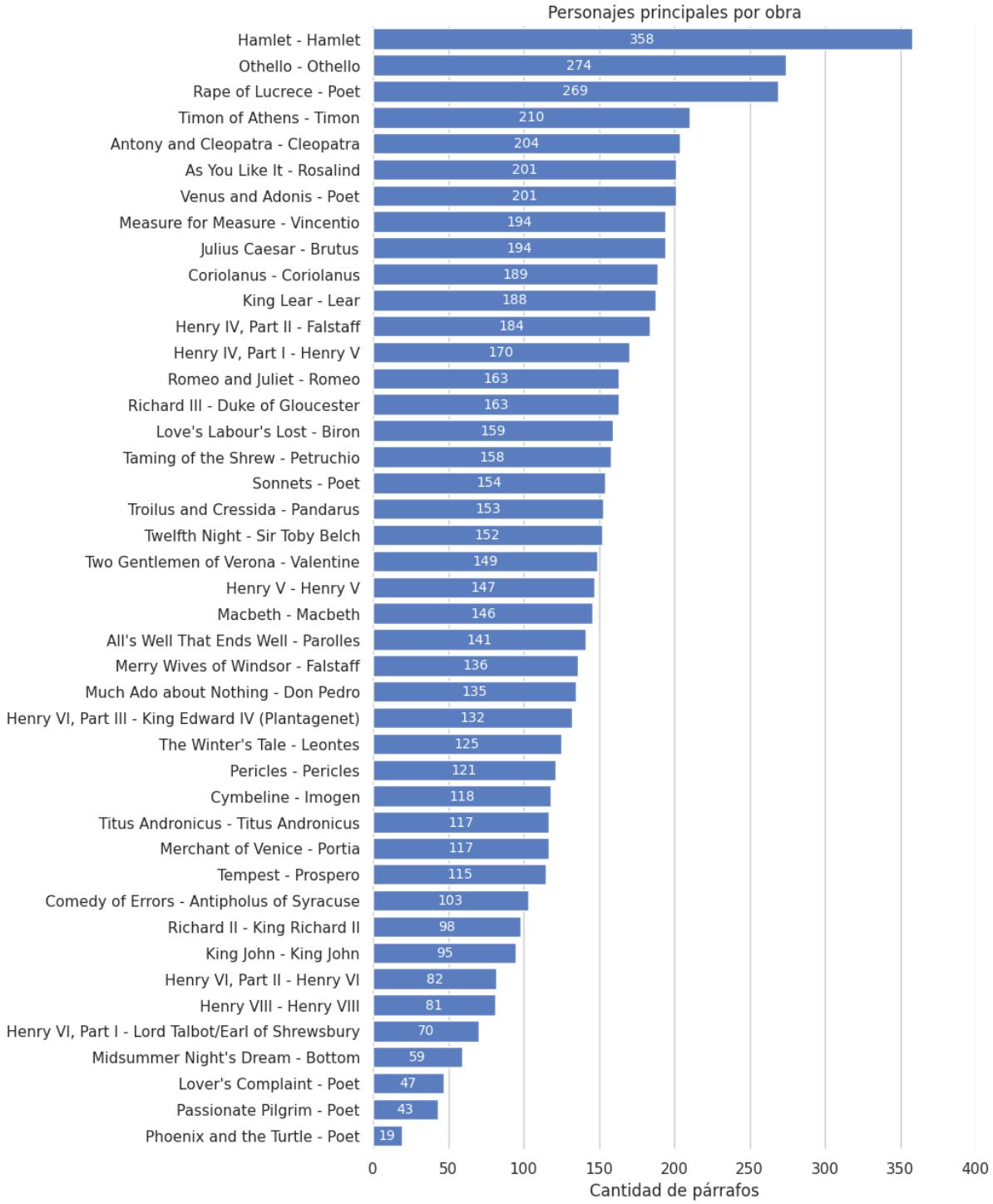


Figura 6: En la imagen puede observarse un gráfico de barras comparativo de la cantidad de diálogos (párrafos) dichos por los personajes con más intervenciones en cada obra de William Shakespeare. En azul oscuro se muestra la cantidad de diálogos asignados al personaje de nombre y obra que se indica a la izquierda (el largo de la barra se escribe en el centro de la misma).

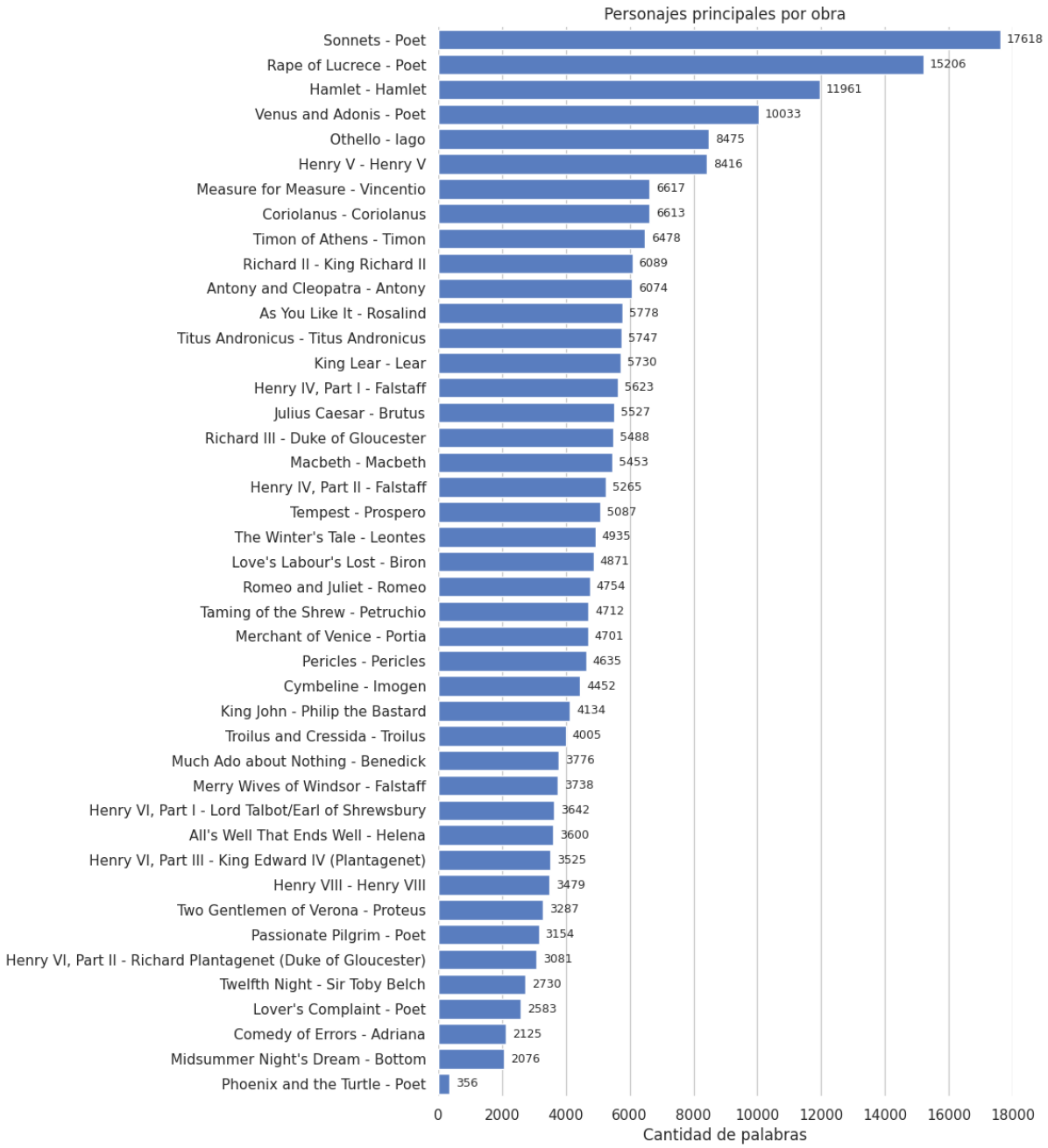


Figura 7: En la imagen se muestra un gráfico de barras comparativo de la cantidad de palabras asignadas a los personajes con más intervenciones más verbosas en cada obra de William Shakespeare. En azul oscuro se muestra la cantidad de palabras asignadas al personaje de nombre y obra que se indica a la izquierda (el largo de la barra se escribe a la derecha de la misma).

Otra manera de ver gráficamente la relevancia de cada personaje es mostrando la cantidad de párrafos agrupados por personaje, pero aplicando las barras según si son párrafos de obras distintas. Esto permite distinguir el aporte a cada obra en particular. Dado que eran varios personajes y varias obras, se optó por no añadir una leyenda (ya que la combinación de colores iba a ser confusa) sino por escribir el nombre de cada obra sobre su barra correspondiente. Esta alternativa facilita la asociación de cada barra con su obra, pero “ensucia” la visual en ciertos puntos donde los títulos son demasiado largos. En esos casos se optó por girar 15° el texto. Para algunos personajes se incluye, en un color gris translúcido, una barra que cuenta el total de párrafos de ese personaje en las obras donde

es el principal y añadiendo los párrafos de las demás obras donde aparece (pero no tiene el mayor número de diálogos). En este caso, por límites de espacio e información considerada relevante, no se incluyó el largo de cada barra.
Es interesante observar que la Figura 8 deja en evidencia que, a pesar de que el personaje con más diálogo en una única obra es Hamlet, en realidad el Poeta es quien acumula más diálogos sumando todos sus papeles protagónicos.

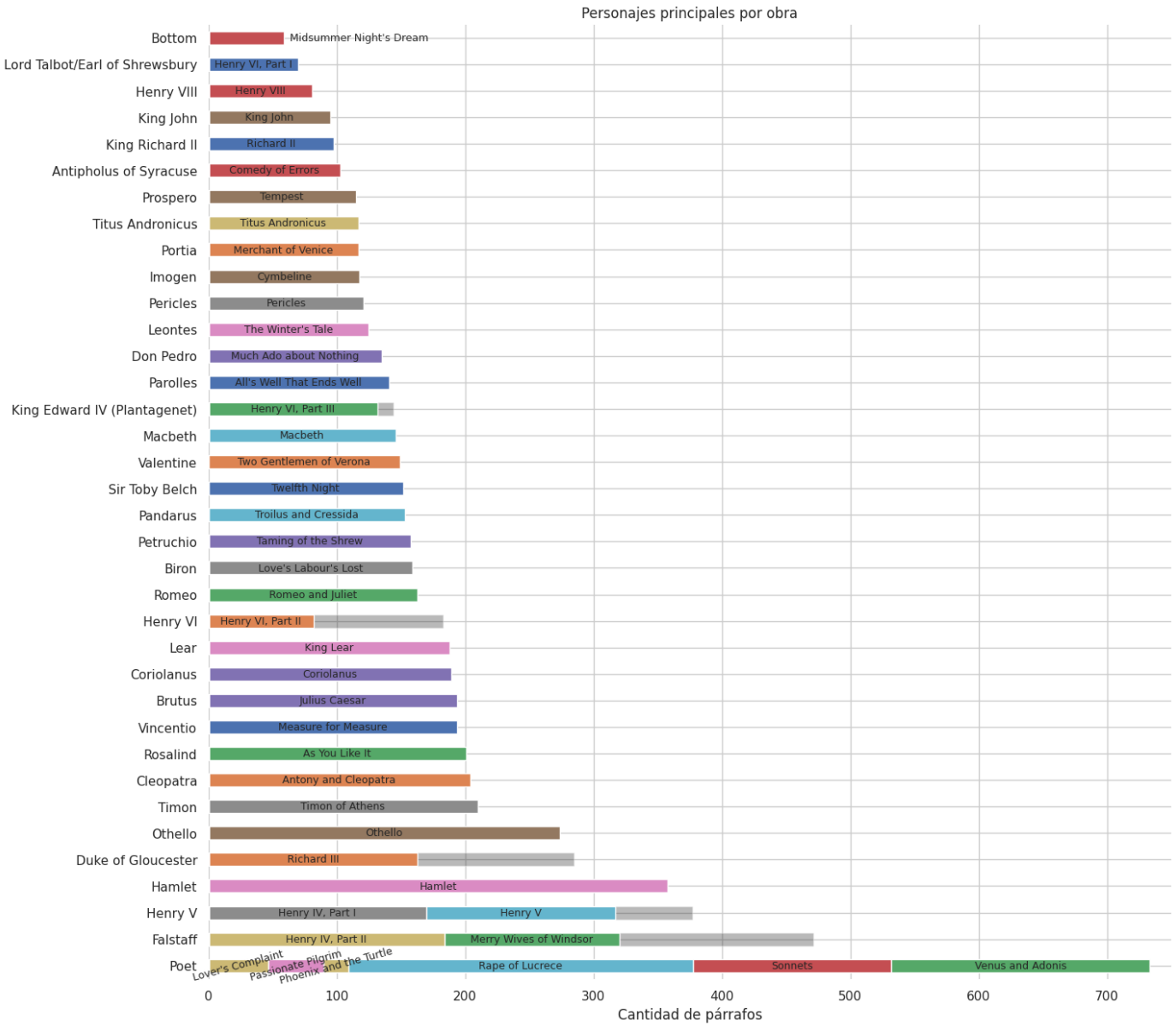


Figura 8: Gráfico de barras comparativo de la cantidad de diálogos (párrafos) dichos por los personajes con más intervenciones en cada obra de William Shakespeare. La cantidad de diálogos asignados al personaje de nombre que se indica a la izquierda se indica de acuerdo al largo de la barra asociada a cada obra (el título de la obra se escribe en el centro de su barra). Un personaje puede ser protagonista de más de una obra, en cuyo caso las barras se grafican apiladas. En gris se indica el total de párrafos acumulados contando todos los trabajos de WS (sea protagonista o no).

Cronología de las obras

Con los datos de la tabla *works* se decidió contabilizar la cantidad de obras que fueron creadas por año y de esta manera ver si existía algún tipo de tendencia.

Al calcular la cantidad de obras por año se obtuvo que la máxima cantidad de obras publicadas fue de 4 en el año 1594. Al listar la cantidad de obras por año se encontró que en el año 1603 no se realizó ninguna obra. Por lo general la producción artística de William Shakespeare osciló entre una y dos obras por año, desde que comenzó a publicar en 1589 hasta su última obra en 1612.
En la Figura 9 se muestra gráficamente la cantidad de obras realizadas por año. A simple vista, parecería que su producción literaria se concentra mayormente en la primera mitad del período, tendiendo a decrecer hacia la segunda mitad (parece haber un quiebre en el ritmo de escritura luego de la pausa de 1603).

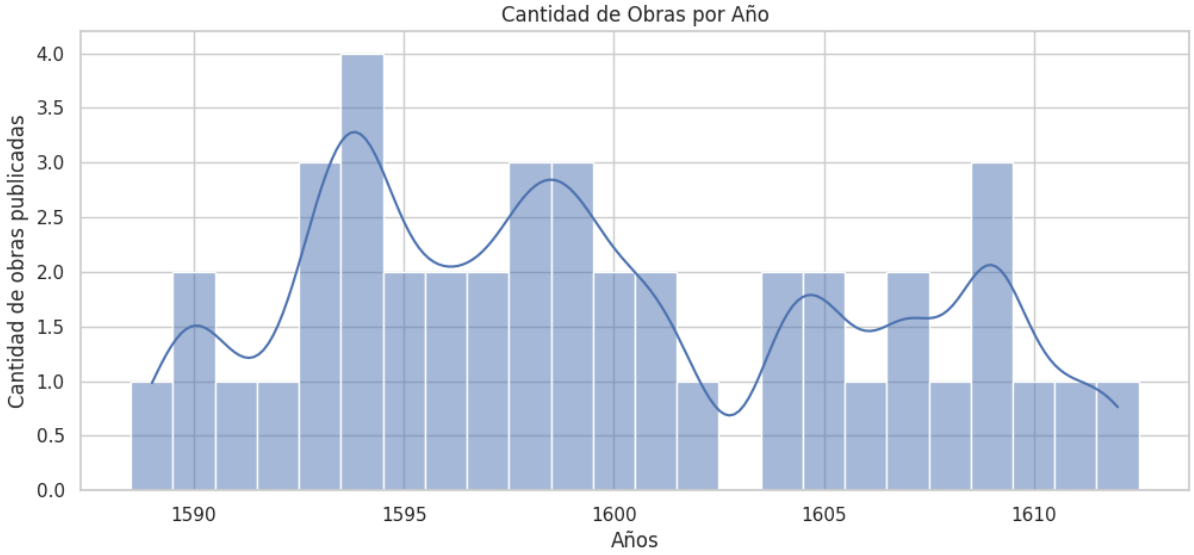
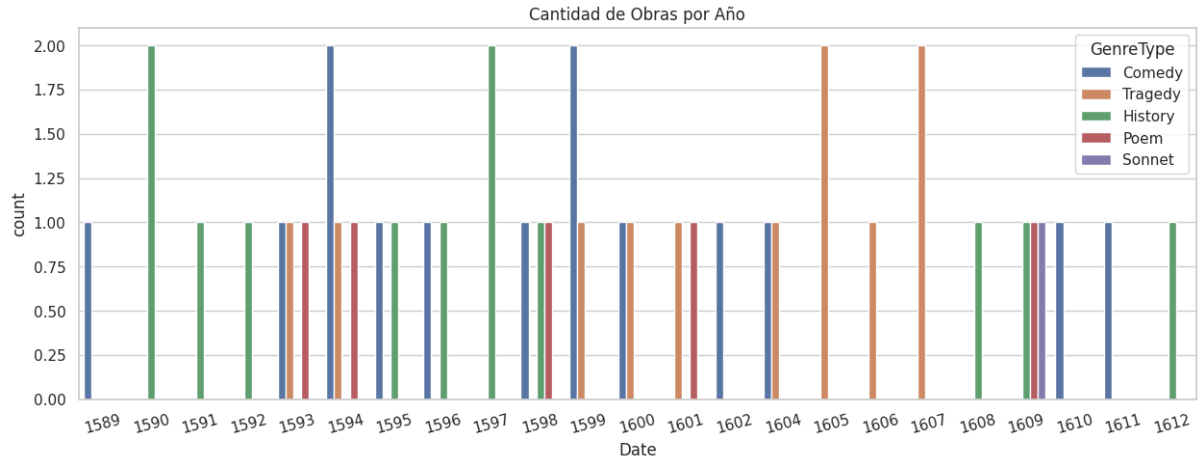


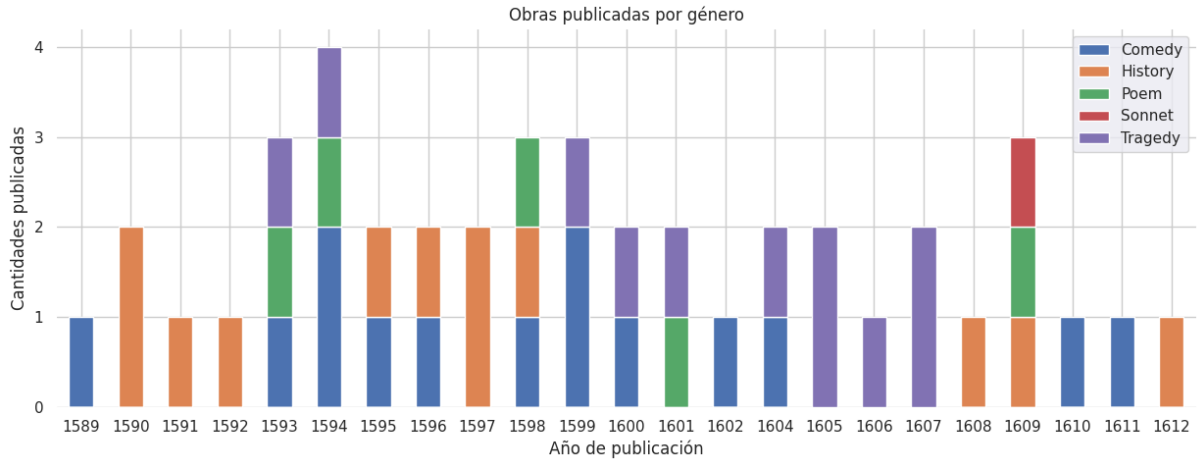
Figura 9: Gráfico de barras comparativo de la cantidad de obras publicadas por William Shakespeare entre 1589 y 1612. El máximo número de trabajos alcanzados fue 4 en 1594, mientras que el mínimo se dio en 1603, no habiendo publicado ninguno.

Otro aspecto interesante que se obtuvo de los datos fue el género literario de sus obras. William Shakespeare se dedicó a escribir en 5 géneros distintos: Comedia, Tragedia, Historia, Poemas y Sonetos. De la totalidad de sus obras conocidas 14 pertenecieron a la comedia, 12 fueron históricas, 11 tragedias, 5 poemas y 1 soneto. En la Figura 10 se pueden apreciar dos visualizaciones de las obras según su género a lo largo de los años.
En dicha imagen, se utilizaron dos abordajes distintos para explorar formas de representar la producción por género de Shakespeare: una que realizaba un histograma por cada género, reuniendo un máximo de 5 barras por año (1 por género), y otra que distinguía lo mismo pero aplicando horizontalmente las barras.
La primera visualización, de la Figura 10a, resulta ventajosa para comprender ya que se indica el género literario de acuerdo a un código de color y todas las barras comienzan desde el cero. Por eso, si una barra finaliza en la altura 2, significa que se publicaron 2 obras ese año bajo esa categoría. Sin embargo, dado que se tienen datos de 23 años de creación literaria, la cantidad de barras dibujadas es muy grande, y el ancho de cada barra, pequeño (no es una visualización escalable al aumentar los años). Adicionalmente, la ausencia de producción literaria para ciertos géneros en varios de los años aumenta la dificultad de separar visualmente a qué año corresponde cada barra.
En tanto, la segunda visualización, de la Figura 10b, es más “limpia” visualmente, ya que hay una sola barra por año, con colores cambiantes según a qué género pertenece cada creación. Además, esta forma de visualizar permite obtener fácilmente el valor total de

obras publicadas para un año dado (se corresponde con la altura total de la barra). No obstante, hay algunos detalles que pueden hacer la lectura menos intuitiva: las barras indicando publicaciones en cierto año se “apilan” en lugar de superponerse, lo que implica que el “cero” de cada barra está por encima del límite de la barra anterior (para el mismo año), por lo que no es lineal extraer la cantidad de publicaciones para un género a simple vista. Por otro lado, puede resultar difícil agrupar visualmente las barras del mismo género, ya que para algunos años éstas comienzan desde el cero y para otros comienzan desde arriba de otras barras. Se considera que ésta no es una visualización que permita escalar en cantidad de géneros literarios.
En general, se cree que si bien ambas visualizaciones necesitan más trabajo, el uso de una u otra depende en gran parte de a qué se quiere hacer referencia al mostrarlas y no hay una que sea intrínsecamente mejor que la otra. Otra visualización que se deseaba explorar, pero no fue posible, fue crear una figura con 5 subfiguras apaisadas (una fila por género) donde cada una tuviera un histograma de un género específico.



(a)



(b)

Figura 10: Gráfico de barras comparativo de la cantidad de obras publicadas por William Shakespeare entre 1589 y 1612 según el género literario al que pertenecen. En (a) puede observarse una visualización a partir de un histograma con las barras de los géneros yuxtapuestas para un mismo año, mientras que en (b) puede verse lo mismo pero con las barras apiladas verticalmente.

Respecto a la producción de trabajos de acuerdo al género literario, parecería haber una mayor concentración de obras de historia y comedia al principio de su creación, con mayor

abundancia de tragedias hacia el final. Sin embargo, estas tendencias no están claramente delineadas (no hay un quiebre entre una y otra).

Conclusiones

Durante el trabajo se utilizaron los datos de la base de datos relacional Open Source Shakespeare para un primer análisis exploratorio de los datos y contenidos de las tablas. Se evaluó el estado de los datos, encontrando algunas entradas faltantes en la tabla de personajes, además de realizarse una limpieza de datos sobre el texto de los diálogos recopilados.
Finalmente, se exploraron diversas visualizaciones que mostraran las distribuciones de las palabras más usadas, los personajes más prominentes o con más líneas, y las características de producción de obras a lo largo de los años. Los gráficos resultantes asistieron en el entendimiento de qué contenían las tablas, y facilitaron la formulación de interrogantes que sería posible explorar mediante herramientas estadísticas o de aprendizaje automático.

Referencias

[1] George Mason University. *Download the source code and OSS database*. (n.d.). Open Source Shakespeare. Disponible en: www.opensourceshakespeare.org.

[2] Información del proyecto Jupyter, https://jupyter.org/.

[3] Repositorio en Gitlab del Notebook implementado. Disponible de forma pública en: https://gitlab.fing.edu.uy/lucia.lemes/introcd.