Tarea 1:
Introducción a la Ciencia de Datos

Mayo 2023

Integrantes:

Federico Matonte (3.832.168-8)

Luciana Vidal (5.147.188-8)

Introducción 2
1. Procesamiento de datos 3
1.1. Carga de datos 4
1.2. Calidad de datos 4
2. Análisis de datos 4
2.1. Personaje con más párrafos 4
2.2 Visualización de la obra a lo largo de los años 5
2.3. Conteo de palabras 7
2.3.1. Normalización del texto 7
2.3.2. Conteo de palabras 7
2.3.3. Personajes con mayor participación 9
2.4. Preguntas a responder 11
Referencias bibliográficas 12

Introducción

El objetivo del presente documento es conectarse a una base de datos relacional y analizar los datos de la misma, realizando la limpieza correspondiente en caso de ser necesario.

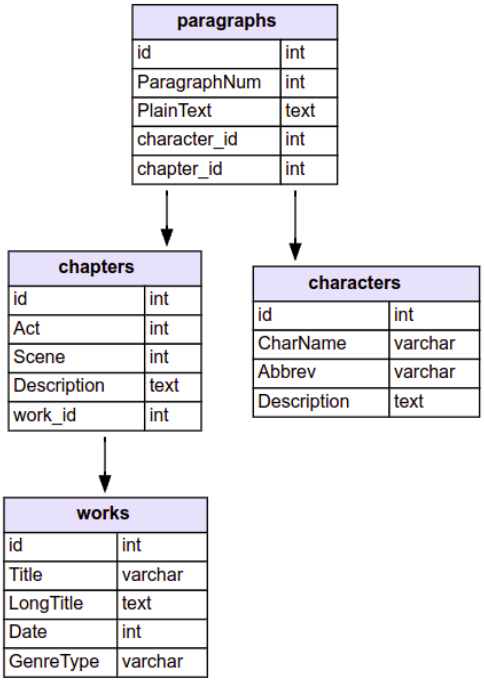
La base de datos contiene las obras de William Shakespeare [1], por lo que el objetivo es utilizar los conocimientos adquiridos en el curso de Introducción a la ciencia de datos para analizar dichos datos. Para esto, se trabajó con un notebook desarrollado en Python, el cual quedó subido en el repositorio público de GitHub llamado introcd_grupo_17 [5].

El documento se organiza de la siguiente forma: en la Sección 1 se explica el procesamiento de datos, detallando la carga y calidad de los mismos, y en la Sección 2 el proceso de análisis de datos: estudiando el personaje con más párrafos, visualizando la obra a lo largo del tiempo, trabajando sobre el conteo de palabras en distintos escenarios, y generando una lista de interrogantes que sería interesante analizar en un trabajo a futuro.

1. Procesamiento de datos

La base de datos a analizar contiene todas las obras del dramaturgo William Shakespeare [1], está formada por cuatro tablas relacionadas entre sí que contienen la información de las obras, el texto de cada una de ellas y los personajes. En la Figura 1 se muestra el esquema de dicha base de datos.

Figura 1. Diagrama de la base de datos de las obras de William Shakespeare [1].



La tabla de párrafos (**`paragraphs`**) contiene el texto de cada párrafo de las obras, indica el personaje que lo interpreta y el capítulo al que pertenece dicho pasaje. Un párrafo pertenece a un capítulo y es interpretado por un único personaje.

La tabla de capítulos (**`chapters`**) contiene la información de cada capítulo: descripción que indica en su mayoría al escenario, el número de acto al que pertenece y el número de escena, así como la referencia a la obra a la que pertenece. Por lo que la tabla posee una jerarquía de pertenencia de tres niveles: empezando por la obra, siguiendo por el acto y finalizando por la escena. Un capítulo pertenece a una única obra y puede estar relacionado a varios párrafos.

La tabla de obras (**`works`**) contiene la información sobre la obra: título, título largo, fecha y género literario. Una obra tiene un único género y puede tener varios capítulos.

Finalmente, la tabla de personajes (**`characters`**) contiene la información sobre el personaje: un nombre, una abreviación y una descripción. Un personaje puede interpretar varios párrafos, por lo que puede estar en varios capítulos y obras.

1.1. Carga de datos

Siguiente el ejemplo provisto por los docentes se utiliza la biblioteca sqlalchemy para crear una conexión a la base de datos Shakespeare [1], alojada en relational.fit.cvut.cz con la que se creará un motor de Base de datos, que resolverá la consulta sql que traerá las tablas para ser almacenadas en variables a utilizar como dataframes.

1.2. Calidad de datos

Para analizar la calidad de los datos se revisó si las tablas tenían datos faltantes.

En la tabla de personajes se descubrió que más de la mitad de los personajes no tienen una descripción (51.029% tienen una descripción vacía). Por otro lado, cinco personajes no tienen abreviatura (0.395).

2. Análisis de datos

2.1. Personaje con más párrafos

Se utiliza como línea base una versión recortada de A LIST of the 20 longest Shakespearean roles incluida en el libro The Book of Lists 2 (1980) obtenida en la pregunta de Quora “Which character has the most lines in a Shakespeare play?” [2].

Tabla 1. Top 10 de los personajes según sus intervenciones.

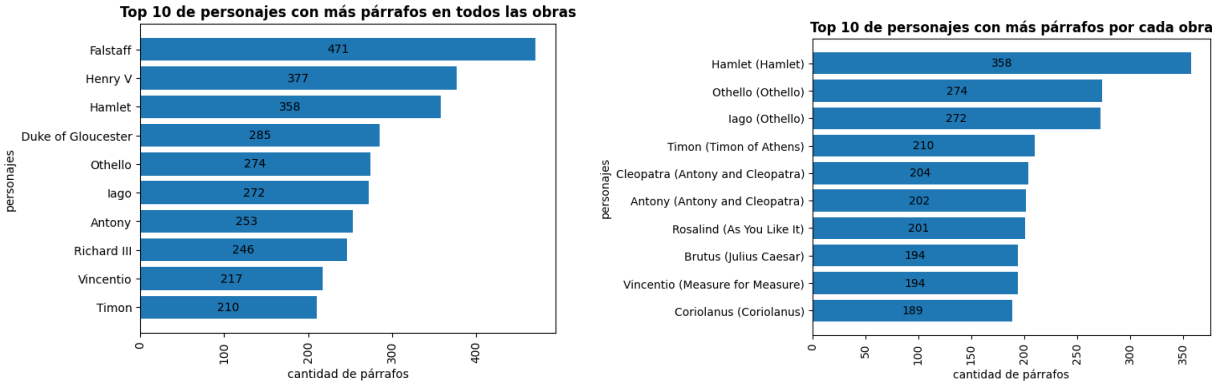
Posición	Personaje	Cantidad de líneas	Obra
1	Hamlet	1,422	Hamlet
2	Falstaff	1,178	Henry IV (parts one and two)
3	Richard III	1,124	Richard III
4	Iago	1,097	Othello
5	Henry V	1,025	Henry V
6	Othello	860	Othello
7	Vincentio	820	Measure for Measure
8	Coriolanus	809	Coriolanus

9	Timon	795	Timon of Athens
10	Antony	766	Antony and Cleopatra.

La pregunta no solo nos brinda una línea base sino también plantea que algunos de los personajes aparecen en más de una obra.

Preparamos entonces 2 listas de los 10 personajes con mayor cantidad de párrafos, considerando en la primera que los personajes puede aparecer en más de una obra (y sus nombre puede identificarlos) y en la segunda que los personajes son locales a las obras (y es necesario tanto sus nombres como los de las obras para identificarlos).

Figura 2. Listas de personajes con más párrafos. A la derecha se hace distinción de los personajes por obra, y a la izquierda no.

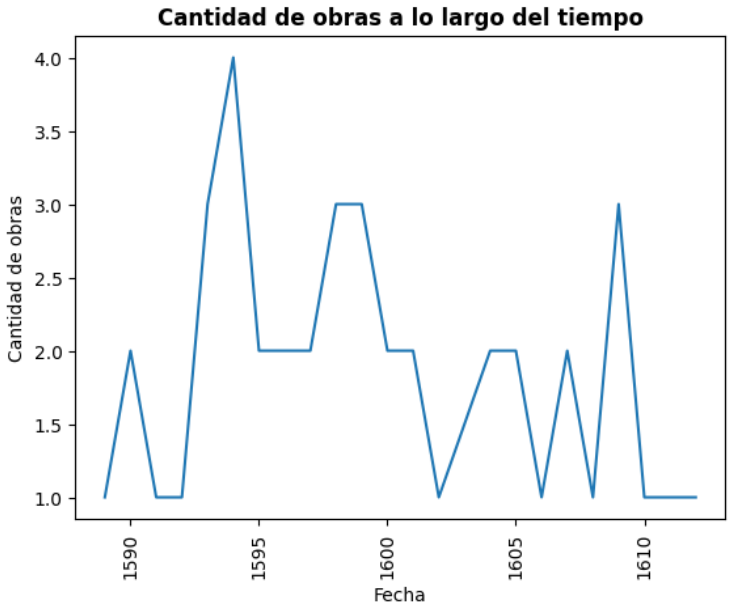


Si consideramos que los personajes son los mismos en distintas obras Falstaff es el personaje con más párrafos, lo que es consistente con el hecho de que aparece en tres obras distintas presentes en el juego de datos: Henry VI Part I, Henry VI Part II y Merry Wives of Windsor. En cambio si diferenciamos los personajes por obras, Hamlet es el personaje con más párrafos en el juego de datos, lo que es consistente con el hecho de que es el personaje principal de la obra más extensa del dramaturgo.

2.2 Visualización de la obra a lo largo de los años

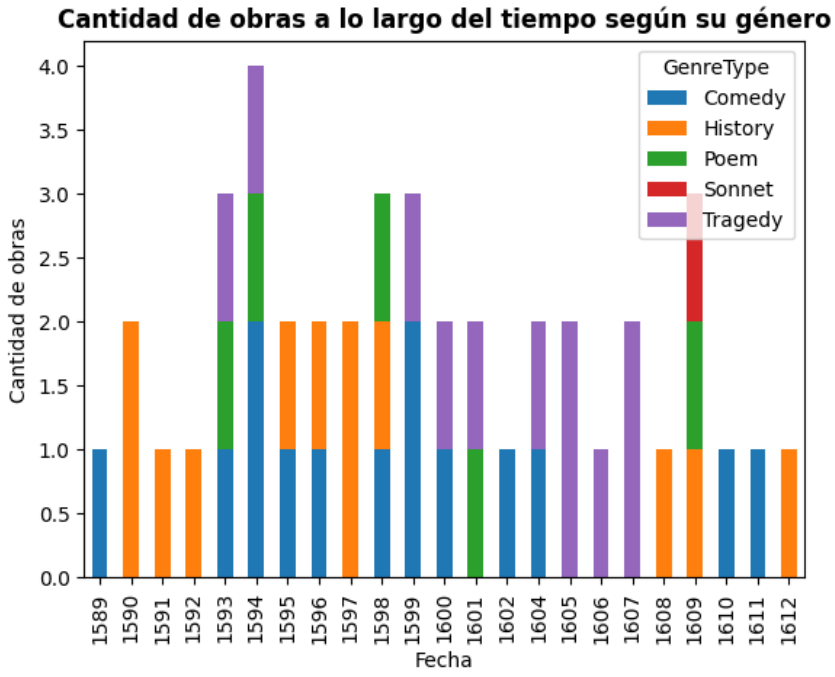
Como primer acercamiento a la información graficamos la cantidad de obras a lo largo de los años, pudiendo observarse que desde 1589 hasta 1612 el artista produjo obras de forma ininterrumpida, pero variando su productividad.

Figura 3. Gráfica mostrando la cantidad de obras a lo largo del tiempo.



Buscando algo que explique la variación se vuelve a graficar la cantidad de obras a lo largo del tiempo, pero esta vez como un gráfico de barras. Pudiendo notar períodos de mayor interés en ciertos géneros, como el que va de 1590 a 1592 en que sólo compuso ficción histórica, o el de 1604 a 1607 en que seis de las siete obras producidas fueron tragedias.

Figura 4. Gráfica sobre la cantidad de obras a lo largo del tiempo según su género literario.



2.3. Conteo de palabras

2.3.1. Normalización del texto

Antes de proceder a realizar el conteo de palabras, se debió normalizar el texto, por ejemplo para que una palabra en mayúscula representa lo mismo que en minúscula. Sobre el mismo se hicieron distintas transformaciones:

- Conversión de letras a minúscula.
- Eliminar los indicadores de escena, y aquellas frases que estén dentro de las misma dado que es una descripción utilizada para explicarle al lector sobre el contexto en el que se desarrolla la escena.
- Reemplazar los signos de puntuación, por espacios: ("\n", ",", ";", ".", "?", "!", ":", "--", "\"", "(", ")","&c" y "[","]").
- Expandir las palabras contraídas con una pequeña lista de ejemplos.
- Eliminamos las contracciones no presentes en la línea y su apóstrofe.
- Eliminar las palabras denominadas stop words dado que no aportan un significado relevante, para así enfocarse en las palabras que son importantes.

2.3.2. Conteo de palabras

Con el fin de analizar cuáles son las palabras más utilizadas en las obras de Shakespeare, se graficaron las diez más mencionadas en la Figura 5 y se pudo constatar que son las denominadas stop words (no aportan significado por sí mismas).

En la Figura 6 se eliminaron las stopwords obtenidas de las listas de NLTK [3] y se graficó el top 10, y el fin de mejorar la visualización en la Figura 7 se graficó una nube de palabras (en la que el tamaño de la palabra representa la cantidad de ocurrencias).

Figura 5. Top 10 de las palabras más utilizadas en las obras.

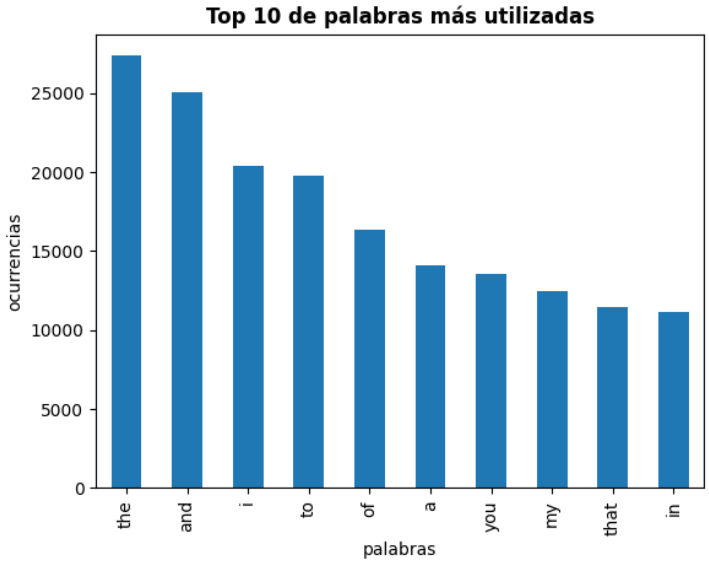


Figura 6. Top 10 de las palabras más utilizadas en las obras, eliminando las stop words.

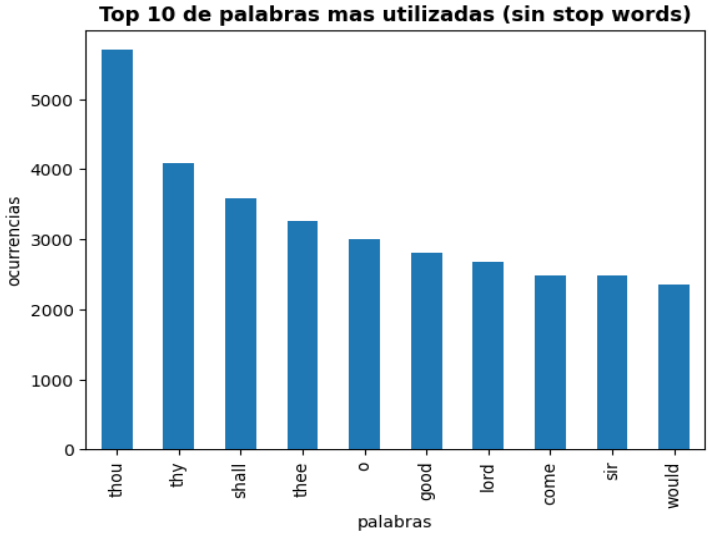
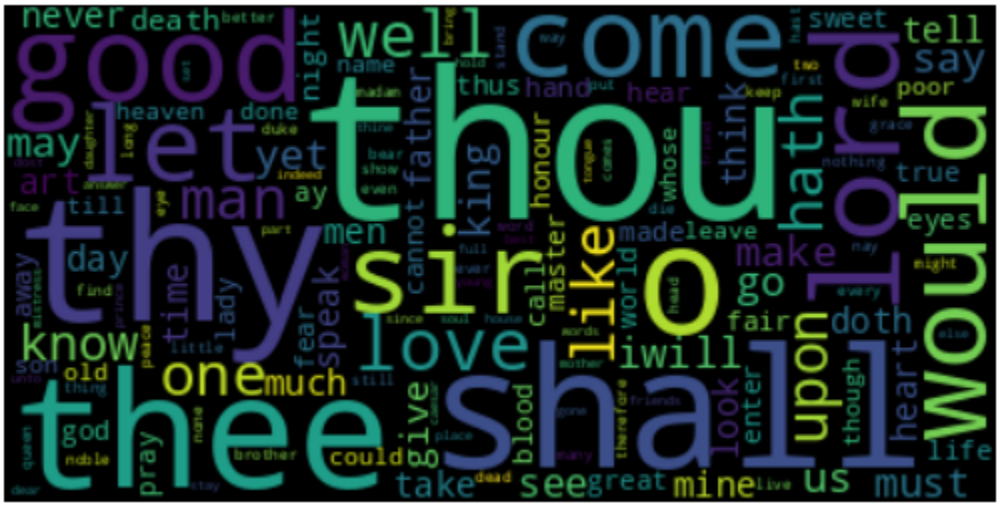


Figura 7. Nube de palabras con las más utilizadas en las obras, eliminando las stop words.



Buscando encontrar diferencias entre géneros o personajes, se podría utilizar una gráfica de barras apiladas con las palabras en el eje de las x, la cantidad de ocurrencias de una palabra en el eje de las y, usando los géneros o personajes para separar la cantidad de personajes (colores en las barras).
Una opción más refinada sería establecer que cada sujeto (personaje o género) es representado como un vector en n dimensiones, con n la cantidad de palabras, pero al no poder graficar en tantas dimensiones se utilizará un algoritmo de reducción de dimensiones (por ejemplo PCR) para así convertir esta presentación en un vector que sí pueda ser graficado: en dos dimensiones sería graficarlos como puntos; o en cuatro dimensiones utilizar dos coordenadas, una para el radio (haciendo que un círculo sustituya al punto) y una cuarta dimensión que represente un color en una paleta de colores.

2.3.3. Personajes con mayor participación

Al analizar los personajes con mayor cantidad de palabras, se pudo constatar que hay párrafos dichos por un personaje llamado “Poet” que hace referencia a la representación del yo lírico y por lo tanto no es un personaje. Por otro lado, también se encontró indicaciones de dirección (personaje: “(stage directions)”), que nuevamente son palabras que no corresponden a un personaje.

Finalmente, eliminando al yo lírico y a las indicaciones, se observó que hay personajes que tienen diálogos en más de una obra por lo que su participación es mayor aunque

no necesariamente lo es dentro de cada obra. En la Figura 8 se graficaron los 10 personajes con más palabras inter obra y en la Figura 9 los de intra obra.

Figura 8. Top 10 de los personajes con mayor cantidad de palabras inter obra.

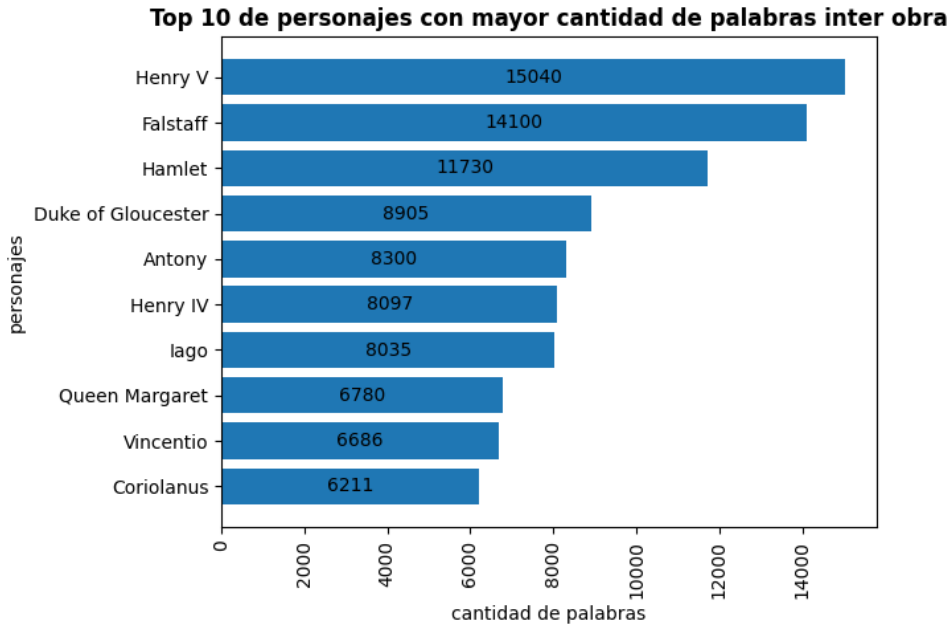
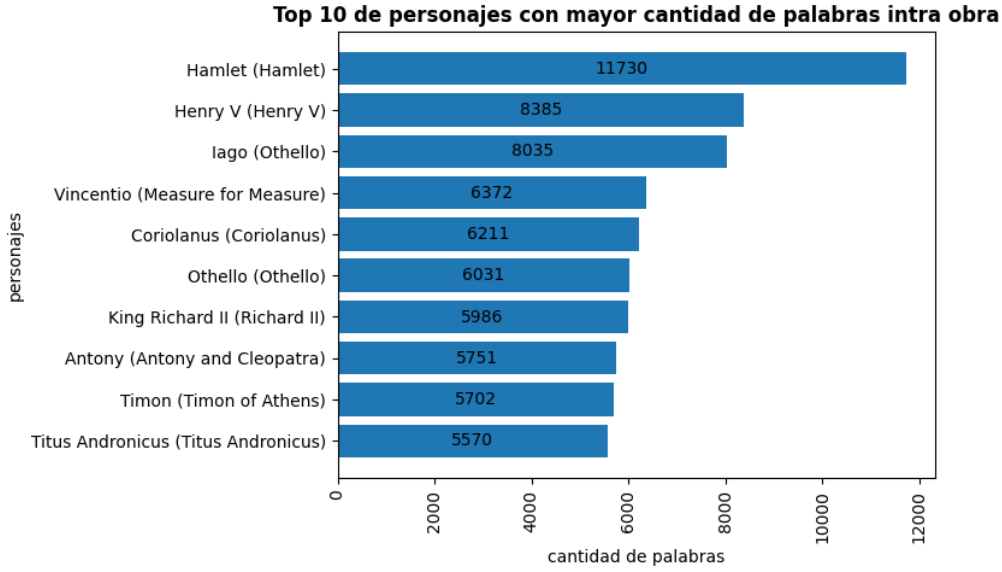


Figura 9. Top 10 de los personajes con mayor cantidad de palabras intra obra.



Analizando ambas gráficas, se constata que si bien un personaje puede tener mayor cantidad de palabras a los largo de todas las obras, esto no asegura que su participación intra obra también sea mayor, este es el caso de Henry V que si bien en el análisis inter obra está primero, si se lo analiza a lo largo de las obras en las que

participa no está en la misma posición. El personaje aparece en tres obras, pero únicamente en una está como personaje con mayor cantidad de palabras, esto se puede observar en la Tabla 2.

Tabla 2. Ranking de las apariciones de Henry V en sus obras.

Obra	Ranking
Henry IV, Part I	2
Henry IV, Part II	3
Henry V	1

2.4. Preguntas a responder

A continuación se listan preguntas que podrían ser interesantes para responder:

- Buscar frases repetidas: ¿es posible detectar frases repetitivas (conjuntos ordenados de palabras) que se utilicen de forma regular?, ¿algunos personajes tienen mayor predisposición a hablar con frases prearmadas que otros?, ¿algunos géneros son más propensos a usar frases repetidas?
- Buscar arquetipos de personajes: ¿es posible entrenar un clasificador de arquetipos de personajes utilizando este corpus?, ¿qué tipos de arquetipos de personajes se encuentran más presentes en sus obras?, ¿la capacidad de detectar arquetipos es local al género literario o a modo de ejemplo es posible entrenar un clasificador de arquetipos con comedias y que sea bueno detectando arquetipos en tragedias?
- Predecir características: ¿es posible entrenar un clasificador de personajes que prediga el sexo o la edad de estos a partir de sus líneas?
- Test de Bechdel: ¿dada la estructura de los datos es posible encontrar si una obra cumple con el Test de Bechdel [4] (al menos dos personajes femeninos, con nombres, que hablan entre sí de un tema que no sea un hombre)?
- Palíndromos: ¿qué palíndromos (cadena de palabras o letras simétricas, que pueden ser leídos igual de derecha a izquierda o izquierda a derecha) se encuentran presentes en las obras?
- Predictor de muertes: ¿es posible entrenar un clasificador que utilizando las primeras líneas de un personaje prediga si este se muere o no?, ¿cuántas líneas son necesarias para que el predictor sea eficiente?, ¿cuán dependiente del género literario es este clasificador?

Referencias bibliográficas

[1] Sitio web Relational Database Repository, accedido por última vez 21/05/2023.
https://relational.fit.cvut.cz/dataset/Shakespeare

[2] Sitio web Quora, accedido por última vez 21/05/2023.
https://www.quora.com/Which-character-has-the-most-lines-in-a-Shakespeare-play

[3] Sitio web GitHub, accedido por última vez 21/05/2023.
https://gist.github.com/sebleier/554280

[4] Test de Bechdel, accedido por última vez 21/05/2023.
https://feministfrequency.com/video/the-bechdel-test-for-women-in-movies/

[5] Sitio web de GitHub, accedido por última vez 21/05/2023.
https://github.com/lupish/introcd_grupo_17