Introducción a la Ciencia de Datos

Edición 2024

Informe de Laboratorio 1

v1.0.0
21 de mayo 2024

Grupo 15
Matias Rolando
Emiliano Viotti

Índice

1. Introducción

William Shakespeare, nació el 23 de abril de 1564 en Stratford-upon-Avon, Inglaterra y en lo que hoy en día parecería una vida corta (52 años), se transformó en una figura titánica del mundo de la literatura. Este dramaturgo y poeta inglés dejó un legado imborrable con sus más de 39 obras literarias, existen al menos dos corrientes que discuten incluso hoy en día la atribución de ciertas obras, entre entre las que se destacan sus tragedias y comedias, obras como "Hamlet", "Romeo y Julieta" y "El rey Lear". Ya sea si has leído alguna obra de William Shakespeare o no, es muy probable que reconozcas algunas frases con origen en su obra como "Ser o no ser, esa es la cuestión" o "El amor es un humo hecho con el vapor de suspiros". Estas líneas no solo demuestran su maestría lingüística, sino que también reflejan las intrigas universales sobre el amor, el poder y la tragedia, manteniendo su relevancia a través de los siglos.

En este trabajo llevado adelante en el contexto del primer laboratorio del curso Introducción a la Ciencia de Datos de la Facultad de Ingeniería, UdelaR, edición 2024 [2], nos proponemos adentrarnos en la obra de William Shakespeare con un enfoque de ciencia de datos, analizando sus principales obras utilizando algunas técnicas sencillas de análisis de datos.

Esperamos que disfrutes este viaje a través de los datos, el tiempo y principalmente, de la lengua inglesa, tanto como nosotros lo hemos disfrutado.

"*Ten más de lo que muestras habla menos de lo que sabes."*

William Shakespeare

2. Metodología

Para este trabajo de investigación tomamos como referencia las etapas de la metodología CRSIP-DM [11], mencionadas en el curso. Vale la pena aclarar que no seguimos al detalle la metodología ya que algunas etapas no aplicaban al contexto de este trabajo. En particular nos basamos fuertemente en: i. entendimiento de los datos, ii. procesamiento de los datos y ii. modelado, etapa que en nuestro caso reformulamos como análisis descriptivo en lugar de análisis predictivo.

1. Iniciamos nuestro viaje entendiendo los datos para lo cual primero nos hicimos de los mismos mediante la descarga desde el sitio web. A su vez, implementamos funciones auxiliares para leer los datos localmente desde archivos .csv. En la sección Obtención de los Datos se ahonda sobre este paso.
2. Una vez nos hicimos con los datos emprendimos la segunda etapa de la metodología para lo cual nos apoyamos en una combinación de inspección manual de los datos, un EDA (Exploratory Data Analysis) implementado con Python y Pandas y las preguntas enunciadas en la pauta de este laboratorio. Los detalles se presentan en la sección Entendimiento de los Datos.
3. Apoyándonos en los resultados de la etapa anterior y teniendo presente los objetivos de análisis, iniciamos la etapa de Preparación de los Datos. Aquí realizamos diferentes modificaciones a los mismos entre los que se incluyen normalizaciones, filtrados y diferentes pre-procesamientos para mejorar los resultados del análisis. Se pueden ver todos los detalles en la sección Procesamiento de los Datos.
4. Una vez concluído el procesamiento de los datos, continuamos con el análisis de los mismos. Para esto nos planteamos los objetivos descritos en la pauta del laboratorio, tales como analizar la evolución de la obra de Shakespeare a través de los años, determinar las palabras más frecuentes, etc. Los detalles se pueden encontrar en la sección Análisis.
5. Finalmente, a modo de conclusión y resumen del trabajo llevado adelante, formulamos diferentes conclusiones a partir de los datos y los análisis realizados. Las mismas se pueden encontrar en la sección Conclusiones.

A nivel tecnológico nos apoyamos en el lenguaje Python para codificar y en herramientas y bibliotecas del ecosistema Python como Pandas para la carga, transformación y agregación de los datos, Matplotlib y Plotly para realizar gráficos y visualizaciones, spacy para utilizar técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN), entre otras. La lista completa de dependencias se puede encontrar en el archivo *requirements.txt* dentro del repositorio del proyecto.

El código fuente del proyecto se encuentra público y disponible en un repositorio Git (ver [3]). Siguiendo buenas prácticas de código open-source, el proyecto cuenta con un archivo README.md que profundiza en aspectos técnicos no cubiertos por este documento como los pasos para la creación de un ambiente virtual para ejecutar el análisis realizado, entre otros.

Finalmente, se realizaron varios experimentos y análisis que se pueden encontrar como notebooks (archivos *.ipynb*) en el repositorio (carpeta notebooks). Dicho esto, a los efectos de brindar un análisis completo y consolidado, se generó un Jupyter Notebook de nombre `laboratorio_1.ipynb` en la raíz del repositorio que utilizando como hilo conductor la misma estructura presente en este documento.
Este notebook contiene todo el código que implementa los análisis, pre-procesamientos y diferentes visualizaciones incluidas aquí. Se recomienda altamente la lectura del mismo a modo de complemento, ya que este documento presenta un resumen de toda la información allí disponible.

3. Obtención de los Datos

La etapa de obtención de datos se desarrolló sin grandes inconvenientes ya que buena parte de las funciones para acceder a los datos, fueron proporcionadas por el equipo docente. Los únicos inconvenientes que merecen ser mencionados son:

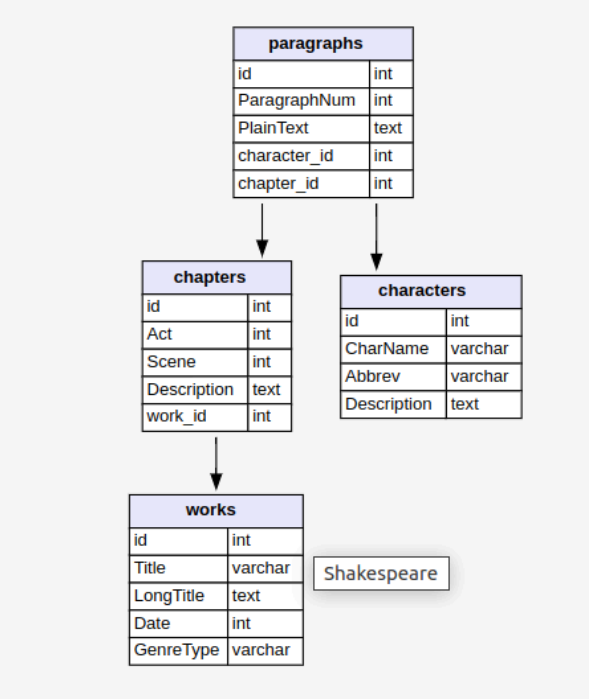
1. El link que figuraba en la pauta del laboratorio se encontraba “roto”, pero lo solucionamos rápidamente cambiandolo por este nuevo link.
2. Las versiones más recientes de Pandas inicializan de forma diferente la conexión a la base de datos, utilizando SQLAlchemy. Como en el archivo requirements.txt proporcionado por los docentes las versiones no se encuentran “fijas”, se descarga la versión más reciente de Pandas y el código proporcionado por los docentes no funciona correctamente. Con un poco de ayuda de la comunidad encontramos como solucionarlo rápidamente.

A nivel técnico, para la obtención de los datos nos conectamos directamente a la base de datos pública, para leer los datos directamente de 4 tablas: works, chapters, paragraphs y characters. Realizamos un “full-scan” en cada una de las tablas (`SELECT * …`) y cargamos los datos en un data frame, por último cada data frame se guarda localmente en un archivo .csv a los efectos de permitir posteriores cargas de datos desde los archivos locales. Para esto utilizamos las bibliotecas SQLAlchemy y Pandas. Este proceso se encuentra implementado en la función `load_table()`. A su vez, orquestamos la descarga de las cuatro tablas en la función `load_dataframes()`.

4. Entendimiento de los Datos

Como hemos mencionado anteriormente, los datos de este trabajo se corresponden a las obras de William Shakespeare, incluyendo para cada obra su título, la fecha de publicación, el género, así como la obra en sí misma: capítulos, párrafos e información de los personajes. En resumen el dataset contiene las obras de Shakespeare e información relevante asociada a las mismas (o metadata) que facilitarán nuestro análisis.

Los datos se encuentran distribuídos en cuatro tablas: works, chapters, paragraphs y characters. El modelo físico de tablas puede verse en la imágen a continuación.



Modelo físico de la BD - imágen de relational-data.org

EDA

Existen muchas técnicas para llevar adelante un análisis exploratorio de datos, algunas dentro de la intuición y otras de una naturaleza más estadística. En este trabajo, nos vamos a limitar a ejecutar algunos análisis básicos e intuitivos para principalmente ganar mayor conocimiento sobre los datos y validar algunas hipótesis sobre la calidad de los mismos. A su vez, nos vamos a apoyar en la librería ydata-profiling para realizar de forma rápida algunos de estos análisis.

Sobre el Análisis:

Para cada tabla/dataframe vamos a ejecutar principalmente el mismos análisis:

1. Vistaso rápido de los datos
2. Revisión de tipos
3. Características macro
4. Conteo de missing-values
5. Conteo de duplicados
6. Revisión de valores inválidos

Todos los detalles de estos análisis así como el código que los implementa se puede encontrar en el notebook que acompaña este análisis. Además, en la carpeta `assets/reports/` se encuentran archivos `.html` con el reporte generado utilizando la biblioteca ydata-profiling.

Obras (Works)

En esta tabla se encuentra la información básica de una obra como Título, Título Extendido, Fecha de publicación y Género. No incluye el contenido de la obra en sí. Además, notar que cada obra tiene un número de identificación único. Por otro lado la fecha de publicación de la misma se reduce al año y por último el género es un valor categórico que varía de obra en obra entre los valores: *Comedy, Tragedy, History, Poem, Sonne*t

Veamos una muestra de los datos de esta tabla para fijar ideas:

	id	Title	LongTitle	Date	GenreType
0	1	Twelfth Night	Twelfth Night, Or What You Will	1599	Comedy
1	2	All's Well That Ends Well	All's Well That Ends Well	1602	Comedy
2	3	Antony and Cleopatra	Antony and Cleopatra	1606	Tragedy
3	4	As You Like It	As You Like It	1599	Comedy
4	5	Comedy of Errors	The Comedy of Errors	1589	Comedy

A continuación, resumimos las principales características de esta tabla, obtenidas a partir de diferentes consultas realizadas sobre los datos (ver notebook por detalles técnicos).

	id	Title	LongTitle	Date	GenreType
Tipo	int64	object	object	int64	object
# Missing Values	0	0	0	0	0
# Duplicados	0	0	0	20	38
# Rows	43				
# Duplicados Row	0				

Notar que un “missing value” puede corresponderse a diferentes valores dependiendo del tipo de la columna e incluso a veces de la semántica del orígen del dato. Por ejemplo en una columna numérica u object, un missing value puede ser `NaN` ya que pandas, automáticamente convierte los valores `None` a `NaN`. Por otro lado, una columna string podría tener el valor string vacío (“”) como un missing value también. A su vez, en una columna numérica que sólo toma valores positivos `-1` podría representar un “missing value” también. Contar los missing value es muy importante ya que nos dice qué proporción de los datos disponibles son realmente útiles para buena parte de los análisis que queremos realizar. En el notebook, implementamos la función `count_empty_values()` que se encarga de contar valores faltantes usando diferentes mapeos entre tipos de columna y valores faltantes.

Por otro lado, contamos duplicados a nivel de fila (combinación de todas las columnas), ya que es un error bastante común en datasets de poca calidad y suele estropear cualquier

análisis cuantitativo. En este caso no hay duplicados. También contamos duplicados a nivel de cada columna, para empezar a obtener mayores insights sobre los datos.

A continuación resumimos las principales observaciones.

Observaciones

- Los tipos de datos inferidos por pandas son correctos para procesar los datos
- No vemos problemas grandes de calidad de datos:
  - No hay filas enteramente duplicadas
  - La columna `id` identifica correctamente a la obra (valores únicos para la columna)
  - Tanto `Title` como `LongTitle` presenta valores únicos en la columna, lo cual descarta por ejemplo una misma obra presente con diferente `id`
  - No hay missing values en ninguna columna
- Además, todas las obras tienen fecha entre 1589 - 1612 (fechas comprendidas en el período en que vivió William Shakespeare)

Por último, vale la pena mencionar que por tratarse de un autor que vivió hace 400 años, existen corrientes divergentes sobre la atribución de algunas de sus obras (teoría Oxfordiana[12]). Sin embargo, escapa al alcance de este trabajo, determinar si todas las obras de Shakespeare se encuentran en este dataset así como validar si los datos son correctos. Nos restringimos a asumir que los datos son correctos y sacar conclusiones en base a este dataset sobre la obra de Shakespeare.

Capítulos (Chapters)

En esta tabla se encuentra la información básica del capítulo de una obra como Acto, Escena, Descripción y una referencia a la obra a la que pertenece. Al igual que en la tabla anterior, esta tabla no incluye el contenido perse del capítulo a excepción de la descripción. Por otro lado, la referencia a la obra nos permite cruzar la información de esta tabla con la tabla anterior. Por último, cada capítulo tiene un id que lo identifica de forma única.

Veamos una muestra de los datos de esta tabla para fijar ideas:

	id	Act	Scene	Description	work_id
0	18704	1	1	DUKE ORSINO’s palace.	1
1	18705	1	2	The sea-coast.	1
2	18706	1	3	OLIVIA’S house.	1
3	18707	1	4	DUKE ORSINO’s palace.	1
4	18708	1	5	OLIVIA’S house.	1

A continuación, resumimos las principales características de esta tabla, obtenidas a partir de diferentes consultas realizadas sobre los datos (ver notebook por detalles técnicos).

	id	Act	Scene	Description	work_id
Tipo	int64	int64	int64	object	int64
# Missing Values	0	0	0	0	0
# Duplicados	0	939	790	426	902
# Rows	945				
# Duplicados Row	0				

Observaciones

- Los tipos de datos inferidos por pandas son correctos para procesar los datos
- No vemos problemas grandes de calidad de datos:
  - No hay filas enteramente duplicadas
  - La columna id identifica correctamente al capítulo
  - No hay missing values en ninguna columna
- Todos los capítulos referencian a una obra (`work_id`) válida en la tabla de obras (works)
- Vemos duplicados a nivel de Acto (`Act`) y Escena (`Scene`) pero esto podría ser porque casi todas las obras tienen actos 1, 2, 3..
- Por otro lado, vemos duplicados a nivel de la Descripción (`Description`) y esto seguramente se deba a casos como las filas 2 y 4 en la muestra de datos, donde en el acto 1 las escenas 3 y 5 se desarrollan en el mismo lugar: *"OLIVIA'S house."*

Párrafos (Paragraphs)

En esta tabla se encuentran los párrafos de una obra (`PlainText`), así como información adicional de los mismos entre los que encontramos: un número que identifica al párrafo (id), un número de secuencia que indica la posición del párrafo dentro de la obra (`ParagraphNum`). Además cada párrafo está asociado a un capítulo particular de una obra de Shakespeare a partir de la columna `chapter_id`. También cada párrafo está asociado a un personaje a partir de la columna `character_id`.

Veamos una muestra de los datos de esta tabla para fijar ideas:

	id	ParagraphNum	PlainText	character_id	chapter_id
0	630863	3	[Enter DUKE ORSINO, CURIO, and other Lords; Mu...	1261	18704
1	630864	4	If music be the food of love, play on;\nGive m...	840	18704
2	630865	19	Will you go hunt, my lord?	297	18704
3	630866	20	What, Curio?	840	18704
4	630867	21	The hart.	297	18704

A continuación, resumimos las principales características de esta tabla, obtenidas a partir de diferentes consultas realizadas sobre los datos (ver notebook por detalles técnicos).

	id	ParagraphNum	PlainText	character_id	chapter_id
Tipo	int64	int64	object	object	int64
# Missing Values	0	0	0	0	0
# Duplicados	0	31495	1849	34245	34520
# Rows	35465				
# Duplicados Row	0				

Observaciones

- Los tipos de datos inferidos por pandas son correctos para procesar los datos
- No vemos problemas de calidad de datos relevantes:
  - No hay filas enteramente duplicadas
  - La columna id identifica correctamente al párrafo
  - No hay missing values en ninguna columna
- Todos los párrafos referencian a una capítulo (`chapter_id`) válida en la tabla de capítulos (chapters)
- Todos los párrafos referencian a un personaje (`character_id`) válido en la tabla de personajes (chapters)
- Llama la atención la repetición a nivel de PlainText.

Si observamos los párrafos en conjunto con los capítulos, podemos ver que hay varias repeticiones que son interesantes de observar en profundidad.

	chapter_id	PlainText	count
240	18708	[Exit]	7
21981	19220	[Exit]	6
28341	19370	[Exit]	6
8777	18901	[Within] Francis!	5
6618	18863	[Exit]	5

Para un mismo capítulo (mismo `chapter_id`) podemos ver varios párrafos que son similares y de la tabla anterior que no es más que un conteo de parejas `<chapter_id, PlainText>` repetidos ordenados descendentemente, podemos ver que no son cualquier tipo de párrafo. Son indicaciones de escena, en particular para que el personaje deje el escenario. Más adelante en el análisis veremos que este tipo de indicaciones tienen un impacto bien interesante en nuestro análisis.

En principio podemos asumir que tiene sentido entonces, encontrar repetición a nivel de la columna `PlainText` de esta tabla.

Personajes (Characters)

En esta tabla se encuentra la información básica sobre los personajes en las diferentes obras, como Nombre (`CharName`), Abreviación (`Abbrev`) y Descripción (`Description`). Además, todo personaje tiene un número de identificación único. Notar además que los personajes se relacionan directamente con los párrafos pero la referencia se encuentra presente en la tabla de párrafos y no en esta tabla.

Veamos una muestra de los datos de esta tabla para fijar ideas:

	id	CharName	Abbrev	Description
0	1	First Apparition	First Apparition	NaN
1	2	First Citizen	First Citizen	NaN
2	3	First Conspirator	First Conspirator	NaN
3	4	First Gentleman	First Gentleman	NaN
4	5	First Goth	First Goth	NaN

A continuación, resumimos las principales características de esta tabla, obtenidas a partir de diferentes consultas realizadas sobre los datos (ver notebook por detalles técnicos).

	id	CharName	Abbrev	Description
Tipo	int64	object	object	object
# Missing Values	0	0	5	646
# Duplicados	0	309	302	799
# Rows	1266			
# Duplicados Row	0			

Observaciones

- Los tipos de datos inferidos por pandas son correctos para procesar los datos
- No hay filas enteramente duplicadas
- La columna id identifica correctamente al personaje
- Para la cantidad bastante pequeña de obras analizadas (43) llama la atención la cantidad de personajes (1266) con un promedio de 29 personajes por obra.
- Encontramos algunos problemas de calidad:
    - Hay 5 personajes sin abreviación
    - Hay 646 personajes sin descripción ~ 51%
    - Hay 309 valores duplicados en la columna `CharName` ~ 24.4%

La cantidad elevada de missing values en la columna Description es una advertencia fuerte para cualquier tipo de análisis que queramos realizar posteriormente utilizando la misma, ya que contamos con datos solamente para un 49% de los personajes.
Por otro lado, la cantidad de filas repetidas en las columnas `CharName` y `Abbrev` también es una advertencia fuerte para cualquier tipo de análisis. Exploremos un poco más en profundidad este fenómeno, agrupando las filas por la columna `CharName:`

	id	CharName	Abbrev	Description	count
67	68	All	All	NaN	23
778	779	Messenger	Mess	NaN	23
768	769	Messenger	Messenger	NaN	23
85	86	All	ALL	NaN	23
772	773	Messenger	MESSENGER	NaN	23
1048	1049	Servant	Servant	servant to Diomedes	21
1059	1060	Servant	SERVANT	NaN	21
680	681	Lord	Lord	NaN	9
675	676	Lord	LORD	NaN	9
848	849	Page	PAGE	to Falstaff	8

Como podemos ver en la tabla anterior hay varios personajes especiales que cumplen un rol más bien genérico en una obra como: Messenger (mensajero), Servant (sirviente) o Lord. Es esperable que estos personajes a su vez se repitan tanto dentro de la obra (por ejemplo múltiples mensajeros anónimos) como entre obras.

Otro de los personajes que llama la atención es el de All que a su vez aparece dos veces en el ranking: id=68 e id=86. Esto tiene una explicación en la propia obra de Shakespeare. En varias obras hay diálogos que son recitados por múltiples personajes en la escena. Tal es el caso por ejemplo, del siguiente diálogo extraído de la obra Macbeth:

“(...) Fair is foul and foul is fair: Hover through the fog and filthy air. (...)”

Como podemos ver en el siguiente fragmento de la obra original disponible en Google Books, el diálogo se corresponde a tres brujas en escena y la dirección de escena “All” para que las tres brujas repitan al mismo tiempo la frase, termina reflejándose en esta tabla como que el párrafo pertenece a un personaje de nombre “All”. Esta definición tomada quizás por quienes armaron este dataset, es bien interesante al momento de analizar la relación entre personajes y párrafos en la obra.

ACT I.

SCENE I. *A desert place.*

Thunder and lightning. Enter three Witches.

First Witch. When shall we three meet again
In thunder, lightning, or in rain?
Second Witch. When the hurlyburly's done,
When the battle's lost and won.
Third Witch. That will be ere the set of sun.
First Witch. Where the place?
Second Witch. Upon the heath.
Third Witch. There to meet with Macbeth.
First Witch. I come, Graymalkin!
Second Witch. Paddock calls.
Third Witch. Anon. 10

B

Extracto 1 de la obra Macbeth - Google Books

2 *MACBETH.*

All. Fair is foul, and foul is fair:
Hover through the fog and filthy air. [*Exeunt.*

Extracto 2 de la obra Macbeth - Google Books

5. Procesamiento de los Datos

De la sección anterior no surgen problemas de calidad de datos relevantes que impliquen la implementación de pre-procesamientos específicos para solventar problemas como descartar columnas, imputar datos automáticamente o transformaciones importantes para normalizar los datos. A continuación se explican las pocas transformaciones que implementamos en función de las cosas que detectamos y los análisis que realizamos.

Normalización de PlainText

Para obtener resultados más precisos en los análisis que vamos a realizar más adelante, es necesario normalizar las palabras de cada obra, de forma que por ejemplo: “The” sea la misma palabra que “the” o “THE”. Notar que en algunos casos de procesamiento de lenguaje natural esto puede ser algo no deseado. Sin embargo, en nuestro caso vamos a contar las frecuencias de palabras en la obra de Shakespeare o encontrar los personajes más relevantes, por lo que nos interesa tratar estas variantes como una sola palabra (normalizar valores).

Para esta tarea utilizamos la función `clean_text()` implementada por el equipo docente, a la que agregamos separadores adicionales. La lista completa de separadores que utilizamos es la siguiente:

```
["[", "\n", ",", ":", ";", ".", "]", "(", ")", "?", "!", "'", "-",
"\"", "{", "}"]
```

Por otro lado, eliminamos espacios en blanco innecesarios al principio y al final de cada párrafo con la función `strip()`, además de pasar todo a minúsculas con la función `lower()`. Ambas funciones de la biblioteca `strings` de Python.

Por último, a partir de la columna `PlainText` de paragraphs, obtenemos las palabras normalizadas separando el texto resultado de invocar a la función `clean_text()` mediante la función `split()` también de strings. Esta función separa un texto en tokens utilizando por defecto el espacio en blanco como separador.

Remoción de Stopwords

Para el análisis de frecuencia de palabras, nos interesa centrarnos en palabras que de alguna forma describen a la obra de William Shakespeare. Sin embargo, todos los idiomas tienen palabras frecuentes que se usan para armar sentencias y que no caracterizan a ningún texto. A estas palabras se las suele denominar “stopwords”

Como parte del preprocesamiento previo al análisis de palabras frecuentes, eliminamos de la lista de palabras de la obra de Shakespeare, las stopwords del idioma inglés. Para esto nos apoyamos en la biblioteca de NLP Spacy que tiene una lista de stopwords para cada

modelo de lenguaje. En particular para este trabajo utilizamos el modelo `en_core_web_sm`. La lista completa de modelos de lenguaje se puede consultar aquí.

En particular este modelo incluye 326 stopwords entre las que se incluyen:

```
while', 'herein', 'make', 'anyway', 'together', 'down', 'throughout',
'our', 'after', 'among', 'me', 'though', 'whom', ''d', 'least', …
```

La implementación de la remoción de stopwords se puede encontrar en el notebook de este trabajo. Por otro lado, vale la pena mencionar que el modelo `en_core_web_sm` es un modelo contemporáneo para el idioma inglés. Por lo tanto, puede que algunas de las palabras propias del lenguaje en la época de Shakespeare, no se encuentren entre las mismas.

6. Análisis

Obras de Shakespeare a Través de los Años

Uno de los primeros análisis realizados sobre los datos, siguiendo la letra del laboratorio punto 1.B, es el estudio de la obra de Shakespeare a través de los años. Para esto trabajamos con los datos de la tabla works limitándose a estudiar tendencias y características en la producción del autor de forma anual. Para esto agrupamos los datos de works por la columna `Date` sumarizando la cantidad de obras escritas para cada año. A su vez, podemos agruparlas por género (columna `GenreType`).

En la siguiente gráfica (realizada con la librería matplotlib) se muestran las obras publicadas año por año de acuerdo a los datos de works.

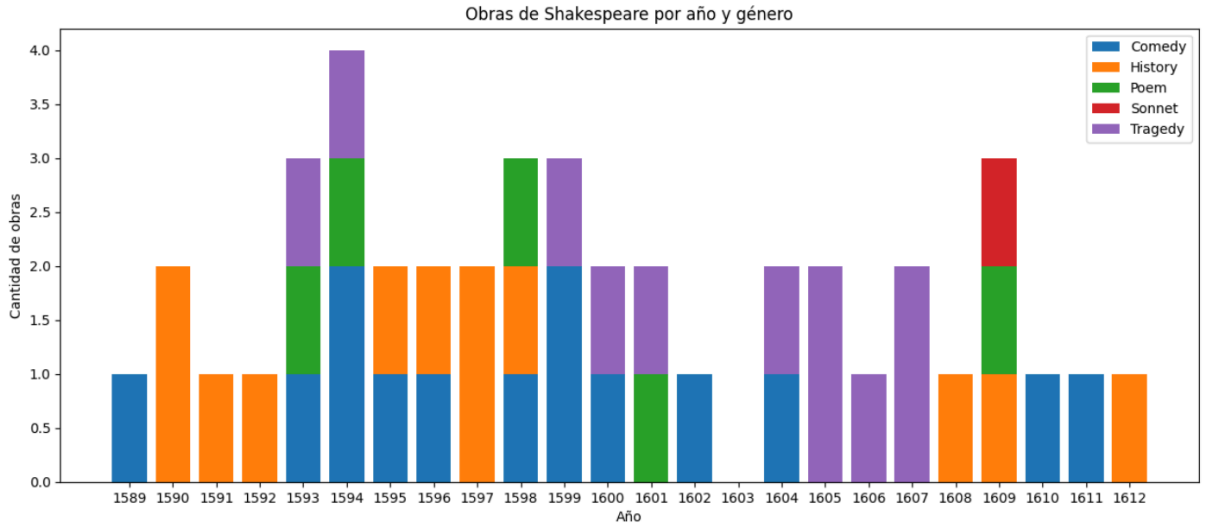


Figura 1 - Cantidad de Obras por Año y Género

A simple vista evidenciamos las siguientes características:

- Entre 1589 y 1612 William Shakespeare escribió casi ininterrumpidamente, publicando al menos 1 obra por año, con un máximo de 4 obras en 1594 y un promedio de 1.8 obras por año.
- En 1603 no tiene obras publicadas.
- Parecería que el apogeo en producción de Shakespeare es entre los años 1593 - 1601.
- Pese a fallecer en el año 1616, no hay registros de obras más allá del año 1612.

Breve nota al margen, tras una corta investigación no encontramos ninguna teoría aceptada que explique por qué no existen obras publicadas en el año 1603. Puede deberse a varios factores, de todos los posibles parecen relevantes para este trabajo mencionar dos:

1. La Peste Bubónica en Londres 1603 podría haber forzado a los teatros a estar cerrados, lo que podría haber restrasado algún estreno.

2. La muerte de la última monarca Tudor, la reina Isabel (1558 - 1603) con la consecuente conmoción socio-política que pudo haber ocasionado en Londres, podría haber afectado de alguna forma al autor.

Por otro lado, todos los autores tienen diferentes períodos o épocas en su obra, caracterizados por diferentes factores internos como la propia madurez y evolución del autor a factores externos como la estabilidad política, guerras, hambrunas, pestes y enfermedades, la influencia de otros artistas, etc. En el caso de William Shakespeare, en la literatura se identifican cuatro períodos bien diferenciables:

- In the Workshop (1589-1593)
- In the World (1594-1600)
- Out of the Depths (1601-1607)
- On the Heights (1608-1612)

En este trabajo, vamos a apoyarnos en los estudios que evidencian estos períodos en la obra del autor, para profundizar en el análisis. Un enfoque alternativo no abarcado en este trabajo podría ser el de experimentar con diferentes divisiones en el tiempo u observando la obra en diferentes ventanas de tiempo (1 año, 5 años, 10 años, etc.) para llegar a una conclusión similar.

Veamos en el siguiente gráfico, como queda dividida la obra de Shakespeare por estos períodos. Notar además, que para facilitar el entendimiento cambiamos el eje x de año de publicación a la edad del autor.

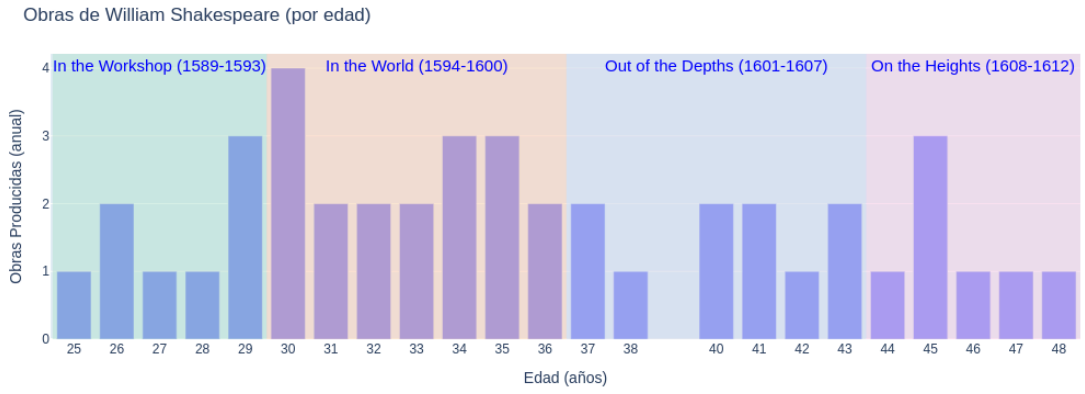


Figura 2 - Cantidad de Obras por Edad y Período

Los primeros años de Shakespeare (25-30 años) se caracterizan por la producción de obras de Historia entre las que se destaca el drama sobre la historia del Rey Enrique IV, con algunas obras de Tragedia, Comedia y Poesía hacia el final de esta etapa.
Su segunda etapa (30-36 años) es quizás la más explosiva en cuanto a producción, continuando con la generación de obras históricas pero desarrollando su faceta de dramaturgo, escribiendo tragedias y comedias. Entre sus obras producidas en este período se destacan *Sueño de una noche de verano* y *La Tragedia de Romeo y Julieta.*

En la tercera etapa (37-43 años) se termina de consolidar como dramaturgo, produciendo principalmente tragedias y comedias. De sus obras más notables en este período podemos destacar *La Tragedia de Macbeth* y *Otelo*.
En su última etapa (44-48 años) introduce a su obra la escritura de sonetos. Vale la pena destacar además, que Shakespeare a lo largo de toda su vida escribió algunos poemas.

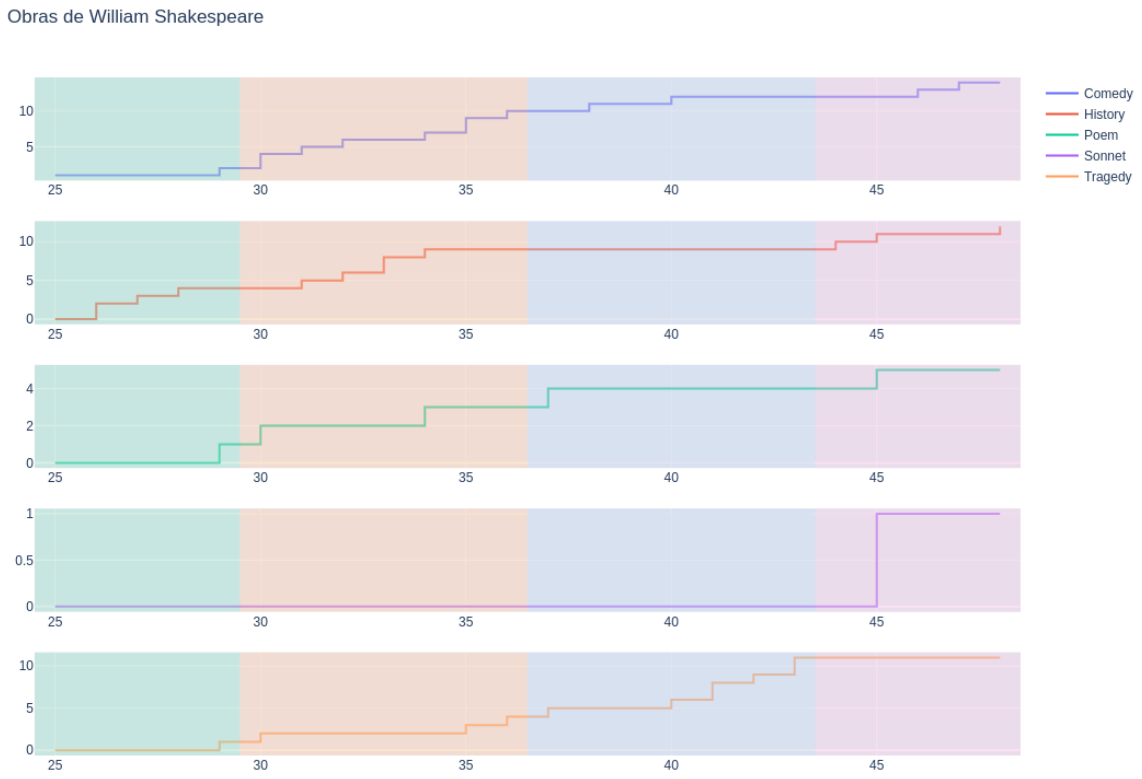


Figura 3 - Obras Acumuladas por Edad y por Género

Para finalizar el análisis y como visualización complementaria, en la siguiente gráfica podemos ver como varía el acumulado de obras del autor a lo largo de su vida, también agrupadas por género.

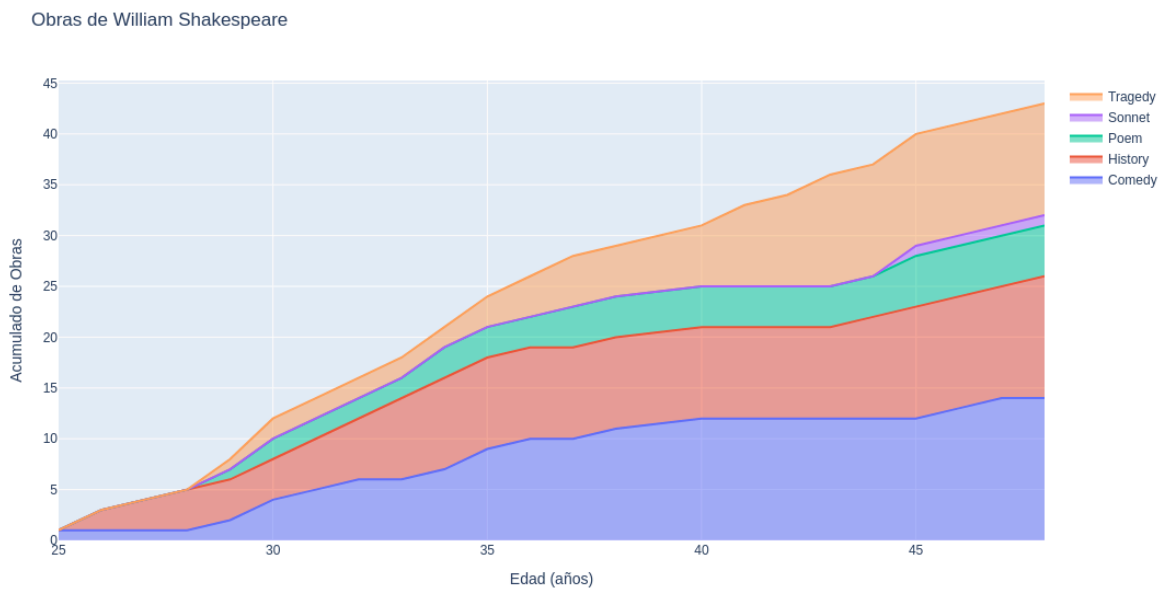


Figura 4 - Obras Acumuladas por Edad y por Género

Análisis de Palabras Frecuentes

Continuando con los análisis propuestos en la letra del laboratorio, específicamente en el punto 2.A, analizamos las palabras más frecuentes en la obra. Mediante este análisis podemos obtener una visión sobre el vocabulario del autor y su uso del idioma.
Para proceder con este análisis se realizó un procesamiento del texto cómo se mencionó en la sección Procesamiento de Datos, esto incluyó:

- Normalizar las palabras, llevándolas a minúscula y remover espacios innecesarios.
- Separar el texto en palabras, basándose en el “ “ como separador.

Una práctica común en procesamiento de lenguaje natural y sobre todo cuando se estudian frecuencias de palabras, es la de ignorar palabras frecuentes del idioma, también conocidas como stopwords (conectores). Estas son palabras muy utilizadas en el lenguaje como "and" y "or", que no aportan mayor información para este tipo de análisis. De incluir estas palabras, le estaríamos quitando peso a palabras menos frecuentes pero más valiosas para el análisis. Es por esto que optamos por remover stopwords antes de analizar las frecuencias de palabras.

Luego de remover las stopwords, las frecuencias se calculan simplemente contando la cantidad de ocurrencias de cada palabra en la obra de Shakespeare.

Posteriormente, utilizando estas frecuencias y mediante la biblioteca wordcloud, construimos la visualización que se puede ver en la Figura 5. La wordcloud (nube de palabras), es un tipo de visualización dónde el tamaño de la palabra está directamente relacionado a la frecuencia de la misma. Notar que algunas de las palabras en la wordcloud

siguen siendo del tipo conectores. Esto se debe a que cómo se explica en la sección Procesamiento de Datos, las stopwords utilizadas son las que maneja el modelo de lenguaje de Spacy, que está basado en un inglés contemporáneo y puede diferir del utilizado por el autor. Otro detalle que se puede observar dentro de la wordcloud es la presencia de las letras, "s", "d" y "o". Luego de una breve investigación esto puede deberse a el uso de contracciones por Shakespeare que no son utilizadas hoy en día, por ejemplo "o'er" siendo esta una contracción de "over".

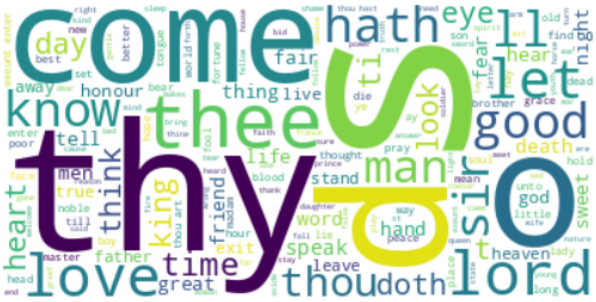


Figura 5: Wordcloud con las palabras más comunes para toda la obra

Una forma de profundizar este análisis es separando el análisis de las palabras más frecuentes por género, a partir de este análisis se obtiene la siguiente figura.

Figura 6: Wordcloud con las palabras más comunes por género.

De la figura anterior se logran ver ciertas tendencias:

- "Will" es una palabra muy usada por Shakespeare sin importar el género, aunque en el soneto es superada por "love" y "time". El uso elevado de will se puede deber a la gran cantidad de usos diferentes que tiene esta palabra y para los diversos contextos para los cuales puede ser utilizada. Esto refleja la riqueza y flexibilidad del idioma inglés durante la era isabelina.
- "Love" es una palabra que resalta en los Sonetos y en los Poemas, contemplando el giro más romántico de estos géneros respecto al resto.

No encontramos otras relaciones significativas en las palabras más usadas separadas por género, esto discrepa de nuestra primera hipótesis respecto a este análisis debido a que suponíamos una marcada diferencia entre las palabras más comunes entre los géneros, por ejemplo, en Tragedia esperábamos palabras relacionadas a emociones negativas.

Personajes más Relevantes

En esta sección, se hará un análisis acerca de la relevancia de los personajes medido en base a la cantidad de palabras y párrafos que tienen en la obra.

Personajes por Palabras

Para este análisis se contabilizó la cantidad de palabras mencionadas por cada personaje a lo largo de toda la obra, este ranking se observa en la Figura 7.

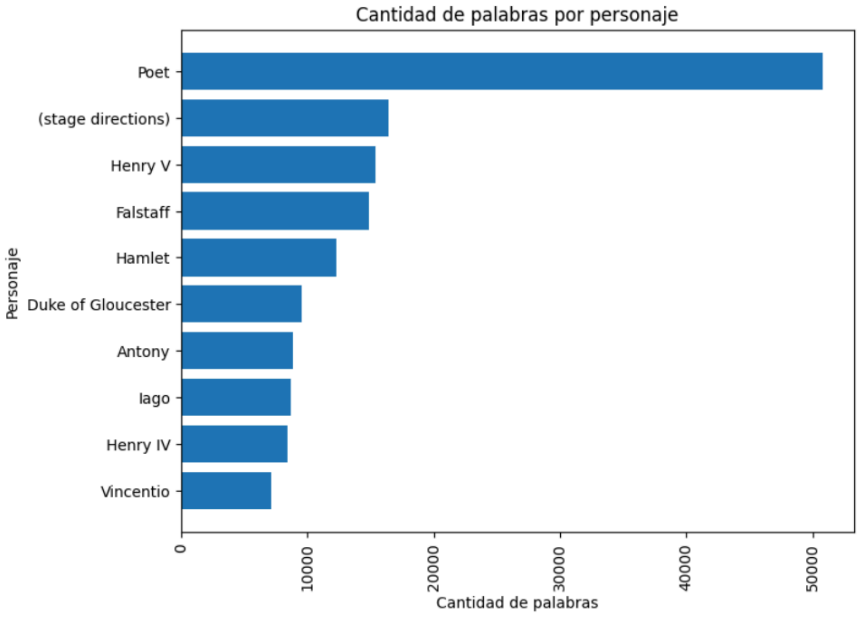


Figura 7. Gráfico de barras representando los 10 personajes con mayor número de palabras mencionadas a lo largo de toda la obra

De esta primera visualizacipon podemos identificar ocurrencias interesantes, por un lado en el top 3 de personajes se encuentran *Poet* y *(stage directions)*. Por un lado, Poet representa el narrador en los poemas escritos por Shakespeare.
Por otro lado, Stage Directions son indicaciones particulares escritas por Shakespeare para facilitar el trabajo de los actores que interpretan un personaje, a la vez que permite una visualización correcta por parte del lector sobre lo que sucede en una escena.

Debido a lo anterior una correcta representación de los personajes más comunes no incluiría Poet y Stage Directions, es por esto que se muestra la Figura 8.

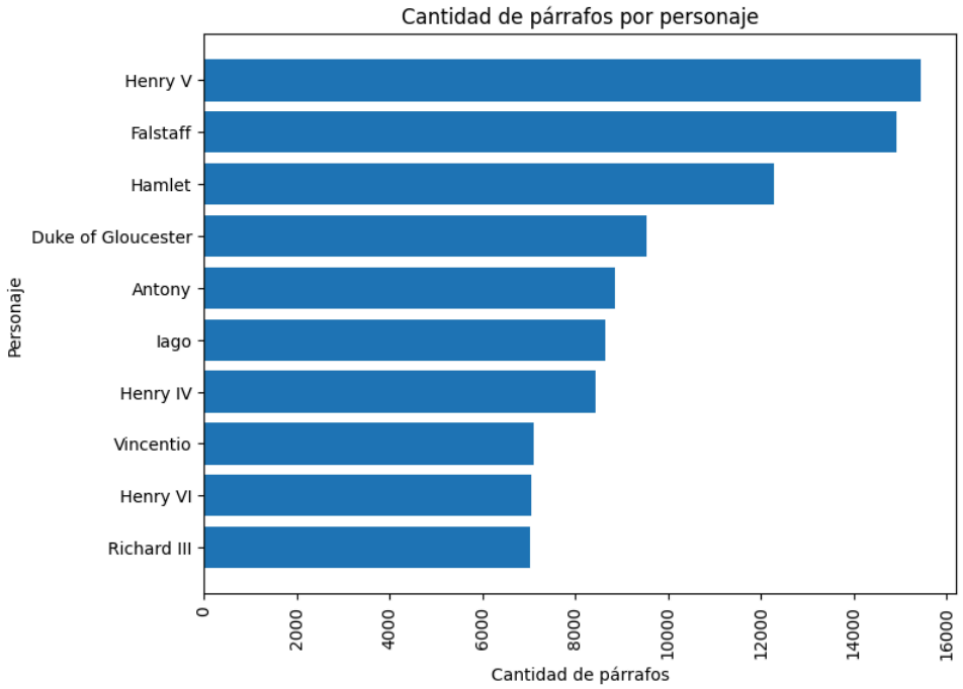


Figura 8. Gráfico de barras representando los 10 personajes con mayor número de palabras mencionadas a lo largo de toda la obra (sin incluir Poet y Stage Directions)

Los personajes con mayor cantidad de palabras a lo largo de la obra son: Henry V, Falstaff, Hamlet y Duke of Gloucester. Dentro de los 10 personajes con mayor número de palabras no encontramos se encuentran los conocidos personajes Romeo y Julieta.

Personajes por Párrafos

Continuando con el análisis podemos hacer una comparación de la cantidad de párrafos mencionados por cada personaje. En este caso también ocurre que *Poet* y *Stage Directions* son los personajes con más párrafos, pero el orden de los mismos está invertido, esto se puede deber a que los párrafos de Poet son más largos pero en menor cantidad.

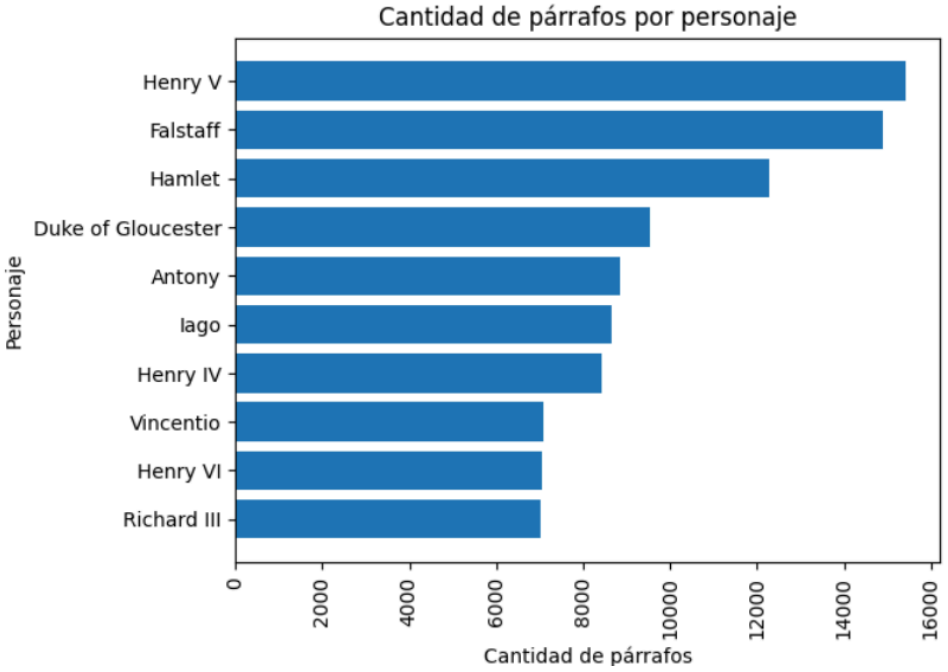


Figura 9: Gráfico de barras que muestra los 10 personajes con mayor cantidad de párrafos en toda la obra (sin incluir Poet y Stage Directions)

De acuerdo a la Figura 9, hay una pequeña diferencia en cuáles personajes se encuentran entre los 10 personajes con mayor cantidad de párrafos, *Henry V* y *Falstaff* intercambian sus posiciones cómo primero y segundo, sugiriendo que Henry V tiene menos intervenciones pero estas intervenciones son más largas.
A su vez hay otros cambios dentro del ranking, aparecen nuevos personajes como "*Timon"* y *"Othello"* , y otros con muchas palabras pero pocos diálogos ya no aparecen listados.

Análisis Complementarios

En esta sección enumeramos algunas preguntas que podrían responderse con este conjunto de datos y los pasos que seguiríamos para llevarlos adelante.

1. Extender stopwords y mejorar tokenizado

Antes de continuar con nuevas preguntas sobre los datos se nos ocurre mejorar el proceso de división de un párrafo en palabras (tokenizado), utilizado por ejemplo para el conteo de frecuencias de palabras. Para esto proponemos, en lugar de utilizar la función `split()` de la librería strings, utilizar el Tokenizer que la librería spacy incorpora con su modelo de lenguaje inglés. Esto debería de ser más efectivo, sobre todo para palabras como `that's` que son separadas automáticamente en `that` e `is`. A su vez podríamos probar con modelos más poderosos como el `en_core_web_lg` o incluso Bert o GPT-4.

A su vez, podríamos investigar en profundidad las frecuencias de las palabras calculadas, para ver si podemos identificar alguna stopword adicional propia del idioma inglés del siglo XVI.

2. Profundizar en el análisis de personajes más relevantes

Luego de realizar el análisis de relevancia de personajes por cantidad de palabras y párrafos, contemplando toda la obra de Shakespeare, nos preguntamos cuán representativo es el mismo. Por ejemplo, se podría analizar la relevancia de un personaje en base a las palabras asociadas al mismo pero en el contexto de una obra, en lugar de toda la obra del autor. Esto nos podría dar un indicio de los personajes principales por obra y analizar cómo varía esta tendencia entre entre obras y géneros.

Otra modificación que podríamos hacer para analizar la relevancia de personajes a lo largo de toda la obra del autor, es calcular la proporción de palabras de la obra asociadas a un personaje, en lugar de contar la cantidad de palabras (normalizar utilizando la longitud de la obra en palabras). De esta forma todas las obras tendrían el mismo peso, sin importar su longitud.

Un caso particular acerca de los datos que descubrimos, es la presencia de personajes especiales que trascienden a una obra cómo "First Citizen", "Servant, "Lord" que no necesariamente son el mismo personaje entre las obras y pueden influir en los diferentes análisis sobre los personajes.
Dependiendo del enfoque del análisis a realizar, se puede comparar los casos en que este tipo de personaje se toman cómo personajes diferentes entre las obras y no como uno común al utilizar toda la obra completa de Shakespeare.

3. Análisis de embeddings por categorías

Una de las preguntas que nos hicimos durante todo este trabajo fue sobre el uso del idioma inglés por parte de Shakespeare a lo largo de sus obras y de aquí surgen más preguntas:

- ¿Cambia de alguna forma el léxico de Shakespeare a lo largo de los años?
- ¿Hay diferencias notorias de vocabulario entre un género y otro? ¿Qué pasa entre una obra y otra?
- ¿Podemos encontrar diferencias notorias en el vocabulario utilizado entre un personaje y otro? ¿Qué pasa por ejemplo entre el vocabulario utilizado por un sirviente y un lord?
- ¿Podemos encontrar algún tipo de firma, que nos permita decir que un fragmento de texto pertenece a una obra de Shakespeare?

Para contestar algunas de estas preguntas pensamos ineludiblemente en el uso de embeddings de un modelo de lenguaje multi idioma como BERT, GPT-4 o cualquier modelo entrenado sobre textos en inglés. Obteniendo los vectores que representan a las palabras de las obras de Shakespeare y agregando algún pre-procesamiento para reducir ruido como remover stopwords, tomar solo las palabras con mayor frecuencia, entre otros, podríamos hacer varios análisis que nos permitan contestar estas dudas. Por ejemplo:

- Reducir la dimensionalidad de los vectores a 2D utilizando técnicas de reducción de la dimensionalidad como PCA o UMAP y graficarlos, para ver si se aprecian agrupaciones. Podemos hacer este ejercicio coloreando los puntos con diferentes

categorías: género, título de la obra o personaje. De allí podríamos ver si las palabras más representativas de un género se encuentran agrupadas y separadas de las palabras de otro género. De forma similar, analizar las palabras por personaje, obra.
- Sobre el análisis anterior, experimentar con modelos de clustering para ver si podemos encontrar agrupaciones que tengan sentido.
- Dado que sabemos a qué obra pertenece cada párrafo, podríamos entrenar un clasificador de texto, basado en estos embeddings que permita determinar la obra más probable a la que pertenecería un párrafo X, que podría ser de una obra de Shakespeare o un texto arbitrario.

4. Misterios no resueltos sobre Shakespeare

Incluso al día de hoy, existe discusión sobre la atribución de la autoría de varias de sus obras. De forma análoga, se teme por la existencia de obras perdidas, no atribuídas al escritor, sobre todo en los años previos a sus comienzos oficiales como escritor.

En este contexto, podríamos utilizar la información existente en este conjunto de datos, para entrenar clasificadores de texto que permitan determinar por ejemplo si alguna de las obras presentes en este conjunto de datos no se ajusta a los patrones de las anteriores, lo cual podría respaldar alguna de las teorías que cuestionan la autoría de Shakespeare en algunos clásicos. A su vez, si contamos con textos de otros escritores contemporáneos y en géneros similares a Shakespeare, podríamos entrenar clasificadores que nos permitan revisar si un texto X de la época podría ser una obra perdida de Shakespeare.

Para esto podríamos usar como ya mencionamos, los embeddings de spacy o utilizar vectores más poderosos como los de BERT o GPT-4.

7. Conclusiones

A partir de los análisis realizados al conjunto de datos con obras de William Shakespeare, podemos formular las siguientes conclusiones a partir de sus obras allí presente y al conjunto de datos en sí mismo.

- El autor comenzó a escribir a la tardía edad de 25 años si se toma en cuenta que muere a los 52 años. Una vida que de todas formas supera la expectativa de vida de un ciudadano londinense del siglo XVI-XVII.
- Pese a esto su obra fue muy prolífera con 43 obras entre Tragedias, Comedias, Poemas, Sonetos y Dramas Históricos.
- De toda su vida como escritor, su período más fructífero fue entre los años 1594 - 1600 en el período que se lo conoce como “In The World” por quienes estudian la obra del autor.
- Llama la atención el año 1603 en donde no vemos obras publicadas, lo cual contradice la tendencia histórica en su producción de obras.
- Su vida como escritor se encuentra claramente marcada por las comedias, 14 obras escritas (33 % de su obra), le siguen los dramas históricos con 12 obras (28% de su obra), y las tragedias con 11 obras escritas (26%). En su mayoría obras de teatro convirtiéndo a Shakespeare en un dramaturgo por excelencia.
- Dentro de la obra de Shakespeare, se utiliza un lenguaje un inglés diferente al actual y esto se ve claramente reflejado en las palabras más utilizadas por el escritor.
- Los personajes que mayor diálogo tienen entre las obras del escritor son Henry V, Falstaff y Hamlet. Lejos de este ranking quedan los conocidos personajes Romeo y Julieta.

Por otro lado, analizando el conjunto de datos y su calidad, podemos concluir que:

- La calidad de los datos presentes es muy buena, mostrando una ausencia de problemas de calidad como datos faltantes, salvo por las columnas Abbrev y Description en la tabla characters en donde se ve una buena presencia de valores faltantes.
- No se ven datos duplicados o incoherentes a simple vista. Esto a su vez tiene mucho sentido, tomando en cuenta de que se trata de un conjunto de datos que posiblemente se haya realizado de forma exhaustiva a partir de la obra de William Shakespeare.

8. Bitácora

La presente sección tiene por objetivo funcionar como una bitácora del proyecto, registrando de forma muy superficial diferentes sucesos que se fueron dando en el desarrollo del proyecto.

- 07-05-2024: Presentación del Laboratorio 1, el link a los datos en la pauta estaba “roto” por lo que tuvimos que acceder a los datos mediante un nuevo link diferente que encontramos en la web. Posteriormente el equipo docente comunicó que este link era el nuevo link. Luego se volvió a romper pero ya teníamos los datos.
- 07-05-2024: Durante la presentación del laboratorio también se detectó un problema con el código de ejemplo para la descarga de los datos, utilizando SQLAlchemy y Pandas. Esto quedó resuelto utilizando un workaround encontrado en Stackoverflow que luego se incorporó como modificación del repositorio de referencia.

Referencias

1. Base de Datos pública de la obra de Shakespeare con datos para este laboratorio https://relational-data.org/dataset/Shakespeare
2. Repositorio oficial del curso Intro-CD con código de referencia para Laboratorio 1 https://gitlab.fing.edu.uy/maestria-cdaa/intro-cd/
3. Repositorio del grupo con código fuente y todos los entregables de esta tarea laboratorio https://github.com/efviodo/mcdaa-intro-cd
4. Towards Data Science introducción a Exploratory Data Analisys https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15
5. Data Camp: Curso Exploratory Data Analysis in Python https://www.datacamp.com/courses/exploratory-data-analysis-in-python
6. Wikipedia William Shakespeare https://en.wikipedia.org/wiki/William_Shakespeare
7. Biblioteca Python Wordcloud https://pypi.org/project/wordcloud/
8. Plotly Información sobre Vertical Lines https://plotly.com/python/horizontal-vertical-shapes/
9. Plotly Información sobre Colors https://plotly.com/python/discrete-color/
10. Four Periods of Shakespeare's Dramatic and Poetic Career https://moirabaricollegeonline.co.in/attendence/classnotes/files/1589611082.docx#:~:text=Although%20the%20precise%20date%20of,the%20Fourth%20Period%20from%201608.
11. Cross Industry Standard Process for Data Mining (CRISP-DM) https://es.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining
12. Autoría de Shakespeare Teoría Oxfordiana https://en.wikipedia.org/wiki/Oxfordian_theory_of_Shakespeare_authorship
13. Stage Directions https://shakespearestagedirections.coe.edu/#:~:text=Stage%20directions%20are%20where%20the,or%20directors%20about%20that%20information