Análisis de las Obras de Shakespeare

Presentación Laboratorio 1 - Introducción a la Ciencia de Datos

Emiliano Viotti - Matías Rolando
Grupo 15
https://github.com/efviodo/mcdaa-intro-cd

Para entender los datos hicimos un EDA:

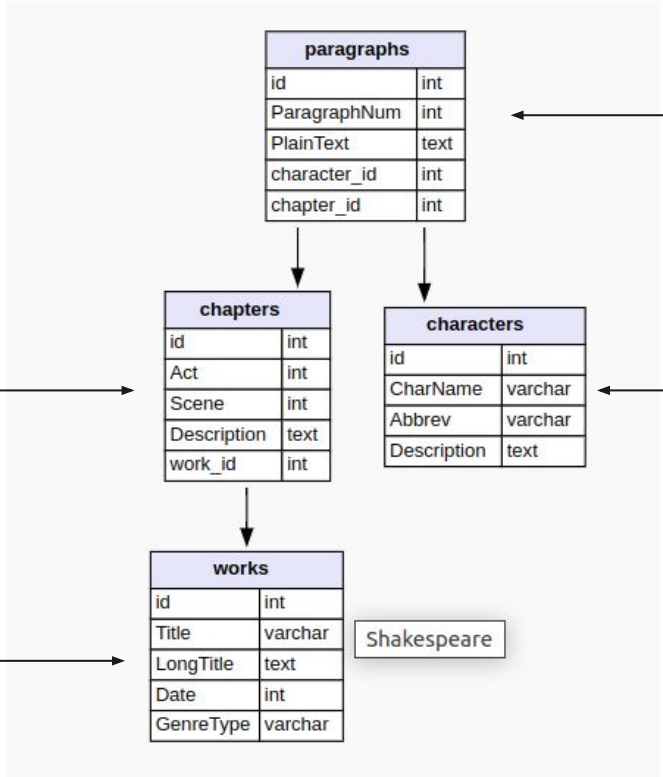
- Analizamos los datos a partir de los dataframes en pandas usando .describe() contando missing values con una función propia y por supuesto viendo los datos.

- También utilizamos ydata-profiling para generar un data profiling en profundidad para cada tabla.

Algunos resultados del EDA:

Capítulos
- 945 capítulos
- 0 duplicados
- 0 missing values
- Varios capítulos con misma descripción

Obras
- 43 obras
- 0 duplicados
- 0 missing values
- 5 géneros



Párrafos
- 35465 párrafos
- 0 duplicados
- 1849 duplicados a nivel de PlainText
- Algunos párrafos son especiales (siguiente slide)

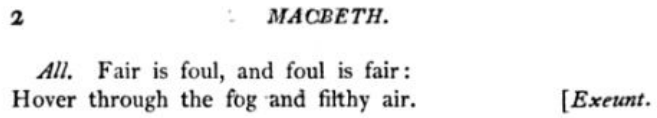
Personajes
- 1266 personajes
- Duplicados a nivel de CharName (24%) y Abbrev (23%)
- Missing Values a nivel de Abbrev (0.39%) y Description (51%)
- Algunos personajes son particulares (ver siguiente slide)

Algunas observaciones sobre los datos:

	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

	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

- Algunos párrafos son indicaciones de escena como “[Exit]”
- Ruido en PlainText: "[", "\n", ",", ":", ";", ".", "]", "(", ")", "?", "!", "'", "-", "\"", "{", "}"
- Además de personajes relevantes como “Romeo” y “Juliet” hay personajes anónimos como Messenger, Lord, etc. Además llama la atención el personaje “ALL” que aparece además como “All”.

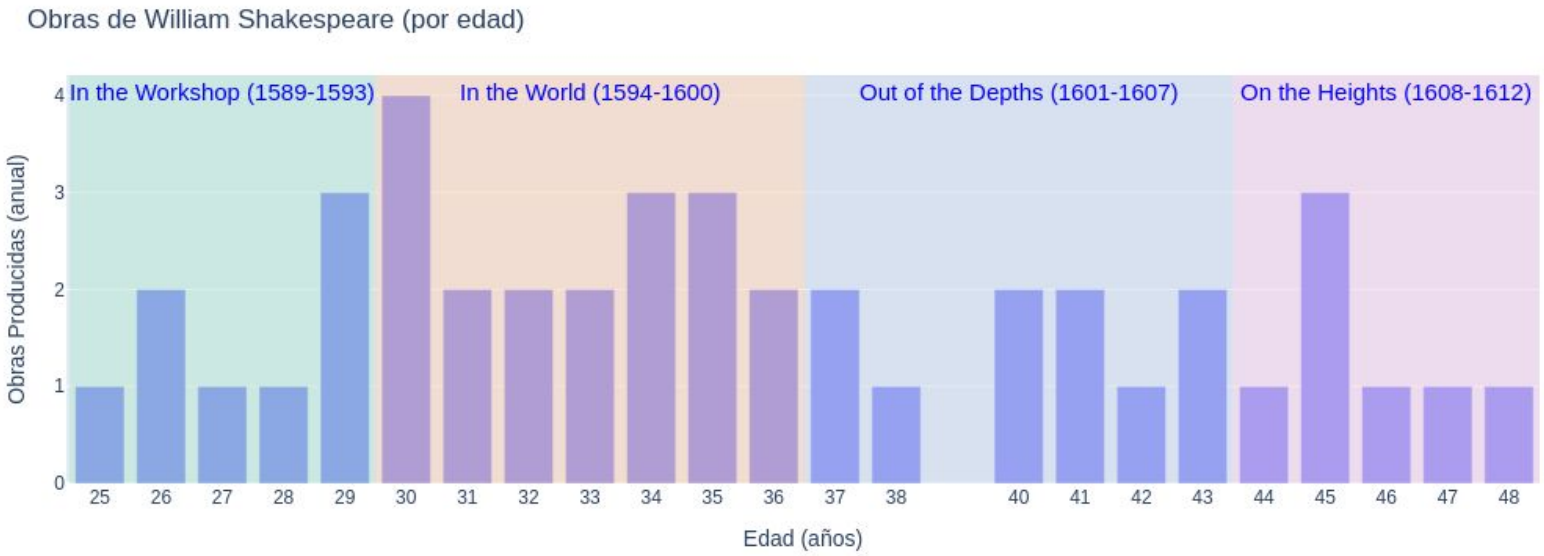


2 MACBETH.

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

Extracto 2 de la obra Macbeth - Google Books

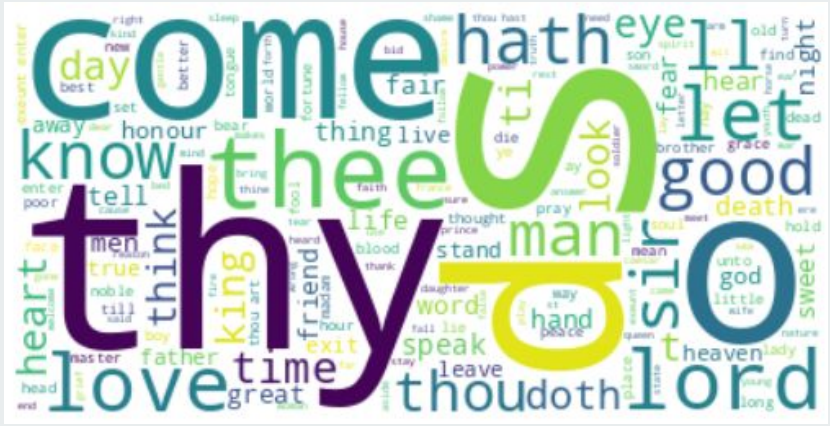
Obras por edad



Algunas referencias hablan de tres a cuatro períodos claros en la obra del autor.

Análisis de palabras más frecuentes

- Utilización de librerías para la remoción de stopwords.
- Normalización del texto.



Toda la obra

Tragedia

Comedia

Soneto

Poema

Historia

Conclusiones

- El dataset presenta muy buena calidad, con la excepción de valores faltantes en dos columnas la tabla Characters.
- No encontramos datos duplicados o inconsistentes.

- Su período más fructífero en producción va entre 1594-1600, conocido como "In the World"
- No publica obras en 1603.
- Obra marcada por la presencia de Comedias.
- Utiliza un inglés diferente al de nuestro tiempo, evidenciado por las palabras más frecuentes.
- Personajes con más diálogo, Henry , Falstaff y Hamplet. Lejos quedan de este ranking Romeo y Julieta.

Análsis complementario

1. Extender stopwords y mejorar tokenizado.
2. Profundizar en el análisis de personajes más relevantes.
3. Análisis de embeddings por categorías.
4. Misterios no resueltos sobre Shakespeare.

Imagen del Globe Theater