shakespeare_propuesta

May 23, 2023

1 Introducción a la Ciencia de Datos: Entregable Tarea 1

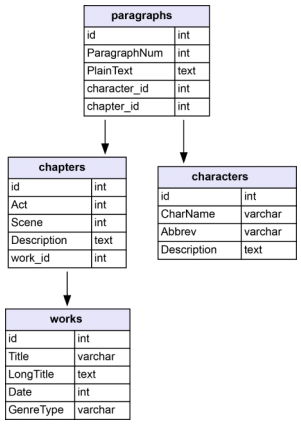
1.1 Parte 1: Cargado y limpieza de datos

a) Se probo las primera celdas del Notebook entregado y funcionaron correctamente. Se cargaron todas las tablas (4) con la función load_table(table_name, engine):

1. DataFrame con todas las Obras:
   - df_works = load_table("works", engine)
2. DataFrame con todos los Capítulos:
   - df_chapters = load_table("chapters", engine)
3. DataFrame con todos los Parrafos:
   - df_paragraphs = load_table("paragraphs", engine)
4. DataFrame con todas los Personajes:
   - df_characters = load_table("characters", engine)

La Base de Datos Relacional de la Obra de Sheakspeare, brindada en el curso se compone de 4 tablas:

1. Tabla Works: Con los registros de cada Obra realizada por Shakespeare.
2. Tabla Chapters: Con los Capítulos de cada una de la Obras de Shakespeare, cada Obra se compone de N Capítulos, y cada Capítulo pertenece a una sola Obra.
3. Tabla Paragraphs: Con todos los Parrafos de la Obra de Sheakespeare, cada Capítulo contiene N Parrafos, y cada Parrafo pertenece a un solo Capítulo.
4. Tabla Characters: Con todos los Personajes de la Obra de Shakespeare, cada Parrafo esta asociado a un Unico Personaje y cada Personaje puede tener N Parrafos, en particular estar incluidos en diferentes Obras.



Se realizo Join entre las diferentes Tablas, para extraer la información relevante y se genero un DataFrame de Pandas con registros por Palabras (Words) de cada Parrafo, de cada Capitulo, de cada Obra, para el Personaje Asociado.

1.1.1 Calidad de Datos

Respecto a Datos Faltantes, se evaluaron las 4 tablas Descriptas con la libreria pandas a través de los DataFrames construidos:

Los únicos valores faltantes o nulos se registran en el campo descripción de la tabla de Personajes.

1. Tabla de Faltantes:

Tabla de Obras: Faltantes

	0
id	0
Title	0
LongTitle	0
Date	0
GenreType	0

Tabla de Capítulos: Faltantes

	0
id	0
Act	0
Scene	0
Description	0
work_id	0

Tabla de Párrafos: Faltantes

	0
id__paragraphs	0
ParagraphNum	0
PlainText	0
character_id	0
chapter_id	0
text_expanded	0
CleanText	0
WordList	0
id__characters	0
CharName	0

Tabla de Personajes: Faltantes

	0
id	0
CharName	0
Abbrev	5
Description	646

Tabla de Palabras: Faltantes

	0
id_words	0
ParagraphNum	0
character_id	0
chapter_id	0
word	0
id_characters	0
CharName	0
id_chapters	0
Act	0
Scene	0
Description	0
work_id	0
Title	0
LongTitle	0
Date	0
GenreType	0
unos	0

2. Tabla de Nulos:

Tabla de Obras: Nulos

	0
id	0
Title	0
LongTitle	0
Date	0
GenreType	0

Tabla de Capítulos: Nulos

	0
id	0
Act	0
Scene	0
Description	0
work_id	0

Tabla de Párrafos: Nulos

	0
id__paragraphs	0
ParagraphNum	0
PlainText	0
character_id	0
chapter_id	0
text_expanded	0
CleanText	0
WordList	0
id__characters	0
CharName	0

Tabla de Personajes: Nulos

	0
id	0
CharName	0
Abbrev	5
Description	646

Tabla de Palabras: Nulos

	0
id_words	0
ParagraphNum	0
character_id	0
chapter_id	0
word	0
id_characters	0
CharName	0
id_chapters	0
Act	0
Scene	0
Description	0
work_id	0
Title	0
LongTitle	0
Date	0
GenreType	0
unos	0

3. Tabla de Duplicados:

Tabla de Obras: Duplicados	Tabla de Capítulos: Duplicados	Tabla de Párrafos: Duplicados	Tabla de Personajes: Duplicados	Tabla de Palabras: Duplicados
0	0	0	0	175270

No se encuentran tampoco registros duplicados en las Tablas originales de la base de datos, y solo se encuentran duplicados en la tabla de Palabras ("words"), pero no son un problema de calidad de datos.

Solamente se trabajo en limpieza y transformación de Datos, para la construcción del DataFrame de Palabras, debido a la cantidad de caracteres incluidos entre las palabras, como signos de puntuación, pregunta, exclamación, separación, etc. Dicho problema fue atacado con la función clean_text() la cual se amplio a los caracteres faltantes. Además de los problemas de puntuación, nos encontramos con las contracciones en Ingles. Para este último problema se resolvio utilizar la librería **pycontractions** creandose la función **expand_contractions()** a partir de aplicar el método "expand_contractions" de dicha libreria. La expansión no es tan simple ya que requiere conocimiento contextual para elegir las palabras correctas de reemplazo en el Párrafo, esto implica bastante tiempo de computo. > La librería usa un enfoque de tres pasos. Primero, las contracciones simples con una sola regla se reemplazan. En el segundo paso, si hay contracciones con múltiples reglas, se procede a reemplazar todas las combinaciones de reglas para producir todos los textos posibles. Cada texto se pasa luego por un corrector gramatical y se calcula la distancia del movimiento de palabras (WMD) entre él y el texto original. Las hipótesis se ordenan por menor número de errores gramaticales y menor distancia del texto original y se devuelve la hipótesis superior como la forma expandida. El conteo de errores gramaticales elimina las peores opciones, pero hay muchos casos que no tienen o tienen el mismo número de errores gramaticales. En estos casos, el WMD funciona como el desempate. El WMD es el costo acumulativo ponderado mínimo requerido para mover todas las palabras del texto original a cada hipótesis. Esto aprovecha el modelo vectorial semántico subyacente elegido, como Word2Vec, GloVe o FastText. Como la diferencia entre cada hipótesis es solo el reemplazo de una contracción con su expansión, la hipótesis "más cercana" al texto original será la que tenga la mínima distancia euclidiana entre el par de palabras de contracción y expansión en el espacio de incrustación. - (AI) Bing-CHAT.

En resumen, se expandio las contracciones con **pycontractions** en cada Párrafo, se convirtio a minúsculas y se reemplazaron los signos de puntuación por el espacio, tal cual fue sugerido en el curso.

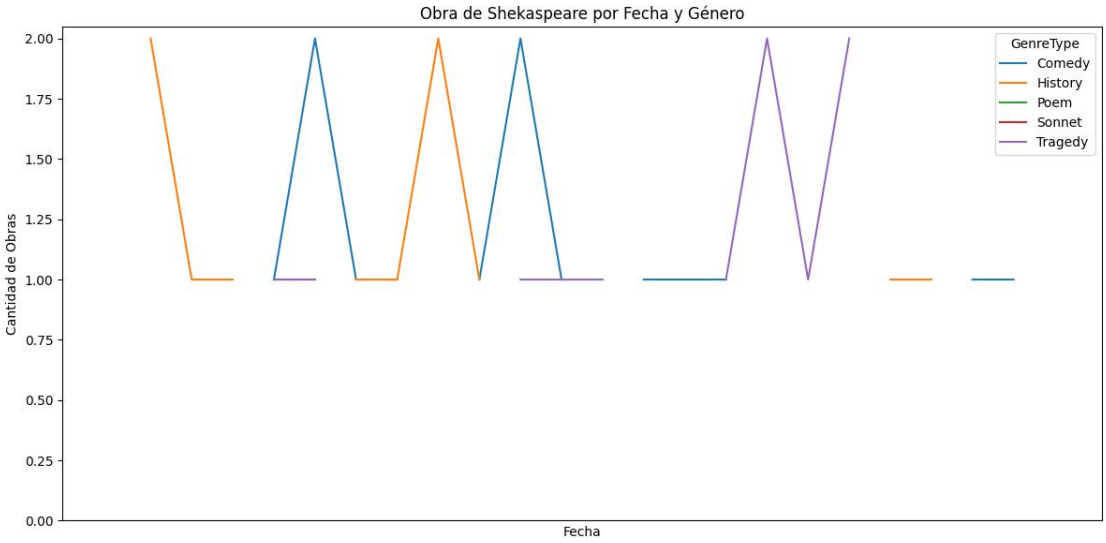
1.1.2 Personajes con más Párrafos:

Respecto a los Personajes con más párrafos, aparecen primero ***"Stage Directions"*** *y* ***"Poet"*** que no son personajes propiamente, el primer caso son las instrucciones de dirección en la mayoria de las Obras que no son Poemas o Sonetos, y el segundo que aparece es justamente quien se asocia a estos 2 últimos generos mencionados.

	Count
CharName	
(stage directions)	3751
Poet	766
Falstaff	471
Henry V	377
Hamlet	358
Duke of Gloucester	285
Othello	274
Iago	272
Antony	253
Richard III	246

El personaje con mas Párrafos, teniendo en cuenta las anteriores aclaraciones es ***Falstaff***. Según Wikipedia, Sir John Falstaff es un personaje de ficción creado por el dramaturgo inglés William Shakespeare. Aparece en tres obras de Shakespeare y recibe un elogio en una cuarta. Su importancia como personaje plenamente desarrollado se formó principalmente en las obras Enrique IV, 1ª parte y 2ª parte, donde es compañero del príncipe Hal, el futuro rey Enrique V de Inglaterra. Un notable elogio de Falstaff se presenta en el Acto II, Escena III de Enrique V, donde Falstaff no aparece como personaje en escena, sino que su muerte es narrada por la señora Quickly en términos que algunos estudiosos han atribuido a la descripción que hiciese Platón de la muerte de Sócrates tras beber cicuta. En comparación, Falstaff es presentado como el bufonesco pretendiente de dos mujeres casadas en Las alegres comadres de Windsor. Luego en un siguiente grupo lo siguen ***"Enrique V"*** y ***"Hamlet"***, luego en un tercer grupo aparecen * *"Duke of Gloucester", "Othello", "Iago", "Antony" y "Richard III"* *.

1.1.3 Obra de Shekaspeare a lo largo del tiempo en cuanto a Géneros.



En el comienzo de la Obra de Shekaspeare se puede encontrar 2 ciclos de creación de Obras en el Género de ***Historia*** alternado con ***Comedia.*** Luego en el final de su Obra se observa un Ciclo de creación de Obras en el Género de la ***Tragedia.***
Hay un Ciclo intermedio en el que su Obra disminuyo en cantidades. Si bien su creación de ***Poemas*** y ***Sonetos*** no fue tan cuantiosa, se observa distribuida a lo largo de su Obra.

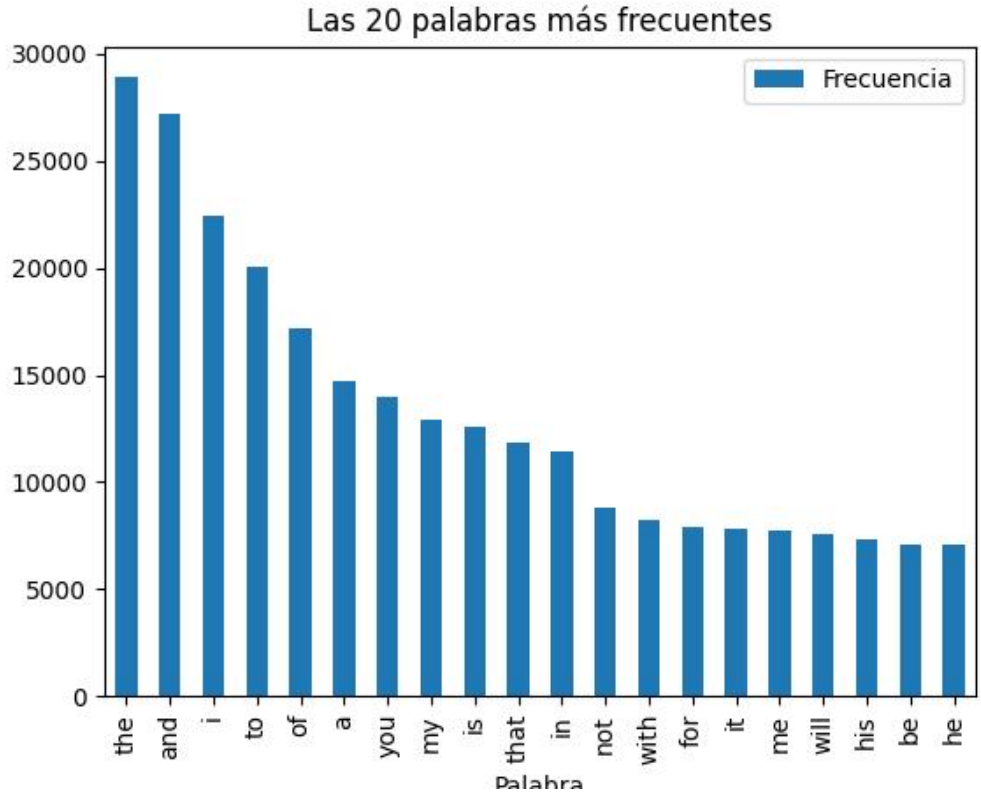
1.1.4 Conteo de Palabras (df_words)

Para la creación de la Tabla de Palabras se partio de la tabla de Párrafos, se expandieron las contracciones del Ingles con la libreria **pycontractions** como se explico anteriormente, luego se pasaron a minusculas todos los caracteres, luego se corrigieron los signos de puntuación con la función clean_text(). Luego se realizo un split sobre el campo "CleanText" asignandolo a la columna "WordList" dividiendo cada elemento de la lista. Por último se hizo un explode() sobre ese campo para crear un registro para cada palabra y en el DataFrame de Palabras se eliminaron los campos; ***"CleanText", "PlainText", "text_expanded"*** y se renombro la columna ***"WordList"*** por ***"word"***.

1.2 Parte 2: Conteo de palabras y visualizaciones

1.2.1 Palabras más frecuentes considerando toda la obra

Si analizamos las palabras mas utilizadas en la Obra de Shekaspeare, encontramos que obviamente la Frecuencia de las 20 palabras mas utilizadas refieren a artículos y palabras de conjunción.



Sin embargo podemos utilizar una visualización de Nube de Palabras, para analizar texto, basada en la librería ***wordcloud*** que permite excluir con el parámetro *stopwords* las palabras mas comunes del idioma, y dejar las palabras con mayor significado:

Con la idea de generar visualizaciones que refieran a las palabras más utilizadas por personajes y géneros podriamos filtrar por estos criterios y mostrar las tablas de Frecuencia y Nubes de Puntos asociadas. Por ejemplo se podrían mostrar subgráficos de Nubes de Punto en relación a los Géneros como la siguiente:

Observando las nubes de Palabras por Género encontramos que ***"Historia"*** y ***"Tragedia"*** se parecen entre sí, mientras que ***"Poemas"*** y ***"Sonetos"*** se parecen entre sí. Mientras que ***"Comedia"*** es una combinación de ambos grupos, mas similar en las palabras más frecuentes al primer grupo mencionado.

Podrían tambien plantearse visualizaciones dinámicas utilizando librerias como **pivottablejs** para explorar en mayor profundidad los datos con *Tablas Dinámicas* o sugerir utilizar Nubes de Palabras

de modo tridimensional, teniendo como ejes las variables relevantes como Género o Personajes. A su vez los personajes podrían ser intentados de clasificar en categorías como *"Nobles", "Plebeyos", "Vasallos",* etc.

1.2.2 Personajes con mayor cantidad de Palabras:

Cantidad de Palabras por Personaje

Palabras por Personaje

('', 'Personajes')	('', 'Cantidades de Palabras')
Poet	49495
(stage directions)	16181
Henry V	15159
Falstaff	14614
Hamlet	12033
Duke of Gloucester	9342
Antony	8667
Iago	8516
Henry IV	8230
Vincentio	6998

Palabras Únicas por Personaje

('', 'Personajes')	('', 'Cantidades de Palabras Unicas')
Poet	7537
Henry V	3214
Falstaff	2907
Hamlet	2799
Duke of Gloucester	2368
Henry IV	2201
Antony	2164
Iago	2040
Queen Margaret	1922
(stage directions)	1917

En este caso podemos observar como en los Párrafos que los Personajes con más palabras son ***"Poet"*** y ***"Stage Directions"***. Que no son Personajes propiamente dichos y refieren a Obras asociadas a los *Poemas* y *Sonetos* en el primer caso y en el resto de las Obras al segundo. Si consideramos la mayor cantidad de palabras únicas por cada Personaje vemos que ***"Stage Directions"*** cae en el ranking debido a que se repiten con mucha frecuencia los mismos tipos de ordenes y textos, para el caso de ***"Poet"*** no se verifica esto debido a que se trata de la voz de los *Poemas* y *Sonetos*. Teniendo en cuenta estas consideraciones se podrian quitar del análisis estos seudo Personajes y encontrariamos en dichos rankings a Personajes como **"Enrique V", "Falstaff",** etc.

1.2.3 Preguntas y análisis posibles a partir de estos datos

Podrian a partir de estos datos textuales, a partir de palabras y párrafos, realizar análisis de sentimientos, predicción de Personajes en base a posibles textos expresados o Párrafos de la Obra de Shekaspeare, caracterizar a los Personajes y Clasificarlos en base a nuevas categorías a partir de análisis de cluster. Podrían responderse preguntas como:

1. ¿Qué obras tienen más personajes y qué relación hay entre el número de personajes y la longitud de la obra?
2. ¿Qué palabras son las más frecuentes en cada obra y qué temas o emociones reflejan?
3. ¿Qué personajes tienen más líneas y qué papel desempeñan en la trama?
4. ¿Qué obras tienen más escenas y cómo se distribuyen las escenas entre los actos?
5. ¿Qué obras tienen más variación lingüística y qué factores pueden influir en ello?

Hay muchas más posibilidades de explorar los datos con técnicas de análisis cuantitativo o cualitativo. Por ejemplo, se podría usar el análisis de frecuencias, el análisis de contenido, el análisis de redes o el análisis de sentimientos.