Tarea 1 - Introducción a la Ciencia de Datos

2023

Brian Britos[1], Juan Pellegrini[2]

Introducción

En este informe se presenta el análisis realizado sobre una base de datos relacional que almacena información sobre la obra completa de Shakespeare.
Primeramente se realiza una exploración y limpieza de datos y luego se analizan la cantidad de párrafos que tiene cada personaje, la cantidad de obras por género que se realizan en el transcurso del tiempo y cómo fue variando el total de obras realizadas por período de tiempo. También se analiza cuáles fueron las palabras más utilizadas.
Para realizar dicho análisis se utiliza un script hecho en python sobre la plataforma *Google Colab*.

Parte 1: Cargado y limpieza de datos

La base de datos está compuesta por 4 tablas:

- **works**: que registra cada obra realizada
- **chapters**: la cual registra cada capítulo de cada obra
- **paragraphs**: muestra cada párrafo de cada capítulo
- **characters**: la cual describe al personaje que interpreta cada párrafo

El siguiente diagrama muestra que entidades contiene cada tabla y mediante cuáles de estas se relacionan. Las mismas se muestran con el mismo color en cada tabla relacionada.

[1] britsimm27@gmail.com

[2] jpellegrini11@gmail.com

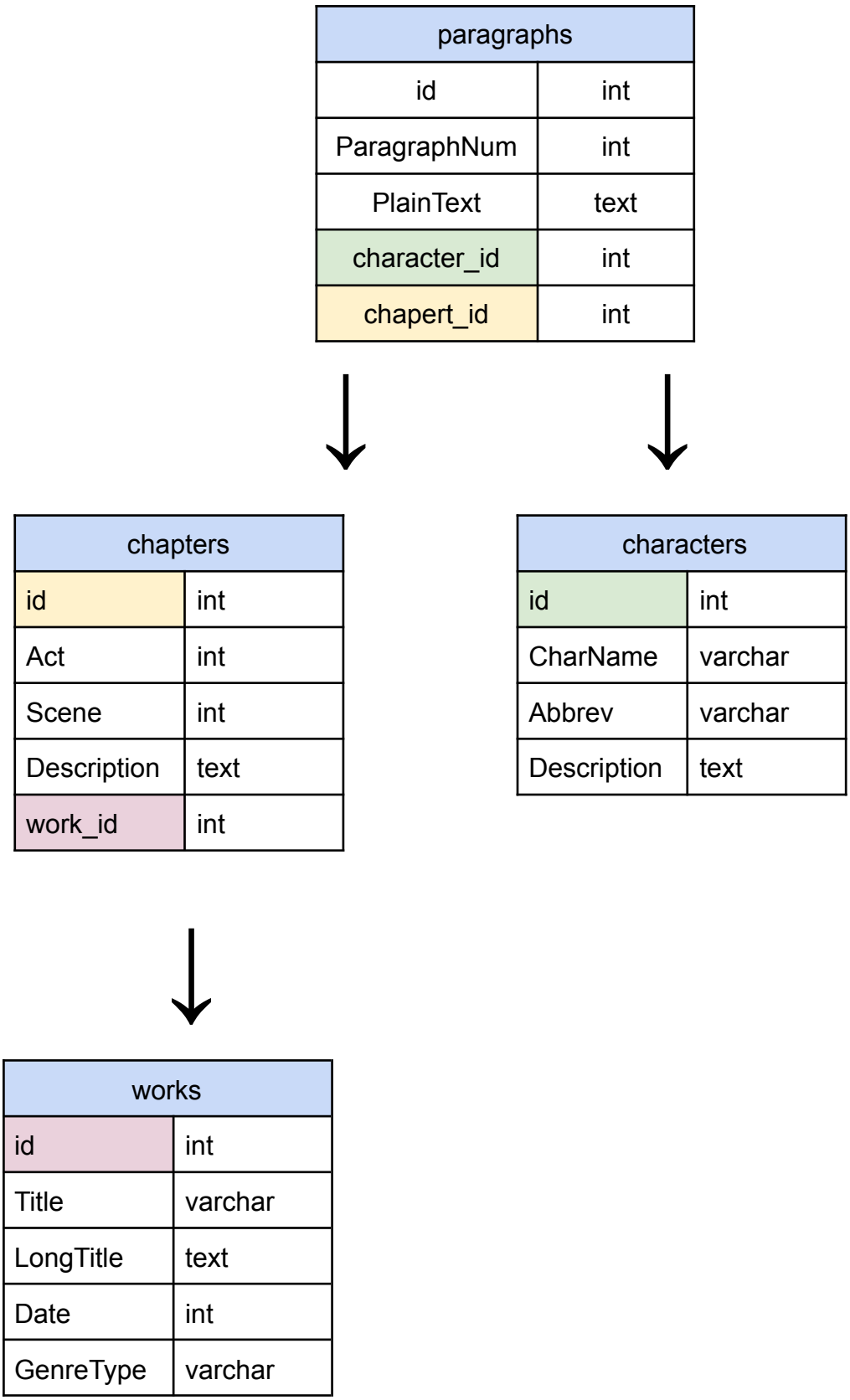

Diagrama lógico de la base de datos.

Carga de datos

En el script las tablas se almacenan respectivamente como:

- df_works
- df_chapters
- df_paragraphs
- df_characters

A lo largo de este trabajo se utilizó la librería *Pandas* para la manipulación de datos.

Limpieza de datos

Para comenzar se buscaron datos faltantes:

Se encuentra que la mayoría de los registros están completos. Únicamente en la tabla characters hay 5 registros que no tienen datos en la columna "*Abbrev*" y 646 no cuentan con datos en el campo "*Description*". De todas formas, no son necesarios para el análisis ya que, de esta tabla, solamente se utilizan las columnas "*id*" y "*CharName*".

En segundo lugar se analizó la tabla "characters" en detalle y se encontró que hay varios nombres duplicados con identificadores distintos. Esto puede deberse a que un mismo personaje puede estar presente en varias obras distintas, y para cada una se le asigna un identificador. Por otro lado, hay personajes que pueden ser "protagonizados" por distintas personas, como por ejemplo "Messenger" y "Lord". A continuación se muestran los cinco personajes que más veces están en la tabla "characters".

Personaje	Recuento
All	23
Messenger	23
Servant	21
Lord	9
Page	8

Por último, se encontró que un personaje no es tal y se tomó la decisión de eliminarlo para el correcto procesamiento en las próximas etapas, el mismo es “(stage directions)”.
Si bien “All” no es estrictamente un personaje, se decide considerarlo en el análisis ya que implica la participación de todos los personajes realizando diferentes acciones.

Recuento de párrafos

A continuación se procedió a realizar el conteo de párrafos por cada personaje, con este fín primero se unieron las tablas “paragraphs” y “characters” mediante su respectivo identificador. Para evitar el problema mencionado al principio de que el mismo personaje puede tener distintos identificadores se decidió realizar el recuento por el atributo “CharName” en lugar del identificador correspondiente.

Se decidió repetir la misma tarea pero realizando el recuento por el atributo “character_id” en lugar de “CharName” para así contrastar la hipótesis planteada al comienzo. La siguiente tabla muestra los 5 personajes con más párrafos en ambos escenarios.

Personaje	Párrafos (“Charname”)	Párrafos (“characters_id”)
Poet	766	733
Falstaff	471	471
Henry V	377	377
Hamlet	358	358
Duke of Gloucester	285	285

Tabla 1: Recuento de párrafos de cada personaje a lo largo de todos los trabajos de Shakespeare.

Se observa que el único cambio sucedió en el personaje “Poet”, que pasa de 766 a 733 párrafos. Se procedió a realizar el conteo de párrafos únicamente del personaje “Poet” distinguiendo por el atributo “character_id”, el resultado fue el siguiente:

character_id	Párrafos
894	733
895	3
896	30

Esto es una prueba de que la hipótesis planteada al comienzo es correcta. Un mismo personaje puede actuar en varias obras, y en cada una de ellas tendrá un identificador distinto.

Cantidad de obras en el tiempo

Es de interés entender la producción de obras de Shakespeare a lo largo del tiempo, tanto en cantidad como en género. Con este fín, se probaron distintos tipos de gráficos para encontrar uno que permita visualizar de manera rápida estos dos atributos. Se decidió que la mejor opción es un gráfico de barras para cada género y realizar el recuento acumulado de las obras cada 5 años. Esto último dado que hacerlo año a año no permite observar de manera sencilla si hay alguna tendencia.

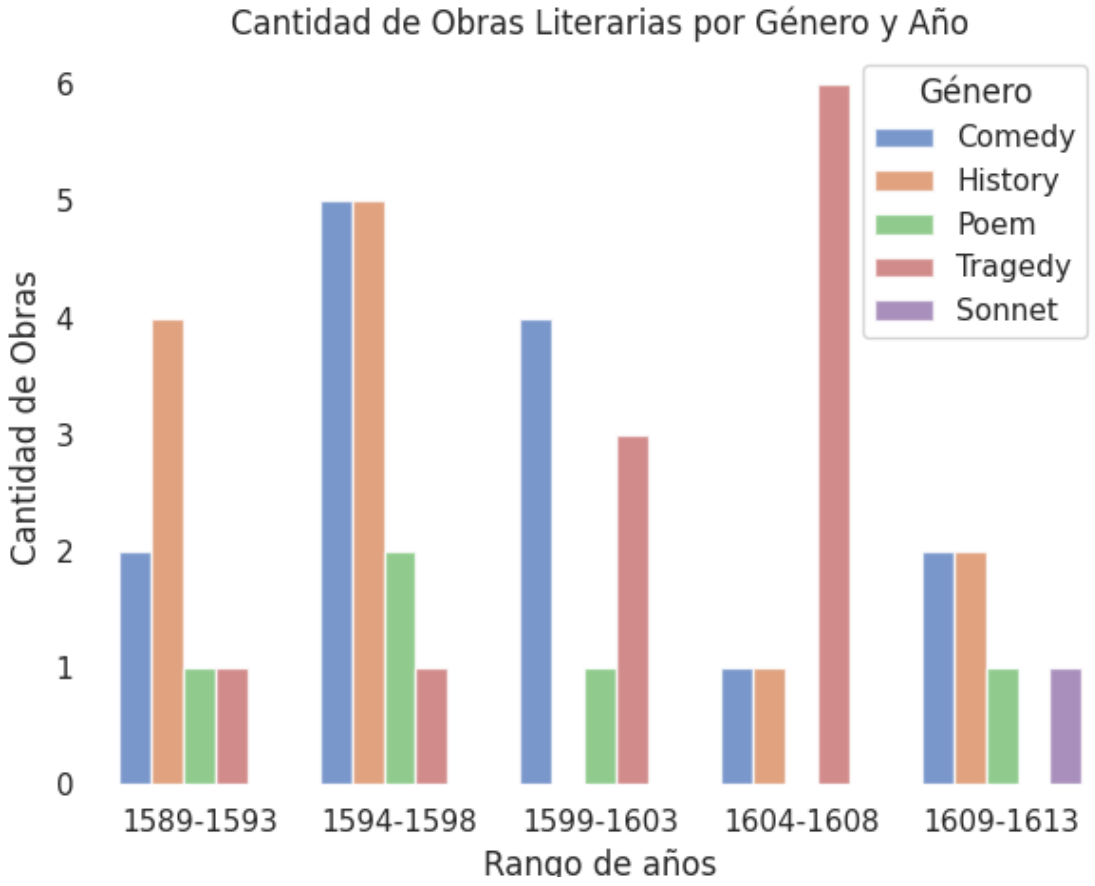

Figura 1: Cantidad de obras por género en grupos de a 5 años.

De la Figura 1 se puede extraer información interesante. Por ejemplo, se aprecia que William Shakespeare comenzó escribiendo mayoritariamente obras de los géneros Comedia e Historia y estos se mantuvieron a lo largo de su vida. Por otro lado, unos años antes de su fallecimiento, incursiona fuertemente en el género Tragedia. Finalmente, se aprecia que escribió 1 obra de género soneto únicamente en sus últimos años de vida.

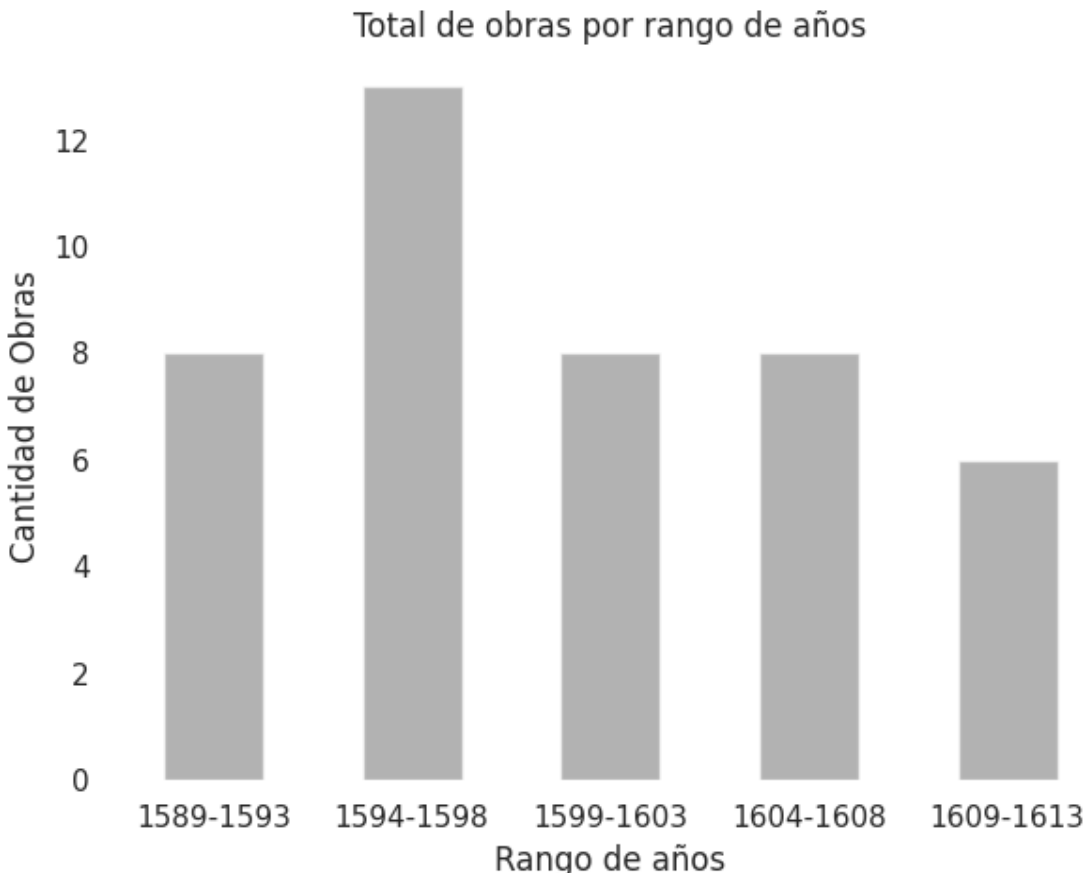

Figura 2: Cantidad de obras en grupos de a 5 años.

A partir de la Figura 2 se deduce que Shakespeare fue aproximadamente igual de productivo a lo largo de su vida, teniendo un máximo de producción entre los años 1594 y 1598. Por otro lado, se observa que en sus últimos años de vida escribió una cantidad menor de trabajos.

Parte 2: Conteo de palabras y visualizaciones

Recuento de palabras

Se decidió dejar el personaje “Poet” dado que se entendió que al no ser un nombre propio varios personajes en distintas obras pueden caer bajo el seudónimo “Poet” sumando así más párrafos y palabras que los demás.

Para comenzar con esta etapa, se implementó una función que normalice el texto, es decir que cambie mayúsculas por minúsculas, que elimine signos de puntuación y que remplace abreviaciones por palabras enteras, por ejemplo “it’s” por “it is”.

Para los signos de puntuación se decidió utilizar *expresiones regulares* para evitar tener que realizar una búsqueda ocular en el texto. A continuación se muestran las 20 palabras más utilizadas a lo largo de todas las obras de Shakespeare.

Figura 3: Recuento de palabras a lo largo de todos los trabajos de Shakespeare.

Con el fin de encontrar diferencias entre géneros o personajes, esta visualización se puede mejorar agrupando el dataframe que contiene las palabras por personaje y género. De esta manera se puede obtener un dataframe que liste cada palabra, la cantidad de veces que la dice un personaje o la cantidad que aparece para un género de obra en particular (o ambas). Para graficar esto se puede crear un gráfico que contenga un conjunto de barras para las 10 palabras más utilizadas, donde cada conjunto represente a distintas personas (podrían ser, por ejemplo, las 5 que utilicen más esa palabra). También se puede realizar lo mismo para graficar por género, estos, al ser menos que cantidad de personajes, se pueden graficar todos y formar grupos de barras, similar a la Figura 1.

Personaje con más palabras

Para el recuento de palabras por cada personaje se procedió a unir las tablas “words” con “characters” y al igual que antes se decidió eliminar el character “(stage directions)” por considerar que no es un verdadero personaje.

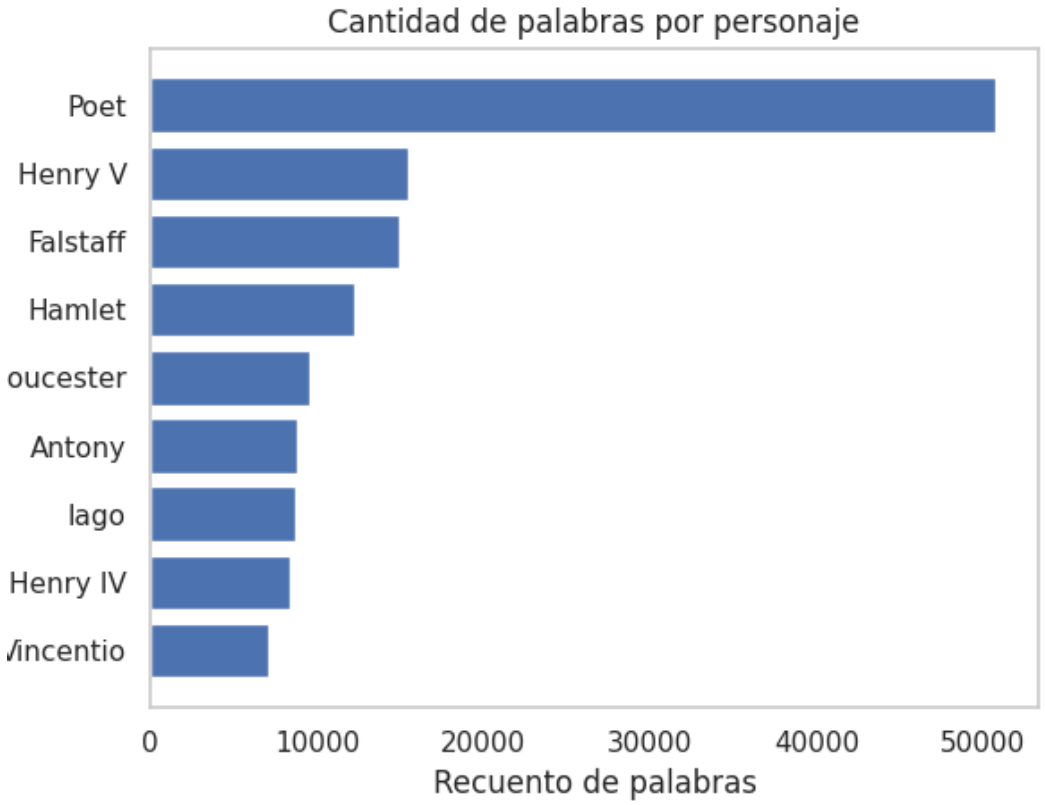

Figura 4: Recuento de palabras por personaje.

A partir de la Figura 4 se puede afirmar que el personaje con más palabras dichas es “Poet”, lo cual no es de extrañar dado que en la Tabla 1 se aprecia que este personaje es quien dice más párrafos, teniendo una gran diferencia con respecto al segundo.

A partir de estos datos se podría responder cuál es la importancia de cada personaje en la obra. Si esto fuera una representación de teatro podría ser utilizado para decidir a qué actores asignarles cada rol (por ejemplo, actores con mayor experiencia o reputación otorgarles papeles que tengan más participación). En el mismo contexto también podría utilizarse como indicador de a quién corresponde pagarle más, quién necesita más atención, etc.

Otro estudio interesante a partir de los datos es entender si hay obras que conviven en el mismo universo, es decir, si hay un conjunto de personajes que se repite en varias obras. Con este objetivo en mente, se podría unir la tabla “works” con “chapters”, está con “paragraphs” y por último con “characters” para así saber cuales son los personajes de cada libro. Luego se podría construir una lista para cada obra que contenga los personajes de la misma y buscar si la intersección de varias de estás listas es distinta de vacío, lo cúal implicaría que uno o varios personajes está en más de una obra.

Hay algunos personajes que es de esperar que estén en varias obras, a saber, los que no tienen nombre propio como “poet” o “messenger”. Se deberían sacar estos tipos de personajes para poder responder de manera certera si hay varias obras que tratan sobre el mismo grupo de personajes.