Introducción a la Ciencia de Datos

Tarea 1

Rossina Primavera
Alina Méndez

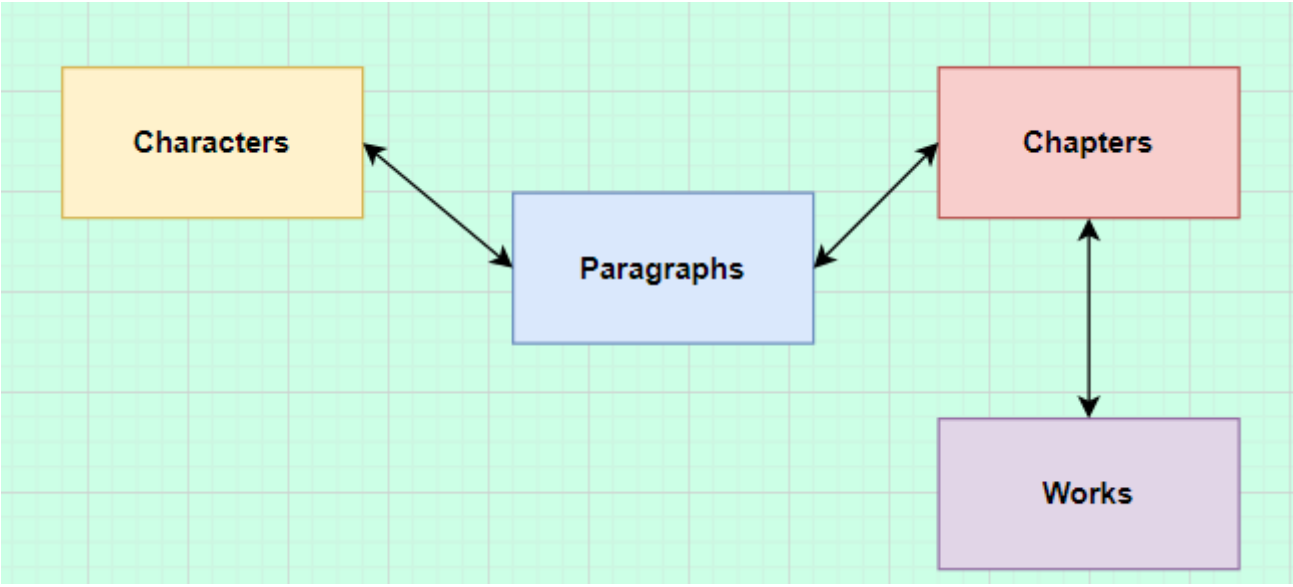
En este informe, se analizará una base de datos relacional abierta cuyo contenido son las obras de Shakespeare. Dentro de la base encontramos cuatro tablas, estas son:

Works (obras)
Paragraphs (párrafos)
Characters (personajes)
Chapters (capítulos)

En cuanto a la función que cumple cada tabla, se puede decir que la tabla *Works* proporciona información sobre las obras realizadas por el autor, incluyendo el año en el que fueron emitidas y el género de las mismas. Luego, se presenta la tabla *Paragraphs* que explicita los párrafos de cada obra, la cual tiene un link a la tabla de *Characters* y *Chapters*. También aparece la tabla de *Characters* que describe los nombres de los personajes de cada obra y, por último, está la tabla de *Chapters* que cumple la función de brindar datos sobre los distintos capítulos de las obras, especificando a qué obra y acto refieren.

Estas tablas se relacionan entre sí de la siguiente manera: la tabla *Paragraphs* contiene el **chapter_id** y **el character_id**, siendo posible identificar a qué capítulo corresponde el mismo (vínculo directo con la tabla *Chapters*), y qué personaje es el encargado de decirlo en la obra (vínculo con la tabla *Characters*). De esta forma se vinculan las primeras tres tablas, restando conocer cómo se relaciona la obra.

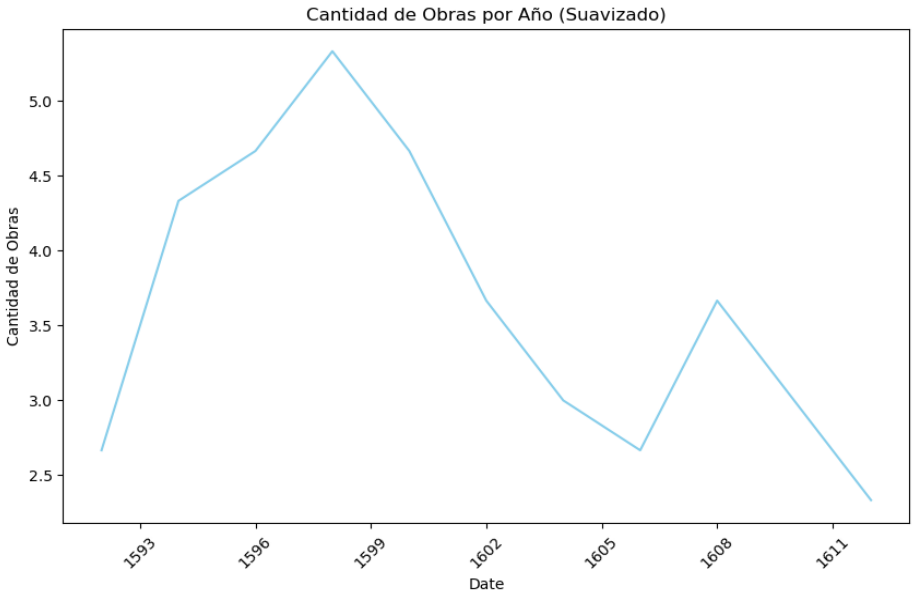
En la tabla *Chapters* podemos encontrar el código que vincula directamente a la obra que corresponde el mismo (lo hace a través del campo **work_id**). Debajo se muestra un diagrama que ejemplifica lo antes expuesto.

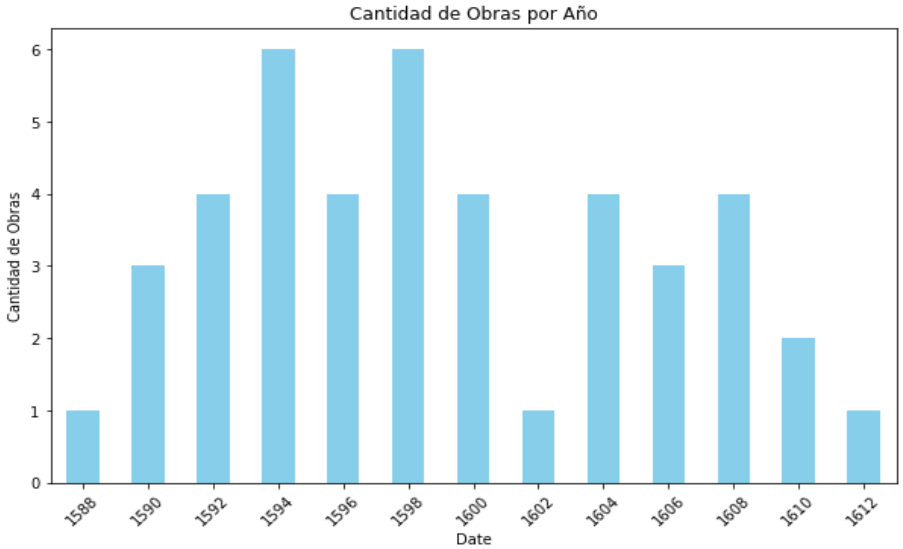


Por otro lado, se visualiza que solamente en la tabla *Characters* existen datos faltantes (null) para los campos de “abreviatura” y “descripción”. En la columna “abreviatura” existen 5 datos faltantes, por lo cual, no es una cantidad significativa sabiendo que la tabla contiene 1266 registros. En cambio, en la columna “descripción” hay 646 datos faltantes, significando el 51% de los registros para ese campo en nulo, lo cual es un porcentaje significativo. Más allá de estos datos faltantes, se considera que este problema no resulta ser grave en términos de calidad de datos ya que las relaciones entre las tablas no se ven afectadas. Adicionalmente, se encuentra que en la tabla Characters la columna de “abreviatura” en varios registros tiene el nombre del personaje de la obra en vez de una abreviatura del nombre, lo cual es inconsistente con la función de la columna.

Cuando se realiza el análisis de la cantidad de párrafos por personaje, vemos que el personaje que tiene más párrafos es inconsistente porque su nombre es “ stage directions”, es decir, no pareciera ser un personaje de una obra. Debido a esto, buscamos el segundo personaje con más párrafos (“Poet”) y sucedió la misma situación, el cual tiene como descripción “The voice of Shakespeare's poetry”. Recién el tercer personaje parece ser consistente, siendo el mismo el de id 393, de nombre Falstaff y de descripción Sir John Falstaff. Mencionamos lo anterior dado que consideramos que es un problema de calidad de datos, la columna “nombre” de la tabla *Characters* estaría trayendo nombres que no involucran personajes de obras. Este personaje aparece en 471 párrafos.

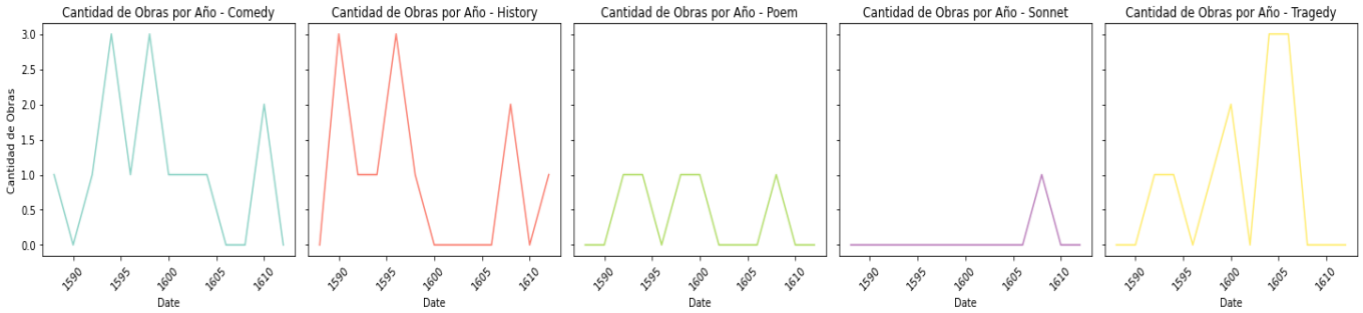
Para visualizar la obra de Shakespeare a lo largo de los años, comenzamos graficando la cantidad de obras realizadas por año. Decidimos agrupar las obras emitidas cada dos años y suavizar la tendencia con el fin de mejorar la interpretación de la misma. Vemos que la cantidad de obras emitidas comienza con una tendencia creciente desde el año 1593 hasta 1598, año a partir del cual se visualiza una caída pronunciada en las obras emitidas por año. En 1606 aprox. se visualiza un leve crecimiento en las obras emitidas, el cual tiene su punto de inflexión en 1608, año en el que las mismas decaen de nuevo hasta el 1613.





Posteriormente, graficamos las obras emitidas por año discriminadas por género, realizamos un gráfico para cada género respectivo. Al igual que en el anterior gráfico, decidimos agrupar las obras en dos años y suavizar los gráficos. A través de los gráficos es posible visualizar que los géneros en los

que el autor generó más obras fueron la comedia, la historia y la tragedia. Estos tres géneros presentan una evolución similar en cuanto a la cantidad de obras emitidas por año. En cambio, las obras del género poesía se produjeron en menor cantidad aunque de manera más significativa que los sonetos, los cuales aparecen en el último tramo de producción de obras del autor (desde 1605 al 1610).



Por otro lado, con el fin de normalizar las tablas y así proceder a realizar el conteo de palabras, decidimos desconsiderar los siguientes caracteres: “.”, “?”, “;”, “!”, “]”, “:”, “ ‘ “.

Un problema que encontramos en el dataset de párrafos es que existen palabras abreviadas, como por ejemplo “there’s”, donde entendemos que para contar correctamente deberemos considerarlo como dos palabras distintas, siendo estas “there” e “is”. Si bien es posible crear funciones que limpien estas abreviaciones, el problema surge cuando la abreviación no sigue una regla unívoca, quiere decir que, la palabra abreviada no es siempre la misma sino que depende del contexto, como por ejemplo “ ‘d”, en este caso puede ser la abreviación de “had” pero también de “would”. Entendemos que en caso de que el modelo que vaya a aplicarse al dataset sea sensible a estas palabras, se debiera hacer la correspondiente limpieza, mientras tanto y para las preguntas que se responderán en este trabajo, se consideran las palabras con apóstrofe como palabras distintas.

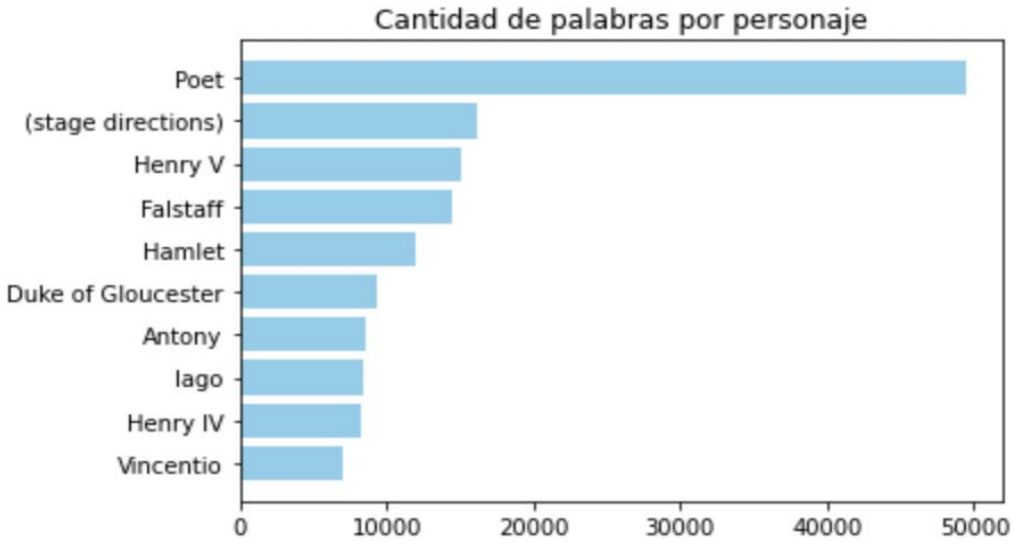
Con el fin de ilustrar las palabras más frecuentes de toda la obra del autor, se plantea la visualización siguiente:

Adicionalmente, si nos proponemos modificar la anterior visualización con el fin de encontrar diferencias entre géneros o personajes, se podría agregar la dimensión de género o personaje y que en cada barra se muestre con colores distintos la proporción en la que la palabra aparece en cada género

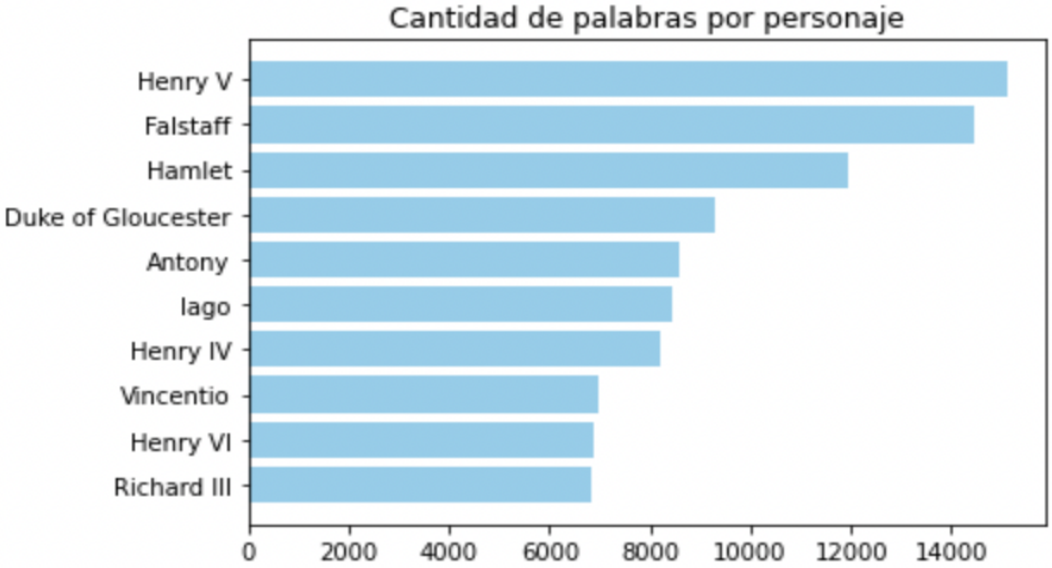
o diálogo del personaje, con el fin de poder discriminar el género o el personaje en la que más aparece esa palabra.

Por otro lado, se realiza una visualización que permite ver los personajes con mayor cantidad de palabras. En la misma, se detecta una inconsistencia dado que aparece que “poet” y “stage directions” son los personajes con mayor cantidad de palabras cuando en realidad no corresponden a verdaderos personajes de las obras. Debido a esto, se decide desconsiderar los respectivos character_id de las obras del autor mediante un drop y volver a crear la visualización. De esta forma, corregimos la anterior inconsistencia dado que ahora solamente se consideran verdaderos personajes.

- Visualización sin modificaciones:



- Visualización con modificaciones:



Al quitar los datos mencionados anteriormente, es posible observar la forma de la visualización varía mínimamente. Principalmente, la palabra “in” que antes estaba en la posición 9 de palabras más frecuentes, ahora pasa a la posición 10 y la palabra “that” que estaba en la posición 10 ahora pasa a la posición 9. También la cantidad de veces que aparece cada palabra disminuye, coherente con haber limpiado los registros del dataset.

Es importante notar que las palabras más frecuentes en la obra son en su mayoría preposiciones o pronombres, coherente con lo que podríamos esperar. Sin embargo, dependiendo del tipo de análisis que queramos hacer si deberíamos considerarlas o no.

Para finalizar, presentaremos una posible pregunta que podríamos responder a partir de los datos, ¿existen personajes que siguen patrones de lenguaje durante la obra? ¿Usan palabras o frases de forma frecuente en sus diálogos, o incluso frecuenta los mismos temas?. Como análisis extra, podría compararse entre personajes, y analizar el contexto histórico de la obra, para encontrar conexiones con la vida del autor.

Para realizar esto, podríamos hacer el conteo de las palabras o frases por personaje, obra, género o fecha. A medida que vamos encontrando tendencias, podemos ir buscando con mayor profundidad. Ejemplo, si encontramos un personaje que repite frases en una cierta obra, podemos analizar si esto sucede dentro del mismo género, y/o si es una tendencia que varía con el tiempo.