Informe Tarea 1
Introducción a la Ciencia de Datos.

Grupo 5
Agustín Porley; Julián Rodríguez

El objetivo principal de esta tarea es analizar una base relacional con la obra completa de William Shakespeare, dicha base de datos cuenta con cuatro tablas:

- *paragraphs*
- *characters*
- *chapters*
- *works*

La tabla *paragraphs* es utilizada para guardar los párrafos de las obras. En su estructura podemos apreciar que existe una columna donde se guarda el párrafo como texto plano (*PlainText*). Cada entrada en esa tabla tiene asignado un número de identificación (*id*), un número correspondiente al párrafo (*ParagraphNum*), un número indicando el personaje a quien corresponde el párrafo (*character_id*) y un identificador de capítulo (chapter_id). El párrafo almacenado se encuentra casi totalmente identificado a excepción de la obra a la cual pertenece.

La tabla *characters* es usada para guardar la información de los personajes. Cada entrada en la tabla cuenta con un número identificador (*id*), un nombre de personaje (*CharName*), una abreviatura del nombre (*Abbrev*) y una descripción (*Description*). Como se puede deducir esta tabla está relacionada directamente con la tabla paragraphs, donde se guarda el número identificador de cada personaje, permitiendo buscar el nombre del personaje en la tabla *characters* a través de su identificador.

La tabla *chapters* tiene la función de guardar la información referente a los capítulos. En ella se puede encontrar la información del acto (*Act*) y la escena (*Scene*) a la que pertenece cada capítulo, así como también su número identificador. Del mismo modo que se relacionaban las tablas *paragraphs* y *characters,* podemos obtener el nombre del capítulo al que pertenece un párrafo, extrayendo el numero identificador del capítulo de la tabla *paragraphs* y buscar su nombre en la tabla *chapters.* En esta tabla también hace referencia a la obra en la que se encuentra cada capítulo, la misma queda determinada por un identificador (*work_id*).

La tabla *works* guarda la información referente a las obras, donde se puede encontrar el número identificador (id), el título (Title), el título largo (LongTitle), el año en que fue escrita (Date), y el género a la que pertenece la obra. Con esta información, podemos extraer el identificador de la obra a la que pertenece cierto capítulo y encontrar toda la información de esa obra en la tabla works. Como es de esperarse, esta última tiene una relación directa con la tabla chapters.

Como resumen se presenta el diagrama de la Figura 1, que indica la relación entre las tablas de la base de datos relacional.

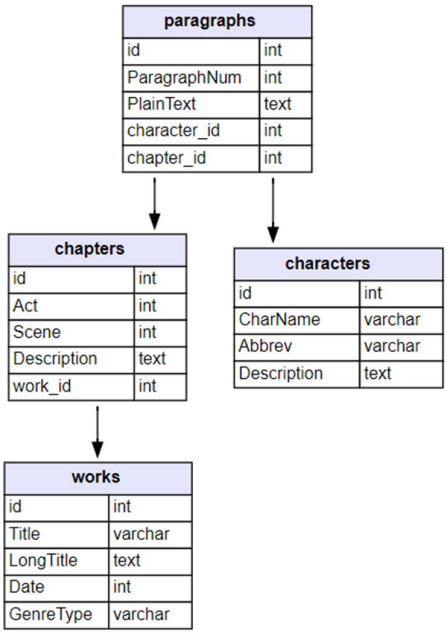

Figura 1: Relación entre las tablas.

Realizando un primer acercamiento a la calidad de los datos se puede analizar la falta de datos. Dentro de la tabla “Characters” existen datos faltantes, específicamente dentro de la categoría “Description” hay 646 espacios vacíos de 1266 y además la categoría “Abbrev” cuenta con 5 datos faltantes del total de 1266. El resto de la base de datos parecería estar completa. También es de interés analizar la cantidad de párrafos por personaje, en este caso se puede apreciar que la mayor cantidad de texto en las obras son direcciones de escena y no diálogos de personajes. Aun así, el personaje con más diálogos es el poeta o “la voz poética de Shakespeare”.

Para visualizar la producción de obras de Shakespeare a lo largo de los años se puede graficar un histograma, dividido por períodos, donde se vea la cantidad de obras generadas en cada período.

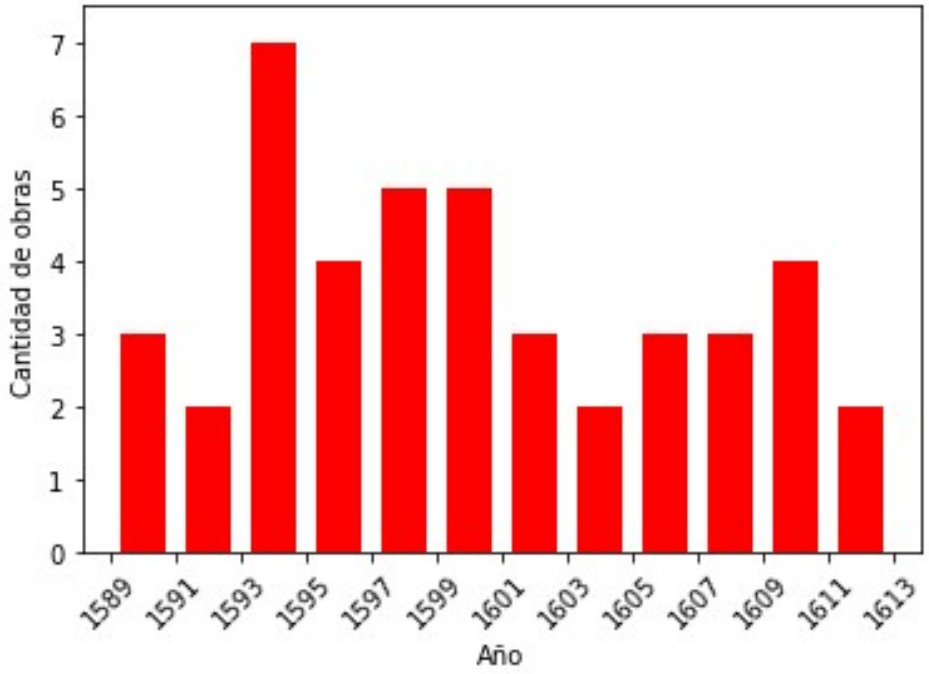

Figura 2: Histograma de obras por períodos.

En la Figura 2 se puede observar cómo su período más productivo fue entre los años 1593 y 1601.
Si realizamos el grafico análogo, pero clasificando la cantidad de obras por género se obtiene la Figura 3.

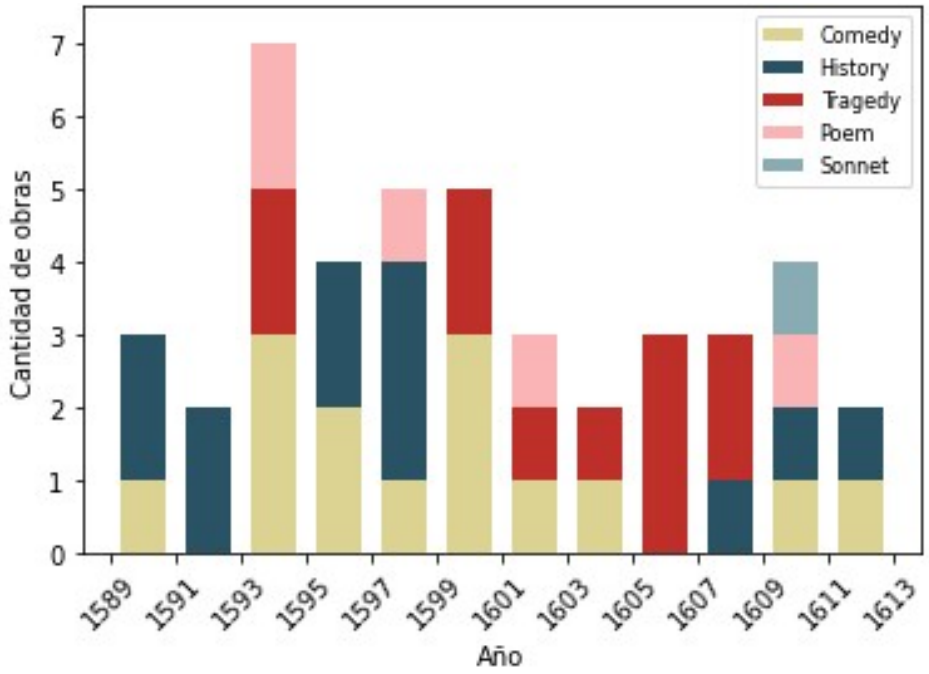

Figura 3: Histograma de obras por períodos apilado por género.

En la Figura 3 se puede observar que inició su carrera escribiendo obras de comedia e historia. Como se puede apreciar, la escritura de comedias fue una constante a lo largo de su producción.

Algo a recalcar es que hubo un período entre el 1599 y el 1607 donde no produjo obras de historia. Por último, al finalizar su carrera se introdujo en el género de los sonetos.

Con el fin de extraer más información de los datos, se decide realizar una limpieza de la base de datos. Para esto, se agregan los siguientes signos de puntuación: ";", "]", ".", ":", "!", "¡", "?", "¿", "-". Adicionalmente se transforman las palabras que comienzan o finalizan con comillas simples agregando a la función “ ‘” y “’ ”. En este caso las expresiones de posesión de palabras en plural serán tratadas como una palabra terminada con comilla simple y por lo tanto no será contada como una palabra diferente. A modo de ejemplo, si en el texto estuviera la frase “My parents' car.”, luego de la transformación quedaría “my parents car”. En este caso “parents’ ” y “parents” serán tratados como la misma palabra. En el caso de las expresiones de posesión en singular o las contracciones gramaticales habituales, serán tratadas como palabras diferentes de las que no tengan contracciones. Si en el texto aparece la expresión “What’s”, luego de la transformación queda como “what’s” y es tratado como una palabra diferente de “what”.
Luego de realizada la limpieza, se puede extraer fácilmente las palabras mas frecuentes. Por medio de un grafico de barras (Figura 4) se logra identificar con facilidad cuales son las palabras más frecuentes considerando toda la obra.

Figura 4: Ocurrencia de palabras en toda la obra.

En caso de querer encontrar diferencias entre géneros se podría implementar un gráfico de barras apiladas, donde se apilarían por género. Por otro lado, si el interés es diferenciar entre personajes se podría tomar los cinco personajes con más palabras y una sexta categoría que englobe al resto de los personajes, para luego implementar el grafico de barras apiladas al igual que con los géneros.

Ya evaluamos que personaje posee mayor cantidad de párrafos, pero algo interesante es observar cual personaje posee mayor número de palabra (fig. 5).

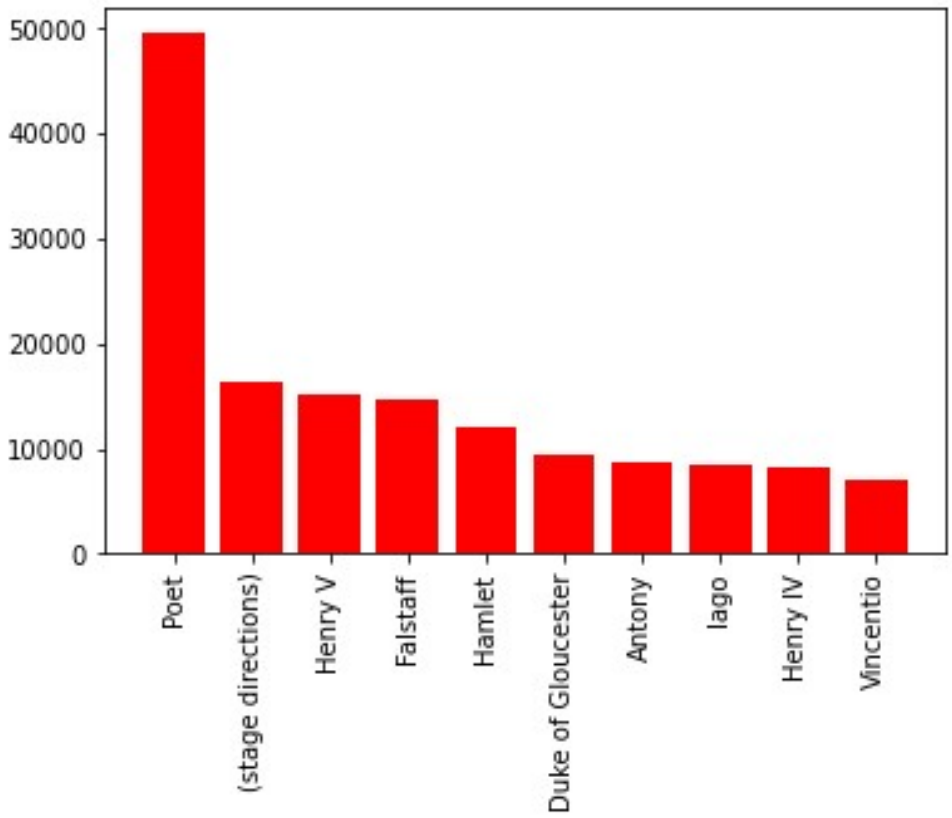

Figura 5: Personajes con más palabras.

Lo que primero llama la atención es que hay un personaje con muchas más palabras que el resto. Como se agrupó por nombre tal vez personajes con el mismo nombre se contaron de manera repetida pero que en realidad tienen un número de identificación diferente.
Lo otro que resalta es que aparece una el nombre "(stage directions)" nuevamente el cual no es estrictamente un personaje, sino que son las explicaciones para el movimiento de los personajes en escena.
En este caso habría que omitir en la visualización a "(stage directions)" y corroborar que el nombre "Poet" corresponde a un solo identificador. En caso de que sea más de uno se debería volver a ordenar filtrando por identificador y luego asignando nombre. Aun así, parece lógico que la voz poética sea el "personaje" con mas palabras debido a que los poemas se construyen de la misma a diferencia de las obras donde las direcciones escénicas tienen un papel secundario.

Al finalizar es importante considerar que otras pregustas se podrían contestar, a continuación, se presentan ejemplos y una posible forma de responderlas:

- ¿Cuáles son las obras que tienen más palabras?
  - Se debería realizar el ranking de palabras agrupados por obras
- ¿Los años que escribió más obras se corresponden a los años que escribió más palabras?
  - Se podría hacer un histograma de cantidad de palabras escritas por períodos.
- ¿Qué tan extensas son las obras de acuerdo con su género?
  - Se podría realizar el ranking palabras agrupando por género.
- ¿Los personajes aparecen en más de una obra?
  - Se podría determinar la ocurrencia de cada personaje en las obras.
- ¿Cuál es la obra con más personajes?
  - Se puede hacer un ranking de cantidad de personajes agrupando por obras.

- ¿Los personajes con mayor cantidad de palabras pertenecen a obras distintas?
    - Se podría realizar un ranking de palabras agrupando por personajes en un gráfico de barras, donde cada barra tiene apilada la cantidad de palabras del personaje en cada obra. Luego se observa si los personajes comparten obras.