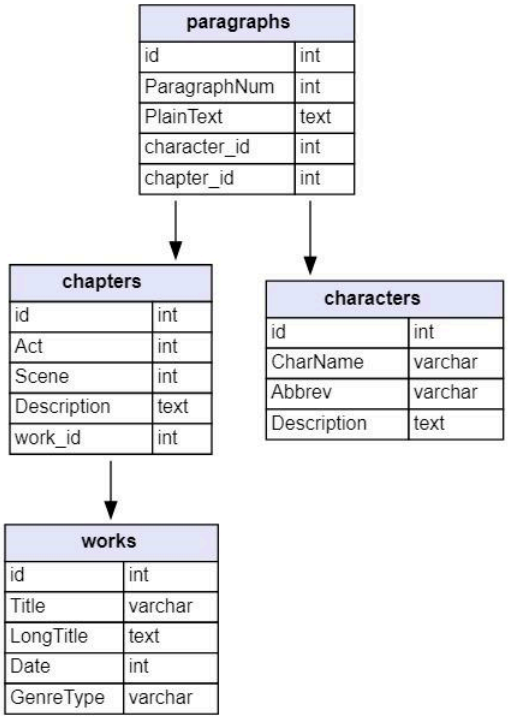
TAREA 1 – Intro CD (Gastón Acosta - Martín Sacco)

PARTE 1:

A -

La base de datos con la que se trabajará en la Tarea 1 se trata de una base de datos relacional. Tenemos cuatro tablas que contienen distinta información respecto a la obra de Shakespeare con el siguiente esquema de relaciones para el modelo de Base de Datos:



- La tabla “works" agrupa las obras escritas durante los años. Las clasifica según su género (columna “GenreType”), año de realización (columna “Date”), título (columna “Title”), título largo (columna “LongTitle”) y además, la tabla cuenta con una columna “id” que corresponde a un identificador numérico para cada obra. Este identificador id se entiende será la clave para relacionar los datos entre tablas según el modelo de la Base de Datos. También se observa un valor natural en una columna sin nombre que en principio no tendría razón de ser más que un contador de filas, comenzando desde cero.

- La tabla “chapters” reúne los capítulos correspondientes a todas las obras. Cuenta con una columna “Act” que se refiere al número de acto, una columna “Scene” correspondiente al número de escena para cada acto detallado, una columna “Description” correspondiente a una breve descripción de cada una de las escenas y también, cuenta con una columna “id” correspondiente a un identificador numérico para cada capítulo/escena. A su vez, cada capítulo se relaciona con su respectiva obra a través de la columna “work_id”, que como se mencionó en la tabla anterior, es el valor que permitirá relacionar las tablas según el modelo de Base de Datos definido. También se visualiza una columna sin identificación, en la cual se detalla un natural consecutivo comenzando desde 0 que en principio no tendría razón de ser más que la de contar filas.

- La tabla “characters” agrupa los personajes de las distintas obras. La misma cuenta con una columna “CharName” que contiene los nombres de los personajes, una columna “Abbrev” correspondiente a las formas abreviadas de dichos nombres, una columna “Description” que contiene una breve descripción de quién es cada personaje y la columna “id” que contiene los identificadores numéricos para cada personaje. Respecto a la calidad de los datos de esta tabla, cabe destacar que, la columna “Description” está incompleta faltando descripciones para varios personajes. Sumado a esto anterior, también se ven varios datos repetidos o con nombres que no parecerían tener valor en principio. Cabe destacar que se sigue manteniendo la existencia de la columna sin identificación como se mencionó en el resto de las tablas.

- La tabla “paragraphs” agrupa todos los párrafos correspondientes a todos los capítulos de todas las obras. Esta contiene la columna “ParagraphNum” correspondiente al número de párrafo dentro de la obra correspondiente, la columna “PlainText” que es el texto que contiene el párrafo correspondiente, la columna “character_id” que se refiere a la columna “id” de la tabla “characters” y, por lo tanto, relaciona cada párrafo al personaje correspondiente. También, la columna “chapter_id” que se refiere a la columna “id” de la tabla “chapters” y relaciona cada párrafo con el capítulo en el cual se encuentra el mismo. Por eso se observan valores de “chapter_id” repetidos, al igual que pasa con los valores de “character_id”, porque un personaje puede estar asociado a más de un párrafo y un

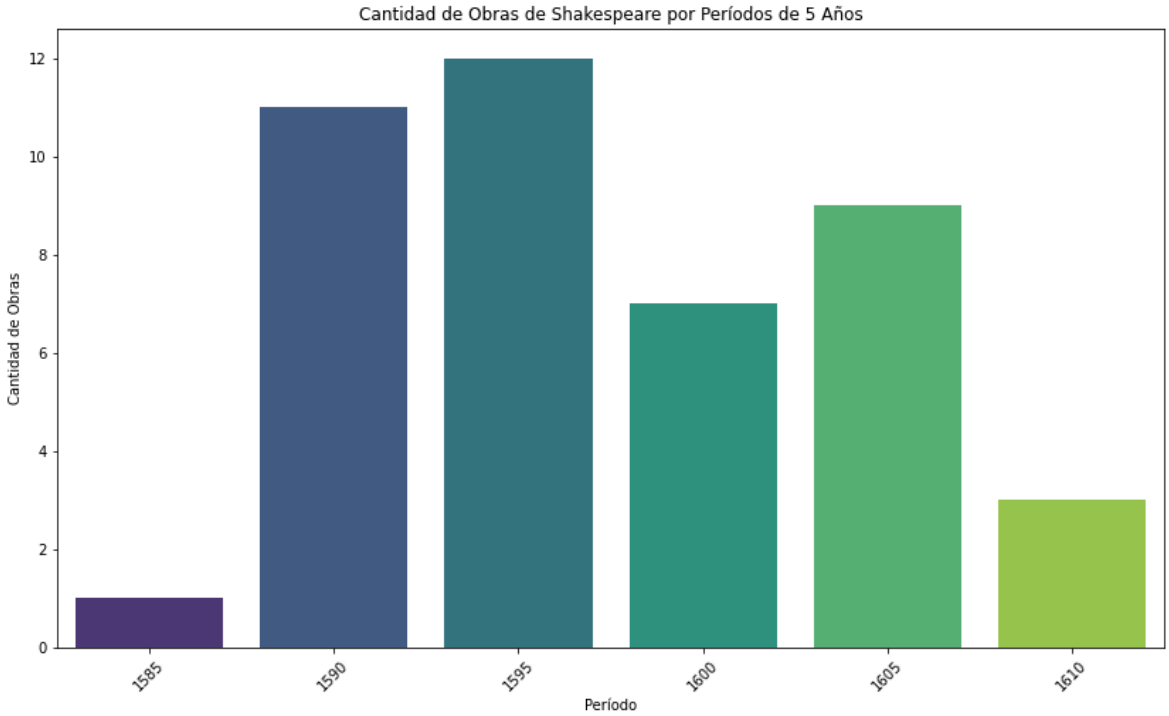
capítulo contiene varios párrafos. Por último, la columna “id” que de igual manera que sucede con las demás tablas es un identificador numérico, en este caso, de cada párrafo de la totalidad de las obras. Se observa en este caso también la existencia de la columna sin identificación.

Un aspecto a comentar es que el valor numérico de la columna “id” comienza con el número 1 en las tablas “works” y “characters” y no comienza con el valor 1 en las tablas “paragraphs” y “chapters”.

Contestando la pregunta de cuál es el personaje con más párrafos. Al ejecutar el código para obtener dicha información el resultado es que el personaje con más párrafos corresponde al “character_id” 1261 con 3.751 párrafos asociados. Si observamos a quién corresponde dicho “id” en la tabla “characters” vemos que no corresponde a ningún personaje sino a “stage directions”, es decir, a acotaciones que se realizan durante el guión de las distintas obras. Depurando el código para obtener los 5 personajes con más párrafos (incluyendo el “id” 1261) tenemos que el segundo personaje con mayor cantidad de párrafos es aquel que tiene el “id” 894, con 733 párrafos. Nuevamente nos fijamos a quién corresponde dicho “id” en la tabla “characters” y vemos que es el “Poet” que según la “Description” asociada es “the voice of Shakespeare´s poetry”. Una observación de este resultado es que al inspeccionar la tabla “characters” vemos que el personaje “Poet” se repite 3 veces con los “id” 894, 895 y 896. De todas formas, en el DataFrame “paragraphs_by_character” generado en el archivo python se observa que todos los “id” mencionados tienen párrafos asociados. Por lo tanto, asumimos que son personajes diferentes. El tercer personaje con más párrafos es aquel que tiene el “id” 393, con 471 párrafos. Nos fijamos a quién corresponde dicho “id” en la tabla “characters” y vemos que es “Falstaff” que según la “Description” asociada es “Sir John Falstaff”.

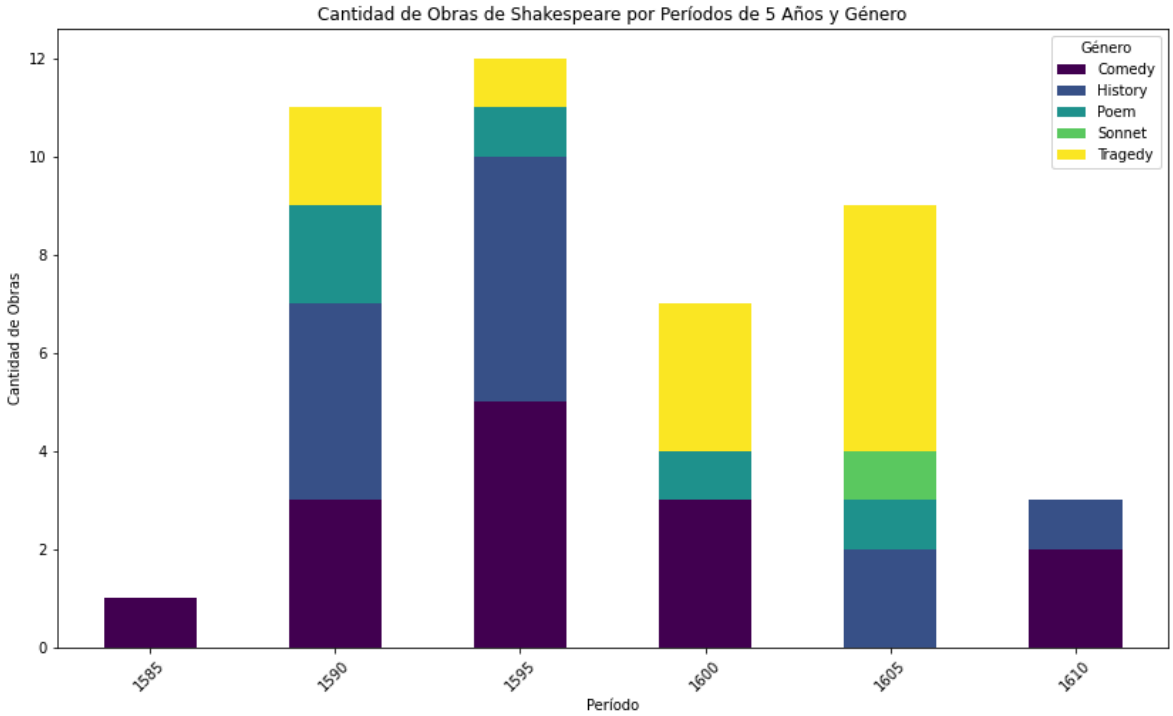
B -

A continuación se presenta el gráfico que muestra la evolución de la cantidad de obras en el tiempo. Se agruparon las mismas en períodos de 5 años:



En este gráfico se puede ver la evolución de la cantidad de las obras realizadas a lo largo del tiempo. Se observa que en los períodos 1590-1595 (representado como 1590) y 1595-1600 (representado como 1595) es donde realizó la mayor cantidad de obras con 11 y 12 obras, respectivamente. Seguidos por el período 1605-1610 (representado como 1605) con 9 obras totales. Ya en el último período 1610-1615 (representado como 1610) la cantidad de obras baja considerablemente, resultando ser 3.

Puliendo el código para obtener además de la cantidad de obras por períodos de 5 años los géneros de las mismas tenemos el siguiente gráfico:



En este nuevo gráfico se pueden observar los géneros de las distintas obras realizadas por Shakespeare a lo largo del tiempo. Se aprecian aspectos interesantes al analizar este gráfico. En primer lugar, en sus inicios los géneros preferidos por el dramaturgo fueron la “Historia” y la “Comedia”. En los primeros 3 períodos de las 24 obras que escribió 9 fueron del género “Comedia” y 9 fueron del género “Historia”, 18 obras de las 24 totales se distribuyeron entre estos géneros. Luego, en los siguientes dos períodos se observa claramente un incremento en las obras del género “Tragedia”, tal vez, reflejando etapas difíciles de la vida del escritor. Finalmente, se ve una merma en la cantidad de obras del último período respecto al penúltimo, 3 y 9, respectivamente. Y no se observan obras del género “Tragedia” en este último período.

C -

En función de realizar conteo de palabras, realizamos algunas modificaciones sobre la columna “PlainText” de la tabla “paragraphs”. Primero, se definió la función clean_text que tiene como objetivo transformar a minúscula todas las palabras de la columna mencionada. Luego, reemplazamos todos los signos de puntuación identificados por espacios y adicionamos una nueva columna en el DataFrame “df_paragraphs” llamada “CleanText” que contiene el mismo texto que “PlainText” pero en minúscula y sin signos de puntuación. Lo anterior se realizó

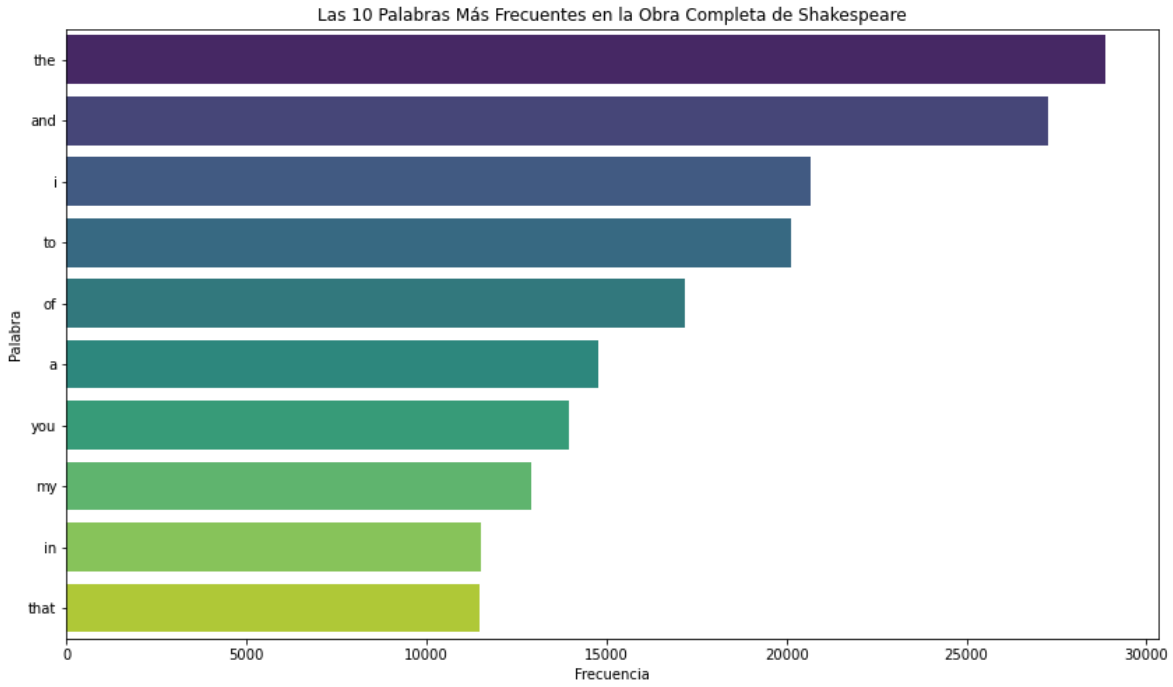
con el objetivo de, en el paso siguiente, transformar la columna “CleanText” en listas de palabras (columna “WordList” que se agrega al DataFrame “df_paragraphs”). Esto es posible al tener espacios y no signos de puntuación entre las palabras ya que simplifica la utilización de la función str.split() sin tener que añadir argumentos. A continuación, se genera un nuevo DataFrame llamado “df_words” donde cada fila ya no es un párrafo sino una sola palabra. Por último, se eliminan las columnas “CleanText” y “PlainText” del DataFrame “df_words” y se renombra la columna “WordList” con el nombre “word”. Cabe destacar que como resultado se obtienen una cantidad considerablemente mayor de filas, dado que cada palabra queda asociada a un único párrafo, por lo que se generan tantas filas como palabras contiene cada párrafo (35465 a 885575).

Como reflexión, observamos que el signo de puntuación (´) que en el idioma inglés se utiliza como forma de abreviar ciertas palabras (por ejemplo: I will = I´ll) aparece en varias ocasiones. Tomamos la decisión de no incluir dicho signo de puntuación a la hora de sustituirlos por espacios ya que de haberlo hecho iban a quedar letras “sueltas” o similar (tomando como referencia el ejemplo citado, quedaría “ll” en lugar de “wil”) y esto implicaría realizar un paso extra de depuración de las palabras para que tomara a “will” y “ll” como la misma palabra.

PARTE 2:

A-

A continuación se presenta un gráfico para visualizar comparativamente las palabras más frecuentes considerando toda la obra:



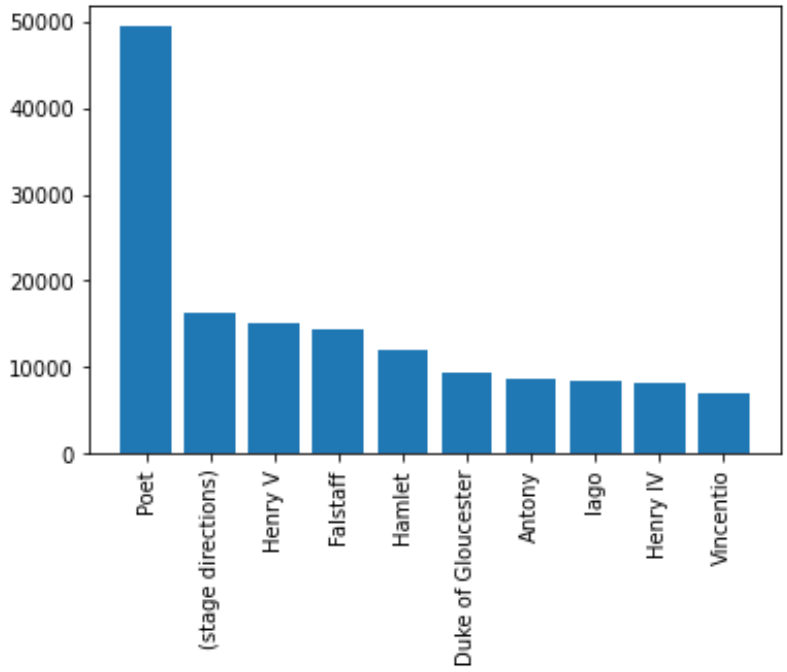
Se observa que la palabra más frecuente considerando toda la obra es “the” con casi 30.000 repeticiones, seguida muy de cerca por “and” con casi 30.000 también.

Para encontrar diferencias en cuanto a la frecuencia de las palabras teniendo en cuenta los géneros o personajes se podría generar un gráfico donde se visualice la palabra más frecuente en cada uno de los géneros. De esta manera se puede comparar entre géneros, si existe alguna similitud a los efectos de la palabra más frecuente. Esto mismo se podría hacer para el resto de las palabras. Utilizar el gráfico de barras con colores parecería ser una buena opción para comparar entre géneros. Otra opción posible sería generar una barra por género, y dentro de la barra que se visualicen en porcentaje, las palabras más frecuentes para cada género. De esta manera, comparativamente por género se puede visualizar de manera relativa la frecuencia por palabras, y entre géneros, el orden de magnitud de cada una de las palabras. Para los personajes se puede generar una visualización similar dado que se cuenta con la identificación de cada personaje respecto a las palabras, generado en el DataFrame del apartado C, Parte 1.

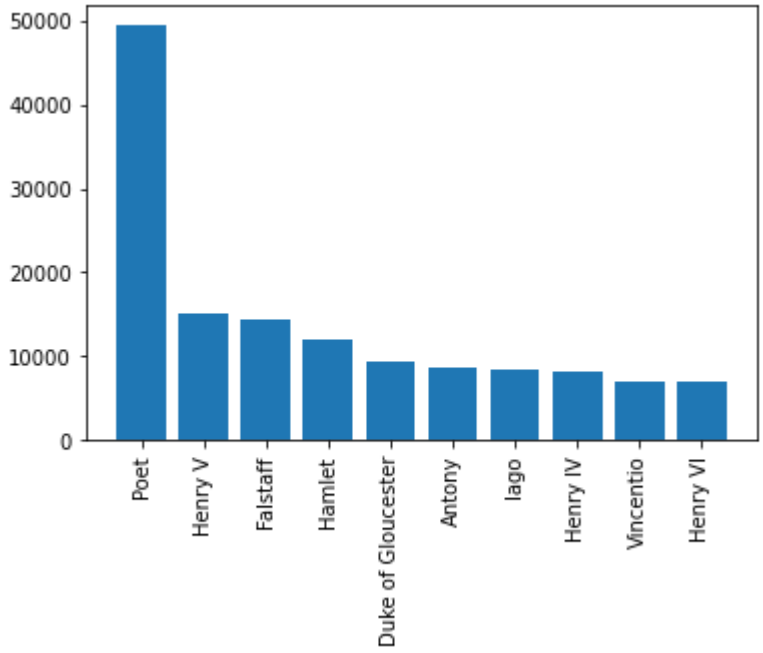
B-

Al ejecutar el código del jupyter notebook en el archivo python para encontrar los personajes con mayor cantidad de palabras sucede que el segundo “personaje” con mayor cantidad de palabras corresponde a “stage directions”. Como comentamos anteriormente “stage directions” no es un personaje de las obras de Shakespeare sino que se refiere a acotaciones que se realizan durante el guión de las distintas obras. Por lo tanto, si queremos considerar sola y exclusivamente a personajes para el conteo de las palabras más frecuentes debemos reflejar en nuestro código que no se tenga en cuenta al “personaje” con el “id” 1261 que corresponde a las mencionadas “stage directions”. A continuación se muestra el antes y después de la visualización gráfica con “stage directions” como “personaje” y luego siendo eliminado.

ANTES:



DESPUÉS:



C-

Preguntas que se podrían responder a partir de estos datos podrían ser las siguientes:

1- ¿Existe alguna relación entre las palabras más frecuentes y el pasar de los años, que se pueda relacionar con la madurez del autor? ¿O se puede decir que fue bastante constante en las palabras que utilizó para escribir, independientemente del género y los años transcurridos?

Teniendo en cuenta el DataFrame generado con la columna “words” y la relación entre las tablas “paragraphs”, “chapters” y “works”, a través de la columna “chapter_id” en la tabla “paragraphs” y la columna “work_id” en la tabla “chapters”, es posible relacionar ambas tablas y llegar al dato entre la fecha “Date” de la tabla “works” y las palabras por fecha. Posteriormente con un gráfico de barras se visualizaría los datos y se aventuraría la respuesta.

2- ¿Existe alguna tendencia de preferencia de palabras por género que sea visiblemente diferenciable entre géneros? ¿Y si se incluyera el pasar del tiempo también? ¿Estas preferencias cambian? ¿Son las mismas?.

Similar a la relación anterior, con las columnas mencionadas en lugar de utilizar el dato “Date” de la tabla “works”, se utilizaría el dato “GenreType” de la misma tabla y se llegaría a la relación entre palabras y género. Con el mismo tipo de gráficos se podría también aventurar la respuesta.

3- ¿Cuántas palabras en total dejó publicadas el autor a lo largo de su carrera en base a sus obras?

Para este caso se mostraría el dato de cantidad de palabras como una sumatoria de la columna “word” en el DataFrame “df_words”.

4- ¿Cuántos personajes diferentes creó a lo largo de sus obras? ¿Existe algún personaje que haya perdurado en el tiempo y haya vivido en diferentes obras?

Para responder la cantidad de personajes sería similar a la respuesta anterior, mostrando el dato del total con una sumatoria de personajes, teniendo la salvedad de contarlos una vez sola a cada uno de los personajes en la tabla “characters". Para responder la segunda pregunta es necesario relacionar los personajes con los títulos de las obras y el tiempo (relacionar las tablas “characters” con “works”). En este caso el camino estaría siendo desde la tabla “characters” a través del “id”, pasando por la tabla “paragraphs” con la columna “character_id”. Una vez generada esta relación, se vincularía con la tabla “chapters”, a través del “chapter_id”, llegando a los títulos, a través del “work_id” presente en la tabla “chapters”, y pudiendo relacionar así los datos de la columna “CharName”, con los de la columna “Date” y “Title” de “works”. La visualización en este caso debería ser una tabla en la cual, por obra (ordenada temporalmente en algún sentido) se detallen los nombres mencionados en cada una, y se resaltan con colores los que se repiten a lo largo de las 43 obras que el autor generó en el tiempo.

5- ¿Existe algún personaje que haya aparecido en una sola obra?

Con la misma tabla anterior es posible responder a esta pregunta, resaltando los valores únicos que se vean comparando entre columnas de los datos resultantes (ya que cada columna sería una obra).

6- ¿Existe alguna preferencia de selección de género mujer u hombre que cambie al pasar el tiempo, que se relacione con los géneros de las obras?

Primero que nada, para responder a esta pregunta habría que generar una relación nueva entre los personajes e identificarlos como "mujer" u "hombre". Una vez generada esta relación, y con los datos relacionados según la respuesta a la pregunta 4, sería posible contar, por año, cuántos personajes mujeres hubo en las obras correspondientes a ese año, y lo mismo para los personajes hombres. Para relacionarlo con el tipo de obra, bastaría utilizar la variable género, y sumar la cantidad de personajes mujeres en cada obra del mismo género. La visualización para este caso parecería ser de vuelta gráfico de barras apiladas, donde se muestran los años, y el conteo de personajes hombres y mujeres. Lo mismo para los géneros de las obras.

7- ¿Cuántas veces aparece la palabra “she” a lo largo de sus obras? ¿Y cuántas la palabra “he”?

Para responder a esta pregunta bastaría con realizar una suma, desde la columna “word” generada en la tabla “paragraphs”, de la palabra "she" y "he" y visualizar el resultado.