Introducción a la Ciencia de Datos

Tarea 1

2024

Leandro Cantera, Miguel Paolino

1. Cargado y Limpieza de Datos

1. A

Función de cada tabla y relación entre ellas

La base de datos “shakpespeare” tiene 4 tablas:

i. **Works:**

Esta tabla incluye 43 obras de Shakespeare, donde los atributos de cada una son: id, título (Title), Título largo (LongTitle), Fecha (Date) y Tipo de género (GenreType)

ii. **Chapters:**

Esta tabla incluye los capítulos de las obras, segregados según acto y escena. Tiene 945 entidades y cada una de ellas tiene 5 atributos: id, acto (Act), escena (Scene), descripción (Description) y la id de la obra (work_id).

Esta tabla está relacionada a la tabla Works a través de ese último atributo (work_id), de esta manera, se puede saber a qué obra pertenece cada uno de esos capítulos, ya que se vincula al atributo determinante “id” de la tabla “Works”.

iii. **Characters:**

Esta tabla incluye los personajes de las obras de Shakespeare, en total contiene 1266 entidades con 4 atributos: id, nombre del personaje (CharName), abreviación (Abbrev) y descripción (Description). El atributo determinante id va de uno en uno desde el 1 hasta el 1266.

iv. **Paragraphs:**

Esta tabla incluye 35.465 entidades, cada una es un párrafo de una de las obras de Shakespeare. Los atributos de cada entidad son: id, número de párrafo (ParagraphNum), el texto de dicho párrafo (PlainText), el id del personaje (character_id) y el id del capítulo (chapter_id). Estos últimos dos atributos referencian a los atributos determinantes de las tablas “Characters” y “Chapters”, respectivamente. Además, como la tabla “Chapters” nos dice a qué obra pertenece cada capítulo, podemos entonces saber a qué personaje, capítulo y obra pertenece cada párrafo de esta tabla.

Estructura y vínculos entre las tablas

En la Figura 1. Esquema de tablas de la base de datos y los campos que lsa conectan. vemos un esquema donde se muestran los campos que conectan las tablas y nos permiten hacer consultas para sacar información o filtrar como ser aquellas que son poemas o sonetos que no tengan un capítulo asociado.

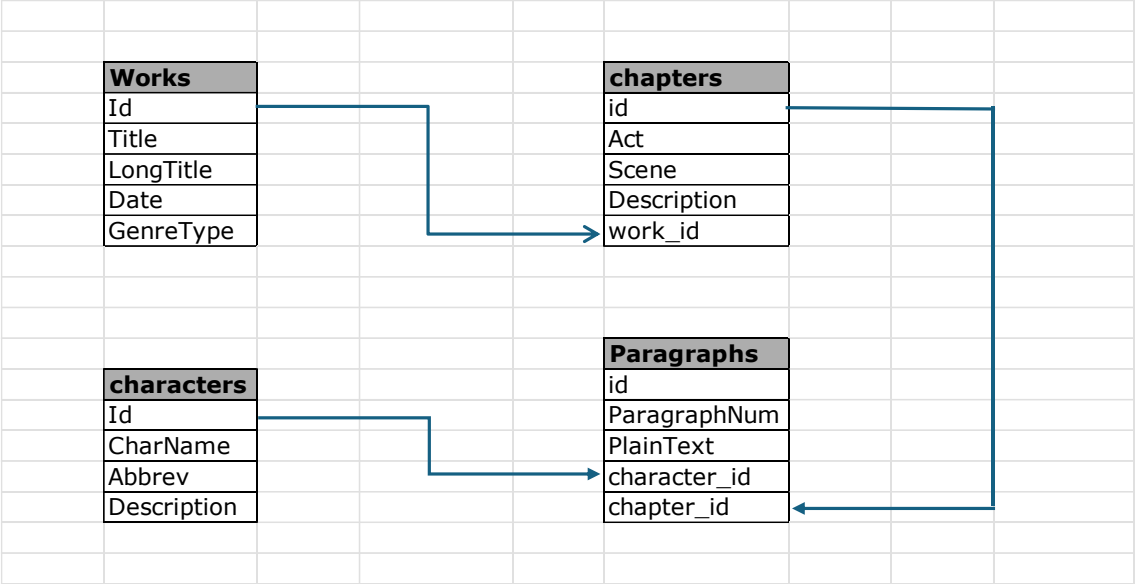

Figura 1. Esquema de tablas de la base de datos y los campos que lsa conectan.

Datos faltantes y otros problemas de calidad

i. **Works:**

En la tabla "Works" no parece haber datos faltantes. La única particularidad que se podría comentar es que no están ordenadas cronológicamente. Sí están ordenadas por título, salvo el primer individuo, obra de título "Twelfth Night", que si el "Twelfth" estuviera escrito en formato de número ("12th"), entonces sí estaría correctamente ordenado.

Por otro lado, ninguna fecha parece estar errada, ya que todas las fechas (del 1589 al 1610) están comprendidas en el período en que Shakespeare estuvo vivo (del 1565 al 1616).

ii. **Chapters:**

Esta tabla tiene un total de 169 entidades con valor = "---" en el atributo "Description". Todas esas entidades corresponden al work_id 35, que son sonetos según la tabla "Works" y por lo tanto no tienen capítulos. Varios de esos valores en realidad están como "---" en el archivo csv, pero Pandas lo interpreta como "---/n". En cualquiera de los dos casos, eso no es información real, es más bien la forma de dar a entender que es un atributo vacío que fue dejado así adrede.

iii. **Characters:**

En este caso, tenemos información faltante en 646 individuos en el atributo "Description", Pandas lo interpreta como "NaN", pero en los datos originales directamente no hay información. A juzgar por el nombre de varois de estos personajes, parecería que no hay descripción porque

son personajes secundarios (ej: Sailor, Roman). En algunos otros casos, no es posible explicar la razón por la que no tienen información (ej: Prince Edward).

Por otro lado, aparecen 23 individuos cuyo "CharName" es "All", además de algunos personajes repetidos, como "First Musician", que aparece 2 veces (id's 7 y 8).

iv. **Paragraphs:**

En esta tabla no se encuentran datos faltantes. Sí se encuentran 3751 individuos cuyo atributo "PlainText" son en realidad acciones que tienen lugar en las obras, ejemplo: [Enter DUKE ORSINO, CURIO, and other Lords; Musicians attending] con id: 630863. En todos estos casos, el nombre del personaje es "(stage directions)", cuyo id en la tabla "Characters" es 1261.

Cantidad de párrafos por personaje

El personaje "(stage directions)" es, además, el "personaje" que más párrafos tiene asignados. Si nos concentramos en personajes que refieren a personas (ej: Orsino, Curio, Maria, etc), es "Falstaff" el personaje que más párrafos tiene asignados, con 471.

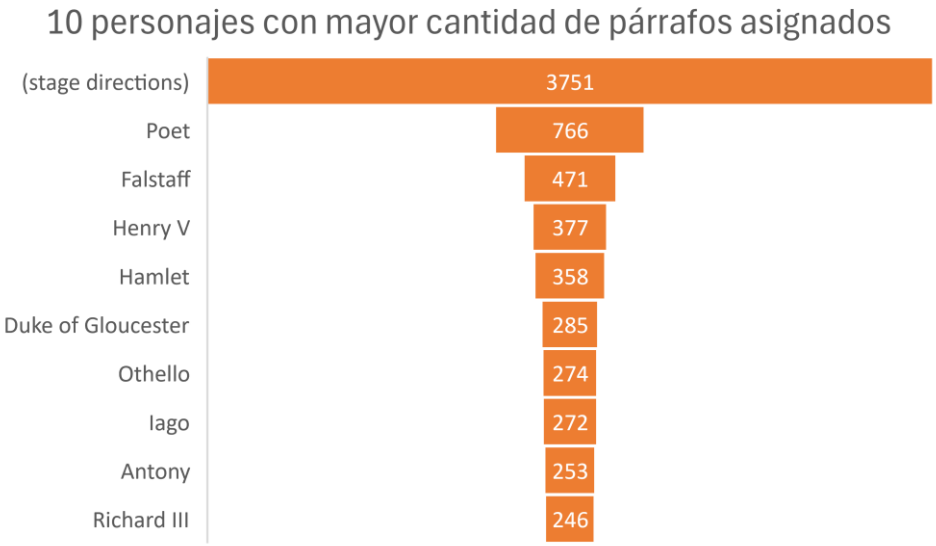

Gráfico 1. 10 personajes con mayor cantidad de párrafos asignados.

1. B

Obra de Shakespeare a lo largo del tiempo

Gráfico 2. Obras de Shakespeare publicadas por lustro a partir de su primera publicación en 1589. Su última publicación fue en 1612.

Período	Comedia	Historia	Poema	Tragedia	Soneto	Total
1589-1593	2	4	1	1		**8**
1594-1598	5	5	2	1		**13**
1599-1603	4		1	3		**8**
1604-1608	1	1		6		**8**
1609-1613	2	2	1		1	**6**

Figura 2. Obras de Shakespeare publicadas por lustro a partir de su primera publicación en 1589 y segregadas por género.

Según la base de datos con la que se está trabajando, Shakespeare publicó un total de 43 obas entre los años 1589 y 1612. La primera mitad de su producción (21 obras) fue publicada es sus primeros 10 años como escritor, entre el año 1589 y el 1598. El resto (22 obras) fue publicada en los siguientes 14 años, entre el 1599 y el 1612.

Si evaluamos su producción por lustro (Gráfico 2. Obras de Shakespeare publicadas por lustro a partir de su primera publicación en 1589. Su última publicación fue en 1612.) podemos observar que entre el primer y el segundo lustro desde que comenzó su producción bibliográfica, fue siendo cada vez más prolífico, hasta llegar a un máximo de publicaciones en 5 años de 13. Luego, su producción fue bajando hasta que, en los últimos 4 años de producción (entre el 1609 y el 1612 – gráfico va hasta 1613 para completar el lustro, aunque no hubo producción en ese último año), la cantidad de obras publicadas llegó al mínimo de 6.

En lo que respecta a los géneros de estas publicaciones, la Figura 2. Obras de Shakespeare publicadas por lustro a partir de su primera publicación en 1589 y segregadas por género. muestra cómo, sobre el inicio de su carrera (desde 1589 a 1603), su producción se centró en los géneros historia y comedia. Sin embargo, con el paso del tiempo pasó a escribir más tragedias, al punto que en el período comprendido entre el 1604 y el 1608, 6 de sus 8 obras fueron tragedias. Finalmente, su único soneto fue escrito en el 1609 sobre el final de su carrera como escritor.

1. C

Para el conteo de palabras, lo primero que se hizo fue agregar los signos de puntuación que faltaban en el código hasta completar la siguiente tabla:

Tabla 1. signos de puntuación incorporados al texto para limpieza.

]	[	\n	,	.	;	:	}	{	-	!	?	‘

Un análisis exploratorio de la tabla “df_paragraphs” una vez que la columna "CleanText" fue anexada a ella, nos muestra que las líneas con signos de puntuación y letras en mayúscula fueron sustituidas por líneas sin signos de puntuación y con todas las letras en minúscula (Tabla 2. Primeras y últimas 5 filas de las columnas “PlainText” y “CleanText” de la tabla “df_paragraphs”.).

Tabla 2. Primeras y últimas 5 filas de las columnas “PlainText” y “CleanText” de la tabla “df_paragraphs”.

	PlainText	CleanText
0	[Enter DUKE ORSINO, CURIO, and other Lords; Mu...	enter duke orsino curio and other lords mu...
1	If music be the food of love, play on;\nGive m...	if music be the food of love play on give me...
2	Will you go hunt, my lord?	will you go hunt my lord
3	What, Curio?	what curio
4	The hart.	the hart
...	...	...
35460	That she is living,\nWere it but told you, sho...	that she is living were it but told you shou...
35461	You gods, look down\nAnd from your sacred vial...	you gods look down and from your sacred vials...
35462	There's time enough for that;\nLest they desir...	there's time enough for that lest they desire...
35463	O, peace, Paulina!\nThou shouldst a husband ta...	o peace paulina thou shouldst a husband tak...
35464	[Exeunt]	exeunt

2. Conteo de Palabras y Visualizaciones

2. A

Visualización de las palabras más frecuentes, considerando toda la obra

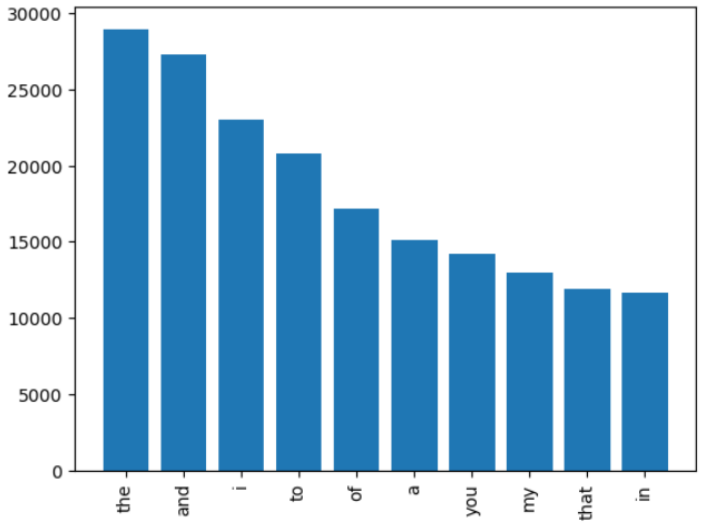

Figura 3. Las 10 palabras más usadas en la obra de Shakespeare ordenadas de forma descendente. En el eje horizontal están las palabras. En el eje vertical está la cantidad de veces que aparece cada palabra en la obra.

Considerando toda la obra, la palabra que más aparece es el artículo “the”, seguida de la conjunción “and”. Según el diccionario de Oxford, las 5 palabras más usadas en la obra de Shakespeare (“the”, “and”, “I”, “to”, “of”) pertenecen a la banda de palabras más usadas en el inglés moderno, con una frecuencia de uso mayor a 1.000 veces por cada 1 millón de palabras típicamente usadas en el inglés moderno[1].

Ideas para modificar la visualización anterior con el fin de encontrar diferencias entre géneros o personajes.

Tomado la idea del código propuesto referente a la agrupación de la cantidad de palabras dichas por cada personaje y la implementación en esta parte 2.A, serían interesantes las siguientes visualizaciones:

- Agrupar las palabras por personaje y luego ver cuáles son las palabras más usadas por cada personaje.
- Hacer lo mismo, pero agrupando por género.
- La cantidad media de palabras por obra según género, de manera de ver si hay géneros cuyas obras son más largas en promedio.
- Siguiendo la categorización por bandas del Diccionario de Oxford, sería posible implementar una nueva tabla que incluya las palabras más usadas en inglés (por ejemplo. Las que tienen frecuencia de uso mayor a 1.000 veces por cada 1 millón de palabras)[1]. Con esa nueva información se podría filtrar las palabras usadas, eliminando estas que son muy usadas y pudiendo tener lago más de luz sobre qué palabras verdaderamente usaba Shakespeare y no qué palabras se usan en el idioma en el que Shakespeare escribía. Esto podría ser aplicable tanto para la visualización original como para las 3 primeras propuestas de esta lista.

2. B

Luego de correr el código que permite encontrar los personajes con mayor cantidad de palabras, comentar los problemas encontrados y las posibles formas de resolverlos.

Como se explica en el código, la particularidad de los resultados es que el personaje con más palabras asignadas es "Poet", que desconocemos si es el mismo personaje a lo largo de todas las diferentes obras. Además, el "personaje" que aparece como segundo en cantidad de palabras es "(stage directions)", que son en realidad indicaciones del escenario como entradas o salidas de personajes (ej: "[Enter VIOLA, MALVOLIO following]").

Una forma de solucionar esto sería revisar la tabla “Characters” y eliminar este tipo de personajes antes de agrupar por personajes y relevar los que dicen más palabras.

[1] Oxford English Dictionary: https://www.oed.com/information/understanding-entries/frequency/

2. C

Preguntas que se podrían intentar responder a partir de estos datos, y posibles caminos para responderlas.

Además de las preguntas que ya hemos respondido, se mencionan a continuación algunas de las preguntas que podríamos intentar respondernos analizando esta base de datos:

- **¿En qué obra aparece cada personaje?**

En este caso, cada personaje tiene su id (en la atabla "Characters") y cada obra tiene su id (en la tabla "Works"), y es posible referenciar la información de los personajes y las obras desde la tabla de párrafos ("Pharagraphs"), como lo muestra la Figura 1.

En este caso, es posible que algunos personajes aparezcan en varias obras.

- **¿Cuál fue el año más prolífico para Shakespeare en términos de producción bibliográfica**

Esta pregunta podría responderse agrupando las obras por año y viendo qué año tiene el máximo de obras publicadas. De una forma análoga, se podría analizar el año menos prolífico desde su primera publicación en 1589 hasta su última publicación en 1612.

- **¿En qué géneros se concentró en escribir ese año?**

Una vez que las obras estén agrupadas por años, se podría seleccionar solamente las obras correspondientes al año más prolífico (o el menos), y agruparlas, esta vez por género, para luego ordenarlo en forma descendente y visualizarlo. Al ser solamente 5 géneros, una visualización sencilla en forma de tabla o de gráfico de distribución podría darnos una buena idea.

- **¿Cuántos actos tiene cada obra? y escenas?**

Para esto, la tabla "Chapters" incluye todos los actos y las escenas que tiene cada obra, e incluye el id de cada obra, por lo que sería posible agrupar las entidades por obra y luego contar la cantidad de actos y la cantidad de escenas. Para esto, una vez agrupadas las entidades por obra, lo más fácil sería buscar el máximo valor dentro de los atributos "act" y "scene", ya que al tener orden ascendente se está contando a sí mismo y no es necesario incluir el código de conteo de individuos. Además, el conteo de individuos no sería del todo útil para los actos ya que hay varias escenas por acto, en este caso se deberían contar cuántos valores diferentes hay en cada subgrupo de la tabla "Chapters" dividida según las obras.

- **¿Cuál es la obra más larga en términos de escenas y palabras?**

Para esto, el conteo de escenas de la pregunta anterior nos serviría para ver cuál es la obra con más escenas.

Para el caso de las palabras, se podría primero ver en la tabla "Paragraphs" cuántas palabras por escena hay, teniendo en cuenta la referencia a la tabla "Chapters" a través del atributo "chapter_id". Luego se podría agrupar ese conteo de palabras por escena según la obra a la que pertenece cada escena, usando el atributo "work_id" de la tabla "chapters" para hacer referencia a la obra a la que pertenece cada escena.

- **¿Con el paso de los años, fue escribiendo obras más largas o más cortas?**

Aquí, podríamos primero agrupar las obras por año, para luego ver la media de palabras por obra y visualizarlo en forma cronológica. Siendo solamente 23 años los que estuvo escribiendo, es posible sacar conclusiones en forma visual.