Tarea 1: Shakespeare

Navadian | Robaina

May 2023

Presentación

En este informe, abordaremos la obra de William Shakespeare a través del análisis de datos y no de forma literaria. Para ello, tomamos una base de datos relacional que contiene 4 tablas.

Una primera tabla a comentar es 'Paragraphs', que es la que contiene la totalidad de párrafos de las obras disponibles en el dataset. La misma cuenta con 35.465 instancias. Cada instancia tiene un texto asociado, que le pertenece a un personaje en un capítulo determinado.

En segundo lugar, tenemos un desglose de 1.266 personajes en la tabla 'Characters', para los cuales se cuenta con su nombre, una abreviación y una breve descripción.

En tercer lugar los capítulos, entendidos como la escena dentro de cada acto. Están contenidos en la tabla 'Chapters', que incluye una breve descripción y la obra correspondiente.

Por último, las 43 obras de Shakespeare se encuentran en la tabla 'Works', donde se describe su título, la versión larga del título, la fecha de publicación y el género al que pertenece.

Calidad de los datos

En términos generales el dataset no presenta datos faltantes, mencionando como excepciones la ausencia en el campo *description* para el 51% de los personajes, lo que podría ser irrelevante y deberse a que dichos personajes tienen un rol secundario que no merece una descripción. En segundo lugar, mencionar que 5 personajes no cuentan con el campo abreviatura.

En base a investigación acerca del dominio, se encontró un dato acerca de la completitud de la obra de Shakespeare y resulta que la cantidad total de obras que se le adjudican al autor es un debate aún abierto entre los académicos. Es por esto que si bien no se cuenta con *missings* en el resto de tablas del dataset, se puede poner en tela de juicio la calidad de datos de la tabla Works.

Además, para la comprensión del dominio, se consultó a una actriz teatral, la cual marcó que algunas obras podrían no tener la totalidad de capítulos, poniendo de ejemplo la obra Hamlet la cual debería poseer un número mayor de escenas en el segundo acto.

Por otra parte, se analizó la cantidad de párrafos por personaje, habiendo en el dataset un personaje llamado *stage directions*, al cual se le asigna el 11% del total de los párrafos. Estos 3.751 párrafos son acotaciones, las cuales no entrarían en los diálogos propios de la obra y tampoco serían de relevancia dependiendo del análisis. En este caso se considera que el dato no es relevante para el análisis y a partir de aquí se descartará.

Otro personaje al que se cuestiona su inclusión es el segundo con mayor cantidad de párrafos llamado *Poet*. La descripción de este personaje indica que es la voz poética de Shakespeare. Consultado con la actriz, se tomó la decisión de tampoco incluir a este personaje en el análisis, entendiendo que no sería el mismo tipo de personaje que aquellos que son representados en escena.

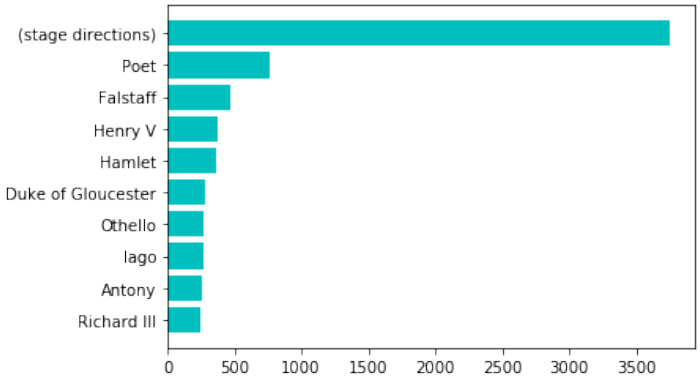

Figure 1: Presenta un gráfico de barras con los 10 personajes con mayor cantidad de párrafos. Se puede visualizar que "stage directions" y "poet" tienen unas apariciones desproporcionales en comparación a los que les siguen.

La obra de Shakespeare

A modo de generar una idea de la obra completa se muestran algunas visualizaciones que permiten ver la producción a lo largo de los años. Además, se visualiza algunos de los personajes más recurrentes[1] para tener noción de la magnitud de la obra.

Considerando la producción por géneros se observa que salvo algunas excepciones en las que no hay registro de publicaciones, el autor mantuvo su producción activa durante más de dos décadas. En la figura 2 se observa como en la primera década predominaron los géneros comedia e historia (en rojo y verde respectivamente). Entrado el siglo 17 se percibe una menor cantidad de publicaciones de todos los géneros que parecen dar lugar a una racha tragedias, lo que podría asociarse a algún sentimiento del autor en dicho período. Sobre

[1]Recurrentes por su cantidad de apariciones pero algunos también lo son por su popularidad.

el final de su carrera produce el primer y único soneto, y ese mismo año alcanza el máximo de publicaciones (2 historias, 1 poema y 1 soneto).

Para transmitir la popularidad del autor se incluye un gráfico con los personajes del conjunto de datos que tienen mayor cantidad de párrafos asociados. Es probable que el lector reconozca más de uno de los nombres ahí presentados.

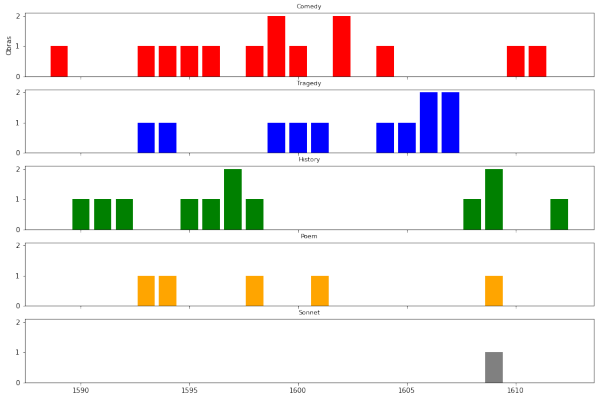

Figure 2: Este gráfico de barras expone la producción de obras del autor en toda su carrera según sus géneros.

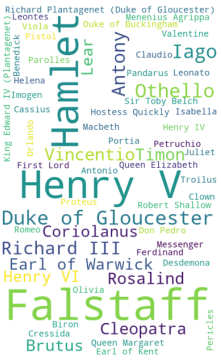

Figure 3: Nube de personajes.

Palabras

A continuación se hará especial foco en uno de los aspectos del que se puede sacar mayor provecho dentro de este conjunto de datos: las palabras. Es necesario normalizar los datos de texto de los párrafos para poder llevar a cabo un análisis.

Para ello se procesan los párrafos de varias formas. En primer lugar se quitan los signos de puntuación, esto se debe a que lo que se busca obtener de este procesamiento es una nueva tabla que tenga información de las palabras, en esta nueva tabla no son relevantes los caracteres que no sean letras.

También se pasan todas las mayúsculas a minúsculas, esto evita que dos palabras aparezcan como distintas solo por tener una diferencia en caracteres que representan una misma letra. Por ejemplo la palabra *Labra* es la misma que la palabra *labra* aunque caractér a caractér sean distintas.

En el ingles existen las contracciones, esto significa que en algunas situaciones al tener dos palabras juntas se escriben en una sola sacando algunas letras (como puede ser *She'll* para expresar *She will*). Otra posible normalización podría ser extender estos casos a las dos palabras que compongan la contracción a modo de tener todas las palabras consideradas, pero esto no siempre es posible por varios motivos. El primero es que la obra de Shakespeare fue escrita a fines del siglo 16 y comienzo del 17, y al ser el idioma algo tan cambiante sería necesario no solo tener un experto del idioma, sino que además experto del idioma en esos años. Otro motivo es que la forma de sistematizarlo no es evidente tampoco, la contracción *'s* en algunos contextos refiere a *is* pero en otros refiere a *has*. A pesar de la falta de conocimiento del dominio optamos por resolver las contracciones con *'ll* pues todas las palabras en los datos que contaban con esos caracteres entendíamos que hacían referencia a la palabra *will*. Se entiende que para este análisis el trabajo de extender las contracciones no le aporta demasiado valor al resultado y por ese motivo no se agregan más cambios de esta índole.

Luego de realizada esta limpieza se consideró que se estaba en condiciones de visualizar las palabras de toda la obra por cantidad, para lo cual se eligió como visualización una nube de palabras. En este proceso se interpusieron las stop words, que son el conjunto de palabras usualmente habladas en cualquier idioma, tales como pronombres, conectores, etcétera. Se intentó crear una lista con las stop words pero era un trabajo bastante tedioso e inagotable, por lo que se recurrió a la librería **nltk**, la cual ya cuenta con un conjunto bastante amplio de stop words en varios idiomas. Al sacarlas se obtuvo un mejor resultado ya que permitió identificar de manera más clara las palabras más frecuentes.

Visualizando las figuras 4 y 5 se puede apreciar la ventaja de filtrar estas stop words, sin embargo una dificultad que presenta el dataset es que debido a la época, el inglés del autor utiliza palabaras antiguas que no fueron contempladas por la librería como podría ser "thou" (que es el pronombre singular de la segunda persona y hoy en día está en desuso).

En la línea del análisis de palabras, resultó de interés la intersección de las mismas con el personaje que las dice. Para ello se utilizó el merge entre ambas tablas proporcionado en el código base y se decidió proseguir con el

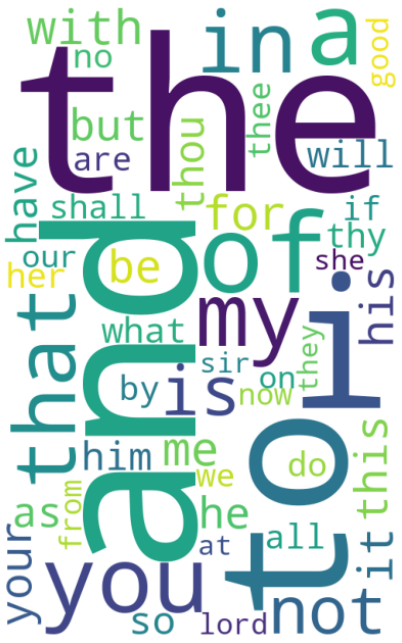

Figure 4: Nube de palabras con *Stopwords*

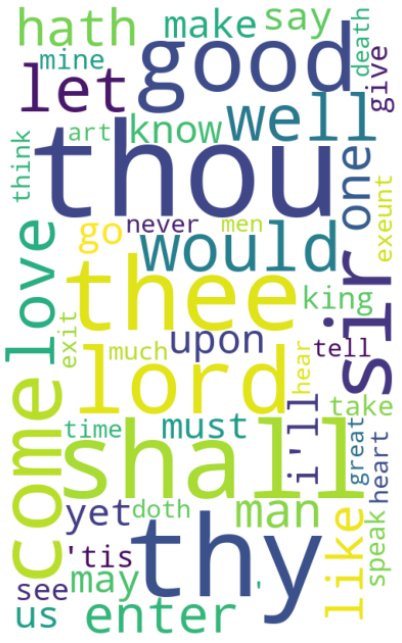

Figure 5: Nube de palabras filtradas

gráfico de barras ya que es apropiado para representar cantidades. Sin embargo, se modificó la orientación de la visualización por barras horizontales, ya que permiten una mejor lectura del personaje en cuestión, además de la eliminación de los personajes *stage directions* y *poet* como se mencionó al principio del informe.

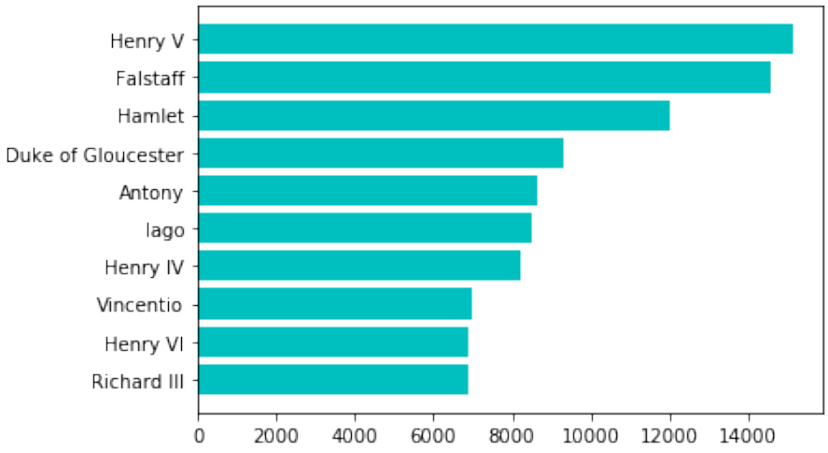

Figure 6: Caption

Futuras líneas de trabajo

Un punto que puede llamar la atención observando la frecuencia de las palabras es que “he” aparece una cantidad significativamente mayor de veces que “she”. Se podría realizar un análisis más detallado sobre esta apreciación para intentar comprobar si el mayor uso de pronombres masculinos que femeninos en la obra corresponde a una predominancia de dicho género, y ver la evolución de esta predominancia en las obras con el pasar de los años o si hay alguna tendencia entre los distintos géneros.

Otro posible tarea que podría ser interesante abordar es la de reconocer patrones del uso de palabras por personajes, reconocer personajes que hagan usos similares del lenguaje podría aportar cierta noción de paralelismo entre personajes que aparecen en distintas obras.