Universidad de la República
Facultad de Ingeniería

Introducción a la ciencia de datos

Tarea 1 - William Shakespeare

Mayo de 2024

Juan Montesano
Manuel Padín

Descripción de tablas y limpieza de datos

Works: Obras de Shakespeare
Chapters: Capítulos dentro de las obras
Paragraphs: Párrafos o líneas de cada personaje.
Characters: Personajes.

Una obra tiene varios capítulos, los cuales poseen varios párrafos. Cada párrafo está asociado a un personaje.

Para limpiar el texto buscamos en todo el texto cada carácter que no sea alfanumérico y lo quitamos.
En la visualización de las palabras más frecuentes realizamos una limpieza más profunda, en esa sección se describe con detalle.

Personaje con más párrafos

Los personajes con más párrafos ordenados de mayor a menor:
(stage directions): Es por lejos el que tiene más párrafos pero no es un personaje real, son direcciones de escena.
Poet: Tampoco es un personaje real (ver análisis en **Personaje con mayor cantidad de palabras**).
Falstaff, **Henry V** y **Hamlet**: Estos personajes, que son los reales que tienen más párrafos, también coinciden con mayor cantidad de palabras.

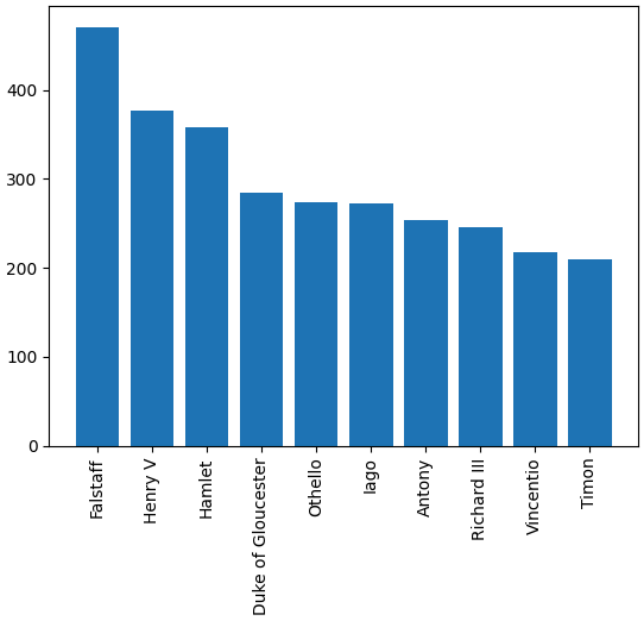

Figura 1 Cantidad de párrafos por personajes “reales”

Personajes con mayor cantidad de palabras. Problema y solución propuesta

Se encontró que el personaje **Poet** es el de mayor cantidad de palabras, analizando en qué obras aparece vemos que es personaje único en 6 trabajos y tiene asignados todos los párrafos, seguramente esos trabajos son solo poemas.
No consideraremos a Poet en ciertos análisis pero en otros sí, por ejemplo no se considera cuando analizamos la interacción de los personajes ya que no es un personaje real, en cambio lo consideramos cuando analizamos palabras utilizadas u otros análisis relacionado a la escritura de Shakespeare.
El segundo personaje es **(stage directions)** que tampoco es un personaje real, el siguiente es **Henry V** (15.428 palabras) y muy cerca **Falstaff** (14.906 palabras)
Obras donde aparece Henry:

- Henry V
- Henry IV Part I
- HenryIV Part II

Obras donde aparece Falstaff:

- Henry IV Part I
- Henry IV Part II
- Merry Wives of Windsor

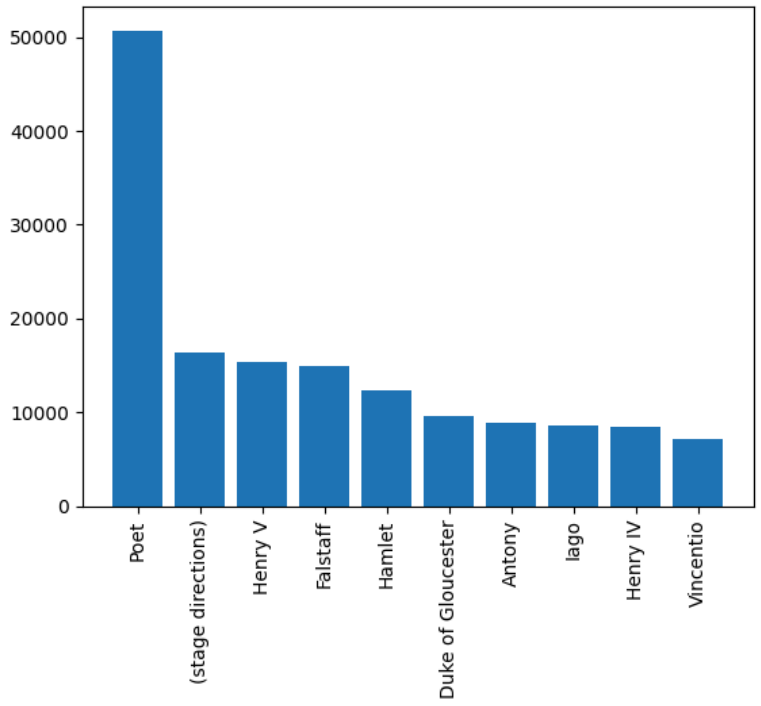

Figura 2 "Personajes" con más palabras

Visualización para comparar palabras más frecuentes de toda la obra

Utilizamos dos algoritmos para calcular palabras más frecuentes, el conteo de palabras estándar y TF-IDF, este segundo asigna un score a cada palabra con más pesos a aquellas que son de importancia y menos a las más comunes (como **the**, **a**, **for**). Creemos que capturar las palabras más importantes es mejor para realizar un análisis comparado con solo contabilizar la frecuencia.

A continuación detallamos la limpieza de palabras que aplicamos ya que encontramos varias palabras que reflejan una realidad pero nuestra intención es analizarlas a fondo. Sin embargo es interesante ver en estos filtros algunos patrones, por ejemplo el uso del inglés antiguo:

- Primero quitamos todos los párrafos del personaje **(stage directions)** porque agregaba muchas palabras que no reflejan el contenido propio de la obra, por ejemplo palabras como **enter**, **exit**, **exeunt** que son propias de indicaciones de una obra cuando entra o sale un personaje de escena.
- Contiene muchas palabras frecuentes que son del inglés antiguo y hoy no se utilizan como **thou**, **thy**, **thee**. En base a la siguiente tabla reemplazamos palabras antiguas por las modernas.

	2ND PERSON SINGULAR ARCHAIC	2ND PERSON PLURAL + MODERN SINGULAR
SUBJECT	thou	you
OBJECT	thee	you
POSSESSIVE	thy, thine	your(s)
REFLEXIVE	thyself	yourself (sg), yourselves (pl)

Figura 3 Tabla de equivalencias entre inglés arcaico y moderno

- Luego quitamos las palabras asociadas a títulos honoríficos (**lord**, **sir**), palabras asociadas a verbos auxiliares (**ll**, **shall**, **would**, **did**, **like**), y otras palabras del dialecto antiguo:
    - Tis: it is
    - Hath: have
- Por último y antes de realizar las visualizaciones aplicamos un filtro para quitar palabras comunes del inglés (**stop words**).

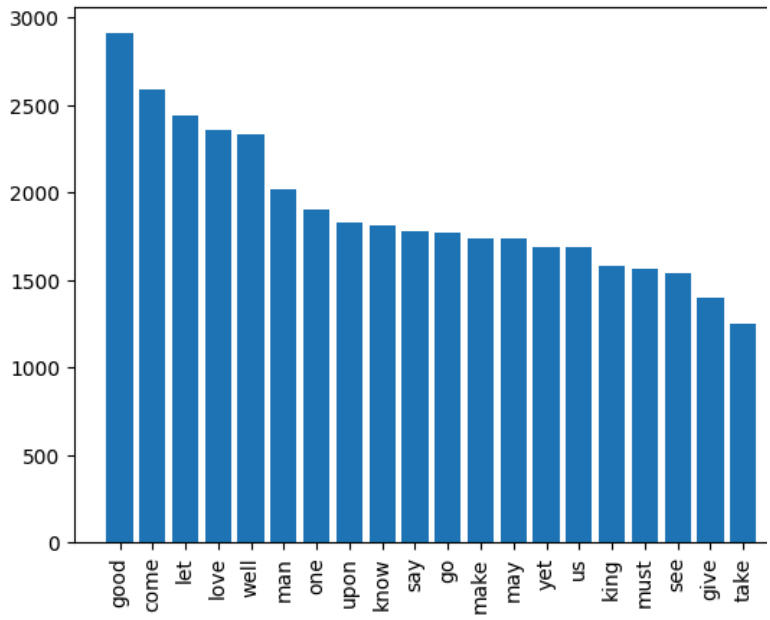

Figura 4 Frecuencia de palabras

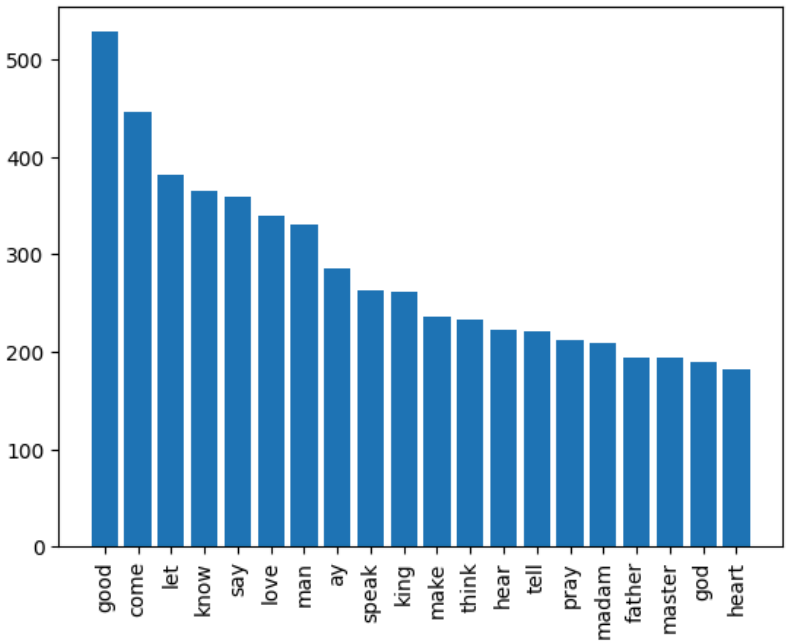

Figura 5 Score de palabras método TF-IDF

Figura 6 Nube de palabra TF-IDF

Palabra Más frecuente	TF-IDF
good	good
come	come
let	let
love	know
well	say
man	love
one	man
upon	ay
know	speak
say	king
go	make

Figura 7 Lista ordenada de palabra más usada por método

Se podría realizar el mismo análisis para cada género y personaje por separado y ver las variaciones de qué palabras son más usadas. A su vez sería interesante encontrar las palabras con más variación a lo largo de los distintos géneros.

Obras a lo largo de los años, tendencias en género y cantidad de obras

Publicó obras desde 1589 al 1612, durante 24 años, sin embargo en 1603 no publicó ninguna. Analizando los géneros podemos ver que al inicio y al final de su carrera publicó sobre Historia. En un periodo de 7 años en la mitad de su carrera se inclinó a la Tragedia, donde coincide el año que no publicó, ¿habrá pasado por un mal momento esos años? ¿Muerte de algún familiar?. Durante toda su vida escribió Comedia (su primera obra y la antepenúltima fue Comedia). Poemas y sonetos escribió muy poco, dos de ellos durante la plaga de 1593 y 1594.

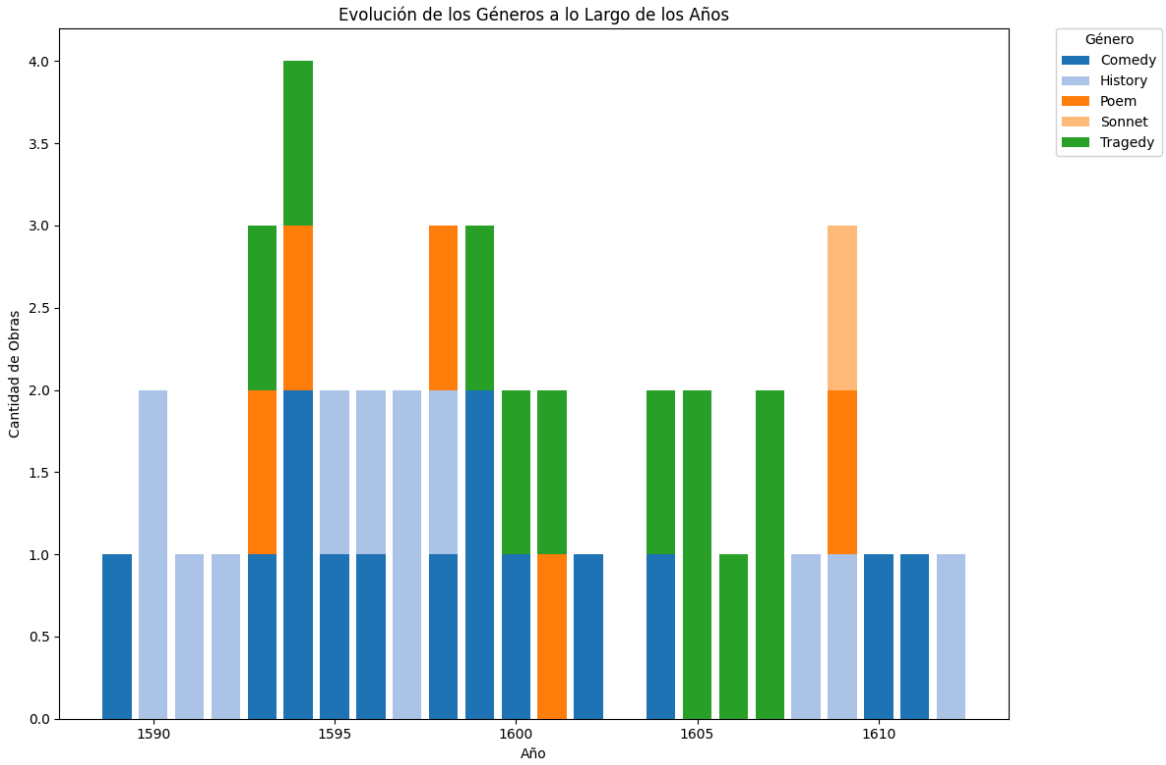

Figura 8: Evolución de los géneros de la producción literaria

Proponer preguntas que se puedan responder con estos datos

Existe un rumor que Shakespeare plagió algunas obras de otros autores, sería interesante buscar diferencia de estilos de escritura en las obras para analizar coherencia de escritura o existe una diferencia y contestar a **¿Shakespeare plagió sus obras?**. Con un modelo especializado en capturar estilo de escritura[1], que lleve los párrafos o capítulos a un vector y luego clusterizar, si vemos más de un cluster definido nos daría indicios de que hay más de un escritor.

¿Cómo evoluciona la forma de escribir durante su carrera?
Por ejemplo, analizar qué palabras utilizaba al inicio que luego dejó de usar, o al contrario, palabras nuevas que incorporaba.
No solo palabras, también analizar frases o modismos de la época.

Podría ser interesante comparar la forma de escribir de shakespeare con otros escritores de la época para poder identificar su estilo único, tendríamos que agregar datos de otras obras de otros escritores, las preguntas serían:
¿Cuál es el estilo único de shakespeare?
¿Qué palabras utiliza más habitual comparado con otros escritores?

[1] https://towardsdatascience.com/finding-a-writers-style-using-neural-networks-a1c3efcb186b