# Tarea 1

## Introducción a la Ciencia de Datos

Victoria Luz
23 de mayo del 2023

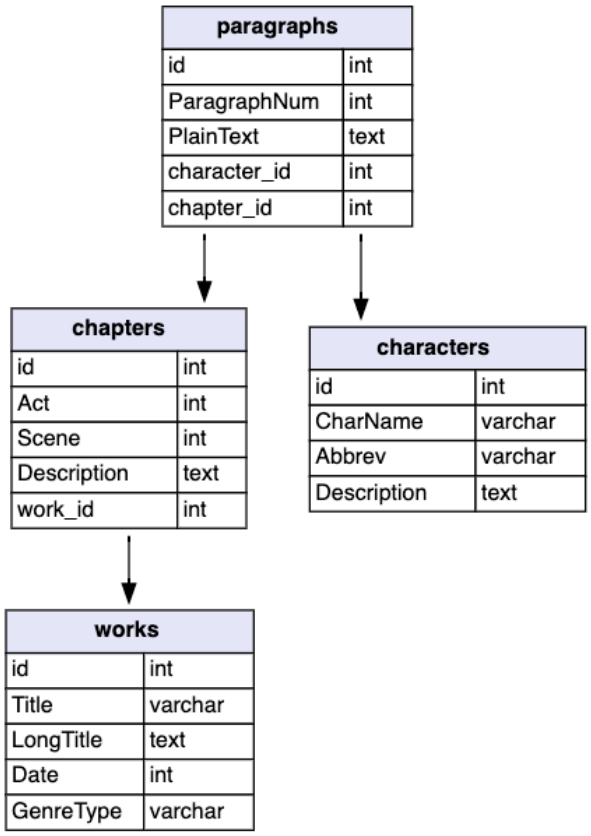
Introducción

El presente trabajo tiene como finalidad analizar las obras completas de William Shakespeare desde una perspectiva de análisis de datos, utilizando herramientas como Python para recabar información relevante a partir del texto.

Se explorarán las obras a lo largo del tiempo, analizando el protagonismo de cada personaje a través de sus apariciones, las palabras más utilizadas y los gèneros más desarrollados.

Composición Base

Para llevar a cabo esta investigación, se utiliza una base de datos relacional abierta con la obra completa del escritor. La misma se compone de cuatro tablas, “paragraphs”, la cual contiene los párrafos de las obras, “characters”, que contiene los nombres de todos los personajes de todas las obras, “chapters”, que contiene todos los capítulos, escenas y descripción de las mismas, y “works”, que detalla el título de todas sus obras. Como se muestra en el flujo a continuación, “paragraphs” se relaciona con “chapters” a través del campo chapter_id y con “charapter” a través del campo chapter_id. Asimismo, “chapters” se relaciona con la tabla “works” a través de work_id.



Limpieza de los datos

Para realizar esta investigación, se debió realizar una limpieza de datos que permitiera llevar a cabo los objetivos planteados. Este paso es fundamental para el trabajo, ya que garantiza no solo la calidad sino también la confiabilidad de los resultados. A continuación, se detallan los pasos realizados.

Inicialmente se realiza una preparación y normalización del texto de la tabla párrafos, convirtiendolo todo a minúsculas, asegurando consistencia y uniformidad, evitando que palabras como "Thou" y "thou" se tomen como palabras diferentes. Asimismo, se eliminan todos los símbolos y signos de puntuación, para que el texto solo contenga palabras.

Posteriormente, se eliminan abreviaciones de conjugaciones de verbo (s, r, ll, etc), sujetos e inicios de preguntas en inglés, ya que al contabilizarlas tienen un peso relevante y no aportan un valor significativo al análisis de las obras.

| word | |
|---|---|
| the | 28933 |
| and | 27312 |
| i | 23006 |
| to | 20820 |
| of | 17179 |
| a | 15084 |
| you | 14227 |
| my | 12951 |
| that | 11910 |
| in | 11656 |
| is | 9723 |
| not | 8862 |
| with | 8296 |
| for | 8075 |
| me | 8046 |
| it | 8038 |
| s | 7796 |
| his | 7328 |
| be | 7206 |
| he | 7109 |

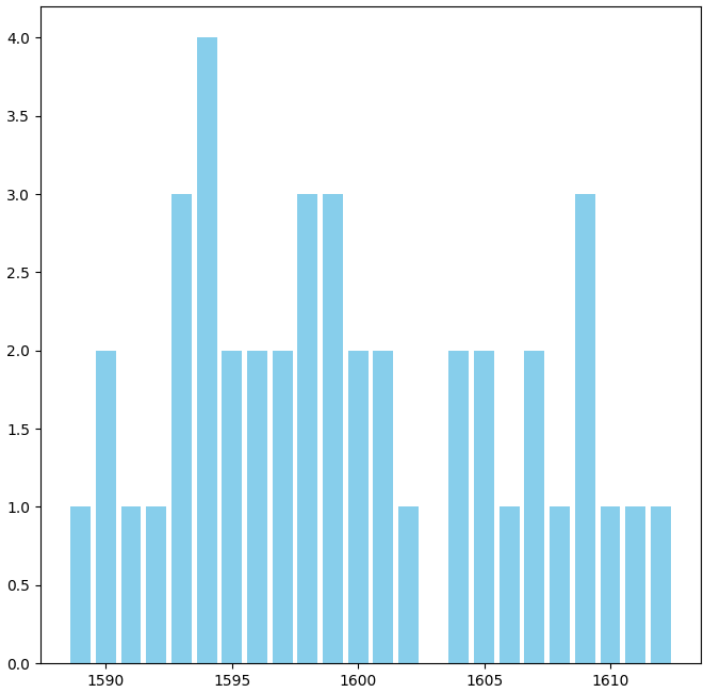
Finalmente, se remueven los personajes “Poet” y “(stage directions)”, ya que si bien se encuentran dentro de la tabla “characters”, no son personajes en si mismo y no son relevantes para el objetivo de la investigación.

| CharName | |
|---|---|
| Poet | 50762 |
| (stage directions) | 16443 |
| Henry V | 15428 |
| Falstaff | 14906 |
| Hamlet | 12291 |
| Duke of Gloucester | 9526 |
| Antony | 8849 |
| Iago | 8643 |
| Henry IV | 8426 |
| Vincentio | 7094 |

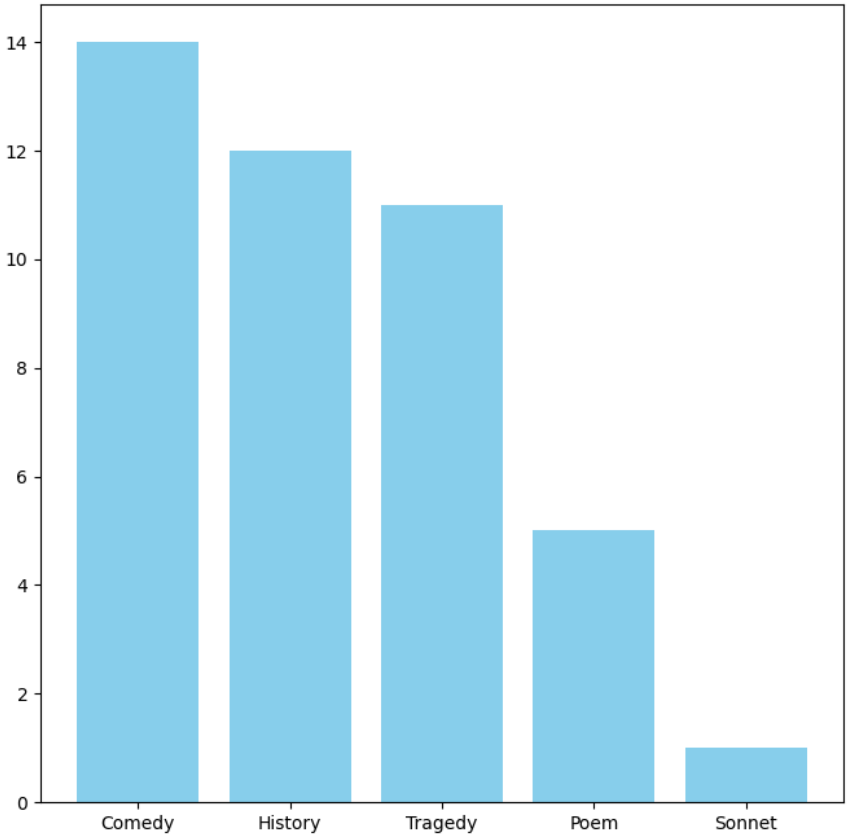
Con respecto a los datos faltantes, se observa que tanto en la descripción de los personajes como en la de los capítulos aparecen vales vacíos. Si bien para el alcance de este trabajo no afecta, si podría ser un inconveniente de limpieza de datos si en un futuro se quisiera profundizar el análisis.

## Principales hallazgos

Durante el proceso de investigación de las obras de William Shakespeare, se han realizado descubrimientos significativos. Un análisis a lo largo de los años revela que un número importante de sus obras (23 en total) fueron escritas en un período concentrado entre 1593 y 1601, cuando el autor tenía entre 29 y 37 años. Durante este periodo, Shakespeare produjo algunas de sus creaciones más destacadas, tales como "Henry VI", "Henry V", "Romeo y Julieta" y "Hamlet", entre otras obras de renombre. Esto demuestra la capacidad del escritor para producir nuevas y exitosas obras.



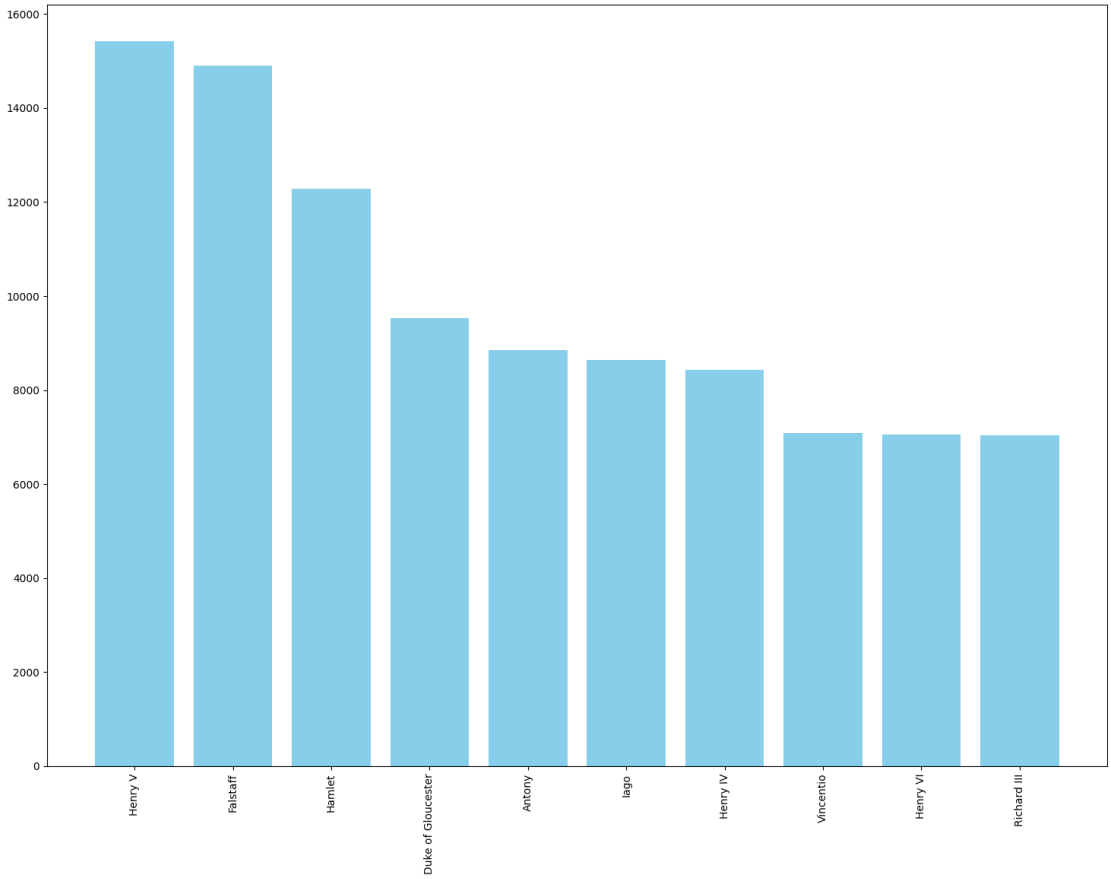
En relación a los géneros abordados por este autor, se destaca que la Comedia, la Historia y la Tragedia son los más predominantes en su obra. Esto se condice con el género de sus obras más exitosas.




Al analizar la cantidad de apariciones de los personajes, se observa una consistencia notable en el hecho de que Henry V y Falstaff se encuentren entre los primeros puestos. Esto se debe en parte a que estos personajes aparecen en más de una obra de Shakespeare. Es importante destacar que la diversidad de estilos de escritura utilizados por el autor tiene un impacto en la frecuencia de aparición de los personajes.

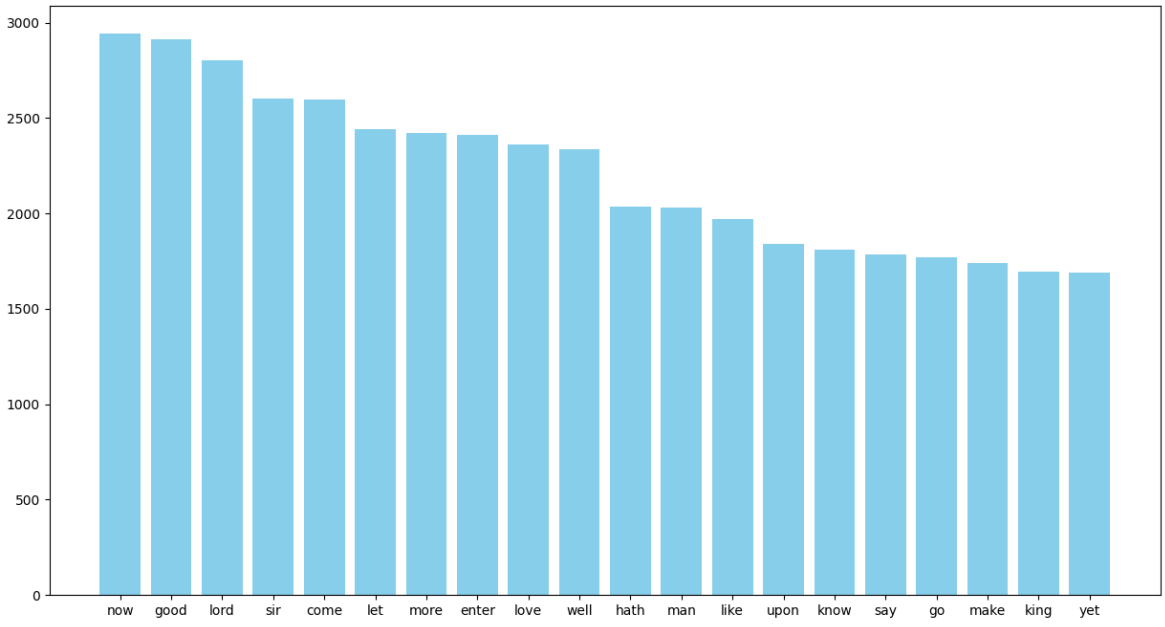
| CharName | |
|---|---|
| Henry V | 15428 |
| Falstaff | 14906 |
| Hamlet | 12291 |
| Duke of Gloucester | 9526 |
| Antony | 8849 |
| Iago | 8643 |
| Henry IV | 8426 |
| Vincentio | 7094 |
| Henry VI | 7049 |
| Richard III | 7034 |

Obras como Henry V y Hamlet se centran en personajes históricos, lo que puede influir en la cantidad de veces que dichos personajes aparecen. Por otro lado, obras como Romeo y Julieta abordan una historia de amor y el conflicto entre familias, lo que resulta en una distribución más equitativa del protagonismo entre varios personajes. A pesar de ser una de las obras más conocidas del autor, la atención se reparte entre un mayor número de personajes en lugar de centrarse en uno o dos protagonistas principales.




En el análisis de las palabras más frecuentes utilizadas por el escritor, se pueden identificar patrones que revelan información significativa. Por ejemplo, se observa una presencia importante de palabras como "Now" y "Good", las cuales hacen referencia al momento en que se desarrolla la acción. Por otro lado, términos como "Lord", "Sir" y "King" reflejan el estatus de los personajes y el contexto histórico en el que se sitúan. Asimismo, la aparición frecuente de diversos verbos denota las acciones y movimientos de los personajes a lo largo de las obras.

| word | |
|---|---|
| now | 2941 |
| good | 2913 |
| lord | 2801 |
| sir | 2603 |
| come | 2598 |
| let | 2442 |
| more | 2420 |
| enter | 2412 |
| love | 2359 |
| well | 2337 |
| hath | 2035 |
| man | 2029 |
| like | 1972 |
| upon | 1840 |
| know | 1812 |
| say | 1783 |
| go | 1771 |
| make | 1740 |
| king | 1696 |
| yet | 1688 |




## Conclusión

En conclusión, el análisis de datos sobre las obras de William Shakespeare proporciona una gran visión sobre sus obras, incluso sin haber leído algunas de ellas. Al estudiar las apariciones de los personajes, se observa que personajes como Henry V y Falstaff son prominentes, posiblemente debido a su presencia en múltiples obras. Sin embargo, es interesante destacar que los nombres de Romeo y Julieta, a pesar de ser reconocidos mundialmente, no figuran entre los más frecuentes.

Además, al examinar la cronología de sus obras, se evidencia un período en particular donde se concentran la mayor cantidad de obras creadas, entre 1593 y 1601, cuando Shakespeare tenía entre 29 y 37 años.

En cuanto a la evolución de los géneros en sus obras, se observa que el escritor abarcó principalmente la Comedia, la Historia y la Tragedia. Esta versatilidad temática y géneros demuestra las grandes habilidades de escritura que éste poseía y facilidad para destacarse en cualquiera de ellos.

En resumen, el análisis de datos en las obras de Shakespeare nos brinda una comprensión más profunda de la frecuencia de los personajes, la evolución cronológica de sus obras y la diversidad de géneros que abordó. Esto nos permite conocer con mayor profundidad su trabajo y a él como persona.

Preguntas a futuro

- ¿Cómo fue la evolución de género de las obras a lo largo de los años?
- ¿Qué proporción de apariciones tienen los personajes hombres vs las mujeres?
- ¿Cómo se relacionan los personajes a lo largo de las diferentes obras?
- ¿Se observa una relación entre las tramas de las diversas obras?