# Laboratorio 1.

# Introducción a la Ciencia de Datos, 2023.

# Grupo 1.

## Introducción.

El presente trabajo se enmarca en el Laboratorio 1 de la materia de posgrado “Introducción a la Ciencia de Datos”, dictada por la Maestría en Ciencia de Datos y Aprendizaje Automático de la Facultad de Ingeniería (UDeLaR), en su edición del año 2023.

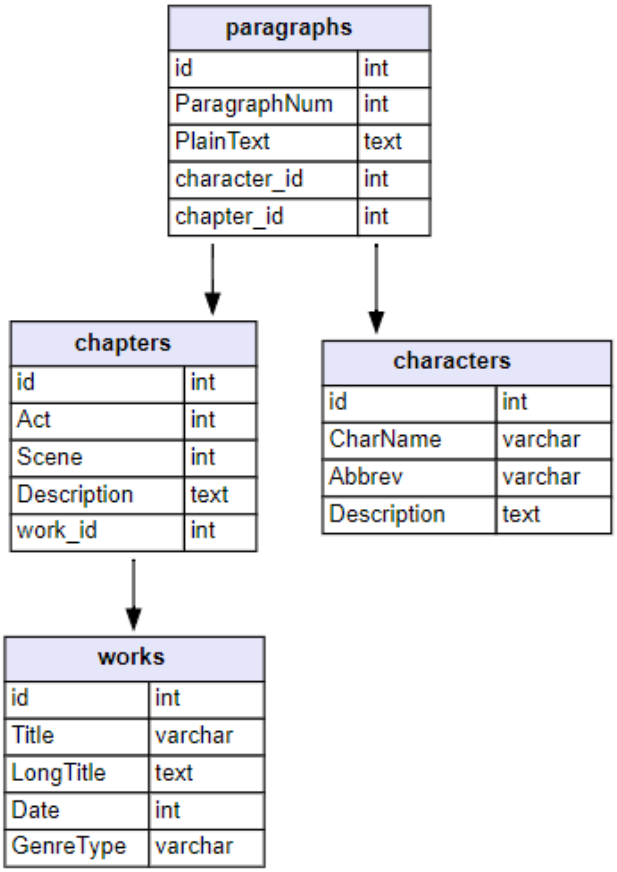
El mismo tiene por objeto hacer una exploración visual de los datos hallados en una base de datos relacional. Dicha base contiene, de manera estructurada, la obra producida por el renombrado escritor inglés, William Shakespeare.

## Base de datos.

La base analizada, es una base pública, disponible en un servidor de la Facultad de Tecnología de la Información de la CTU de Praga. Los datos de acceso para la referida base son los siguientes:

**Host:** fit.cvut.cz
**Puerto:** 3306
**Usuario:** guest
**Password:** relational
**Nombre de la base:** Shakespeare
**Motor:** MySQL

## El diagrama relacional de la base es el siguiente:




## Entorno de trabajo.

Para el presente análisis se usó la herramienta de Jupyter Notebook, sobre la cual se ejecutó código escrito en Python 3.10.

Así mismo se usaron las siguientes librerías de Python:

**pandas**
**sqlalchemy<2.0**
**pymysql**
**seaborn**
**pillow**

# Carga de Datos.

Para un primer acercamiento a los datos, se partió del Notebook subido por los responsables del curso en el siguiente repositorio git:

https://gitlab.fing.edu.uy/maestria-cdaa/intro-cd/-/tree/master/Tarea_1

En el mismo se agregó el código necesario para cargar todas las tablas de la base de datos propuesta, a saber:

- **works**: tabla que contiene todas las obras escritas por Shakespeare, con su título, año de edición y género.

- **chapters**: tabla con los distintos capítulos que componen las obras, con su respectivo acto y escena y una descripción.
- **characters**: tabla que contiene los personajes que aparecen en las obras, su nombre, abreviación y descripción.
- **paragraphs**: tabla que contiene los párrafos contenidos en cada capítulo de las obras. Cada registro contiene el texto del párrafo y el número (orden del párrafo dentro del capítulo).

A su vez, existen relaciones, dadas por claves foráneas entre dichas tablas:

- Existe una relación 1 a N entre works y chapters, representado mediante el campo work_id, dentro de la tabla chapters.
- Existe una relación 1 a N entre chapters y paragraphs, representada por la inclusión del campo chapter_id, dentro de la tabla paragraphs.
- Existe una relación 1 a N entre characters y paragraphs, representada por la inclusión del campo character_id, dentro de la tabla paragraphs.

## Limpieza de datos.

Una vez cargado los datos en dataframes de panda, se hizo una limpieza de los mismos utilizando funciones que esta librería brinda.
La secuencia de pasos para limpieza fue la siguiente:

- Pasaje a minúscula de todas las palabras del campo Plaintext de los párrafos mediante función str.lower().
- Eliminación de directivas para interpretación, eliminando todo lo que estuviese entre paréntesis rectos mediante el uso de expresiones regulares: str.replace('\[.*\]', '', regex=True).
- Se eliminaron de los párrafos, todos los siguientes signos de puntuación: \n , . ? ! : ; - ' “ [ ] (  ) &
- Se filtraron aquellos párrafos que, luego de las sustituciones anteriores, resultaran vacíos.
- Se filtraron los párrafos vinculados a los personajes de ID 894 y 1261. (se explica esta decisión a continuación.

### Conteo de palabras.

Siguiendo con el flujo propuesto por el Notebook inicialmente, donde se hace un conteo de palabras por personaje, se obtuvo lo siguiente:

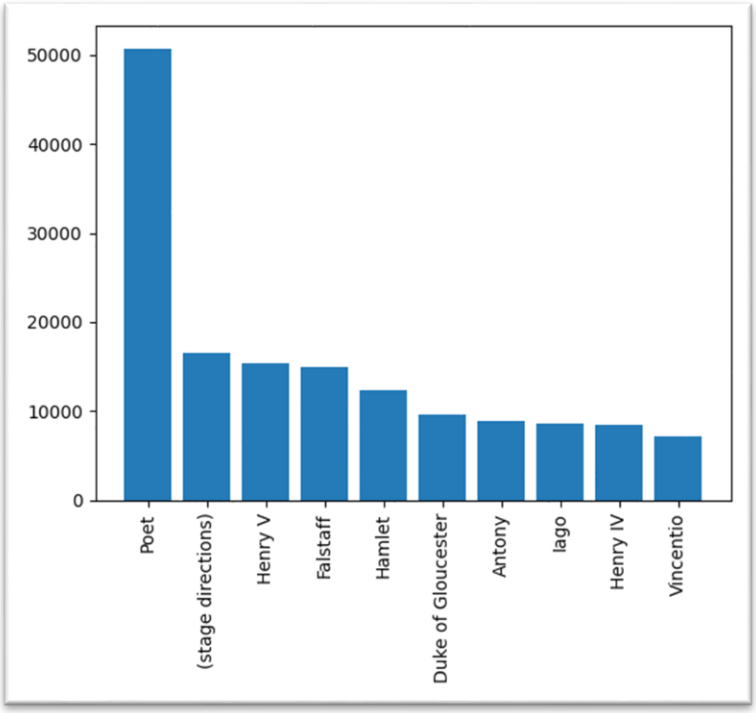



*Ilustración 1: Cantidad de palabras por personaje (original)*

Notamos en dicha gráfica, dos inconsistencias en la representación:

1- Todas las obras (works) cuyo tipo es Poem o Sonnet, carecen de personajes explícitos, por ende, todas las palabras contenidas en esas obras, se vinculan a un personaje "dummy", llamado Poet (ID 894). Esto hace que dicho personaje inexistente, sea el que más palabras tiene, cuando en realidad no es representativo de lo que se desea visualizar con esta gráfica. Por eso se decidió filtrar al personaje, de ID 894, de los resultados. (Se filtra por ID, por la posibilidad de que, en el resto de las obras, existiera otro personaje al cual se referencia como Poet y que si interesa representar).
2- Existe otro personaje '(stage directions)', ID 1261, que tampoco hace al objeto del ejercicio ya que su función es agrupar todas aquellas líneas que no son pronunciadas por ningún personaje, sino que son instrucciones para la representación en escena de la obra. También existía en medio de los diálogos, instrucciones que aparecían como texto entre paréntesis rectos. Por esto último, no bastaba con filtrar el personaje como en el punto anterior, sino que la decisión que se tomó fue eliminar de todos los párrafos, todo el texto que apareciera entre paréntesis recto.

El nuevo resultado obtenido, realizados estos dos cambios, fue el siguiente:

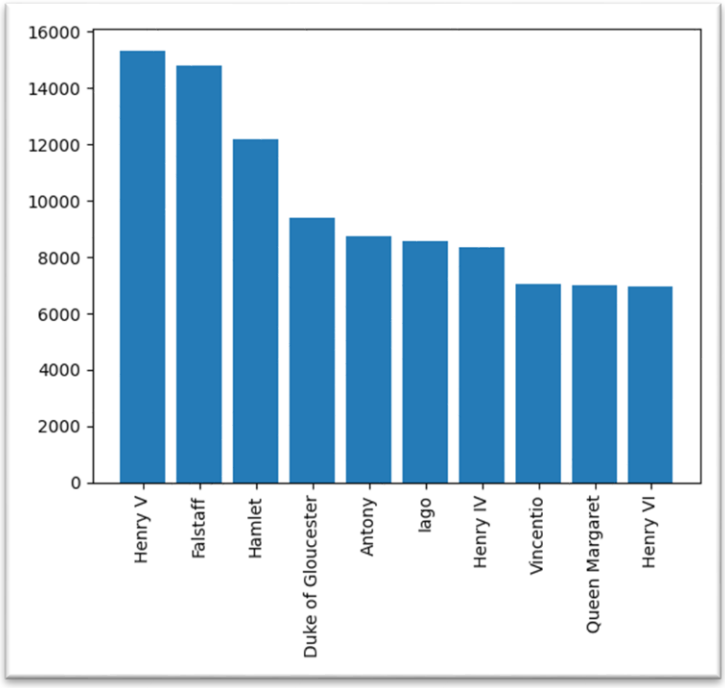



*Ilustración 2: Palabras por personaje luego de la depuración de datos.*

## Párrafos por personaje:

Análogamente, se calcularon la cantidad de párrafos por personaje, también aplicando los criterios y filtros que se mencionaron en el punto anterior.

Lo obtenido fue lo siguiente:

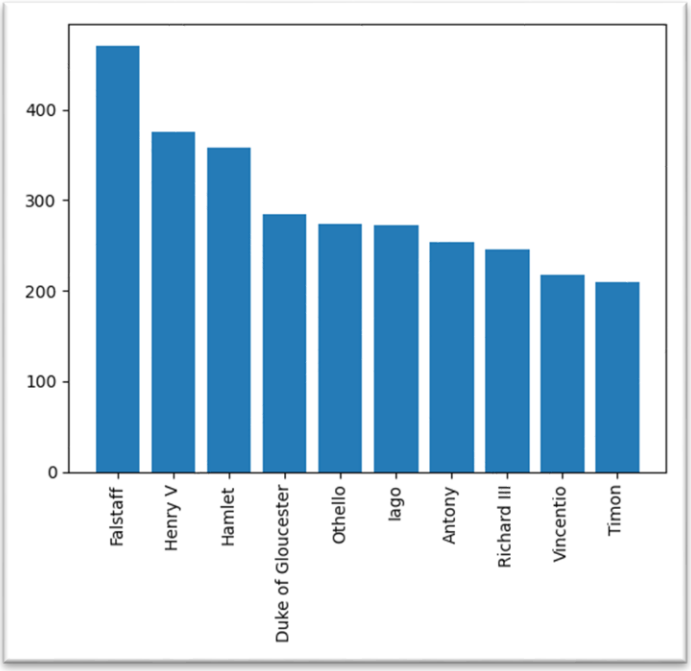



*Ilustración 3: Párrafos por personaje.*

Sir John Falstaff aparece como el personaje con mayor cantidad de párrafos (ya habíamos visto que era el segundo con mayor cantidad de palabras). Esto parece razonable si tomamos en cuenta que dicho personaje tiene diálogos en tres obras distintas:

- Henry IV, Part 1
- Henry IV, Part 2
- The Merry Wives of Windsor

Además, en todas ellas tiene un rol importante, ya que es compañero inseparable del Príncipe Hal, quién ha de convertirse en el mismísimo Rey Henry V de Inglaterra.

## Evolución anual de la producción Shakesperiana.

Para el presente análisis, se agruparon todas las obras según género, partiendo su producción en períodos de tres años. La gráfica que se obtuvo fue la siguiente:

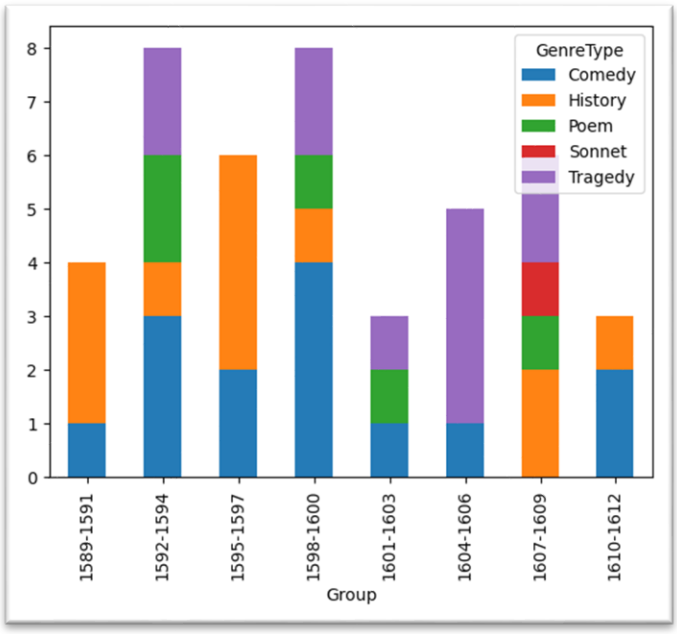



*Ilustración 4: Producción trianual según género*

A golpe de vista, observamos que su período más productivo en cuanto a cantidad de obras, se dio entre 1592 y 1600, período al que siguieron tres años de magra producción en comparación con toda su trayectoria.

En cuanto a la evolución de géneros, podemos decir que los poemas y sonetos aparecen esporádicamente a lo largo de su carrera artística, pero no han significado una producción cuantiosa ni constante en el tiempo.

Al comienzo de su carrera, los géneros más abordados por William eran la comedia y la historia, mientras que, sobre el final de su carrera, su producción se basaba principalmente en Tragedias e Historias, perdiendo la Comedia parte de su relativa importancia.

También parece interesante medir la producción artística, no sólo mediante la cantidad de títulos publicados, sino también en cantidad de palabras, ya que poemas o sonetos pueden ser más cortos y sin embargo contabilizar lo mismo que una tragedia según la métrica anterior.

Navegando por las relaciones, de manera de unir las obras (works) con las palabras que la componen, y agrupando por los mismos períodos trianuales, obtuvimos lo siguiente:

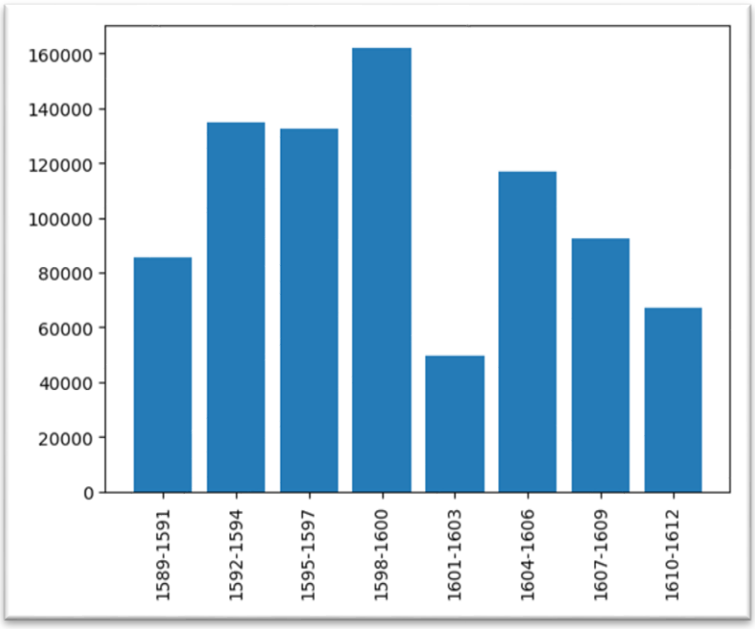



*Ilustración 5: Palabras escritas por período*

En general observamos un comportamiento similar en cuanto a períodos de mayor y menor producción, tal vez con la salvedad del período comprendido entre 1607 y 1609, que en cantidad de obras había sido bastante prolífero, pero medido en cantidad de palabras, resulta un tanto menos productivo, denunciando una menor cantidad de palabras por obra en promedio. Un análisis estadístico con mayor detalle, podría arrojarnos más información en este sentido, pero el alcance de este trabajo, el tiempo requerido para el mismo, sumado a nuestras limitaciones, no nos lo han permitido.

## Palabras más frecuentes.

Se agrupó la lista de palabras calculada en pasos anteriores, por la palabra en sí, para luego contabilizar sus ocurrencias.

Al graficar los resultados, se obtuvo la siguiente representación:

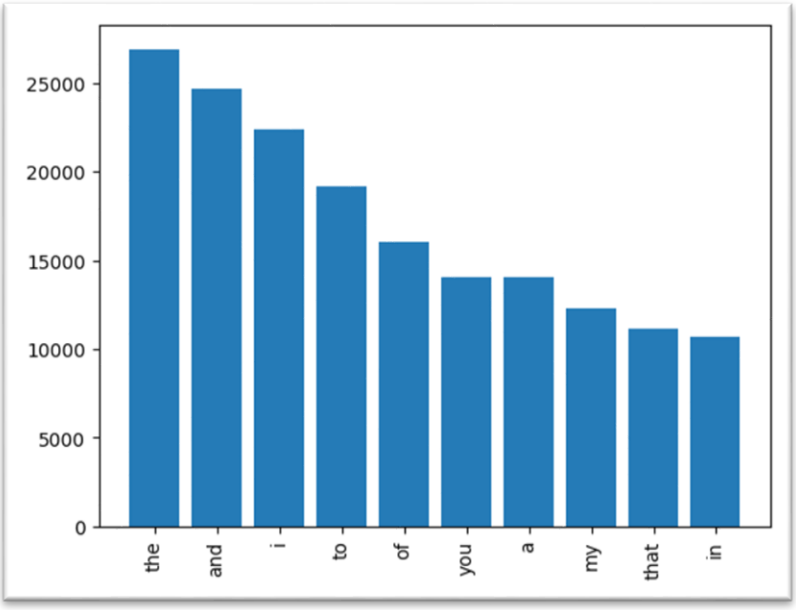



*Ilustración 6: Palabras con mayor ocurrencia.*

Si bien el resultado parece lógico, estaría bueno tener un agrupamiento de palabras según su naturaleza (sustantivo, adjetivo, verbo, preposición, artículo, etc.) de manera de poder hacer distintos cortes a los datos y poder presentar más y mejor información.

Así mismo, una red ontológica que relacione palabras según su cercanía semántica, también puede brindar otra perspectiva e información escondida entre tantos datos.

Otro camino que parecería interesante recorrer, sería el de hacer un análisis estadístico de las ocurrencias de las palabras en función del género de la obra en la que aparece y tratar de vincular la frecuencia relativa de una palabra (apariciones de esa palabra en función del total de palabras de la obra) con el género.

También se podría explorar la frecuencia de una palabra en los distintos párrafos y tratar de generar un modelo de clasificación, por ejemplo, para deducir el personaje que esboza dicho párrafo.

## Posibles preguntas a partir de los datos.

Algunas de las preguntas que se nos ocurren son las siguientes:

**¿Es posible generar un modelo de predicción que reciba un párrafo y deduzca qué personaje lo dice?**

Podríamos tratar de generar un modelo de clasificación que tome como input, la obra y el texto de un párrafo, que saque métricas de frecuencias de palabras en el texto analizado y trate de predecir quién dice esas palabras.

Un buen ajuste de ese modelo, hablaría bien de Shakespeare y su capacidad narrativa para diferenciar distintos personajes mediante distintos modismos, léxicos, estilos oratorios y sentimientos.

**¿Fue incorporando Shakespeare nuevo léxico con el pasar de los años?**

Tratar de detectar palabras o frases que aparecieron con frecuencia alta en su obra a partir de cierto año, sin que aparecieran con anterioridad.

**¿Qué participación tiene en la obra de Shakespeare el género femenino?**

Si bien en lo ya realizado en el presente trabajo, vemos que todos los personajes con mayor oratoria son del género masculino, estaría bueno hacer un análisis similar, pero considerando verbos que denoten acciones trascendentes para cada obra, de manera de cuantificar la importancia de los distintos personajes a través de sus acciones dentro de la obra. Tal vez lady Macbeth no es quién más habla en la obra Macbeth, pero nadie puede negar su protagonismo total en el desarrollo de los hechos.

**¿Se puede deducir el género de una obra con un modelo de predicción?**

Se podría procurar entrenar un modelo, por ejemplo, un árbol de decisión, que, a partir de los capítulos, personajes y palabras de una obra, trate de clasificarla en su debido género. El desafío más grande, lo presenta la baja cantidad del conjunto de entrenamiento, pero parece, a priori, factible.

**¿Es posible hacer un análisis de sentimientos a través de las obras de William Shakespeare?**

Existe mucha teoría y algoritmos prácticos para llevar adelante análisis de sentimientos a partir de textos. Tal vez, por ser la obra tan extensa, se podría partir el problema en uno más pequeño, como analizar la evolución de los sentimientos de cada personaje a través de cada obra.

**¿Se puede generar un *stochastic parrot* que produzca nuevas obras al estilo de Shakespeare?**

Hoy que los loros estocásticos están tan de moda, parece un lindo desafío entrenar un modelo que sea capaz de escribir con prosa similar a Shakespeare, obras a partir de pautas ingresadas por los usuarios.