# INTRODUCCIÓN A LA CIENCIA DE DATOS

## Tarea 1

## Análisis inicial de la base de datos *The Open Source Shakespeare* (OSS)

**Agustina Añasco - Horacio Solé**

**Maestría en Bioinformática (PEDECIBA)**

**2024**

## Introducción

William Shakespeare, nacido en 1564 en Stratford-upon-Avon, es considerado uno de los escritores más importantes de la literatura universal. Durante su trayectoria escribió alrededor de 195 trabajos en total, de los cuales 39 son obras de teatro, 154 sonetos y 2 son poemas narrativos.

Su influencia en la literatura y la cultura es inmensa y duradera por varias razones. En primer lugar, Shakespeare es conocido por su innovadora contribución al desarrollo del inglés moderno. Inventó muchas palabras y frases que todavía se usan hoy en día. Por otra parte, las obras de Shakespeare abordan temas universales como el amor, la traición, la ambición, el poder, la justicia y la identidad. Estos temas siguen siendo relevantes y resonando con audiencias de todas las épocas y culturas.

Además, es célebre por la creación de personajes que reflejan su comprensión de la naturaleza humana. Personajes como Hamlet, Lady Macbeth, Othello y Lear son estudiados por su complejidad psicológica y su capacidad para reflejar aspectos universales de la condición humana.

Se desempeñó como escritor de varios géneros, a saber, comedias, tragedias, historias y poesía, demostrando una gran capacidad para explorar una amplia gama de emociones y situaciones humanas.

Asimismo, las obras de Shakespeare han sido adaptadas innumerables veces en teatro, cine, literatura y otros medios artísticos. Directores de cine como Orson Welles, Akira Kurosawa y Kenneth Branagh, entre muchos otros, han adaptado sus obras, ampliando su impacto cultural.

Aunque sus obras reflejan el contexto social y político de la Inglaterra isabelina y jacobina, también trascienden su época. Shakespeare capturó tanto las inquietudes de su tiempo como cuestiones atemporales, permitiendo que sus obras sean continuamente reinterpretadas y sigan siendo relevantes.

Parte de dicha información sobre su obra se encuentra organizada en la base de datos "Open Source Shakespeare", que está compuesta por cuatro tablas principales: works (obras), chapters (capítulos), characters (personajes) y paragraphs (párrafos). Cada una de estas tablas se utilizarán en el presente informe para la obtención de datos relevantes, así como se pondrán en práctica mejoras para optimizar su utilización.

## Objetivos

A partir de la consigna planteada en la tarea 1, se espera poder realizar una primera aproximación al análisis de una base de datos:

1. Comprensión y descripción de la base de datos utilizada y su estructura: Examinar la base de datos para comprender su estructura, incluyendo las tablas, los atributos y las relaciones entre ellas para poder describirlas.

2. Exploración de los datos contenidos en la base de datos: Realizar un análisis exploratorio de los datos para identificar patrones, tendencias e inconsistencias. Utilizar técnicas de visualización de datos para representar gráficamente los resultados del análisis exploratorio.
3. Manipulación de los datos para la limpieza del texto y el conteo de palabras: Realizar el conteo de palabras, identificando las palabras más frecuentes.

**Metodología**

Para llevar a cabo esta tarea se utilizó el lenguaje de programación Python, utilizado en el entorno de desarrollo Jupyter Notebook, a través de Conda, una herramienta de gestión de paquetes y ambientes muy utilizado en bioinformática por facilitar la instalación y administración de librerías y dependencias. Para el procesamiento, análisis y visualización de datos se utilizaron los módulos time y pathlib, y las siguientes librerías: pandas, matplotlib, sqlalchemy y pymysql.

Para realizar la limpieza y el conteo de palabras se retiraron varios símbolos y signos de puntuación que puedan dificultar el conteo: paréntesis rectos, paréntesis curvos, saltos de lìnea, puntos, comas, punto y coma, signos de exclamaciòn, guiones, comillas y apóstrofos, estos últimos principalmente debido a que en el idioma Inglés suelen acortarse como “I’m” en realidad son dos palabras: I am.

Para realizar los análisis temporales de las obras se consideró el periodo 1588-1613, el cual posteriormente fue dividido en periodos de 5 años.

La base de datos fue obtenida de la página web *Relational Dataset Repository* (disponible en https://relational-data.org/dataset/Shakespeare).

Los comandos utilizados para este proceso se encuentran en el archivo adjunto en el repositorio de GitHub: Tarea1_icd_notebook.py.

**Descripción de la base de datos y resultados obtenidos**

La base de datos original está compuesta por 4 relaciones: *works*, *chapters*, *characters* y *paragraphs*, vinculadas entre sí como se muestra en la figura 1.

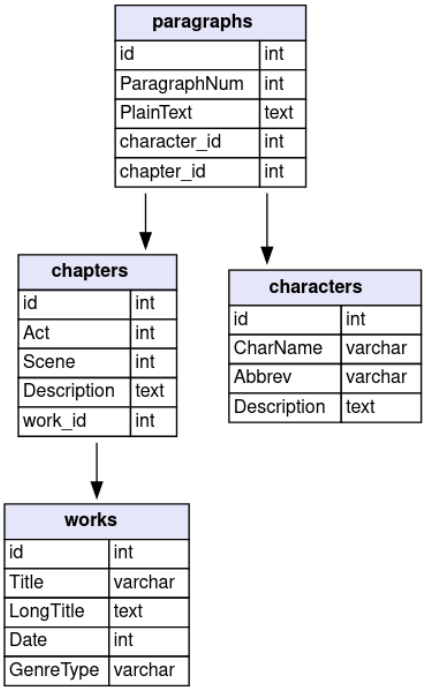


**Figura 1.** Esquema conceptual de la Base de Datos utilizada, se muestran las relaciones, los atributos y el tipo de variable correspondiente a cada uno.

La tabla *works* contiene información detallada sobre algunos de los trabajos realizados por el autor. En total, se compone de 43 filas (tuplas) y 5 columnas (atributo), los cuales se describen a continuación:

- id: Un código de identificación específico y único para cada una de las obras, numerado del 1 al 43.
- Title: El título por el cual comúnmente se conoce a la obra.
- LongTitle: La versión extendida del título de la obra.
- Date: El año de publicación de la obra, abarcando desde el año 1589 hasta 1612.
- GenreType: El género al cual pertenece la obra. Las opciones son comedia, tragedia e historia, que son los géneros dentro de los cuales se encuentran las obras de teatro de Shakespeare, aún así, en la tabla también encontramos poemas y sonetos.

Al examinar la tabla, no se detectan valores nulos, información faltante o redundante. Sin embargo, se observa cierta inconsistencia en la estructura de la columna LongTitle, particularmente en aquellos que comienzan con la palabra “The”. Por ejemplo, se puede ver que algunas obras estàn registradas como “The Passionate Pilgrim” y otras como “Tragedy of Hamlet, Prince of Denmark, The”, lo que podría implicar dificultades en la bùsqueda de datos o el filtrado, así como problemas en caso de querer ordenar los datos de la columna en orden alfabético, por ejemplo. Sería conveniente estandarizar y normalizar los datos de la columna para mantener un mismo formato o convención de los mismos.

La tabla *chapters* contiene información sobre los capítulos de cada obra registrada en la tabla *works*. En total presenta 945 tuplas y 5 atributos, donde encontramos:

- id: Un código de 5 dígitos (comprendido entre 18704 y 19648) que permite identificar a cada uno de los capítulos.
- Act: Número del acto al cual pertenece el capítulo.
- Scene: número de la escena a la cual pertenece el capítulo.
- Description: Presenta el título del capítulo o escena, tal cual como aparece en la obra original. Normalmente informa sobre el escenario donde tiene lugar la escena de la obra.
- work_id: Este campo actúa como una clave foránea, relacionando cada capítulo con su obra correspondiente en la tabla *works*.

No se observó falta de datos ni inconsistencias en los valores de la tabla.

La tabla *characters* contiene información detallada sobre los personajes de las obras de Shakespeare presentes en la tabla *works*. Presenta un total de 1266 tuplas y 4 atributos, detallados a continuación:

- id: Un código numérico del 1 al 1266 que es específico para cada personaje y permite identificarlo.
- CharName: El nombre de cada personaje.
- Abbrev: La forma abreviada del nombre de cada personaje, puesto que suele aparecer en los textos.
- Description: Aparece en muchos casos una breve explicación dando información útil sobre quién es el personaje exactamente.

Se observó que varios registros en la tabla carecen de información en las columnas “Abbrev” (5 nulos en 1266 tuplas) y “Description” (646 nulos en 1266 tuplas). Sería conveniente que todos los personajes cuenten con la información completa en ambos atributos, siguiendo una misma convención, para mantener la consistencia de los datos. Por otra parte, los nombres de algunos personajes aparecen más de una vez, como “First Senator” (id: 11 y 12) y “First Citizen” (id: 409, 410, 411, 412 y 413), y como es el caso de “Earl of Salisbury” (id: 965, 966 y 967) donde vemos una inconsistencia en el registro de sus datos en los distintos atributos, llevando a confusiones. Estos ejemplos, podrían tratarse de un error de información duplicada o bien de personajes pertenecientes a distintas obras. Para resolver esto, sería conveniente agregar un nuevo atributo: work_id, que permita vincular a cada personaje con la obra a la cual pertenece, evitando posibles dificultades a la hora del análisis de los datos en general (ambigüedades, errores de interpretación, problemas al consultar datos y su gestión).

Por último, la tabla *paragraphs* contiene información sobre los párrafos de los textos incluídos en la base de datos con la cual estamos trabajando. La misma presenta 35.465 tuplas y 5 columnas, en las cuales encontramos información de los siguientes atributos:

- id: Un código numérico de 6 dígitos (desde el 630863 al 666326) que permite identificar cada uno de los párrafos.
- ParagraphNum: La línea de la obra en la cual inicia el párrafo específico, esto resulta útil para ubicar cada párrafo en una obra específica.
- PlainText: Presenta el texto correspondiente a la obra (el párrafo, tal como está escrito).

- character_id: Actúa como clave foránea (relación con la tabla *characters*), permitiendo vincular cada párrafo con el personaje que lo interpreta o dice.
- chapter_id: De la misma forma que el atributo anterior, este permite saber en qué capítulo del texto está el párrafo en cuestión (relación con tabla *chapters*).

No se observó falta de datos ni inconsistencias en los valores de la tabla.

A partir de la comprensión de estas tablas, es posible determinar algunas estadísticas sobre la base de datos.

El personaje con más párrafos es aquel con id: 1261, que al buscar en la tabla *characters* aparece como “Stage directions”, es decir, las indicaciones de escena o instrucciones que normalmente se encuentran en el guión de una obra de teatro. Si bien este resultado era esperado, consideramos que no se trata de un personaje, por lo que decidimos realizar un ranking de los 10 personajes que aparecen en la tabla *characters*, ordenados según el número de párrafos en la tabla *paragraphs* (tabla 1).

**Tabla 1.** Lista de los 10 personajes (según la tabla *characters*) con más párrafos atribuidos en las obras de Shakespeare registradas en la base de datos.

| character_id | CharName | ParagraphNum |
|---|---|---|
| 1261 | Stage directions | 3751 |
| 894 | Poet | 733 |
| 393 | Falstaff | 471 |
| 573 | Henry V | 377 |
| 559 | Hamlet | 358 |
| 531 | Duke of Gloucester | 285 |
| 844 | Othello | 274 |
| 600 | Iago | 272 |
| 120 | Antony | 253 |
| 945 | Richard III | 246 |

Respetando el hecho de que, en realidad, “Stage directions” y “Poet” no son personajes realmente, podría decirse que el personaje con mayor cantidad de párrafos en la base de datos es Falstaff. A partir de esta observación, consideramos esencial una actualización de la tabla para su análisis, con el fin de clarificar qué se considera verdaderamente como 'personajes' según la concepción del diseñador de la base de datos.

**Limpieza del texto, conteo de palabras y visualizaciones:**

Se realizó la limpieza del texto y el conteo de palabras, con el fin de conocer cuál es el personaje con mayor cantidad de palabras en las obras. Se obtuvo que se trata de “Poet”, y nuevamente, debido a que no se trata de un personaje propiamente dicho, resulta conveniente ver cuáles son los personajes con mayor cantidad de palabras en estas obras (figura 2).

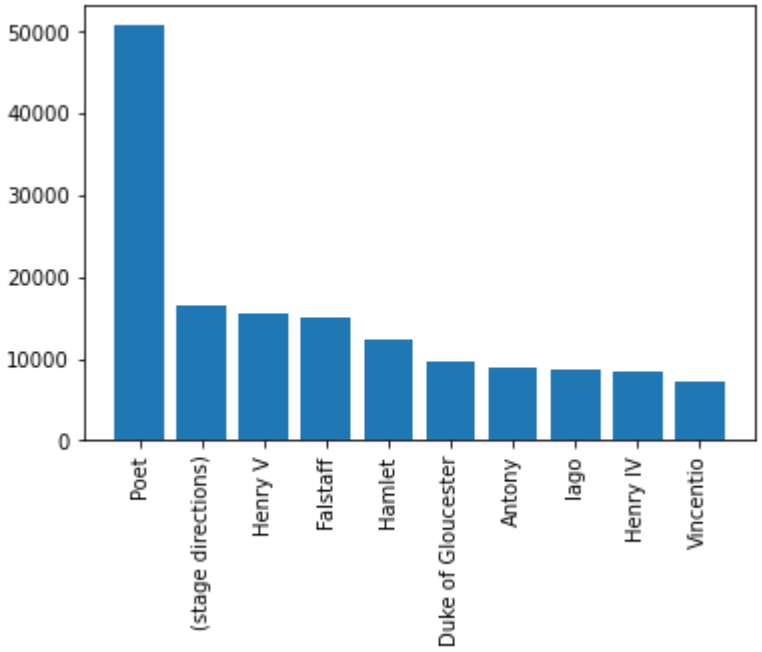


**Figura 2.** Personajes de las 43 obras de Shakespeare con mayor cantidad de palabras. *Eje x*: personajes, *eje y*: cantidad de palabras.

En números generales, se determinó que, según los registros de la base de datos, hay 909.360 palabras en las 43 obras, siendo “the” la palabra más repetida dentro de las obras, con 28.933 registros (figura 3).

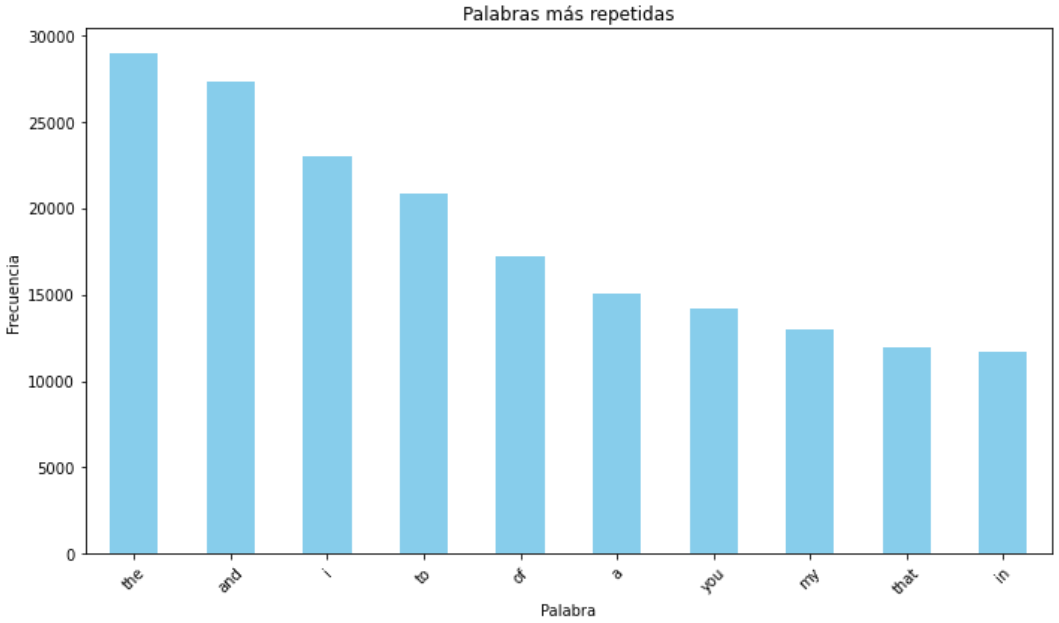


**Figura 3.** Gráfico de las 10 palabras más frecuentes en las obras analizadas.

Por otra parte, se realizaron los análisis de tendencia de las obras incluídas en la base de datos. A partir de esto se pudo observar que la mayoría de las obras pertenecen al género Comedia (n=14). Le siguen en número los géneros Historia (n=12) y Tragedia (n=11). Además encontramos un total de 5 poemas y 1 soneto.

Se realizó un histograma que permite observar la distribución de las 43 obras en el tiempo, según su año de publicación (figura 4). Puede observarse que la mayor parte de las obras (n=13, 30.2% del total de obras) fue publicada entre 1593 y 1597. De estas, 38.5% pertenecen al gènero Comedia; 30.7% al gènero Historia; 15,4% al género Tragedia y 15,4% son poemas (figura 5).

Entre los años 1593 y 1602 el autor publicó más obras del género Comedia que en el resto de los periodos de tiempo analizados, publicando 10 de las 14 obras en total. Mientras que entre 1603 y 1607 publicó principalmente obras del género Tragedia. Respecto al género Historia, puede verse cierta constancia en el número de publicaciones, apareciendo en todos los periodos de tiempo con excepción del periodo 1603-1607, justamente.

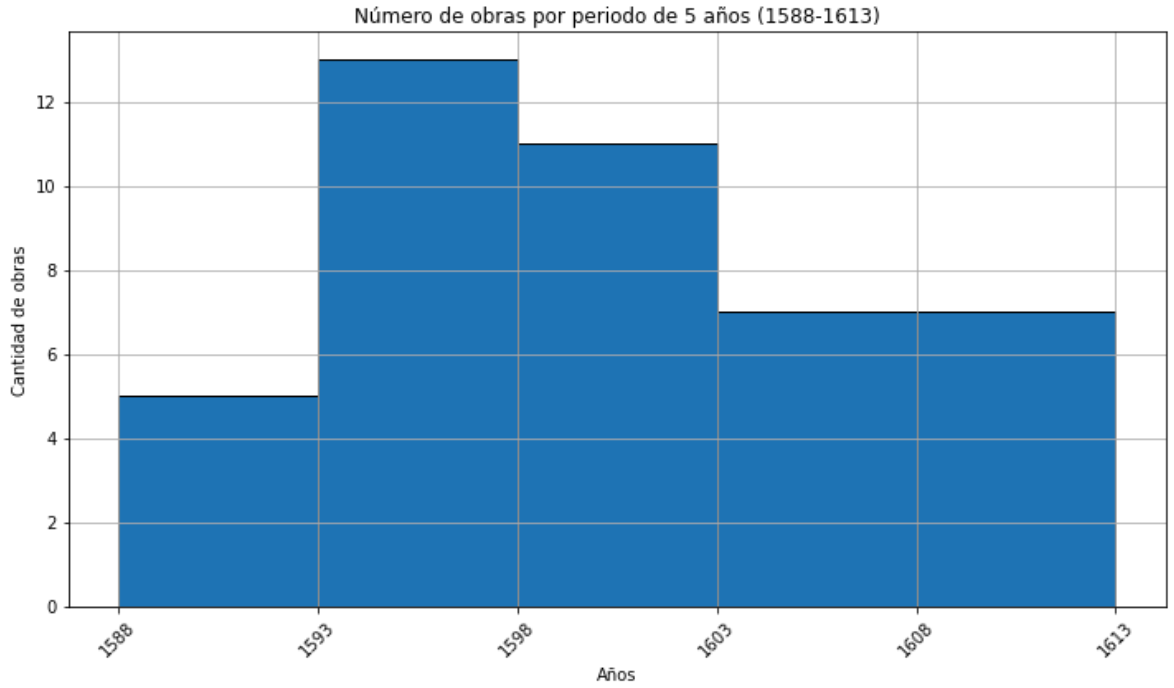


**Figura 4.** Histograma de la distribución de las obras registradas en la base de datos según su fecha de publicación. Se observa el tiempo en periodos de 5 años.

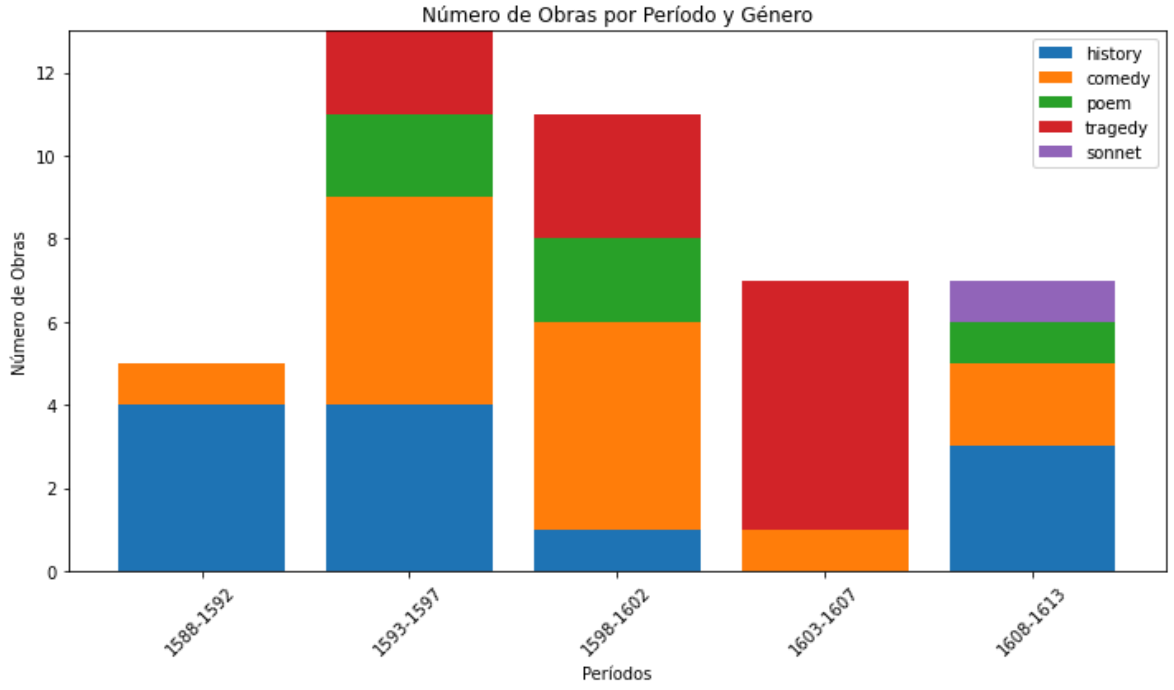


**Figura 5.** Gráfico de barras apiladas de las obras de Shakespeare en el tiempo, según el género.

**Proponga preguntas que se podrían intentar responder a partir de estos datos, y mencione posibles caminos para responderlas (sin implementar nada)**

- Podríamos determinar la cantidad de palabras en cada una de las obras y determinar cuál es la más larga.
- Tambièn podemos determinarlo por género o tipo de texto (Historia, Tragedia, Comedia, Poema y Soneto), utilizando el atributo “chapter_id” en la tabla *paragraphs*, el atributo “work_id” en la la tabla *chapters* y con esto podemos establecer relaciones entre *paragraphs, chapters* y *works* (figura 6, a).
- Se puede determinar la cantidad de personajes aproximados de cada obra si tenemos en cuenta que la tabla paragraphs contiene dos atributos: “character_id” y “chapter_id”, pudiendo ver los párrafos de cada personaje en la obra, podemos ver el “chapter_id” y en la tabla *chapters* ver el “work_id” para saber a qué obra pertenece este capítulo. Con esto, estaríamos vinculando las 4 tablas, aunque sería más fácil si la tabla *characters* incluyera el atributo “work_id” directamente (figura 6, b).

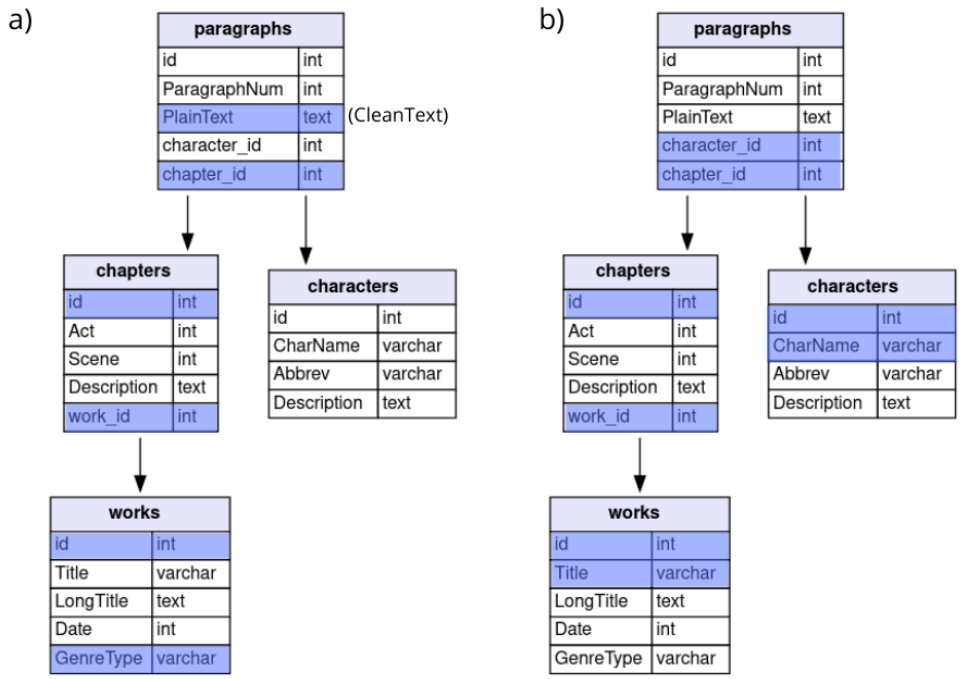


**Figura 6.** Esquemas de los posibles caminos para el análisis de datos. a) Cantidad de palabras por género de las obras, b) cantidad de personajes por obra.

**Conclusiones y recomendaciones**

1. Estandarización de Datos en la Columna LongTitle:

Inconsistencias referentes a las estructuras de columnas que hacen referencia a títulos pueden dificultar el manejo de la BD en cuanto al filtrado y búsqueda específica de obras. A ello se le suma un problema, en la misma tabla existe otro atributo con una categoría similar, como lo es “Title”.

En cuanto a lo anterior, recomendamos: Estandarizar y normalizar los datos de esta columna (LongTitle) para mantener un formato uniforme, mejorando así la eficiencia en la gestión y consulta de los datos. Por otra parte, se podría mejorar usando solo un atributo que haga referencia a los títulos de las obras.

2. Inconsistencia y falta de Información en columnas “Abbrev” y “Description”:

Se detectaron registros incompletos en las columnas "Abbrev" (5 nulos en 1266 tuplas) y "Description" (646 nulos en 1266 tuplas). La falta de información completa afecta la integridad y utilidad de la base de datos.

Se recomienda: Completar la información faltante en ambos atributos y seguir una misma convención para mantener la consistencia de los datos.

3. Duplicación y ambigüedad en nombres de personajes:

Algunos personajes aparecen con el mismo nombre en múltiples registros, como "First Senator" y "First Citizen", lo que puede deberse a errores de duplicación o a personajes distintos en diferentes obras. La inconsistencia en los datos de personajes como "Earl of Salisbury" también fue notable.

Recomendamos: Introducir un nuevo atributo "work_id" para vincular cada personaje con su obra correspondiente, eliminando ambigüedades y facilitando un análisis más preciso.

4. Integridad de la tabla “Paragraphs” :

La tabla paragraphs, que contiene 35,465 tuplas y 5 columnas, no presenta falta de datos ni inconsistencias. Los atributos incluyen: id, ParagraphNum, PlainText, character_id y chapter_id. La integridad de esta tabla es elevada, permitiendo un análisis fiable de los párrafos en las obras de Shakespeare, por tanto, no hay recomendaciones que hacer.

5. Sobre el análisis estadístico:

- Personajes por número de párrafos:

El personaje con más párrafos es "Stage directions", seguido de "Poet", aunque estos no son considerados personajes en el sentido tradicional.

Se podría actualizar la tabla para clarificar qué se considera como 'personajes' y excluir direcciones de escena y otros elementos no pertinentes.

- Conteo de palabras:

La base de datos registra un total de 909,360 palabras en las 43 obras de Shakespeare, siendo la palabra más frecuente “the”, con 28,933 apariciones.

- Análisis de géneros literarios:

La mayoría de las obras pertenecen al género comedia (14 obras), seguido de historia (12 obras) y tragedia (11 obras). Además, hay 5 poemas y 1 soneto.

- Distribución temporal de las obras:

La distribución temporal de las obras muestra que la mayor parte de las obras (13 obras, 30.2%) se publicó entre 1593 y 1597. De estas, el 38.5% son comedias, el 30.7% son historias, el 15.4% son tragedias y el 15.4% son poemas.

- Tendencias de publicación por género en diferentes periodos:

Entre 1593 y 1602, Shakespeare publicó más comedias que en otros periodos, con 10 de las 14 obras de este género.

Entre 1603 y 1607, se enfocó principalmente en tragedias.

Las obras del género Historia se publicaron de manera constante en todos los periodos, excepto entre 1603 y 1607.