FACULTAD DE INGENIERÍA

UNIVERSIDAD DE LA REPÚBLICA URUGUAY

Informe Tarea 1

Presentación y Análisis de Resultados Exploratorios

Curso: Introducción a la Ciencia de Datos

Autores: Enzo Faliveni y Mikaela Lezcano

Fecha: 23/05/2023

1. INTRODUCCIÓN Y ANÁLISIS EXPLORATORIO

El presente trabajo consta de analizar la base de datos relacional abierta con la obra completa de William Shakespeare. Para ello, luego de cargar los datos, lo primero a realizar es un análisis exploratorio de las distintas tablas que componen la base.

La base presenta 4 tablas: "Characters", "Paragraphs", "Chapters" y "Works". La primera contiene información relacionada a los nombres de los personajes de todas las obras y se vincula con la tabla "Paragraphs" a través de la variable "id" y "character_id" respectivamente. A su vez, esta última ("Paragraphs") presenta información sobre el texto y el personaje asociado a los párrafos de los diferentes capítulos de las obras y se relaciona con la tabla "Chapters" a través de los campos "chapter_id" y "id" respectivamente. Por su lado, "Chapters" tiene información sobre los actos de las diferentes escenas en cada una de las obras y se vincula con la tabla "Works" mediante las variables "work_id" y "id". Finalmente "Works" posee información asociada a los títulos y el género de las diferentes obras que hizo Shakespeare en su carrera.

Las dimensiones de la tabla "Characters" son 1266 filas y 4 columnas, dentro de las cuales se encontraron valores faltantes en la variable "Abbrev" (5 casos) y en el 50% de los registros en "Description" (646 casos). Por otro lado, se observa que existen 957 nombres de personajes únicos (309 casos son nombres repetidos los cuales a su vez corresponden a 125 nombres diferentes). El hecho de que existan personajes de distintas obras que comparten el mismo nombre podría significar un problema de calidad de los datos.

Dentro de la tabla "Paragraphs" se hallan 35.465 filas y 5 columnas sin datos faltantes. En la tabla "Chapters" tampoco existen valores faltantes y sus dimensiones son 945 filas y 5 columnas.

Analizando la tabla "Works", se encuentran 43 filas y 5 columnas y ningún dato faltante. Al mismo tiempo, la misma contiene 43 nombres de obras diferentes, publicadas entre 1589 y 1612. Los géneros de las mismas varían entre comedia, tragedia, historia, poema y soneto.

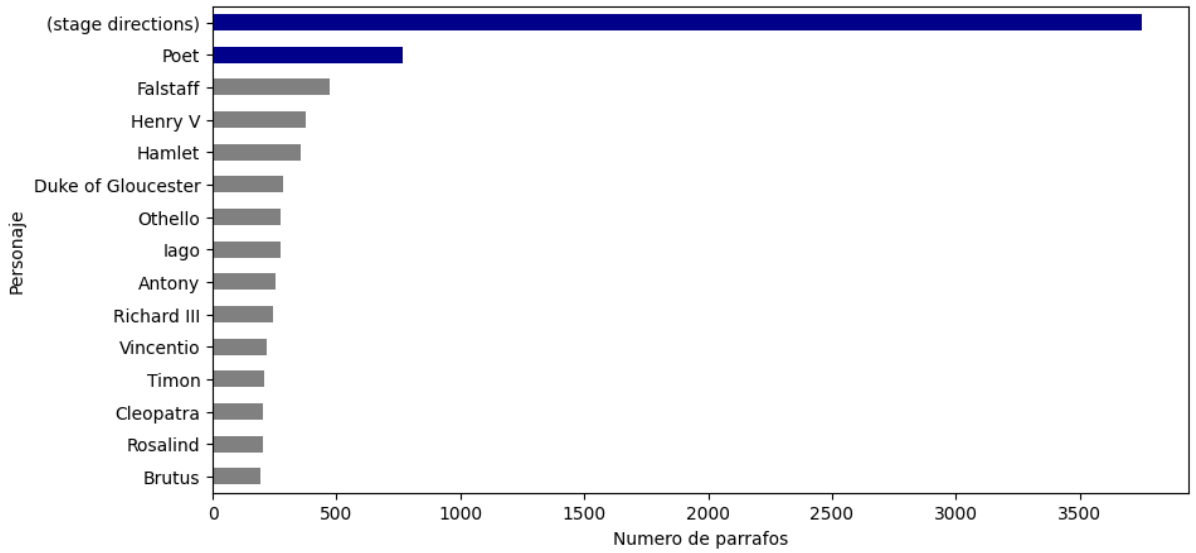
2. PÁRRAFOS POR PERSONAJE

En la actual sección se analiza y visualiza la cantidad de párrafos por personaje. Algo a tener en cuenta es que del total de 957 personajes diferentes, 925 tienen párrafos asociados, es decir que hay 32 personajes que no hablan. Como se puede ver en la Figura 1, el primer puesto lo ocupa “(stage directions)”, en español: direcciones del escenario o acotaciones, con 3.751 ocurrencias. Este término se usa en teatro y dramaturgia para describir una instrucción proporcionada por el dramaturgo o director que indica cómo se debe realizar una acción o movimiento en particular en el escenario. Las direcciones de escena generalmente se escriben en cursiva y se colocan entre corchetes o paréntesis para diferenciarlas del diálogo. Por lo tanto, se puede afirmar que no se trata de un personaje y se comprende porque ocupa el primer lugar.

En segundo lugar se observa “Poet”, el poeta (766 ocurrencias). En las obras de William Shakespeare, el personaje del poeta suele representar al propio dramaturgo. Por otro lado, en varias de las obras de Shakespeare, hay personajes a los que se hace referencia como "poeta" o "el poeta", que a menudo sirven como observadores, comentaristas o

participantes en los eventos dramáticos. Le siguen Falstaff, Henry V y Hamlet con 471, 377 y 358 ocurrencias respectivamente.

Figura 1: Cantidad de párrafos por personaje

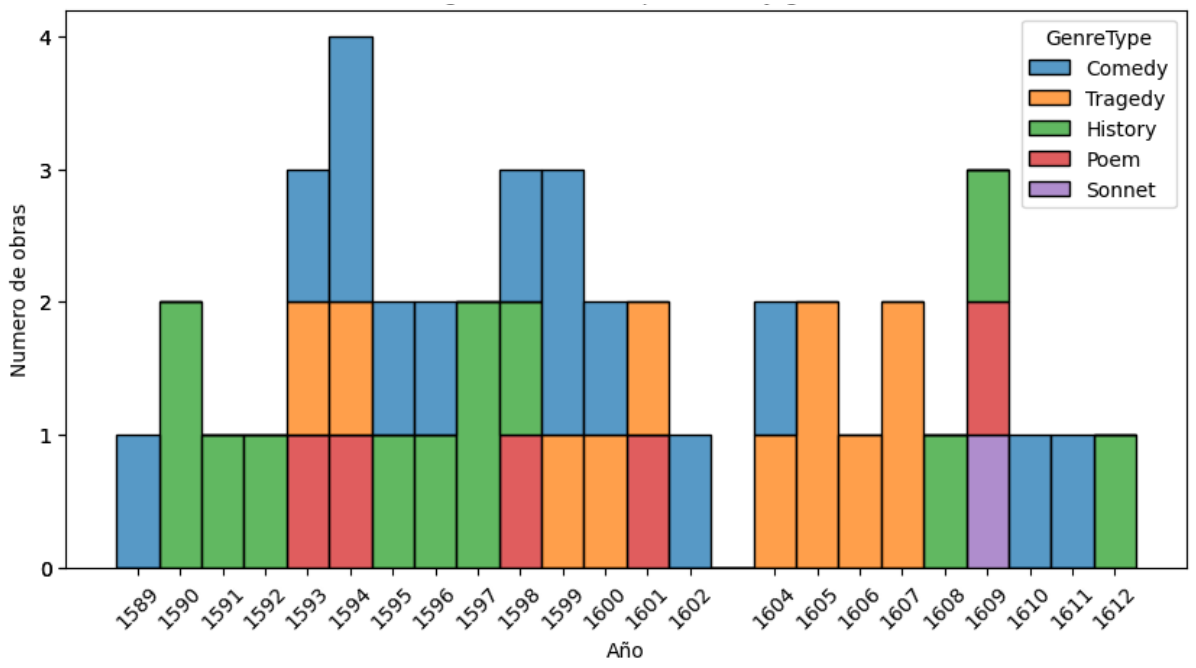


En conclusión, depende de si “poet” se considera un personaje o no, cuál es el personaje con mayor cantidad de párrafos en todas las obras de Shakespeare. En caso de considerarlo, él es el personaje con mayor cantidad de párrafos, de lo contrario es “Falstaff”.

3. OBRAS DE SHAKESPEARE EN EL TIEMPO

En esta sección se investiga cómo fue la trayectoria de trabajo de William Shakespeare a lo largo de los años. Como muestra la Figura 2, el período de mayor producción se da entre 1593 y 1601 aproximadamente. Durante el mismo priman las obras de tipo comedias (9) y en segundo lugar las tragedias e historias (5 de cada una).

Figura 2: Cantidad de obras por fecha y género



Resalta la baja presencia de poemas (5 en toda su carrera) y la producción de un único soneto en 1609 (ver Figura 3).

Figura 3: Cantidad de obras por género

GÉNERO	CANTIDAD
Comedia	14
Historia	12
Tragedia	11
Poema	5
Soneto	1

4. LIMPIEZA DE DATOS TIPO TEXTO

Para poder aplicar técnicas de Procesamiento de Lenguaje Natural (PLN o NLP en inglés) es necesario hacer una limpieza previa de los datos. La misma consta de normalizar el texto (pasarlo todo a minúsculas), eliminar todos los signos de puntuación y si el texto está en inglés, como es en este caso, expandir las contracciones. La preparación de los datos puede ser realizada a distintos niveles de profundidad en función de los objetivos que se quieran lograr con ellos en etapas posteriores. En este caso, se decidió eliminar las “stopwords” o palabras vacías como pueden ser artículos, preposiciones, conjunciones y otros términos muy frecuentes en un idioma en particular, ya que son palabras muy comunes que no aportan un significado semántico importante para el análisis de texto, sino que introducen ruido en el análisis. Eliminar las palabras vacías permite enfocarse en las palabras clave y mejora la eficacia computacional ya que se reduce el tamaño del conjunto de palabras y por ende disminuye la cantidad de datos a procesar, lo cual a su vez puede acelerar los algoritmos de procesamiento de texto y reducir los requisitos de memoria.

Siguiendo la idea de hacer el proceso más eficiente en términos de enfoque y computacionalmente, otra posibilidad podría ser aplicar técnicas de derivación y/o lematización.

La derivación, también conocida como stemming, consiste en eliminar los sufijos de las palabras con el objetivo de obtener su raíz. Es un enfoque basado en reglas heurísticas y no tiene en cuenta el contexto o la semántica de las palabras. Las palabras derivadas pueden no ser lexicográficamente válidas. Por ejemplo, las palabras "corriendo", "corre" y "corrió" se reducirían a la raíz "corr" mediante la derivación.

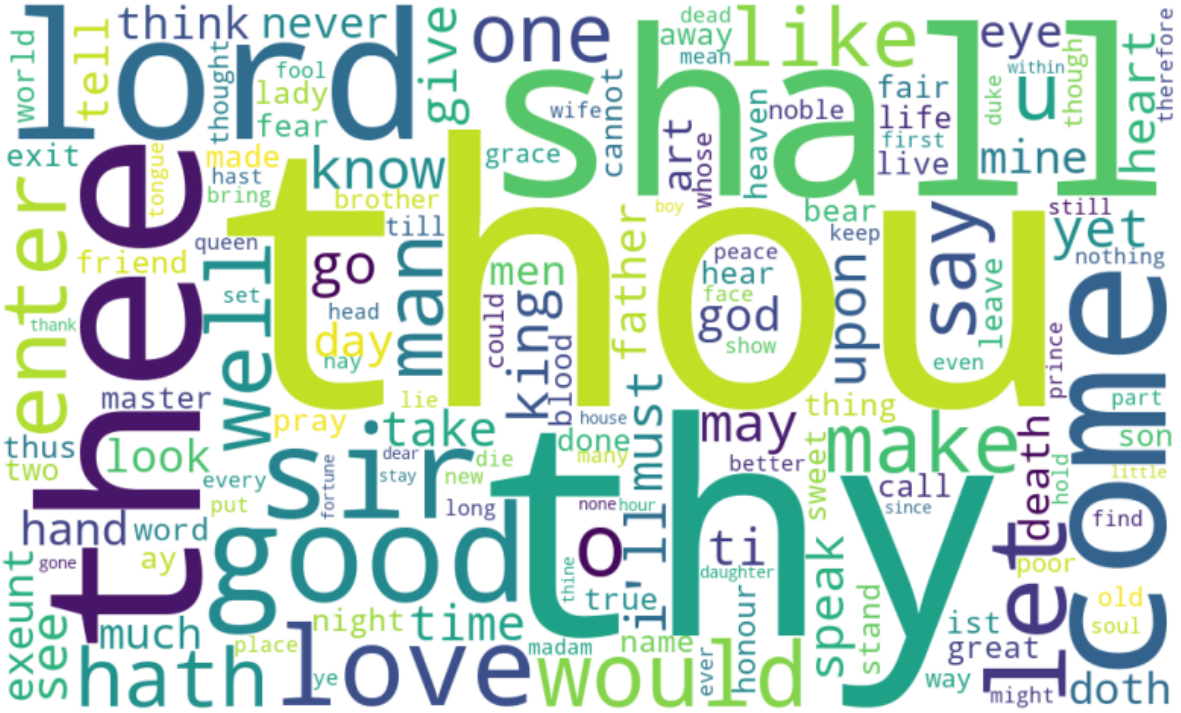
Por otro lado, la lematización implica determinar la forma canónica o lema de una palabra en función de su significado en el contexto. A diferencia de la derivación, la lematización tiene en cuenta el contexto y la categoría gramatical de las palabras. El resultado de la lematización es una palabra válida y reconocible en el idioma. Por ejemplo, las palabras "corriendo", "corre" y "corrió" se lematizarían todas a "correr".

5. PROCESAMIENTO DE LENGUAJE NATURAL

5.1. Palabras más frecuentes en el conjunto de todas las obras

A fin de investigar cuáles son las palabras que más se repiten en las obras de William Shakespeare se muestra la siguiente visualización:

Figura 4: Palabras más repetidas en todas las obras



La palabra "thou" es notoriamente la que alcanza un mayor número de repeticiones (5.768), seguida de "thy" con 4.262, "shall" con 3.721 y “thee” con 3.378.

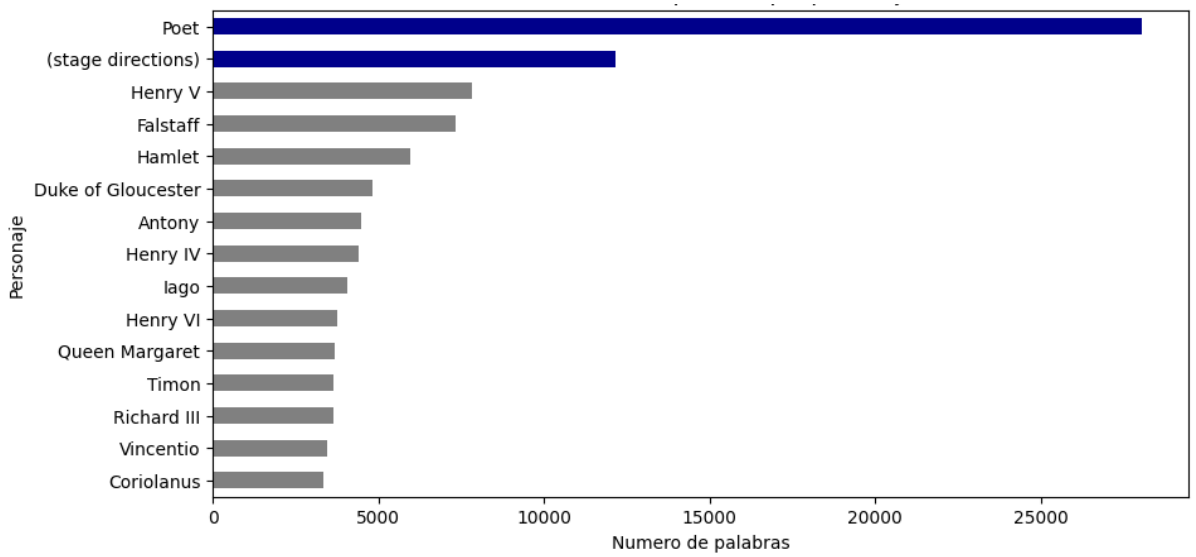
Podría ser interesante hacer este mismo análisis pero diferenciando por personaje y así conocer qué tipo de palabras priman en cada uno; lo mismo si se hace la diferenciación por género o incluso por ambas variables al mismo tiempo, es decir analizar las palabras más frecuentes de los personajes asociados a cada género. A modo de ejemplo, quizás en los personajes de las tragedias priman palabras asociadas a eventos y sentimientos negativos.

5.2. Personajes con mayor cantidad de palabras

A diferencia de lo que sucedía cuando se hizo el estudio de personajes con mayor cantidad de párrafos, en este caso se invierten los primeros dos lugares y el personaje “poet” pasa a ocupar el primer puesto y por notoria diferencia (ver Figura 5), es decir que alcanza una mayor cantidad de palabras a pesar de tener menos párrafos asociados en comparación a “(stage directions)”.

El tercer y cuarto lugar son ocupados por “Henry V” y “Falstaff” con 7.846 y 7.352 palabras respectivamente. Nuevamente se deja planteada la interrogante respecto a si “poet” se debería considerar un personaje o no.

Figura 5: Personajes con mayor cantidad de palabras en todas las obras



6. PREGUNTAS QUE SE PODRÍAN RESPONDER CON LA BASE DE DATOS UTILIZADA

- Pregunta 1: ¿cuántos actos hay por escena en promedio?

Forma de responder: a partir de la tabla “Chapters”, agrupar por escena contando la cantidad de actos en cada una y luego calcular el promedio.

- Pregunta 2: ¿cuántos personajes hay por acto o escena en promedio?

Forma de responder: unir las tablas “Characters” y “Paragraphs”, ese resultado unirlo con la tabla “Chapters” y luego agrupar por acto o escena contando la cantidad de personajes y finalmente calcular el promedio.

- Pregunta 3: ¿es posible identificar el género de una obra a partir del texto de la misma?

- Pregunta 4: ¿es posible identificar el personaje asociado a un párrafo en función del texto del mismo?

Las últimas dos preguntas se responden de manera similar. De la misma forma que se hizo en la sección 5.1. se podría hallar dicha relación mediante heurísticas meramente estadísticas generadas a partir de las palabras más frecuentes por obra y/o por personaje, y finalmente clasificar la obra y/o el párrafo mediante sus N palabras más frecuentes asociadas al tipo de obra y/o personaje.

Este método es una primera aproximación a abordar el problema, el cual proporciona una referencia inicial para la evaluación. Sin embargo, una solución de mayor complejidad y potencial de capacidad predictiva, implica el uso de modelos de Procesamiento de Lenguaje Natural. A continuación, se presenta esta alternativa.

7. LONG SHORT TERM MEMORY MODEL

Como se dijo en la sección anterior, una propuesta interesante a partir de la base de datos disponible, consiste en utilizar los personajes como etiquetas categóricas y el texto procesado, para determinar cuál es el personaje más probable que haya pronunciado un

texto dado. La propuesta se fundamenta en el uso de técnicas más avanzadas de PLN. Es relevante destacar que existen múltiples alternativas para alcanzar este objetivo. A modo de ejemplo, se plantea un enfoque específico que involucra el preprocesamiento del texto previamente etiquetado. A través de la tokenización del texto utilizando el tamaño total del vocabulario y estrategias adecuadas de “padding”, se generan vectores (secuencias de tokens) de longitud fija, con el objetivo de ser utilizados posteriormente como insumos del modelo.

Luego, se procede a la creación del conjunto de entrenamiento y prueba. Es importante mencionar que el problema se tratará como un problema de clasificación múltiple, por lo tanto, se realiza la codificación one-hot en las etiquetas. Esto implica que el modelo generará un vector de probabilidades cuya longitud coincide con la cantidad de categorías, es decir, los personajes considerados en el estudio.

Por último, se procede con la implementación del modelo. Este paso puede ser realizado de diversas formas, pero como propuesta concreta, es posible entrenar internamente los embeddings y utilizar capas de LSTM en forma secuencial. No obstante, también se considera como opción válida el uso de embeddings pre-entrenados en un conjunto de datos más amplio y variado. A modo de demostración, los detalles específicos sobre la implementación de esta solución se encuentran disponibles en el código correspondiente.

8. COMENTARIOS Y REFLEXIONES FINALES

Tras llevar a cabo el análisis del presente informe, se puede afirmar que se ha logrado realizar una evaluación precisa de la calidad de los datos disponibles, así como interpretar su naturaleza y distribución mediante la generación de visualizaciones comprensibles y adecuadas. Estos resultados, han brindado una base sólida para afrontar de manera efectiva las nuevas problemáticas planteadas, desarrollando soluciones coherentes y escalables.

9. REFERENCIAS BIBLIOGRÁFICAS

- Text Normalization for Natural Language Processing (NLP)
- Multiclass Text Classification Using Deep Learning
- Stemming
- Lemmatisation