Introducción a la Ciencia de Datos
Maestría en Ciencia de Datos y Aprendizaje Automático
Facultad de Ingeniería, UdelaR

Tarea 1

Grupo 20: Joana Auriello, Pablo Molina

Índice

Introducción

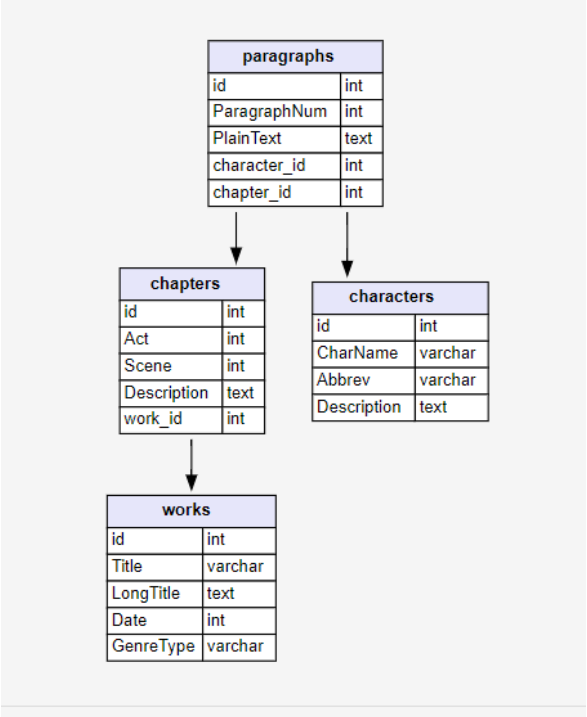
En el presente trabajo se realizará diversos análisis basados en la base de datos relacional que contiene las obras de William Shakespeare y sus atributos. A partir de ello se observarán distintas tendencias y se aplicarán técnicas para análisis de texto que se explicarán más adelante.

Los objetivos principales de este informe serán obtener conclusiones generales sobre tendencias y características de las obras de William Shakespeare utilizando metodologías para el estudio de la calidad de los datos y otras técnicas de exploración de datos sobre la base, mayormente basado en distintas visualizaciones.

El informe se estructura de la siguiente manera: en la primera sección procederemos a realizar la exploración inicial de datos, en la segunda sección se realizará observaciones sobre la calidad de los datos, la tercera sección se focalizará en el análisis de las obras y distintas tendencias, en la cuarta sección se explicarán las transformaciones realizadas a los datos para el análisis de texto y en la quinta sección se indican preguntas para futura investigación sobre la base de datos principal. Se finaliza el informe con las principales conclusiones sobre la totalidad del trabajo.

Exploración inicial de datos (EDA)

Modelo



El modelo de la base de datos de las obras de Shakespeare está diseñado para almacenar información sobre las obras, capítulos, personajes y párrafos de los textos. La base de datos está compuesta por las siguientes tablas: works, chapters, characters, y paragraphs.

Tablas

Works: almacena información básica sobre las obras de Shakespeare. Muestra el nombre de la obra, su fecha de publicación y a qué género corresponde.
Se compone por las siguientes columnas:

- id (int): Identificador único de la obra.
- Title (varchar): Título de la obra.
- LongTitle (text): Título completo o extendido de la obra.
- Date (int): Año de publicación o de primera representación de la obra.
- GenreType (varchar): Género literario de la obra (Ejemplo:comedia, tragedia).

Chapters: almacena información sobre los capítulos de las obras, los cuales pueden estar divididos en actos y escenas. Sabremos qué actos y escenas corresponde a cada párrafo y a que obra corresponde el conjunto de párrafos. También nos brinda una breve descripción de cada combinación Acto - Escena. Se compone por las siguientes columnas:

- id (int): Identificador único del capítulo.
- Act (int): Número del acto al que pertenece el capítulo.
- Scene (int): Número de la escena dentro del acto.
- Description (text): Descripción breve del capítulo.
- work_id (int): Identificador de la obra a la que pertenece el capítulo (llave foránea de works).

Characters: almacena información sobre los personajes que aparecen en las obras. Nos permite comprender el nombre del personaje, su abreviación o sobrenombre y una breve descripción del mismo. Se compone por las siguientes columnas:

- id (int): Identificador único del personaje.
- CharName (varchar): Nombre del personaje.
- Abbrev (varchar): Abreviatura o nombre corto del personaje.
- Description (text): Descripción del personaje.

Paragraphs: Nos brinda el orden y el contenido de cada párrafo del texto, también nos permite saber quién lo dijo y a que capítulo corresponde. Se compone por las siguientes columnas:

- id (int): Identificador único del párrafo.
- ParagraphNum (int): Número secuencial del párrafo.
- PlainText (text): Texto del párrafo.
- character_id (int): Identificador del personaje que dice el párrafo (llave foránea de characters).
- chapter_id (int): Identificador del capítulo donde se encuentra el párrafo (llave foránea de chapters).

Relaciones

Todas las tablas cuentan con un indicador único de línea como clave primaría. Las relaciones entre tablas estarán determinadas por cómo estas claves primarias de las tablas forman parte de otras tablas como clave foránea.

La tabla de Párrafos será la cual conectará personajes con los capítulos y los trabajos ya que es la que reúne las claves foráneas para conectar todas las tablas del modelo.

Relaciones entre Tablas
Relación entre works y chapters

- Tipo de Relación: Uno a Muchos
- Una obra (works) puede tener múltiples capítulos (chapters). Esta relación se establece mediante la columna work_id en la tabla chapters, que actúa como llave foránea refiriéndose a la columna id en la tabla works.

Relación entre chapters y paragraphs

- Tipo de Relación: Uno a Muchos
- Un capítulo (chapters) puede contener múltiples párrafos (paragraphs). Esta relación se establece mediante la columna chapter_id en la tabla paragraphs, que actúa como llave foránea refiriéndose a la columna id en la tabla chapters.

Relación entre characters y paragraphs

- Tipo de Relación: Uno a Muchos
- Un personaje (characters) puede tener múltiples párrafos (paragraphs) asignados. Esta relación se establece mediante la columna character_id en la tabla paragraphs, que actúa como llave foránea refiriéndose a la columna id en la tabla characters.

En resumen, el diagrama ER (Entidad-Relación) de la base de datos muestra claramente estas relaciones mediante flechas que apuntan desde las llaves foráneas hacia las tablas referenciadas. Esto facilita la comprensión de cómo los datos se conectan y se organizan en la base de datos.

Calidad de datos

La calidad de datos es un aspecto crucial en cualquier análisis, ya que determina la precisión y la confiabilidad de los resultados obtenidos. En el contexto de la base de datos que estamos analizando, que contiene información sobre las obras de William Shakespeare, es fundamental asegurarnos de que los datos sean completos, coherentes y exactos.

¿Qué es la calidad de datos?

Se refiere a la medida en que los datos son precisos, completos, consistentes y relevantes para el propósito deseado. Datos de alta calidad permiten tomar decisiones informadas y

realizar análisis precisos, mientras que datos de baja calidad pueden llevar a conclusiones erróneas y decisiones equivocadas.

Existen varias técnicas y prácticas comunes que se utilizan para asegurar y mejorar la calidad de datos. Algunas de las técnicas más básicas incluyen:

- Validación de datos: Asegurarse de que los datos ingresados cumplen con ciertos criterios o reglas predefinidas. Esto incluye la verificación de tipos de datos, formatos, rangos y valores permitidos.
- Limpieza de datos: El proceso de identificar y corregir errores en los datos. Esto puede implicar la eliminación de duplicados, la corrección de valores incorrectos, la normalización de formatos y la eliminación de registros incompletos o irrelevantes.
- Consistencia de datos: Verificar que los datos sean consistentes entre diferentes tablas y fuentes. Por ejemplo, asegurarse de que las relaciones entre las tablas sean correctas y que no haya discrepancias entre ellas.
- Integridad referencial: Mantener la precisión y coherencia de las relaciones entre las tablas de una base de datos. Esto se logra mediante el uso de claves primarias y foráneas, y asegurándose de que las referencias entre tablas sean válidas.
- Completitud de datos: Garantizar que todos los campos necesarios estén completos. Esto implica identificar y manejar valores nulos o faltantes de manera adecuada.
- Análisis de valores atípicos: Identificar y analizar valores que se desvían significativamente del resto de los datos. Los valores atípicos pueden indicar errores en los datos o puntos de datos válidos que requieren una atención especial.

En el análisis de la base de datos de las obras de Shakespeare, hemos implementado varias de estas técnicas para asegurar que los datos sean de alta calidad antes de proceder con el análisis detallado y las visualizaciones correspondientes. Estas técnicas nos han permitido identificar y corregir posibles problemas en los datos, garantizando así la precisión y confiabilidad de nuestros resultados.

Calidad de datos en la base de datos propuesta

Se procede a realizar el análisis de calidad de datos para cada tabla de la base de datos.

La tabla "Works" contiene información sobre las obras de William Shakespeare, incluyendo títulos, fechas y géneros. Se encontraron los siguientes puntos principales: las primeras filas del DataFrame muestran una estructura consistente y apropiada de los datos. No se encontraron valores faltantes en ninguna de las columnas, lo que indica una buena completitud de los datos. Además, no se detectaron registros duplicados, asegurando la unicidad de los datos. La columna "GenreType" contiene cinco géneros únicos (Comedy, Tragedy, History, Poem y Sonnet), lo cual es coherente con las categorías de las obras de Shakespeare.
La tabla contiene un total de 43 títulos de obras. Los tipos de datos en la tabla son apropiados para cada columna: enteros para identificadores y fechas, y objetos para títulos y géneros. Las estadísticas descriptivas de la columna "Date" muestran que las fechas de las obras están dentro de un rango razonable, entre 1589 hasta 1612.

La tabla "Paragraphs" contiene información detallada sobre los párrafos de las obras de William Shakespeare, incluyendo el texto de los párrafos, el número de párrafo, el personaje

y el capítulo al que pertenece cada párrafo. El análisis de calidad de datos reveló los siguientes aspectos clave: las primeras filas del DataFrame muestran que la estructura de los datos es coherente y adecuada para el análisis. No se encontraron valores faltantes en ninguna de las columnas, lo cual indica una completitud total de los datos. Además, no hay registros duplicados, asegurando que cada párrafo es único dentro de la base de datos. Los tipos de datos en las columnas son apropiados: enteros para los identificadores, números de párrafo, identificadores de personajes y capítulos, y objetos para el texto de los párrafos.

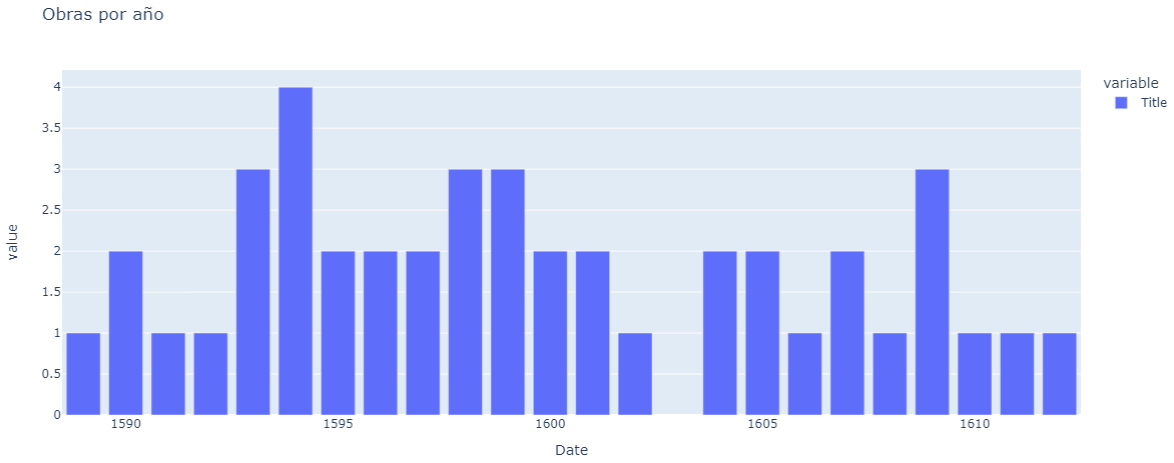
La tabla "Characters" contiene información sobre los personajes de las obras de William Shakespeare, incluyendo el nombre del personaje, una abreviatura y una descripción opcional. Al analizar la calidad de los datos, se observaron los siguientes puntos clave: las primeras filas del DataFrame muestran una estructura clara y coherente de los datos. No hay valores faltantes en las columnas "id" y "CharName", lo cual es crucial ya que estos campos identifican a cada personaje de manera única. Sin embargo, hay valores faltantes en las columnas "Abbrev" y "Description", con 5 y 646 valores faltantes respectivamente. Esto indica que muchas descripciones de personajes no están disponibles, lo cual podría limitar algunos análisis detallados de los personajes. No se encontraron registros duplicados, asegurando la unicidad de cada entrada de personaje. Los tipos de datos son adecuados, con enteros para los identificadores y objetos para los nombres, abreviaturas y descripciones. Finalmente, el análisis muestra que hay un total de 1266 personajes únicos en la base de datos, lo cual refleja la diversidad de personajes en las obras de Shakespeare.

La tabla "Chapters" proporciona información detallada sobre los capítulos de las obras de William Shakespeare, incluyendo el acto, la escena, una descripción breve y el identificador de la obra a la que pertenece cada capítulo. Al revisar las primeras filas, se observa que la estructura de los datos es coherente y clara. No hay valores faltantes en ninguna de las columnas, lo cual asegura la integridad de la información proporcionada. Además, no se encontraron registros duplicados, lo que garantiza la unicidad de cada capítulo registrado.

En resumen, las tablas incluidas en la base de datos presentan una alta calidad de datos con buena completitud, unicidad y consistencia, lo cual es crucial para realizar análisis más profundos y precisos sobre las obras y textos de Shakespeare.

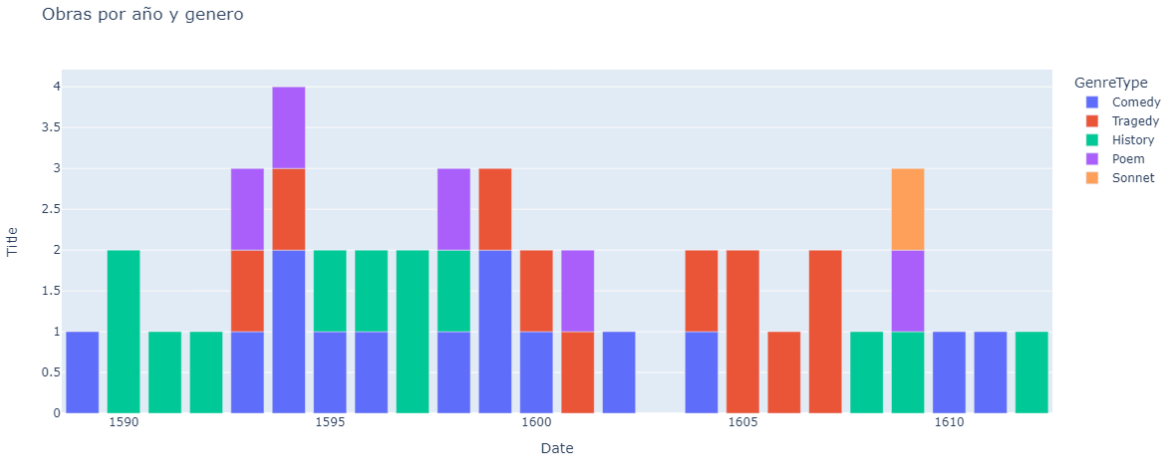
Análisis de obras y tendencias

Obras de Shakespeare por años



Hay ciertos años donde Shakespeare fue particularmente productivo, como 1590, 1594, 1595, 1597 y 1600, donde se observa un aumento notable en la cantidad de obras escritas. Siendo 1594 el año más fructífero.
En los últimos años, específicamente 1613, se observa una ligera disminución en la producción, lo cual es consistente con el final de su carrera creativa antes de su muerte en 1616.

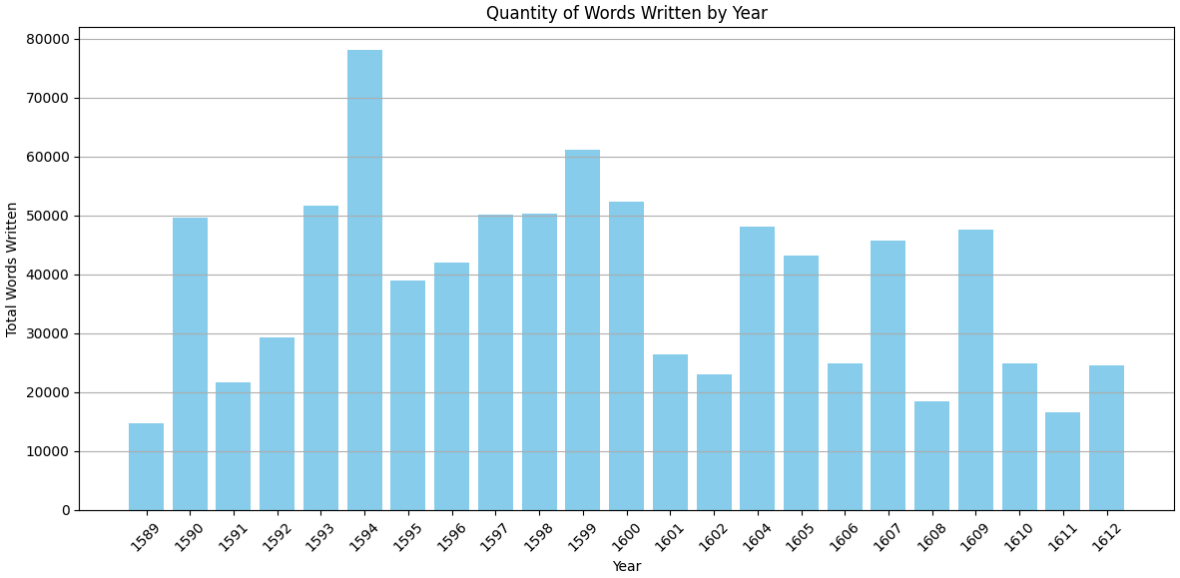
Obras de Shakespeare por año y género



Al observar esta gráfica, podemos interpretar lo siguiente:

1. Obras Históricas: Las obras históricas son más frecuentes en los primeros años de la carrera de Shakespeare, especialmente a principios de la década de 1590 y alrededor de 1597.
2. Comedias: Las comedias aparecen consistentemente a lo largo de los años,pero se concentran mayoritariamente entre 1693 y 1600.
3. Tragedias: Las tragedias se concentran principalmente entre 1600 y 1608, con algunos de los trabajos más notables como "Hamlet", "Otelo" y "Macbeth" en este período.
4. Sonetos y Poemas: Los sonetos y poemas se concentran en un corto período. Los poemas "Venus y Adonis" y "La violación de Lucrecia" fueron publicados a principios de la carrera de Shakespeare (1593-1594). Su único soneto fue publicado en 1609.

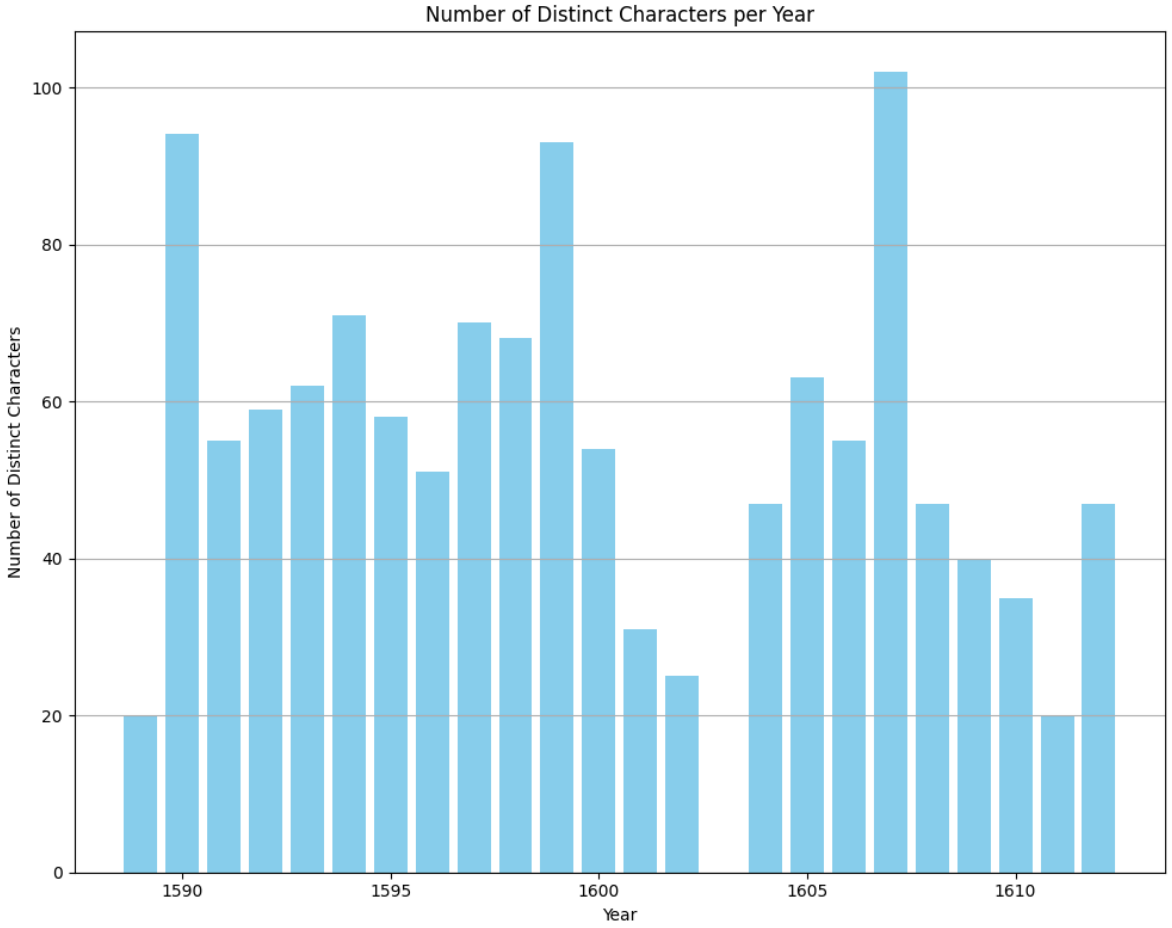
Por palabras



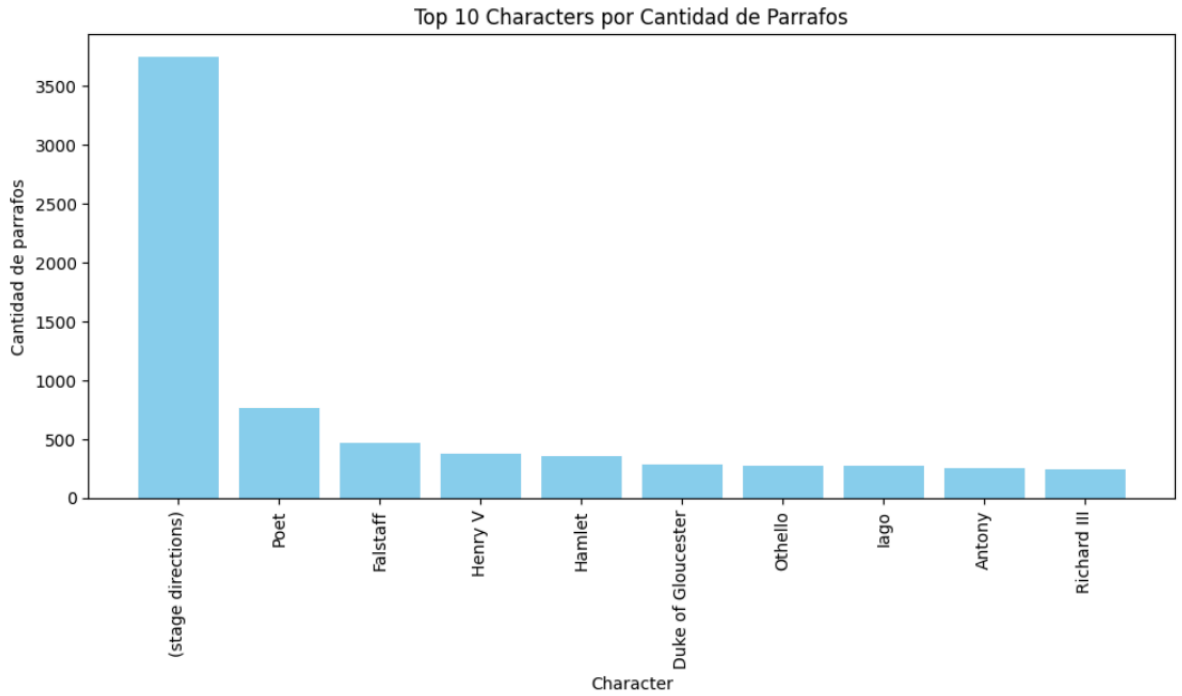
La longitud de las obras varía considerablemente de un año a otro. Por ejemplo, en 1590, aunque se publicaron varias obras todas forman parte de "Henry VI".

Por personajes

No podemos identificar una correlación notoria entre la cantidad de personajes y la cantidad de palabras de una obra.



Cantidad de párrafos por personaje



Antes de explicar la distribución de párrafos por personaje, es fundamental destacar el papel de las "stage directions" en las obras de Shakespeare. Estas instrucciones escénicas, aunque no constituyen diálogo hablado, son relevantes para la comprensión de la obra y su representación teatral. Brindan orientación sobre la ambientación, los movimientos de los personajes y otros aspectos escénicos, enriqueciendo la experiencia teatral. Es por esto que stage directions aparece en primer lugar en el gráfico presentado, dada su predominancia en las obras de Shakespeare a pesar de que no es considerado un personaje en sí.

Aunque no siempre es considerado uno de los personajes principales en las obras de Shakespeare, el Poeta(Poet) tiene una significativa cantidad de párrafos. Este personaje desempeña roles diversos en las obras de Shakespeare, y su función es aportar profundidad y contexto a las tramas. Es una herramienta literaria que Shakespeare utiliza y es esperable que también lo encontramos en el top de párrafos por personaje.

Personajes como Falstaff, Henry V y Hamlet se destacan por la cantidad de párrafos que les son atribuidos. Esto refleja su importancia en las respectivas obras. Por ejemplo, Falstaff, desempeña un papel crucial en el desarrollo de la trama y la exploración de temas relevantes.

Transformaciones de datos y normalización

¿Para qué sirve normalizar datos en general y en particular para análisis de texto?

La normalización de datos es un paso crucial en cualquier análisis de datos, y tiene beneficios significativos tanto en modelos numéricos como en análisis de texto. Normalizar los datos asegura que todas las variables sean comparables entre sí. En modelos numéricos, la variabilidad en las magnitudes de las columnas puede confundir al modelo, dándole un peso excesivo a columnas con valores más altos simplemente debido a sus unidades. La normalización ajusta los valores de las columnas para que tengan un rango similar, evitando que una variable domine a las demás por su escala.

En el análisis de texto, la normalización tiene como objetivo reducir la redundancia y corregir errores que pueden surgir cuando palabras diferentes se usan para expresar el mismo concepto. Esto es crucial porque un modelo de lenguaje podría interpretar palabras como distintas a pesar de tener significados similares, lo que puede llevar a errores y a subestimar la frecuencia y el peso de ciertos términos. Por ejemplo, palabras como "análisis" y "analizar" deben ser reconocidas como variantes de la misma raíz.

Además, en el procesamiento de lenguaje natural (NLP), se utiliza una técnica llamada lematización, que agrupa palabras con significados similares. La lematización transforma palabras diferentes que representan la misma idea en una forma común. Por ejemplo, "correr" y "trotar" pueden considerarse suficientemente similares en significado para ciertos análisis, y la lematización las trataría como una sola entidad. Esto no solo reduce la complejidad del modelo sino que también mejora su precisión al capturar la esencia del texto sin ser distraído por variaciones superficiales en la terminología.

Transformaciones realizadas

Para mejorar la calidad y consistencia de los datos textuales en nuestro análisis, se realizaron varias transformaciones en la columna PlainText del Data Frame df_paragraphs, resultando en la creación de una nueva columna llamada CleanText. Estas transformaciones tienen como objetivo preparar el texto para un análisis de texto más eficiente y preciso. A continuación, se describen las transformaciones aplicadas y su utilidad en el análisis de texto:

1. Conversión a minúsculas: Se convirtió todo el texto a minúsculas. Esta transformación es fundamental para evitar que el mismo término sea tratado como dos palabras diferentes debido a diferencias de capitalización. Por ejemplo, "Amor" y "amor" se consideran la misma palabra, lo que simplifica el análisis y mejora la precisión de las estadísticas de frecuencia de palabras.

2. Eliminación de Signos de Puntuación: Se eliminaron todos los signos de puntuación y se reemplazaron por espacios. Los signos de puntuación pueden introducir ruido en el análisis de texto, ya que no aportan información semántica relevante en la

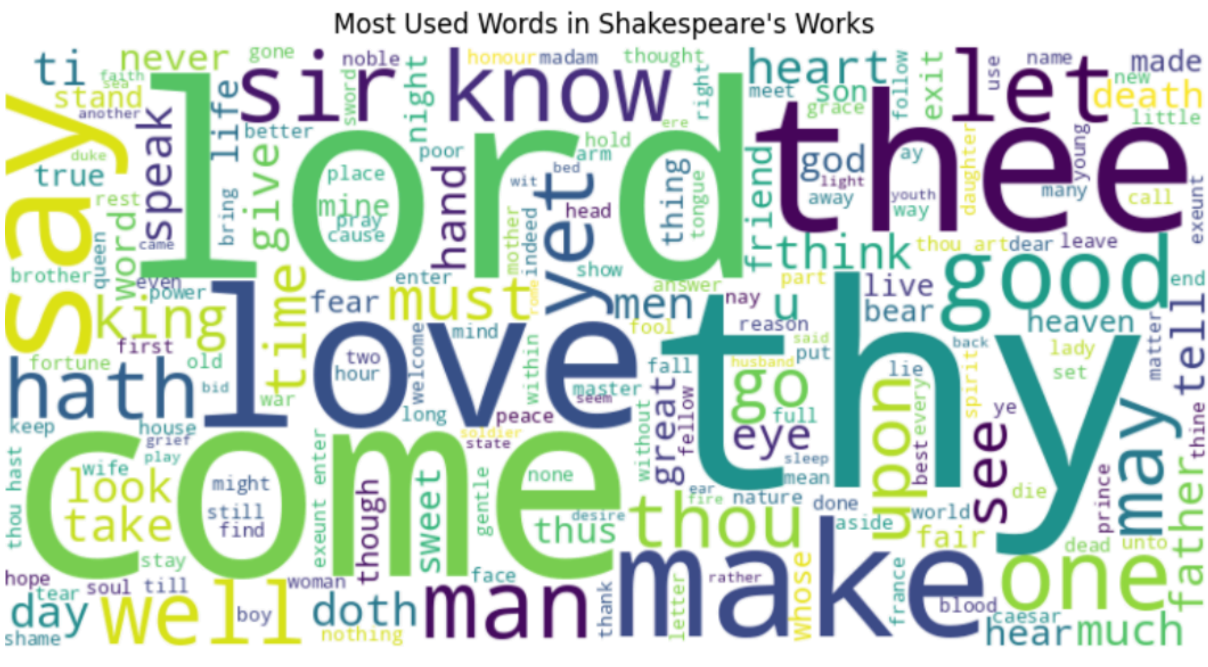
mayoría de los casos. Al eliminarlos, se reducen las variaciones innecesarias en los datos y se facilita la tokenización de las palabras, es decir, la separación del texto en unidades significativas.

3. Expansión de Contracciones: Utilizando un diccionario de contracciones y expresiones regulares, se expandieron las contracciones comunes en el inglés y en las obras de Shakespeare. Por ejemplo, "can't" se expandió a "cannot" y "o'er" se expandió a "over". Esta transformación es esencial para asegurar que todas las formas contractas de palabras sean tratadas uniformemente, mejorando la precisión en el análisis de frecuencia y en otros métodos de procesamiento de lenguaje natural.

Estas transformaciones son esenciales para asegurar que el texto sea procesado de manera consistente y que las palabras se analicen de forma uniforme. La limpieza del texto reduce la complejidad y mejora la precisión de técnicas de procesamiento de lenguaje natural (NLP) como el análisis de frecuencia de palabras, la lematización, la extracción de temas y el modelado de tópicos. Con un texto normalizado y limpio, es más fácil identificar patrones, realizar comparaciones y extraer insights significativos de las obras de Shakespeare.

Análisis posterior a transformaciones

Palabras más frecuentes



En la visualización se muestra una nube de palabras que destaca las palabras más utilizadas en las obras de Shakespeare. El tamaño de cada palabra es proporcional a su frecuencia en el texto. Palabras como "lord", "love", "come", "thy", y "thee" aparecen con mayor tamaño, indicando que son términos recurrentes en las obras del autor. Estas

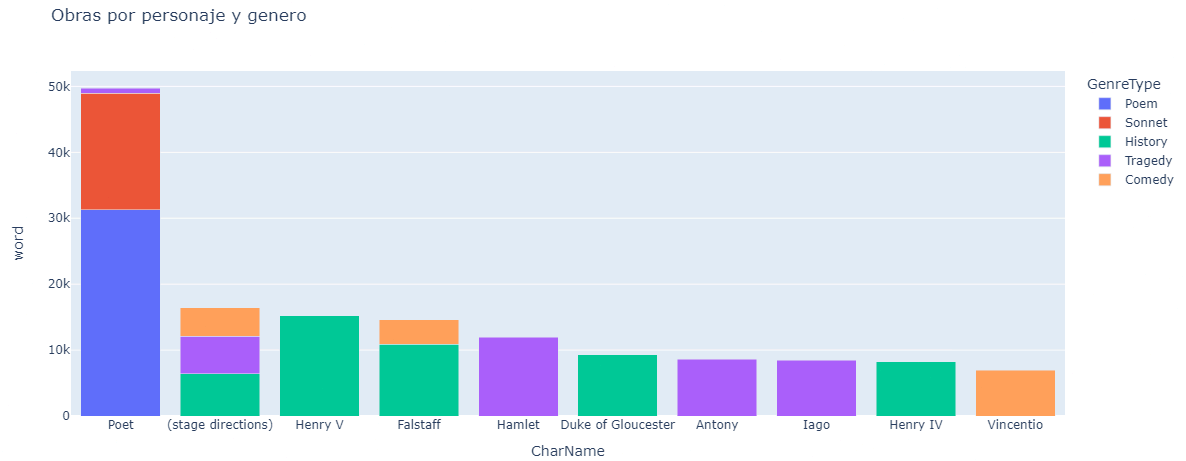
palabras reflejan temas y personajes comunes en la literatura shakesperiana, donde términos relacionados con la nobleza ("lord"), el afecto ("love"), y el lenguaje arcaico ("thy", "thee") son muy utilizados.

Esta visualización es útil para obtener una visión rápida y general de las palabras más comunes, pero tiene limitaciones cuando se trata de análisis más detallados, como la comparación entre géneros (comedias, tragedias, y dramas históricos) o el análisis de vocabulario específico de personajes individuales. Para abordar estas comparaciones, se podrían implementar las siguientes modificaciones:

- Separación por Géneros: Crear nubes de palabras separadas para cada género literario. Esto permitiría identificar palabras y temas que son más usados en comedias, tragedias, o dramas históricos. Por ejemplo, en las comedias, podríamos esperar ver palabras más asociadas con situaciones humorísticas o personajes graciosos, mientras que en las tragedias, palabras relacionadas con el destino y la muerte podrían ser más frecuentes.

- Análisis por Personajes: Generar nubes de palabras para los diálogos de personajes específicos. Esto ayudaría a comprender mejor el lenguaje y el estilo de habla de distintos personajes. Por ejemplo, podríamos comparar el vocabulario utilizado por personajes nobles contra el de personajes plebeyos, o analizar cómo varía el uso del lenguaje en personajes principales versus secundarios.

Personajes más frecuentes por género, por cantidad de palabras



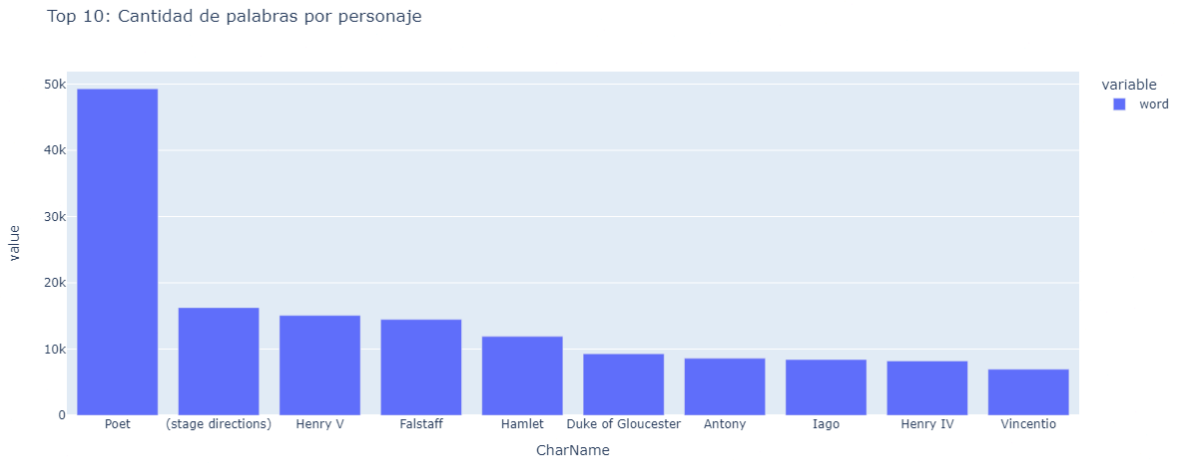
Al analizar la cantidad de palabras por personaje y género, observamos que los dos "personajes" más frecuentes, "Poet" y "Stage Directions", aparecen en más de un género. Como fue explicado previamente, esto se debe a que no son personajes tradicionales, sino anotaciones utilizadas por el autor a lo largo de las obras. Al excluir estos elementos especiales, el personaje con mayor cantidad de palabras es Henry V, lo cual es comprensible dado su rol central en varias obras históricas de Shakespeare. Este personaje

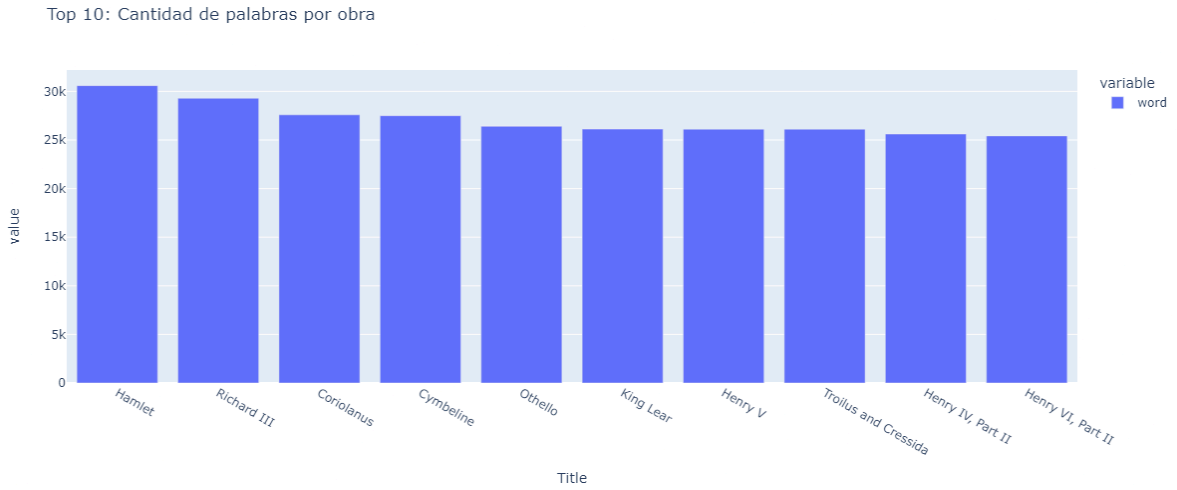
sobresale especialmente en las obras del género histórico, donde su presencia y discursos son predominantes.

Además, personajes como Falstaff y Hamlet también destacan por la cantidad de palabras que pronuncian, cada uno asociado a diferentes géneros como la comedia y la tragedia respectivamente. Esto refleja cómo Shakespeare distribuye el peso de sus personajes principales según el género de la obra, proporcionando un equilibrio entre diálogos extensos y la estructura narrativa de sus diversas piezas literarias.

La gráfica incluye personajes de tragedias (Hamlet, King Lear, Othello, Timon), comedias (Falstaff), y obras históricas (Henry V, Richard III). Esto muestra que personajes con muchos diálogos aparecen en una variedad de géneros.

Personajes con mayor cantidad de palabras



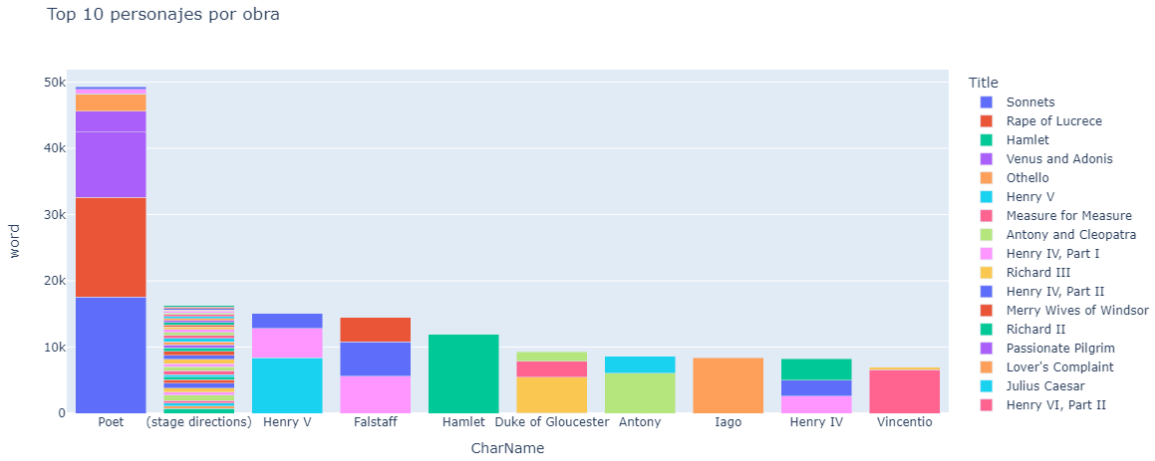


Al analizar la cantidad de palabras por personaje, se observa que los personajes con más palabras suelen aparecer en múltiples obras. Los personajes que más palabras mencionan son "Poet" y "Stage Directions", quienes no son personajes tradicionales de las historias, sino elementos de las anotaciones y apuntes del escritor. Estos elementos adicionales

reflejan un componente estructural significativo en las obras de Shakespeare, y tienen una alta contribución al conteo total de palabras.

En cuanto a la cantidad de palabras por obra, notamos que las obras con mayor número de palabras incluyen "Hamlet", "Richard III", y "Coriolanus". Estas obras presentan una alta densidad de diálogo y monólogos extensos, lo que coincide con la presencia de personajes prolíficos como "Henry V" y "Hamlet" que se destacan como personajes con muchas palabras. La relación entre el volumen de palabras de los personajes principales y la cantidad total de palabras en una obra indica que las obras más verbosas suelen tener personajes centrales muy desarrollados, con roles importantes en múltiples escenas.

Personajes con mayor conteo de palabras por obra



En esta gráfica se observa que Poet es el personaje que tiene más líneas dado que está presente en mayor cantidad de obras en comparación con los demás personajes. Lo mismo sucede con stage directions. Nuevamente, esto está dado por el rol crucial que ambos elementos tienen en las obras de Shakespeare.

Como se observó previamente, existe una relación clara entre la cantidad de palabras dichas por un personaje y su centralidad en la obra, esto resulta lógico si consideramos el dominio de análisis, dado que cuanta mayor relevancia tenga el personaje es más probable que tenga más diálogo.
Esto resulta evidente en casos como Henry V, Hamlet y Henry IV, que presentan obras tituladas con su nombre. A su vez, personajes como Henry V y Richard III tienen muchas palabras debido a sus roles como reyes y figuras centrales en obras históricas.

Generalmente, personajes con más palabras suelen tener un mayor impacto en la narrativa y en el desarrollo de la trama. Sus largos discursos y monólogos pueden ser vehículos para explorar temas profundos o conflictos internos, siendo probablemente uno de los más conocidos “be or not to be” de Hamlet.

Preguntas para futura investigación

- **¿Cómo afecta la vida del artista su capacidad de producción?** Hay años en los que Shakespeare escribió pocas o ninguna obra, como entre 1591-1592 y 1601-1602. Estos períodos de menor producción podrían haber sido influenciados por eventos personales, sociales o políticos, aunque esto requeriría más investigación específica sobre su vida y contexto histórico.
- **¿Cómo afecta la vida del artista los temas sobre los que escribe?** Entender más sobre su vida nos podría ayudar a comprender por qué desde 1693-1600 se concentró en la comedia, pero luego entre 1603-1608 se dedicó a escribir tragedias.
- **¿Cuál es la relación entre el largo de los parlamentos y el rol del personaje en la trama?** Analizar la longitud promedio de los parlamentos de cada personaje y clasificarlos según su rol en la trama (protagonista, antagonista, personaje secundario). Realizar pruebas estadísticas (ejemplo: ANOVA) para determinar si existen diferencias significativas en la longitud de los parlamentos entre diferentes roles o no.
- **¿Cómo varía el uso de ciertas palabras o temas a lo largo de las diferentes etapas de la obra (por actos o escenas)?** Dividir las obras en actos y escenas y realizar un análisis de frecuencia de palabras en cada sección. Visualizar estos cambios a lo largo del tiempo usando gráficos de líneas o heatmaps que muestren la evolución de la frecuencia de palabras clave.
- **¿Hay patrones de lenguaje específicos asociados a ciertos personajes (nobles vs. plebeyos) o situaciones (amor, guerra, traición)?** Se podría clasificar a los personajes y a las situaciones en categorías. En segundo lugar realizar un análisis de frecuencia de palabras y vectorizar los dialogos. Finalmente utilizar técnicas de clustering y análisis de componentes principales (PCA) para identificar patrones y diferencias en el uso del lenguaje. En el caso de PCA se podría reducir la dimensionalidad de los vectores generados y luego visualizar los textos en el espacio de los componentes principales para identificar agrupaciones o diferencias entre nobles, plebeyos, y distintas situaciones.
- **¿Relación entre la cantidad de palabras de una obra y la fama de esa obra?** Ya sea por espectadores o copias vendidas, resulta interesarte conocer si los personajes con más palabras mencionadas son también los más famosos.

Conclusiones

Respecto a los datos podemos mencionar que no se encontraron valores faltantes en la mayoría de las columnas, lo que indica una buena completitud de los datos. Además, no se detectaron registros duplicados, asegurando la unicidad de los datos.

A este buen punto de partida se le realizaron transformaciones con la finalidad de obtener resultados más precisos. Colocar todas las palabras en minúsculas, y transformar las abreviaciones en su versión completa nos permitió contabilizar correctamente la cantidad de apariciones. Como ejercicio para futuro, a estas transformaciones se le podría agregar un proceso de lematización, lo cual permitirá agrupar las palabras por su significado. Al realizarse este ejercicio, probablemente las palabras “sir”, “king” y “lord” sean agrupadas, demostrando que la presencia de un rey es muy frecuente en sus obras.

Analizando las obras de Shakespeare de forma cronológica, encontramos patrones en sus creaciones, resulta interesante a futuro conocer la vida del artista para poder estudiar si es posible establecer una relación con la cantidad de obras producidas, o las temáticas sobre las que escribió. Por ejemplo, con los géneros notamos que no conviven la tragedia y las obras históricas. También tuvo años de inactividad, esto puede deberse a una enfermedad.

Del análisis de palabras por personaje decidimos excluir a “poet” y “stage directions”, ya que entendemos que estos no son personajes reales, con esto en mente vemos que Henry V, Henry IV y Falstaff tienen una fuerte presencia en los personajes con más parlamentos, todos estos forman parte del “universo” de obras del rey Enrique. La mayoría de los personajes con más diálogos aparecen en más de una obra, esta lógica se rompe con Hamlet y Iago (antagonista de Otelo), esto nos indica que la obra gira alrededor de ellos.

Resultaría muy interesante poder establecer la relación entre la cantidad de palabras mencionadas por un personaje y su fama. Aunque Romeo y Julieta puede ser la obra más famosa de Shakespeare, sus protagonistas no forman parte de los personajes con más diálogos, pero sí personajes de Otelo, El Mercader de Venecia y Hamlet. Cruzar los datos actuales con los tickets de obras o libros vendidos nos permitiría saber qué es lo que al público más le gusta de la obra de Shakespeare.

Recursos

- https://pub.towardsai.net/advanced-eda-made-simple-using-pandas-profiling-35f83027061a
- https://plotly.com/python/
- Using pandas and Python to Explore Your Dataset – Real Python