Introducción a la ciencia de datos - Tarea 1

Grupo 20 Auriello - Molina
04/06/2024

Calidad de datos y transformaciones

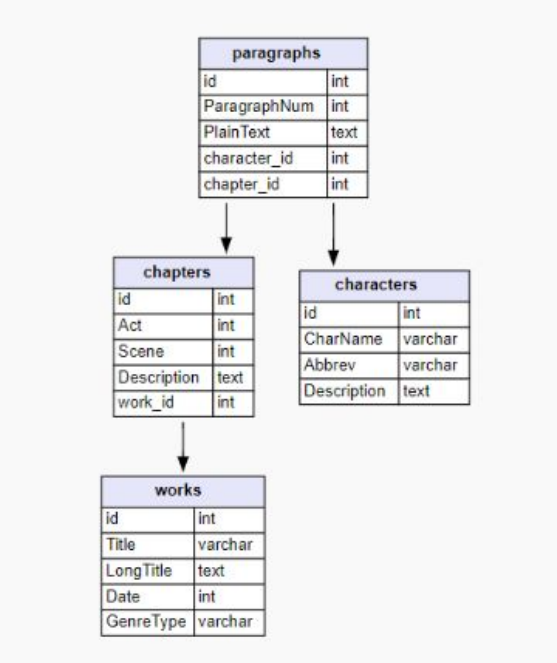
- Completitud de los datos: estructura, datatypes, nulos, duplicados.
- Coherencia y integridad referencial entre los datos: coherencia y relaciones entre tablas.

✓ **Works:** 43 títulos, datatypes apropiados, fechas en rango razonable.

✓ **Paragraphs:** estructura coherente y adecuada.

✓ **Chapters:** estructura coherente y clara.

⚠ **Characters:** valores faltantes en Abbrev(5) y Description(646). 1266 personajes unicos.



Calidad de datos y transformaciones

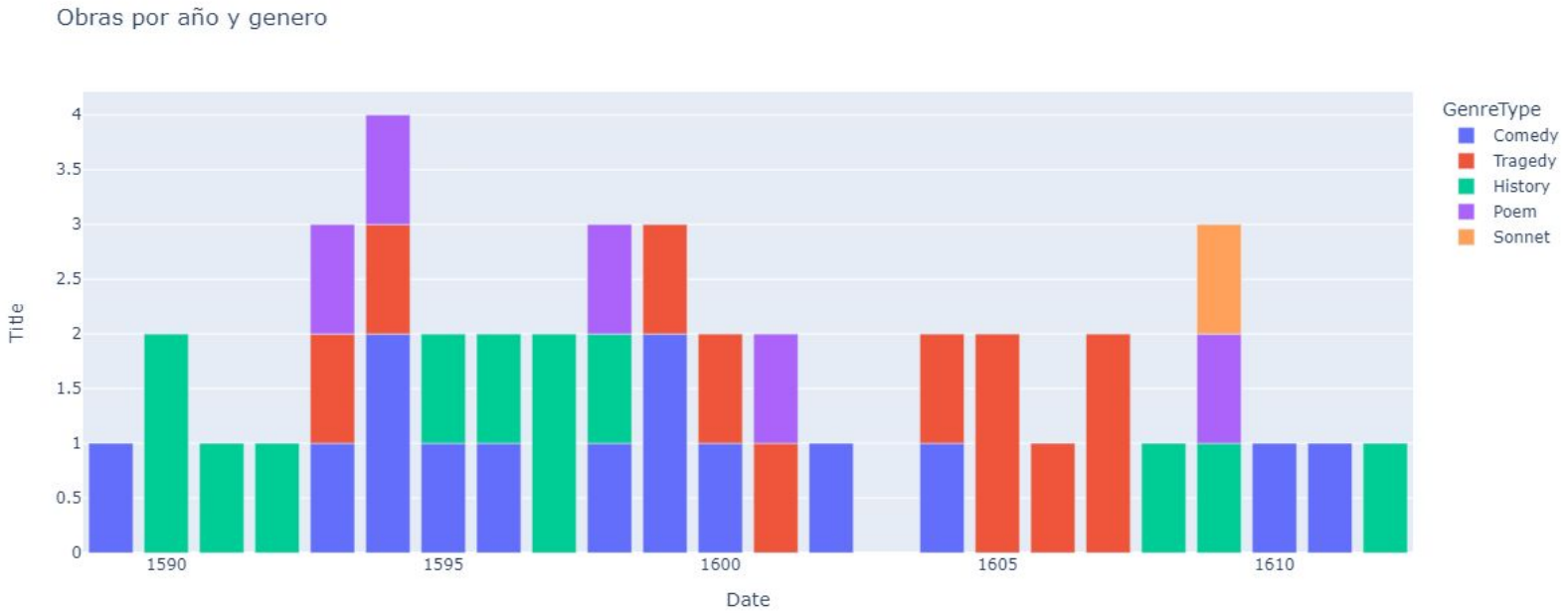
Normalización y transformaciones: ¿Por que es relevante para el análisis de texto?

- ➔ Reducción de redundancia y corrección de errores
- ➔ Interpretación de palabras como distintas, cuando tienen significado similar.

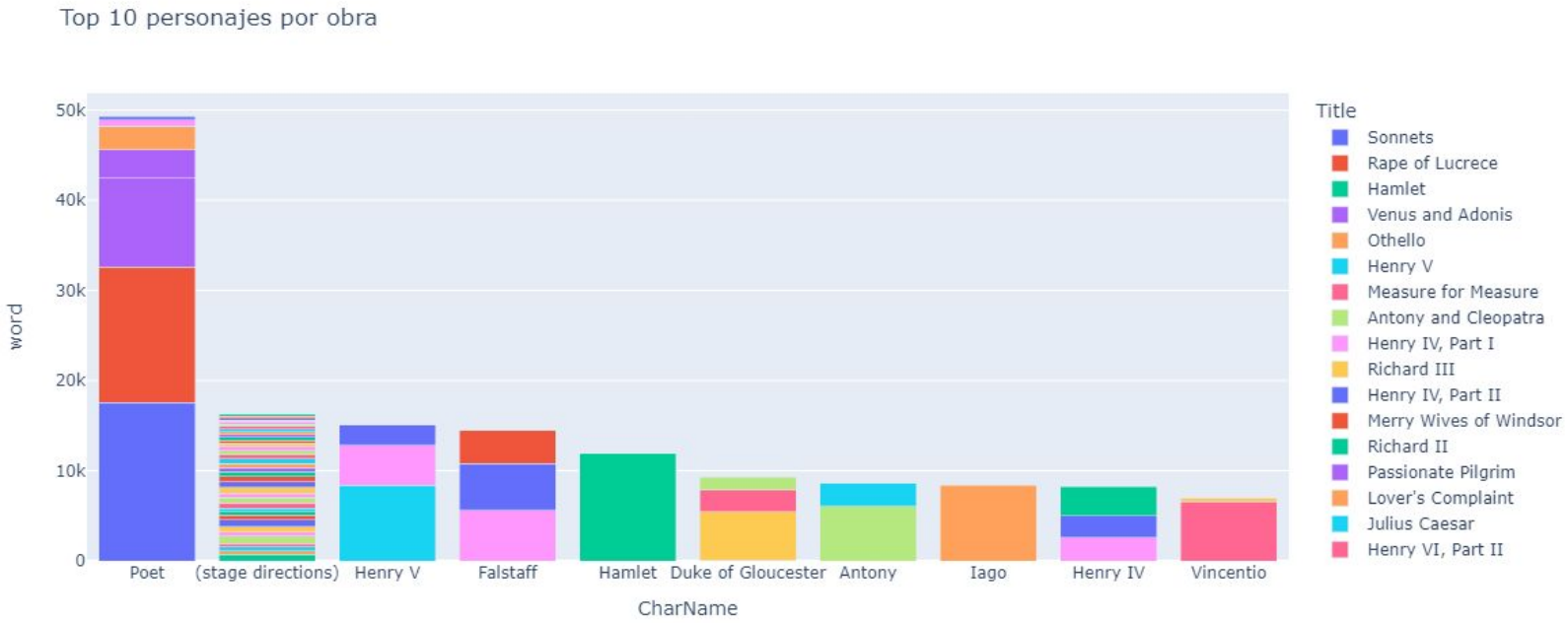
¿Cuál fue nuestro approach?

- Transformar todo a minúsculas
- Expansión de contracciones: diccionario de contracciones y expresiones regulares
- Eliminar signos de puntuación: reemplazo por espacios
- Previo análisis de palabras más frecuentes: eliminamos las stopwords
- Posibles tratamientos a futuro: lematización.

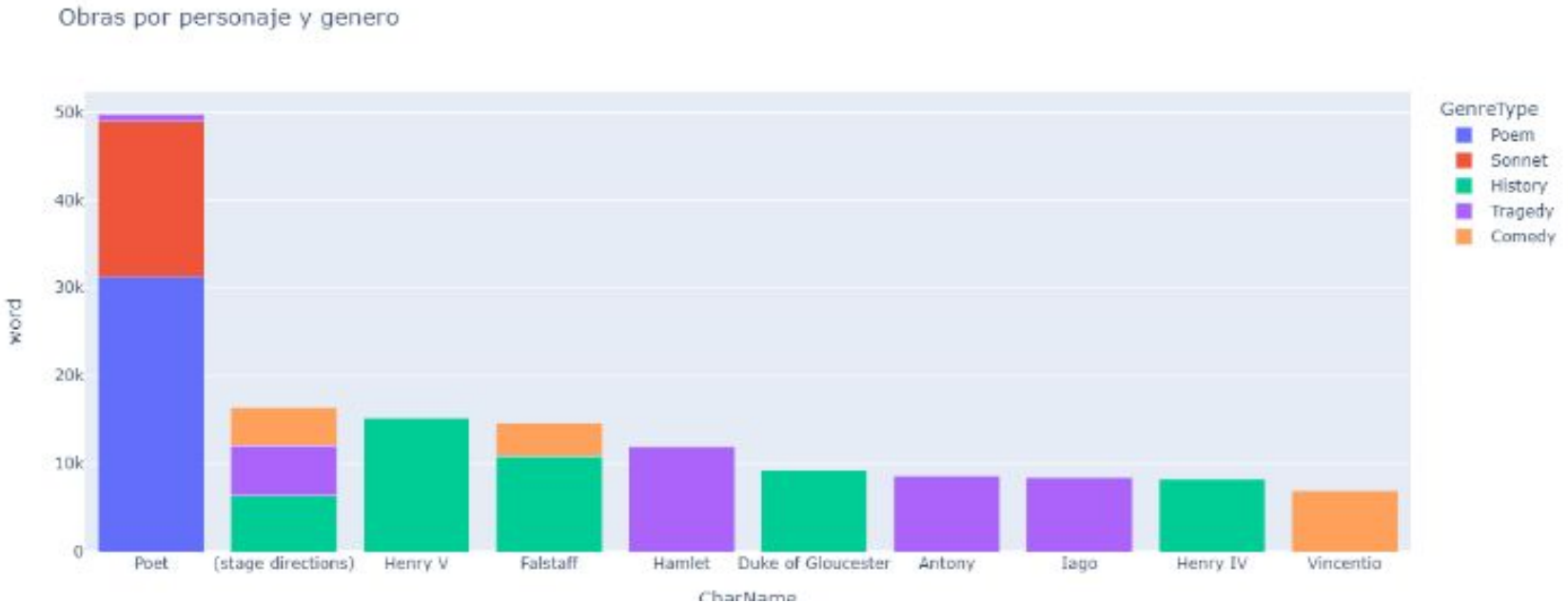
Obras por año y género



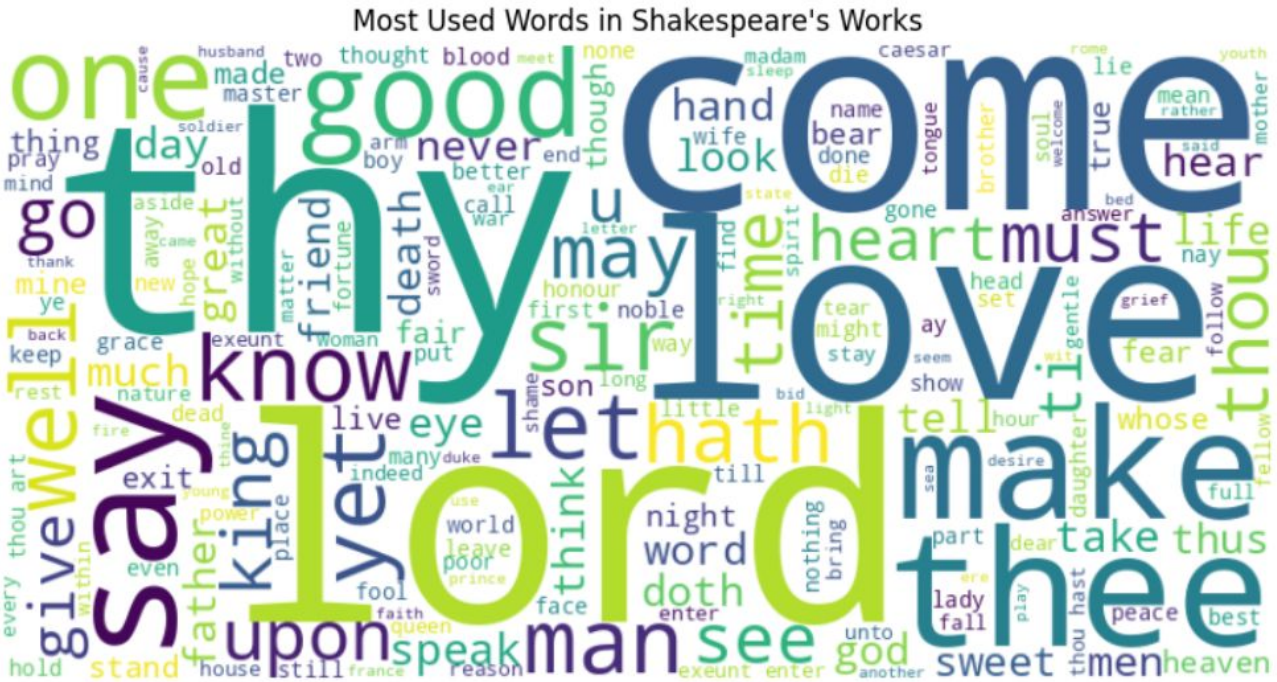
Cantidad de palabras por personajes



Cantidad de palabras por personaje y género



Palabras más frecuentes



Futuras preguntas o investigaciones

- Tratamiento adicional al texto: Lematización
- ¿Cómo afecta la vida del artista su capacidad de producción?
- ¿Cómo varía el uso de ciertas palabras o temas a lo largo de las diferentes etapas de la obra?
- ¿Cómo afecta la vida del artista los temas sobre los que escribe?
- ¿Cuál es la relación entre el largo de los parlamentos y el rol del personaje en la trama?
- ¿Hay patrones de lenguaje específicos asociados a ciertos personajes (nobles vs. plebeyos) o situaciones (amor, guerra, traición)?
- ¿Relación entre la cantidad de palabras de una obra y la fama de esa obra?