Informe Tarea 1

Análisis de las obras de William Shakespeare.

DESCRIPCIÓN BREVE

El presente informe está basado en el análisis del dataset público que contiene todas las obras de William Shakespeare. Pertenece al proyecto “Open Source Shakespeare (OSS)”, que tuvo como fin la creación de un sitio web, que contenga una base de datos de todos las Obras (Obras de Teatro y Poemas), que sea de acceso gratuito, con una interfaz amigable y con herramientas de búsquedas amplias y rápidas.

José Clavijo – Ernesto Bazzano
Curso: Introducción a la Ciencia de Datos – Facultad de Ingeniería – Udelar.

Contenido

Descripción Conceptual de la Base de Datos

Los datos de todas las Obras de Shakespeare están estructurados en cuatro conjuntos:

- **Obras (*Works*)**
- **Capítulos (*Chapters*)**
- **Párrafos (*Paragraphs*)**
- **Personajes (*Characters*)**

Para describir la BD rápidamente decimos que una Obra contiene muchos Capítulos (compuesto por un acto y muchas escenas), un Capítulo cualquiera contiene muchos Párrafos y un Párrafo está vinculado a un Personaje determinado.

Podemos decir también que un Capítulo tiene muchos párrafos y que un Personaje dice muchos párrafos a lo largo de los capítulos de las obras.

En realidad, un párrafo puede involucrar a varios personajes, pero para la creación de la base de datos vimos que se asocia un solo personaje a cada párrafo, por ejemplo, existen personajes con nombres cómo “ALL”, “BOTH”, indicando que más de un personaje en simultáneo pronuncian el mismo párrafo.

En la Figura 1, se puede observar cómo se vinculan los objetos de la base de datos.

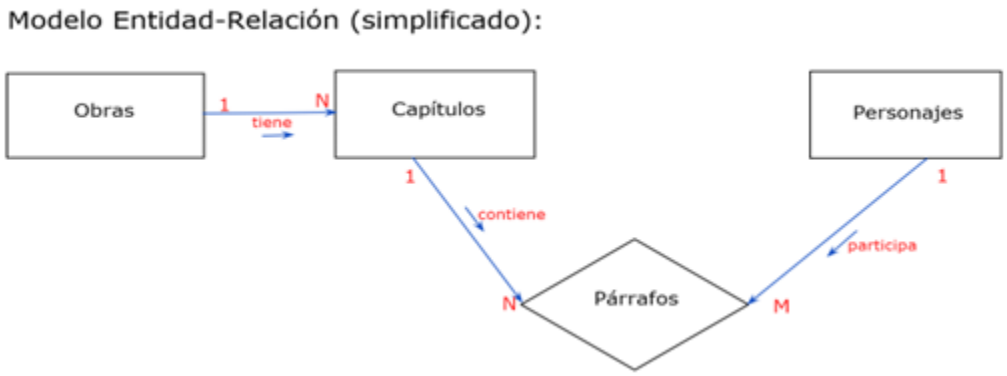


Figura 1: Modelo Entidad - Relación.

Del modelo Entidad-Relación, se puede ver la interrelación entre las tablas y cómo éstas se vinculan, por ejemplo, en la tabla Capítulos tenemos el atributo ***work_id*** como clave foránea para poder vincularlo con la Obra correspondiente y en la tabla de Párrafos tenemos las claves foráneas ***chapter_id*** y ***character_id*** para vincular con el Capítulo y con el Personaje, respectivamente.

De la Tabla 1 a la Tabla 4, se pueden observar los atributos asociados a las **Obras**, **Capítulos**, **Párrafos** y **Personajes**.

OBRAS		
Atributo	**Tipo de variable**	**Descripción**
Id	NUMERICO	Número entero, identificador único de la obra
Title	TEXTO	Título abreviado de la obra
LongTitle	TEXTO	Título de la obra
Date	FECHA	Año de elaboración de la obra
GenreType	TEXTO	Género de la obra.

Tabla 1: Atributos de la Obras

CAPITULOS		
Atributo	**Tipo de variable**	**Descripción**
Id	NUMERICO	Número entero, identificador único del capítulo
Act	NUMERICO	Número entero, asociado al acto del capítulo
Scene	NUMERICO	Número entero, asociado a la escena del capítulo
Description	TEXTO	Breve descripción de la escena.
Work_id	NUMERICO	Número entero, clave foránea para vincular con la tabla OBRAS.

Tabla 2: Atributos de los Capítulos

PARRAFOS		
Atributo	**Tipo de variable**	**Descripción**
Id	NUMERICO	Número entero, identificador único del párrafo
ParagraphNum	NUMERICO	Número entero, asociado al número de párrafo
PlainText	TEXTO	Texto completo del párrafo
character_id	NUMERICO	Número entero, clave foránea para vincular con la tabla PERSONAJES.
Chapter_id	NUMERICO	Número entero, clave foránea para vincular con la tabla CAPITULOS.

Tabla 3: Atributos de los Párrafos

PERSONAJES		
Atributo	**Tipo de variable**	**Descripción**
Id	NUMERICO	Número entero, identificador único del personaje
CharName	TEXTO	Nombre del personaje
Abbrev	TEXTO	Abreviación del nombre del personaje.
Description	TEXTO	Breve descripción del personaje.

Tabla 4: Atributos de los Personajes

Exploración y Calidad de Datos

Datos Faltantes y Palabras más Frecuentes

Para comenzar la exploración se realizó un análisis de datos faltantes en los cuatro conjuntos. Los datos son tratados mediante tablas de la librería Pandas, debido a esto se utilizó la librería ***Pandas-Profiling***[1] para hacer la exploración y el análisis.
En el atributo *Description* del conjunto de Personajes, fue el único lugar dónde se encontraron datos faltantes, en la Tabla 5 se puede ver este resultado. Este atributo no es importante para el análisis a realizar, por lo tanto, la falta de estos datos no es relevante.

Atributo	Datos Faltantes	Datos Presentes	% Datos Faltantes
Id	0	1266	0 %
CharName	0	1266	0 %
Abbrev	5	1261	0.4 %
Description	646	620	51 %

Tabla 5: Resultado Análisis datos Faltantes - Tabla Personajes.

Para el conteo de palabras se creó una nueva tabla en pandas, de nombre *words*, esta tabla también fue analizada con la librería *Pandas-Profiling*, de los resultados del análisis, se puede observar las 5 palabras más frecuentes que aparecen en todas las obras de Shakespeare, esto fue realizado con la información proveniente de la base de datos original, sin ningún tratamiento de datos previo. En las próximas secciones se observará con más detalle el resultado del conteo de palabras de las obras.

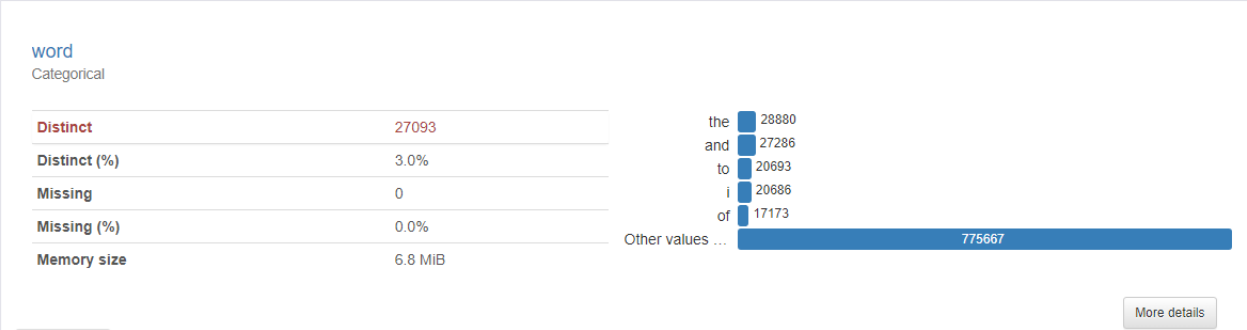


Figura 2: Palabras más frecuentes del atributo word del dataframe words.

[1] pandas-profiling · PyPI

Información Útil sobre la Construcción de la Base de Datos.

De la página web[2] del proyecto que originó la base de datos, se puede obtener información útil sobre la construcción de esta,

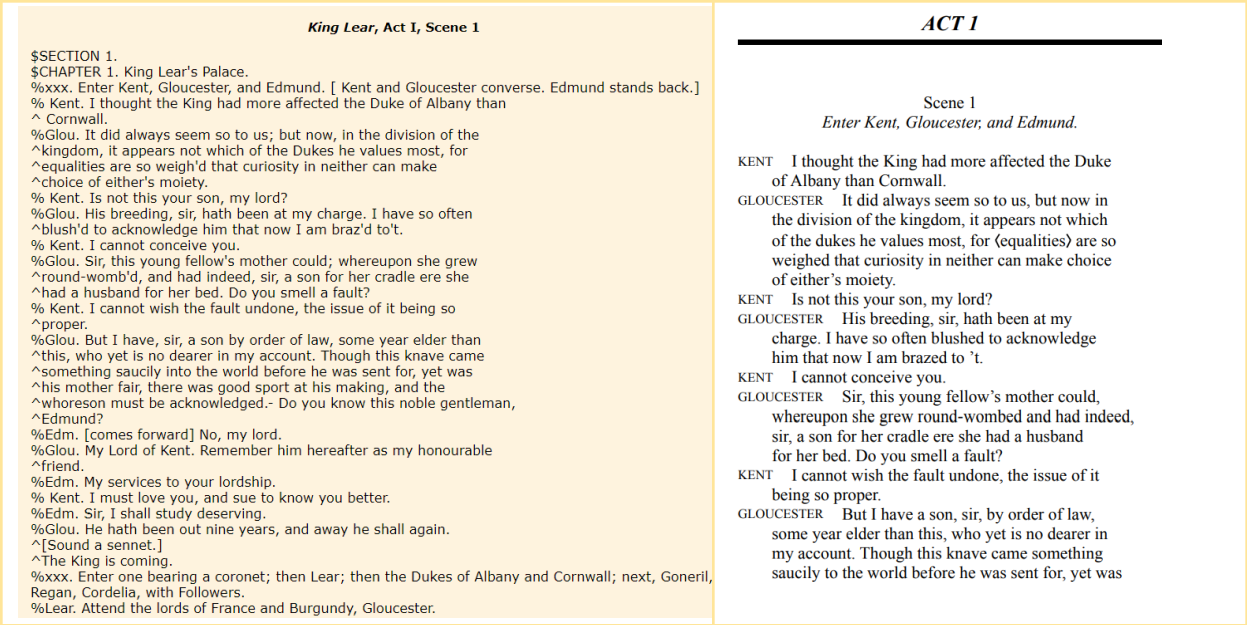
- El texto original utilizado fue preparado para ser interpretado por un analizador de lenguaje, con el objetivo de extraer atributos o información relevante del texto (actos, escenas, personajes, acotaciones de dirección, etc.). Para esto se identificó la información relevante con los siguientes caracteres
    - **$:** Acto Escena.
    - **% xxx:** Acotación para la dirección (del inglés ***stage directions***).
    - **% <Nombre_Personaje>:** inicio de la línea de texto pronunciado por el personaje.
    - **^:** líneas de texto.

  En la Figura 3 se puede observar el texto preparado por el diseñador de la base de datos para ser utilizado como entrada del analizador de lenguaje, además también se observa la redacción original del acto de la obra *King Lear.*
- Acorde a la estructuración de la base de datos, el diseñador tomó algunos criterios para resolver algunos problemas a los cuales se enfrentó:
    a. Las acotaciones para la dirección **(*stage directions*)** son líneas de texto dentro de las obras de Teatro, que no están asociadas a ningún personaje, este texto indica el lugar dónde la acción es tomada o cuando el actor debe ingresar y dejar el escenario, o cuando un musical vocal o instrumental debe sonar. Para incorporar estas líneas a la estructura de la base de datos, el diseñador les asignó, a todas estas líneas de texto, un personaje ficticio de nombre *<stage directions>.* De esta forma se puede diferenciar entre una línea de texto de un personaje y la una línea de texto que es una acotación para la dirección.
    b. Hay líneas de texto que son asignadas a más de un Personaje, que son identificadas como ***<Both>, <All>.*** Para seguir con la estructura original de la base de datos, se generaron nuevos personajes ficticios (al igual que *Stage Directions*), con los nombres <Both>, <All>, por ejemplo.

Las dos soluciones encontradas por el diseñador de la base de datos generan dos problemas de calidad de datos:

a. Hay un personaje ficticio, ***(stage directions)*****,** que no pertenece a la obra, que puede distorsionar el análisis. Para corregir este problema se deberán eliminar, de la Tabla de Datos de Párrafos, todas las líneas de texto que pertenecen al Personaje *(stage directions)***.**
b. La consecuencia de esta decisión tiene la misma consecuencia que el punto anterior. Pero en este caso no se puede eliminar la línea de texto, porque contiene información relevante para el análisis de la obra. Para el análisis de palabras, no genera ninguna consecuencia, pero para el análisis de Personajes si, en la medida que se pretenda tener un conteo exacto de la cantidad de párrafos asociados a cada Personaje. Para solucionar esto, se debería conocer, para cada instante de la obra, que personajes están en escena para asignarles este párrafo a cada uno de ellos.

[2] https://www.opensourceshakespeare.org/



King Lear, Act I, Scene 1

```
$SECTION 1.
$CHAPTER 1. King Lear's Palace.
%xxx. Enter Kent, Gloucester, and Edmund. [ Kent and Gloucester converse. Edmund stands back.]
% Kent. I thought the King had more affected the Duke of Albany than
^ Cornwall.
%Glou. It did always seem so to us; but now, in the division of the
^kingdom, it appears not which of the Dukes he values most, for
^equalities are so weigh'd that curiosity in neither can make
^choice of either's moiety.
% Kent. Is not this your son, my lord?
%Glou. His breeding, sir, hath been at my charge. I have so often
^blush'd to acknowledge him that now I am braz'd to't.
% Kent. I cannot conceive you.
%Glou. Sir, this young fellow's mother could; whereupon she grew
^round-womb'd, and had indeed, sir, a son for her cradle ere she
^had a husband for her bed. Do you smell a fault?
% Kent. I cannot wish the fault undone, the issue of it being so
^proper.
%Glou. But I have, sir, a son by order of law, some year elder than
^this, who yet is no dearer in my account. Though this knave came
^something saucily into the world before he was sent for, yet was
^his mother fair, there was good sport at his making, and the
^whoreson must be acknowledged.- Do you know this noble gentleman,
^Edmund?
%Edm. [comes forward] No, my lord.
%Glou. My Lord of Kent. Remember him hereafter as my honourable
^friend.
%Edm. My services to your lordship.
% Kent. I must love you, and sue to know you better.
%Edm. Sir, I shall study deserving.
%Glou. He hath been out nine years, and away he shall again.
^[Sound a sennet.]
^The King is coming.
%xxx. Enter one bearing a coronet; then Lear; then the Dukes of Albany and Cornwall; next, Goneril,
Regan, Cordelia, with Followers.
%Lear. Attend the lords of France and Burgundy, Gloucester.
```

ACT 1

Scene 1

Enter Kent, Gloucester, and Edmund.

KENT I thought the King had more affected the Duke of Albany than Cornwall.

GLOUCESTER It did always seem so to us, but now in the division of the kingdom, it appears not which of the dukes he values most, for ⟨equalities⟩ are so weighed that curiosity in neither can make choice of either's moiety.

KENT Is not this your son, my lord?

GLOUCESTER His breeding, sir, hath been at my charge. I have so often blushed to acknowledge him that now I am brazed to 't.

KENT I cannot conceive you.

GLOUCESTER Sir, this young fellow's mother could, whereupon she grew round-wombed and had indeed, sir, a son for her cradle ere she had a husband for her bed. Do you smell a fault?

KENT I cannot wish the fault undone, the issue of it being so proper.

GLOUCESTER But I have a son, sir, by order of law, some year elder than this, who yet is no dearer in my account. Though this knave came something saucily to the world before he was sent for, yet was

Figura 3: Párrafos del Acto 1 – Escena 1 de la obra King Lear. A la izquierda se observa el texto preparado para ser utilizado como entrada del analizador de lenguaje (se observan los distintos tipos de caracteres especiales utilizados para identificar información relevante de la obra: $, %xxx, %, ^}). A la derecha se observa el texto original de la obra.

- Posteriormente a estructurar la base de datos para las obras de teatro, el diseñador comenzó a trabajar en los Poemas y Sonetos, tuvo que adaptar su estructura más simplificada a la estructura utilizada para una obra de teatro, la cual está compuesta por Actos-Escenas y líneas de texto (asociados a un personaje de la obra):
    a. Los distintos poemas tienen un nombre, el cual es usado como nombre de la obra. El poema no está estructurado con Actos/Escenas, es una única pieza de varios párrafos. Por lo tanto, el diseñador optó por asignar un único acto y escena a cada poema, cada párrafo como una línea de texto y a cada una de estas líneas le asignó como personaje a ***Poet***.
    b. Los sonetos fueron tratados como una sola obra, de Nombre *Sonnet*s con un único Acto y 154 escenas (una por soneto), el nombre del personaje de cada línea también es ***Poet***.

En este caso, también al introducir el Personaje Ficticio *Poet* puede inducir a errores en el análisis global de las obras de Shakespeare, como se observará en la sección siguiente.

Personajes Ficticios

En la sección anterior se detalló algunas soluciones que encontró el diseñador de la base de datos para solucionar algunos problemas a los cuales se enfrentó, estos están relacionados a la incorporación de Personajes Ficticios a la base de datos.

Para ver el impacto de esta decisión, se puede observar en la Figura 4 que el personaje con más párrafos asociados es el personaje ficticio ***(stage directions)*** y el segundo ***Poet***. Si eliminamos todos los párrafos asociados a ***(stage directions)*** y si solo consideramos las obras de Teatro (no consideramos los Poemas y los Sonetos), el resultado cambia como se observa en la Figura 5.

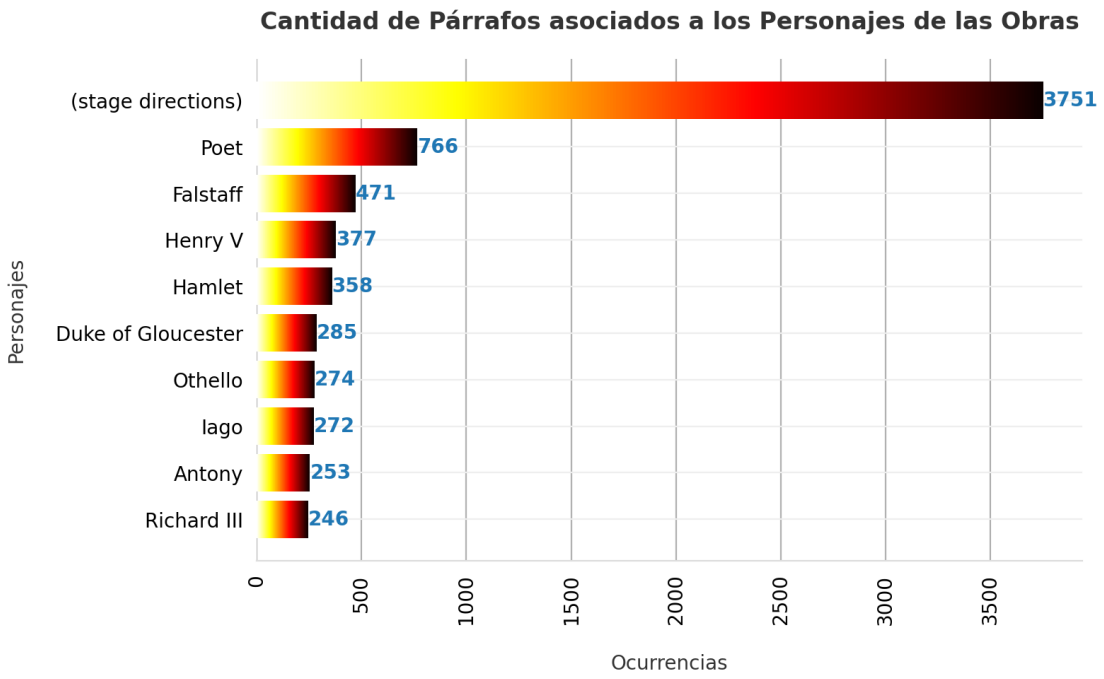


Figura 4: Histograma del conteo de Párrafos asociados a cada Personaje de las obras de Shakespeare.

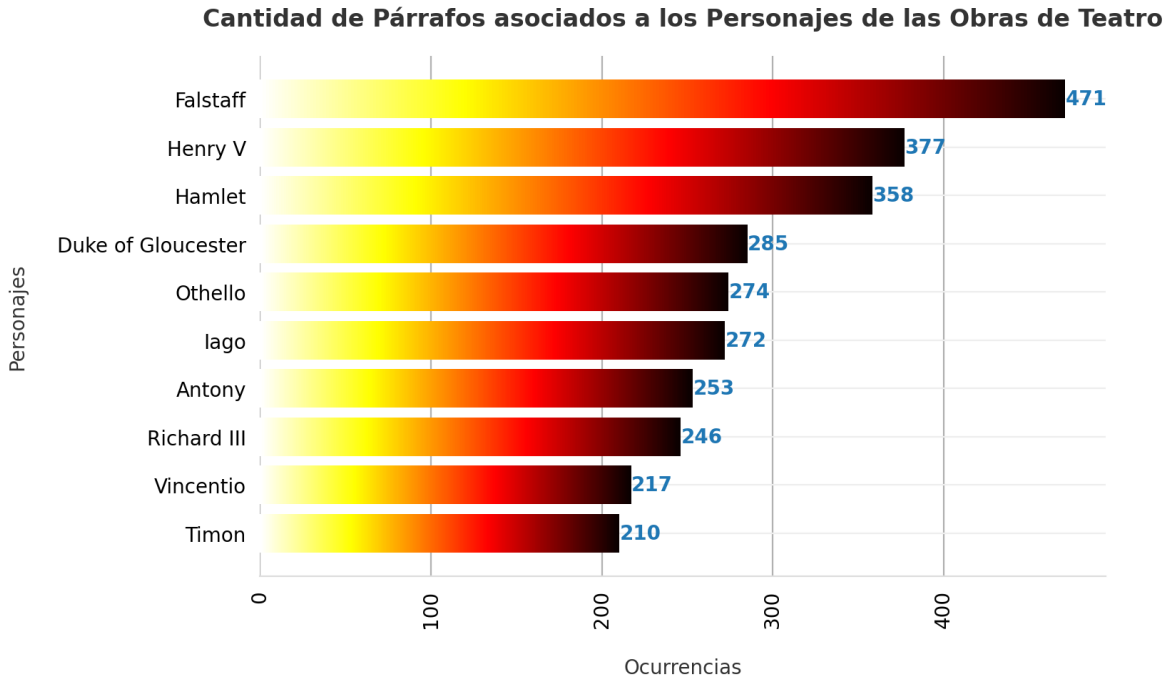


Figura 5: Histograma del conteo de Párrafos asociados a cada Personaje de las obras de Teatro de Shakespeare

Limpieza y acondicionamiento de Palabras: signos de puntuación y caracteres especiales.

Los párrafos contienen una cantidad importante de signos de puntuación y caracteres especiales, que deben ser removidos, para poder realizar una correcta separación de las palabras utilizadas en la obra. En el total de las obras, se pueden encontrar los siguientes caracteres y su ocurrencia asociada:

{ **.** : 24854; **,** : 23353 ; **\n** : 17905 ; **'** : 12719 ; **;** : 9984 ; **:** : 8183 ; **?** : 7940 ; **!** : 6361 ; **-** : 4622 ; **[** : 1954 ; **]** : 1944 ; **)** : 46 ; **(** : 47 ; **&** : 21 }

El análisis se hizo extensivo a los siguientes caracteres, los cuales no fueron encontrados en ningún párrafo de la obra:

{ *****, **+**, **/**, **<**, **=**, **>**, **#**, **$**, **%**, **@**, **^**, **{**, **|**, **}**, **~** }

Los símbolos de la primera lista fueron removidos de los párrafos directamente, excepto los caracteres { **[** , **]** , **'** }, que también fueron removidos, pero previo análisis del contexto dónde estaban ubicados(en las próximas dos secciones se explica esto).

Posteriormente, se generó un listado de palabras para analizar la ocurrencia de estas dentro de las distintas obras. Anteriormente, para un correcto conteo de las palabras, se uniformizaron todas las letras encontradas a minúsculas, de esta forma, por ejemplo: *This* - *this*, se contabilizarán como la misma palabra.

Limpieza y acondicionamiento de Palabras: eliminar paréntesis rectos de acotaciones para la dirección

Las acotaciones para la dirección(***stage directions***), no solo aparecen en líneas de texto exclusivas para esta funcionalidad las cuales el diseñador decidió incorporarlas como párrafos del Personaje Ficticio ***(stage directions)***, sino que también pueden aparecer en párrafos que están asociados a un personaje. Hay un 6% aproximadamente de párrafos que contienen este tipo de texto y se observan únicamente en las obras de Teatro(Géneros Comedia, Tragedia e Historia)

Estas acotaciones para la dirección pueden ser identificadas porque normalmente aparecen entre paréntesis rectos, a continuación un ejemplo de la obra *Twelfth Night*, Acto 1, Escena 1, Párrafo: 22-29 y Personaje Orsino,

Why, so I do, the noblest that I have:
O, when mine eyes did see Olivia first,
Methought she purged the air of pestilence!
That instant was I turn'd into a hart;
And my desires, like fell and cruel hounds,
E'er since pursue me.
[Enter VALENTINE]
How now! what news from her?

Para poder eliminar la frase ***[Enter VALENTINE]***, es necesario identificar los dos paréntesis rectos para eliminar el texto interior, con el objetivo de obtener un párrafo filtrado igual a,

Why, so I do, the noblest that I have:
O, when mine eyes did see Olivia first,
Methought she purged the air of pestilence!
That instant was I turn'd into a hart;
And my desires, like fell and cruel hounds,
E'er since pursue me.

How now! what news from her?

Para esto se utilizarán **expresiones regulares**[3], que son patrones utilizados para encontrar una determinada combinación de caracteres dentro de una cadena de texto. La estructura de la expresión regular para retirar la cadena de texto contenida dentro de los paréntesis rectos es la siguiente:

```
r"\[.+?]"
```

A continuación una explicación de esta estructura:

\[: indica buscar el [(corchete apertura) , se le coloca \ delante para que no lo interprete ya que el "[", tiene un significado especial.
. : Cualquier carácter.
+ : Uno o más caracteres iguales al anterior.
? : El menor número de veces. O sea, la cadena más corta que comience con [y termine con].

Por ejemplo, si la entrada es **"Este es un ejemplo [quiero quitar esto] y me quedo con [esto también se quita] el texto filtrado."** , la salida obtenida será **"Este es un ejemplo y me quedo con el texto filtrado."**

Nota : Ver apéndice Introducción a las Expresiones Regulares para más información.

En un gran porcentaje de párrafos, dónde aparecen acotaciones para la dirección, no se respeta la regla vista anteriormente, a continuación, se observa un ejemplo sobre esto:

I would have said it; you say well. Here comes the king.
[Enter KING, HELENA, and Attendants. LAFEU and]
PAROLLES retire]

En este ejemplo, hay tres paréntesis rectos, y las dos últimas líneas son líneas de texto de ***stage directions***, para poder eliminar este caso, se usó el siguiente procedimiento:

1) el texto entre paréntesis "***[Enter***..... ***and]***", es identificado por la expresión regular **r"\[.+?]"** y es cambiado por un asterisco *****, por lo tanto el texto intermedio queda así:

I would have said it; you say well. Here comes the king.
******

[3] https://docs.python.org/es/3/library/re.html

PAROLLES retire]

2) Se usa una segunda expresión regular **r"*.+?]** para identificar el texto entre el asterisco y el paréntesis recto: **"* PAROLLES retire]"** y sustituirlo por un espacio vacío **" "**, obteniendo el siguiente resultado:

I would have said it; you say well. Here comes the king.

Con estos dos pasos se eliminan este tipo de líneas de texto asociadas a ***stage directions.***

Con estas dos expresiones regulares se eliminan las acotaciones de dirección en el 99,34% de los párrafos dónde aparecen (en total aparecen en 1955 párrafos de 31714 párrafos analizados). El otro 0,66% de los párrafos que no se remueven todos los paréntesis rectos (13 párrafos de todas las obras), es debido a que solo aparece un paréntesis recto, por lo tanto, no se puede discriminar que parte del texto es la acotación para la dirección y que parte es línea de texto del Personaje, a continuación dos ejemplos:

*1) **[standing** forth what says my general*

2) You must be purged too, your sins are rack'd,
You are attaint with faults and perjury:
Therefore if you my favour mean to get,
A twelvemonth shall you spend, and never rest,
*But seek the weary beds of people **sick]***

Es tan baja la cantidad de palabras afectadas por estos últimos casos, que en primera instancia no se tomará ninguna acción para corregirlo. Se eliminará el paréntesis recto que está en el párrafo y se asumirá que todas las palabras son válidas.

Limpieza y acondicionamiento de Palabras: conversión de apóstrofes

Uno de los caracteres especiales que deben ser retirados del texto son los apóstrofes, para obtener un listado de palabras válido. Los apóstrofes son usados para diferentes propósitos, por ejemplo:

1) Para resaltar alguna palabra:
   - *You mistake, knight; '**accost'** is front her, board her, woo her, assail her.*

2) Para resaltar una frase:
   - *give me **'youth whatsoever thou art thou art but a scurvy fellow '***

3) para identificar una contracción corta:
   - ***What's** that? □ **What is** that?*
   - ***'Tis** now struck twelve. Get thee to bed, Francisco. □ **It is** now struck twelve. Get thee to bed, Francisco.*

4) para identificar una posesión:

- *My **niece's** chambermaid ??*

Los casos 1 y 2 de uso son fácilmente identificable y separables, para el caso 3 se usó la librería ***contractions***[4] para transformar las contracciones cortas en dos palabras separadas. Para el caso 4 directamente se eliminó el apóstrofe para desacoplar la palabra principal de la letra s auxiliar, esta última es eliminada posteriormente del listado de palabras.

Se analizó el listado de palabras (866954) de todas las obras, a continuación, un resumen de los resultados del filtrado:

- Palabras con apóstrofes: 27491 (3.17 % del total de palabras que contiene las obras)
  - Palabras entre apóstrofes: 7456 (27 %) - CORREGIDO
  - Palabras cortas con descontracción Satisfactoria: 5560 (20%) - CORREGIDO
  - Palabras cortas con descontracción Fallidas: 14475 (53%) – NO CORREGIDO

De los resultados anteriores, con la librería usada no fue posible transformar todas las contracciones cortas encontradas, ya que el conjunto de contracciones que contiene la librería utilizada es acotado(318) y existen muchas contracciones arcaicas de la época, que actualmente están en desuso, y no están contempladas en la librería utilizada. Como solución se podrían ingresar a la lista de contracciones todas las palabras que faltaron transformar, no se hizo para este informe.

[4] https://pypi.org/project/contractions/

Conteo de Palabras y Visualizaciones

Palabras más frecuentes en las Obras

En la Figura 6, se puede observar las 10 palabras más frecuentes encontradas en todas las obras de Shakespeare, y de la Figura 7 a la Figura 11, se puede observar lo mismo, pero por Género. Si hacemos la unión de las 10 palabras más frecuentes de los 5 Géneros analizados, tendremos como resultado las siguientes 16 palabras más frecuentes:

{i, you, my, his, her, that, thy, thou, the, a, in, of, to, with, and, is}

Del punto de vista gramatical se pueden agrupar estas palabras más frecuentes en las siguientes categorías o clases:

- **Pronombres personales**: {i, you} ; **Pronombres posesivos**: {my, his, her}
- **Pronombres demostrativos**: that ; **Pronombres arcaicos**: {thy, thou}
- **Artículos**: {the, a} ; **Conjunciones** : and
- **Verbo Conjugado**: is , **Preposiciones** {in, of, to, with} ;

Con esta visualización no es posible encontrar diferencias entre géneros, debido a la similitud de las palabras más frecuentes encontradas en cada una de estas. Este tipo de palabras pertenecen al conjunto de palabras comúnmente utilizadas en el idioma inglés y en el procesamiento de lenguaje natural son conocidas como ***Stop Words*[5].** Generalmente, en tareas de procesamiento de texto, es recomendable eliminar este tipo de palabras del texto porque no aportan mucha información, contiene información de bajo nivel que debe ser removida para hacer más foco en la información más relevante o característica del texto.

Dada la observación anterior, para poder encontrar diferencias entre géneros, se deberían eliminar todas las palabras clasificadas como *Stop Words* y volver a visualizar las palabras más frecuentes de cada Género.

Nota : Ver apéndice Análisis de las Palabras más frecuentes en la Obra eliminando *Stop Words*, para una mayor ampliación de los resultados obtenidos después de ejecutar lo referido en el párrafo anterior.

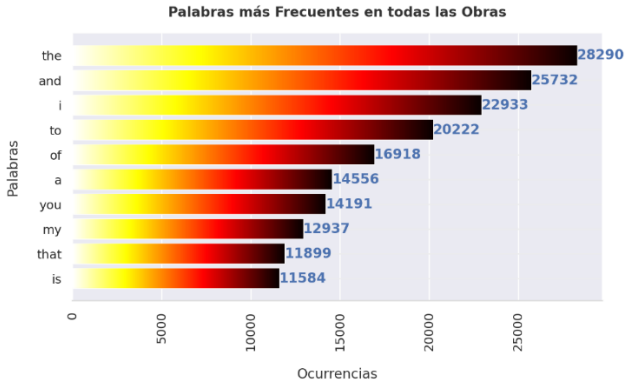


Figura 6: Palabras más Frecuentes en toda la Obra de Shakespeare.

[5] https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a

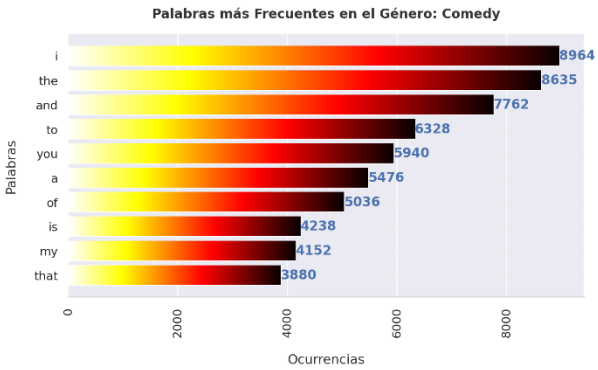


Figura 7: Palabras más frecuentes Género Comedia.

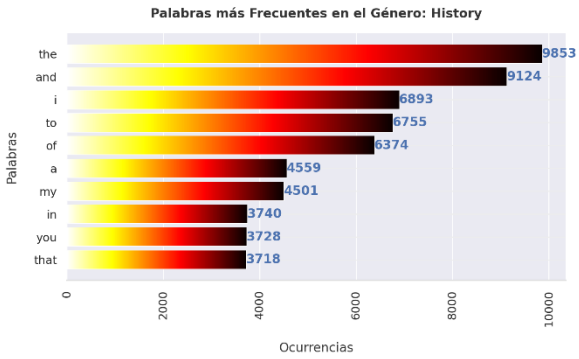


Figura 8: Palabras más frecuentes Género Historia.

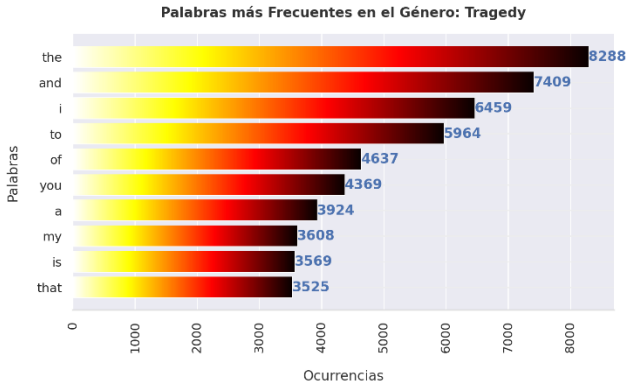


Figura 9: Palabras más frecuentes Género Tragedia

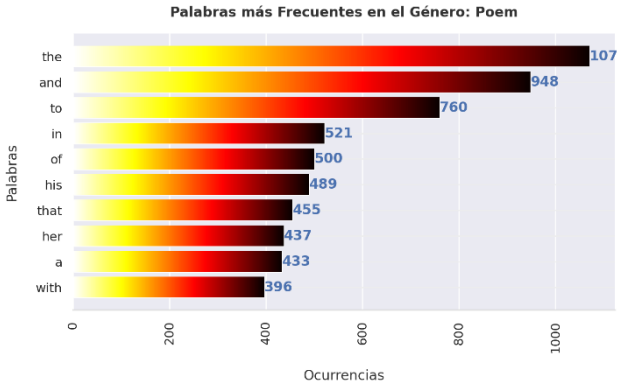


Figura 10: Palabras más frecuentes Género Poema.

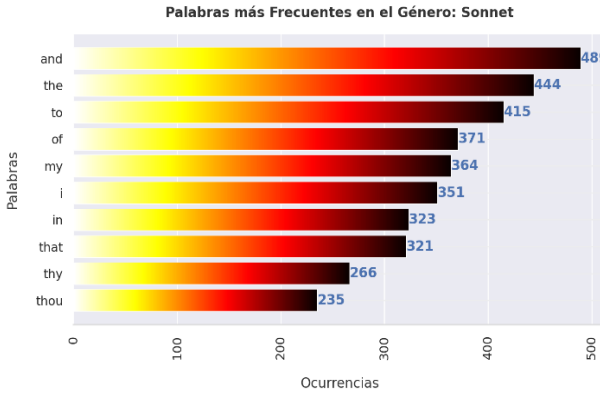


Figura 11: Palabras más frecuentes Género Soneto.

Personajes con más Palabras en las Obras

En la Figura 12, se pueden observar los 10 personajes con más Palabras agrupando todas las obras. Claramente, el personaje ***Poet*** es el que tiene mayor cantidad de palabras asociadas, este personaje ficticio asignado como personaje de todas las líneas de texto de los Poemas y Sonetos (hay que aclarar que existe un personaje *Poet* para las Obras de teatro, pero que tiene baja participación), es la causa de esta ocurrencia tan alta.

En este análisis ya no aparece el personaje ***(stage directions)***, el cual también tiene una cantidad alta de palabras asociadas, ya que los párrafos asociados a este personaje ficticio fueron eliminados anteriormente, cuando se analizó como se distribuían los párrafos entre los personajes.

En la Figura 13, se puede observar la cantidad de Palabras por personajes de todas las Obras de Teatro, no se consideran Poemas y Sonetos, de esta forma el personaje *Poet* desaparece de la lista de los 10 personajes con mayor cantidad de palabras de todas las obras.

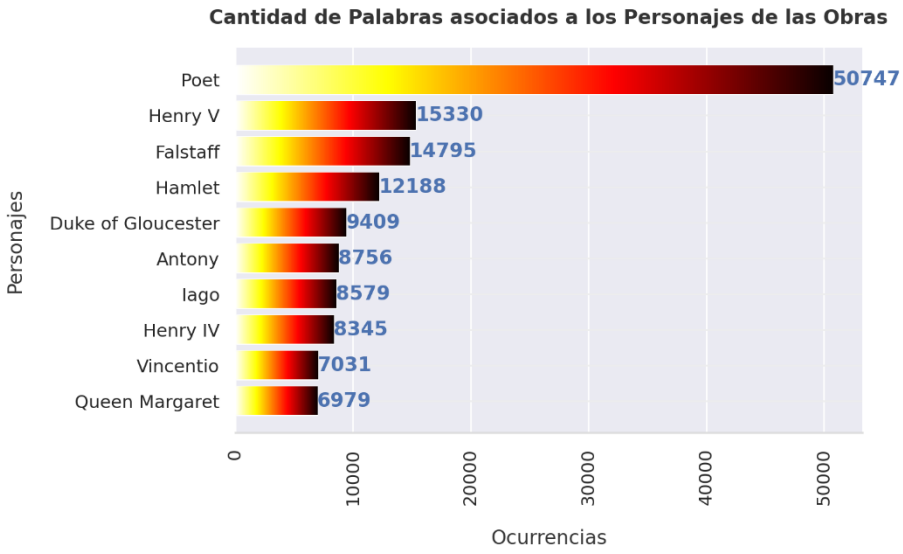


Figura 12: Cantidad de Palabras por Personajes en todas las Obras

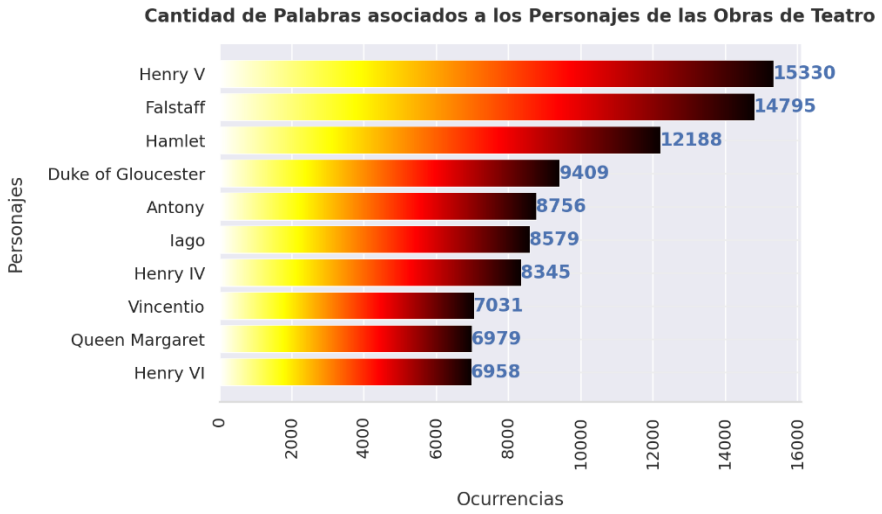


Figura 13: Cantidad de Palabras por Personajes en todas las Obras de Teatro.

Evolución de las Obras a lo largo de los años

Las obras de Shakespeare fueron escritas entre los años {1589-1612}, totalizando 24 años de publicaciones en forma casi ininterrumpida, ya que en el año 1603 no se registran obras. El total de publicaciones en este período de tiempo fueron 43.

En la Figura 14, se puede observar la evolución de la cantidad de obras publicadas por Shakespeare anualmente, se puede observar tres períodos: desde 1589 a 1592 con una tasa de publicación baja con respecto a la media (1.8 publicaciones / año), un segundo período desde 1593 a 1601 con una tasa de publicación mayor a la media y un tercer período desde 1602 a 1612 con una tasa de publicación menor al período anterior, pero mayor al primer período observado. Estos tres períodos pueden observarse más claramente en la Figura 15, dónde se totalizan las obras publicadas cada 4 años.

Para poder visualizar la evolución con el tiempo de los géneros publicados, se realizó una estimación de la densidad de probabilidad de la cantidad de publicaciones por género, utilizando la librería ***Seaborn***[6], con un agrupamiento de 4 años. En la Figura 16, se puede observar esta estimación por género, dónde se observa que las obras de teatro de género Comedia e Historia, al igual que los Poemas, son más frecuentes al inicio del período de publicación y que las obras de teatro de género Tragedia son más frecuentes al final del período. Los Sonetos no tiene una curva de densidad de probabilidad ya que están concentrados en un único año(1609).

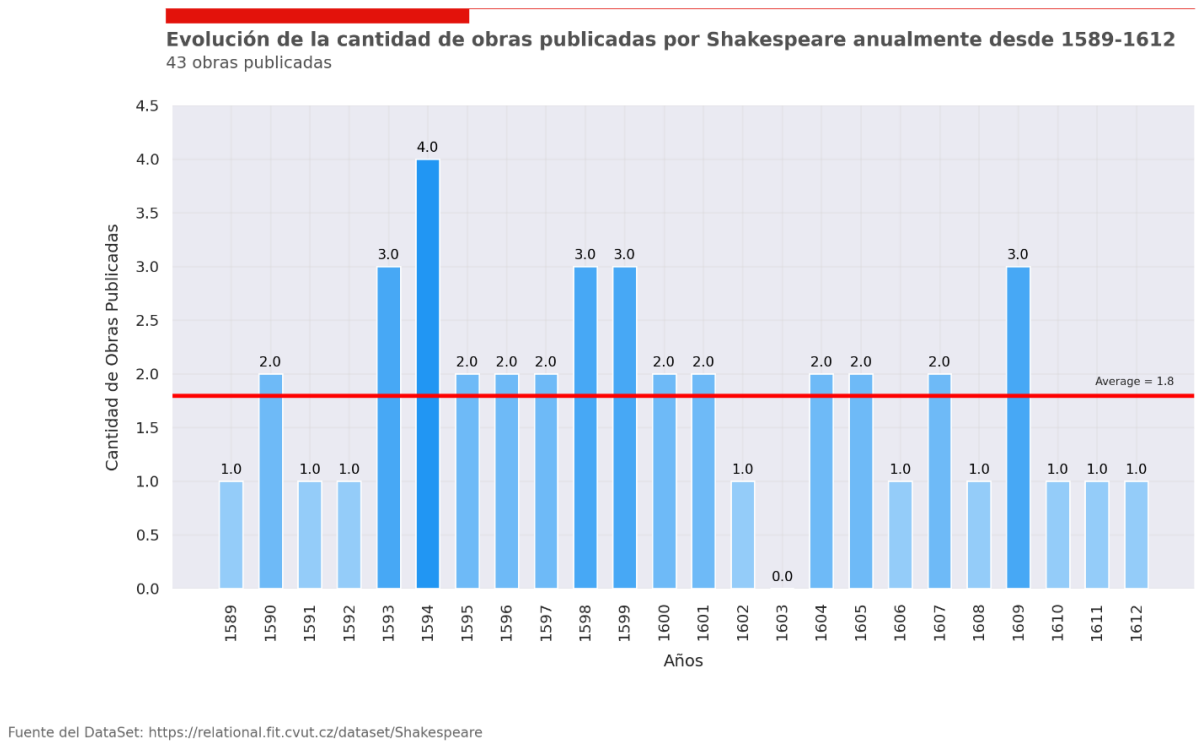
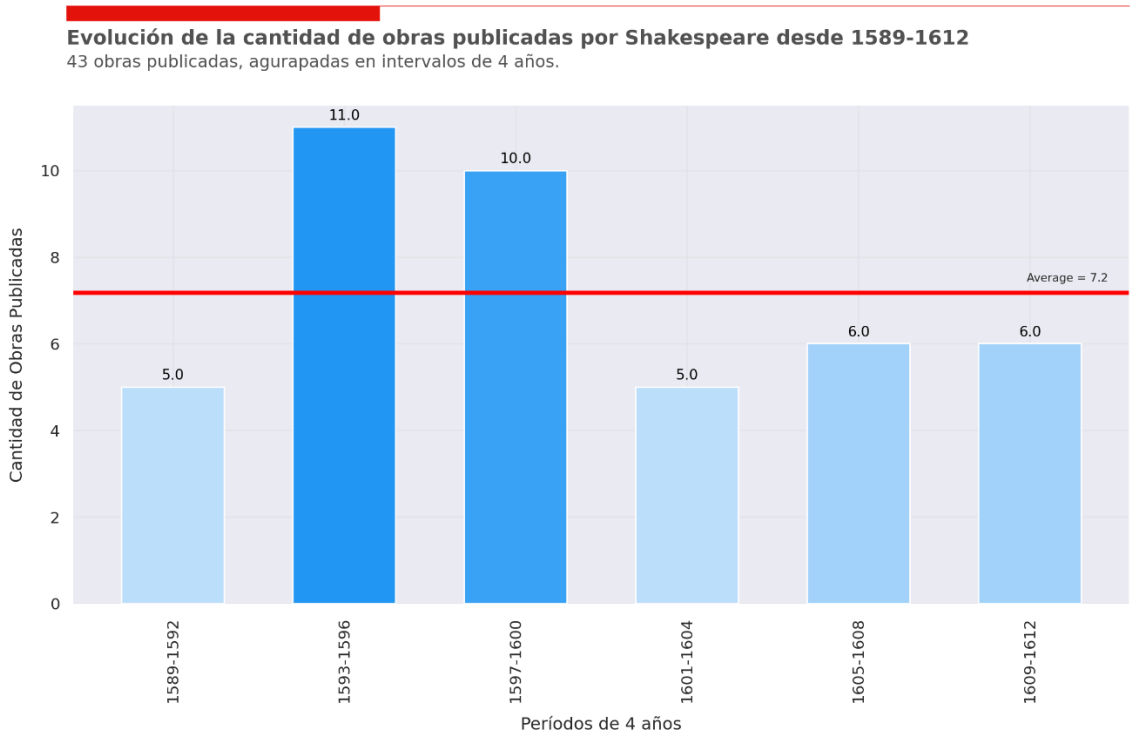


Figura 14: Evolución de la cantidad de obras publicadas por Shakespeare anualmente desde 1589 a 1612.

[6] seaborn.kdeplot — seaborn 0.12.2 documentation (pydata.org)



Fuente del DataSet: https://relational.fit.cvut.cz/dataset/Shakespeare

Figura 15: Evolución de la cantidad de obras publicadas por Shakespeare, agrupadas cada 4 años, desde 1589 a 1612.

Figura 16: Estimación de la Densidad de Probabilidad de la cantidad de Publicaciones por Género, agrupación en períodos de 4 años.

Posibles Preguntas por Responder desde los Datos

Q1 : Dado un párrafo de alguna obra de Shakespeare, a qué personaje está asociado?, a qué obra pertenece, a qué género pertenece?
S1 : Utilizando las palabras asociadas a cada género, a cada obra y a cada personaje, entrenar un algoritmo de clasificación con aprendizaje supervisado para que aprenda cómo diferenciar géneros, obras y personajes, dado un párrafo cualquiera de una obra.

Q2 : Dado un párrafo incompleto de m palabras, cuáles son las próximas n palabras?
S2 : Agrupar el texto de las obras en pares de (m,n) palabras consecutivas, entrenar una red neuronal para que aprenda la tarea.

Q3 : Dado una conversación incompleta entre dos personajes, es posible redactar párrafos de texto libre con el mismo lenguaje utilizado por Shakespeare en esa obra?
S3 : Entrenar una red neuronal que aprenda sobre las obras y el lenguaje particular usado para generar texto libre.

Q4 : Dada una obra de teatro(no necesariamente escrita por Shakespeare, pero con similitudes, por ejemplo escritas en la misma época con un lenguaje similar), es posible determinar si fue o no fue escrita por Shakespeare?
S4 : Entrenar un algoritmo de clasificación con aprendizaje supervisado donde aprenda las características de las obras de Shakespeare y del resto de los autores, para cumplir con la tarea.

Apéndice

Análisis de las Palabras más frecuentes en la Obra eliminando *Stop Words*

Se observó que las palabras más frecuentes de todas las obras pertenecen a un conjunto conocido como *Stop Words* (conjunto de palabras comúnmente utilizadas en un lenguaje en particular), a continuación se presentan resultados visuales y numéricos extraídos del texto, posterior a eliminar este conjunto de palabras. La tarea fue ejecutada utilizando la librería *sklearn*[7] que es pública y genérica.

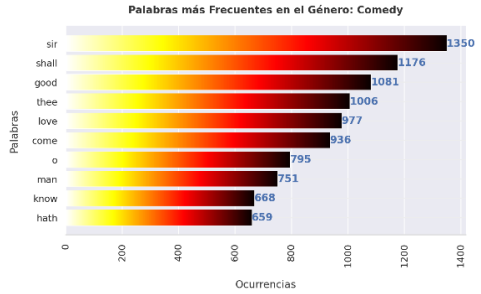


Figura 19: Palabras más frecuentes Género Comedia.

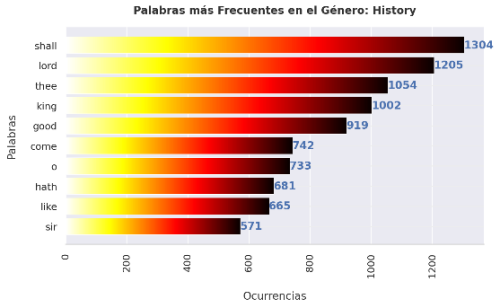


Figura 20: Palabras más frecuentes Género Historia.

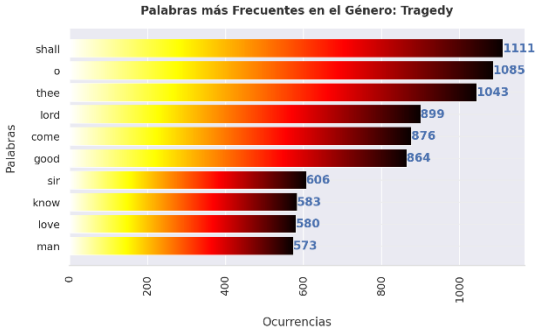


Figura 21: Palabras más frecuentes Género Tragedia.

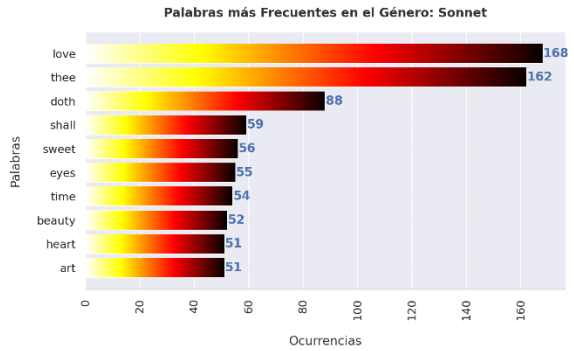


Figura 22: Palabras más frecuentes Género Soneto.

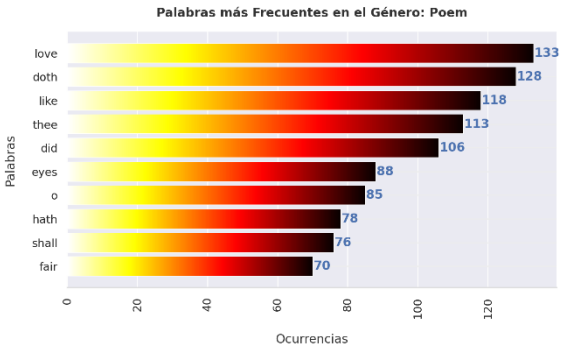


Figura 23: Palabras más frecuentes Género Poema.

[7] https://scikit-learn.org/stable/modules/feature_extraction.html#using-stop-words

Nuevos subconjuntos conteniendo las 10 palabras más frecuentes del género, se puede observar desde la Figura 17 a la Figura 21. La unión de estos subconjuntos resulta en un nuevo grupo de palabras más frecuentes:

1) {sir, shall, good, **thee**, love, come, **o**, man, know, **hath,** lord, king, like}
2) {**doth**, sweet, eyes, time, beauty, heart, art, did, fair}

El primer conjunto son palabras encontradas principalmente en las obras de teatro y el segundo conjunto en sonetos y poemas. Aunque varias palabras se repiten entre géneros, con este primer filtro ya se empiezan a observar palabras frecuentes que son distintas entre las obras de teatro y sonetos/poemas, fortaleciendo la idea de seguir por este camino: reduciendo la cantidad de palabras a analizar de los datos(filtrar para hacer foco en la información relevante y reducir la dimensionalidad del problema) y posteriormente encontrar características únicas de los géneros, para poder diferenciar los mismos(agrupamiento, clasificación).

También se siguen observando palabras del inglés antiguo o arcaico propias de la época(también conocido como Lenguaje *Elizabethan*[8]), como también expresiones particulares utilizadas por el autor para indicar una directiva relacionada a la obra (uso de la letra **O**), quizás estas palabras, en el contexto de las obras, deberían pertenecer al conjunto de *Stop Words.* A continuación una breve descripción de este tipo de palabras:

- **Thee**: Forma arcaica de *you.*
- **Hath**: Forma arcaica de la conjugación del verbo *have* en tiempo presente, tercera persona del singular.
- **O**[9]: expresión utilizada por Shakespeare para indicar que un personaje se dirigirá en ese párrafo al público (parte de un texto teatral conocido como Apartes en español o Direct Address o Aside en inglés), no es característico de un género en particular.
- **Doth:** Forma arcaica de la conjugación del verbo *do* en tiempo presente, tercera persona del singular.

Para tener una segunda visualización de las palabras más frecuentes y empezar a ver diferencias entre los géneros, utilizar esta librería genérica está bien(como se podrá observa más abajo con resultados numéricos), pero quizás no sea lo suficientemente bueno para alcanzar una diferenciación eficiente entre los géneros, dado el lenguaje utilizado por el autor de la obra. Por ejemplo, el lenguaje utilizado contiene palabras del inglés antiguo, que actualmente no son utilizadas en el inglés contemporáneo o moderno, por lo tanto el uso de la librería de *Sklearn* u otra librería moderna no se adecua al lenguaje utilizado en las obras analizadas[10]. Está fuera del alcance de este informe investigar cual conjunto de *Stop Words* sería el adecuado para esta aplicación, pero podría ser objeto principal de estudio si se quisiera avanzar en encontrar diferencias entre géneros, mediante el análisis del contenido de las obras.

[8]https://www.readwritethink.org/sites/default/files/resources/lesson_images/lesson1031/terms.pdf
[9] https://www.shakespeareswords.com/Public/Prices.aspx?ReturnUrl=/Public/Glossary.aspx?letter=o
[10] https://aclanthology.org/W18-2502.pdf

La aplicación del filtro redujo la cantidad de palabras, o dimensionalidad del problema a analizar, de 870813 a 374787(el 57% de las palabras de todas las obras pertenecían al conjunto de Stop Words utilizado).

En el primer análisis realizado a 870813 palabras, si solo consideramos las 1000 palabras más frecuentes por género(estas equivalen aproximadamente al 80% del total de palabras que aparecen en cada género), se cumple que 510 palabras de las 1000(51%) coinciden, este resultado se puede tomar como una medida de similitud entre los géneros. Con la reducción a 374787 palabras(aplicando el filtro de *Stop Words*), pero considerando ahora las 2500 palabras más frecuentes por género(también aproximadamente el 80% de las palabras de las obras por género), se cumple que 1027 palabras de las 2500(41%) coinciden, reduciéndose la similitud entre géneros Vs el caso sin filtrar. Esto último demuestra que las características únicas de los géneros mejoran con la aplicación de la eliminación de *Stop Words*, ya que la coincidencia de palabras frecuentes, como una medida de similitud entre géneros, disminuye.

Introducción a las Expresiones Regulares

Algunos Símbolos en expresiones regulares :

La expresión regular describe un patrón de texto a buscar y en nuestro caso lo sustituimos por un espacio en blanco por ejemplo o cualquier otro carácter o secuencia que sea conveniente. Se cuenta con símbolos especiales para indicar lo que se desea buscar. La siguiente lista muestra un subconjunto de los símbolos especiales a utilizar para especificar un patrón.

1) [ ] : denota un conjunto de caracteres. Entre los [] debe especificarse algún carácter.
2) A-Z : representa un carácter comprendido en el rango A-Z.
3) . : representa cualquier carácter excepto \n, comodín (wildcard).

4) + : busca una o más veces el token anterior.
5) ? : busca coincidencias de cero o una vez del token anterior.
6) * : busca coincidencias de cero o más veces del token anterior.
7) {n,m} : busca n veces como mínimo y m veces como máximo del token anterior.

Ejemplo : /\[.+?]/ <-- Esta expresión regular encuentra todas las secuencias de caracteres entre [].

Explicación:

\[: indica buscar el [(corchete apertura) , se le coloca \ delante para que no lo interprete como en (1).
. : Cualquier carácter.
+ : Uno o más caracteres iguales al anterior.
? : Busca el menor número de veces. O sea, la cadena más corta que comience con [y termine con].

Para un mejor entendimiento de las expresiones regulares se recomienda consultar las siguientes referencias.

https://eead-csic-compbio.github.io/perl_bioinformatica/node18.html

https://regexr.com/

https://docs.python.org/es/3/library/re.html