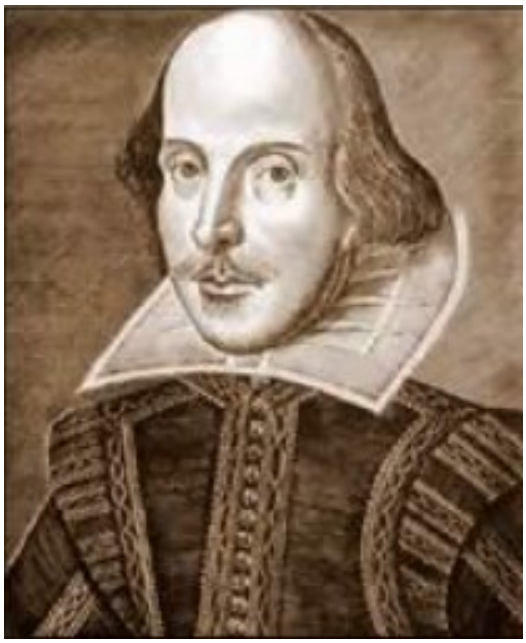


Informe Final

TAREAS 1 y 2.

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

Contenido

TAREA 1

1.1 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*)**

En la Figura 1, se puede observar cómo se vinculan estos objetos.

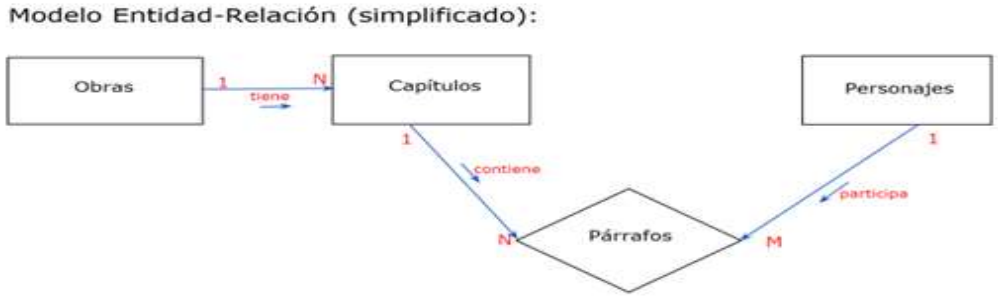


Figura 1: Modelo Entidad - Relación.

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.

En el ***Anexo A1.1*** y ***A1.2***, se pueden observar tablas con los atributos asociados a las **Obras**, **Capítulos**, **Párrafos** y **Personajes**, e información útil sobre la creación de la base de datos.

1.2 Exploración y Calidad de Datos

1.2.1 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 1 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 1 : Resultado Análisis datos Faltantes - Tabla Personajes.

[1] pandas-profiling · Pipis

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.

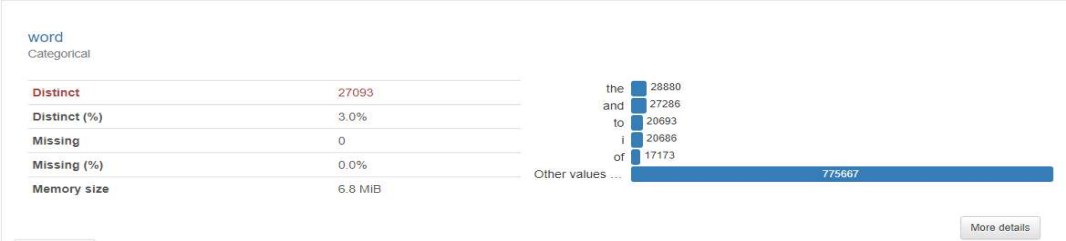


Figura 2 : Palabras más frecuentes del atributo word del dataframe words.

1.3 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 }

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. Los paréntesis rectos son usados como acotaciones para la dirección (***stage directions***), y deben ser removidos junto con las frases que contienen. Los apóstrofes también deben ser removidos, son usados para: resaltar alguna palabra, resaltan una frase, identificar una contracción corta o identificar una posesión. En el ***Anexo A1.3*** se puede observar un detalle de las herramientas utilizadas para eliminar estos caracteres.

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 contabilizan como la misma palabra.

1.4 Conteo de Párrafos, Palabras y Visualizaciones

1.4.1 Personajes con más Párrafos

En el ***Anexo A1.2*** se detalla 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 observa en la Figura 3 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 4. ***Falstaff*** es el personaje con más párrafos asociados, posteriormente en la tarea 2, estos párrafos serán utilizados para entrenar un Clasificador.

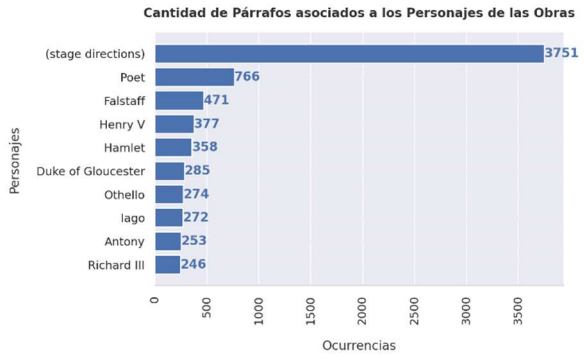


Figura 3 : Histograma del conteo de Párrafos asociados a cada Personaje de las obras de Shakespeare.

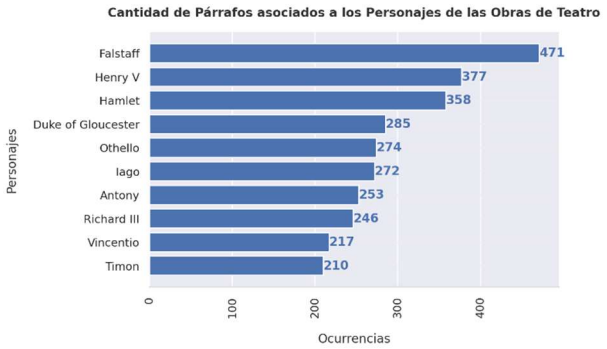


Figura 4: Histograma del conteo de Párrafos asociados a cada Personaje de las obras de Teatro de Shakespeare

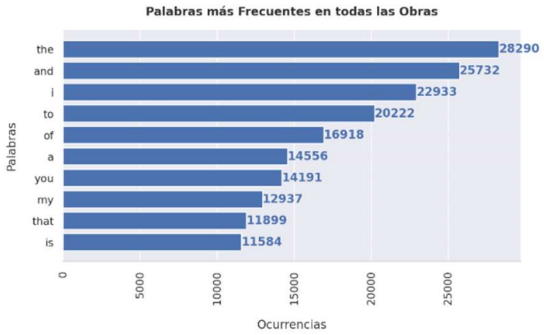


Figura 5 : Palabras más Frecuentes en toda la Obra de Shakespeare.

1.4.2 Palabras más frecuentes en las Obras

En la Figura 5 : Palabras más Frecuentes en toda la Obra de Shakespeare., se puede observar las 10 palabras más frecuentes encontradas en todas las obras de Shakespeare. Si observamos por género, podemos observar que la unión de las 10 palabras más frecuentes resulta en un total de 16 palabras: **{i, you, my, his, her, that, thy, thou, the, a, in, of, to, with, and, is}**

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***[2]***.*** 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. El ***Anexo A1.4*** incluye un análisis de las Palabras más frecuentes eliminando ***Stop Words***, y en el ***Anexo A1.5*** se puede observar un análisis de los Personajes con más Palabras en las Obras.

1.4.3 Evolución de las Obras a lo largo de los años, Preguntas a responder con los datos.

En la tarea 1 se solicitaba presentar una gráfica con la evolución de las obras a lo largo de los años y presentar preguntas que se podrían responder con los datos. Estos dos puntos pueden ser observados en el ***Anexo A1.6 y A1.7.***

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

TAREA 2

2.1 Dataset y representación numérica del texto

2.1.1 Muestreo Estratificado y Aleatorio.

Para el análisis propuesta para esta tarea, una vez que se aplicó la función ***clean_text()*** de la tarea 1(Limpieza de Signos de Puntuación y Paréntesis Rectos usando Expresiones Regulares), se seleccionaron únicamente los párrafos de los tres personajes a analizar en primera instancia(***Antony***, ***Cleopatra*** y ***Queen Margaret***), posteriormente generamos un conjunto de datos para **entrenamiento** de un **70%** del total de los párrafos disponibles y un conjunto de datos para **test** con el **30%** restante, utilizando la función ***train_test_split*** de la librería ***sklearn*** para realizar el **muestreo estratificado**[3] (cada conjunto contiene aproximadamente el mismo porcentaje de párrafos de cada uno de los personajes) y asignación **aleatoria** de los párrafos para cada conjunto(a priori no hay un sesgo en los párrafos seleccionados para entrenamiento y para test).

NOTA: El muestreo estratificado se obtiene con la siguiente línea de código:

(X_train, X_test, y_train, y_test) = train_test_split(X, y, test_size = .3, shuffle = True, random_state = 0, stratify=y)

2.1.2 Balance de Párrafos

En la Figura 6, se puede observar la distribución de los párrafos por personaje en el conjunto de entrenamiento y test, se observa la correcta estratificación de los párrafos.

La distribución no es uniforme, se observa un pequeño desbalance, dónde ***Antony*** es el personaje con más párrafos (253), segundo ***Cleopatra*** (204), y por último ***Queen Margaret*** con la menor cantidad (169), los tres totalizan 626 párrafos.

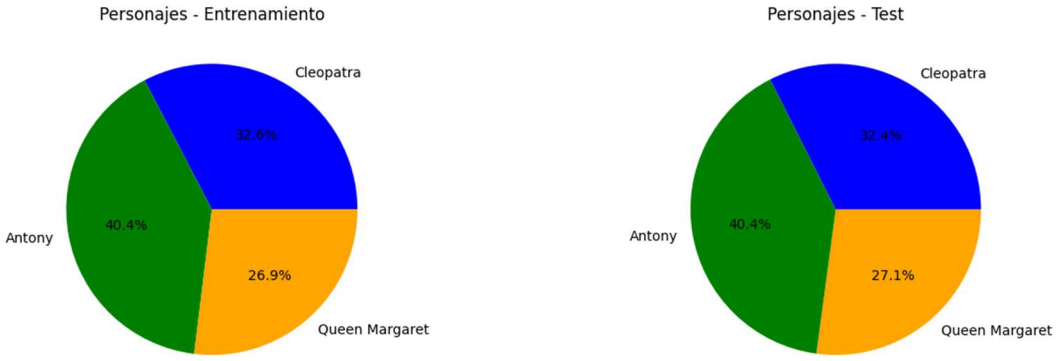


Figura 6 : Distribución de Párrafos en conjuntos para Entrenamiento y Test

2.1.3 Transformación de Párrafos a Representación Numérica

Para poder realizar tareas sobre documentos de texto con herramientas de Aprendizaje Automático, primero es necesario transformar los documentos a vectores numéricos (se denominan vectores de características o ***features*** en inglés), una alternativa para realizar esto es utilizar una técnica de nombre ***bag of words*** (o bolsa de

[3] https://www.sharpsightlabs.com/blog/scikit-train_test_split/
https://keepcoding.io/blog/para-que-sirve-el-train-test-split/#Shuffle_y_random_state
https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

palabras). Para nuestro trabajo los documentos serán párrafos, y la transformación de párrafos a un vector de ***features***, con esta técnica, consiste en dos etapas:

1. Asignar un identificador entero único (**j : 0,1,….,(N-1)**) a cada palabra ($\mathbf{w_j : w_0,w_1,…..,w_{N-1}}$) que aparece en cualquier párrafo del conjunto de entrenamiento (por ejemplo, creando un diccionario dónde la clave es el número entero **j,** asignando como valor asociado a esa clave la palabra $\mathbf{w_j}$).
2. Para cada párrafo (**i : 0,1,…,(M-1)**), hay que contar el número de ocurrencias de cada palabra $\mathbf{w_j}$ y guardar este valor como el elemento **{i,j}** de la matriz de conteo **X**, dónde **j** es el identificador único asociado a la palabra $\mathbf{w_j}$ definido en el punto 1).

A continuación un ejemplo sencillo de cómo aplicar las dos etapas anteriores en un conjunto de 2 párrafos:

Párrafo 1: " Este párrafo es una prueba para contar palabras de prueba".

Párrafo 2: " Este es un segundo párrafo de prueba"

ETAPA 1) Diccionario de palabras:

{ 1: 'Este' : , 2: 'párrafo' , 3: 'es' , 4 : 'una' ,5: 'prueba' , 6: 'para' ,7: ' contar' , 8: 'palabras' , 9: 'de', 10: 'un', 11: 'segundo' }

ETAPA 2) Matriz de Conteo:

Primero creemos los diccionarios de conteo por párrafo:

Párrafo 1: { **1** : 1 , **2** : 1 , **3** : 1 , **4** : 1 , **5** : 2 , **6** : 1 , **7** : 1 , **8** : 1 , **9** : 1 , **10** : 0 , **11** : 0 }

Párrafo 2: { **1** : 1 , **2** : 1 , **3** : 1 , **4** : 0 , **5** : 1 , **6** : 0 , **7** : 0 , **8** : 0 , **9** : 1 , **10** : 1 , **11** : 1}

Mediante estos diccionarios podemos generar una matriz dónde las filas son los párrafos, las columnas las palabras encontradas en todos los párrafos de entrenamiento y el contenido de la matriz es el conteo de palabras. A continuación la representación de los párrafos como matriz de conteo:

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 2 & 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

En el ejemplo anterior se observan pocos ceros en la matriz de conteo, pero para un problema con muchas palabras, la cantidad de ceros que tendrá la matriz será muy grande. Para optimizar el espacio en memoria que ocupará la matriz X, se utiliza una estructura de datos que almacena las posiciones de la matriz dónde el conteo es distinto de cero, en inglés a esta estructura se le denomina **Document-Term matrix** o **dt-matrix**.

Para nuestro problema, la dimensión de la matriz de conteo, incluyendo ***Stop Words***, es de 438x2796 (# párrafos x # palabras) de las cuales solo el 0,863% de las posiciones de la matriz tienen valores no nulos (438x2796x0,00863 = 10568 posiciones), un valor muy bajo. Esto es un ejemplo claro de los beneficios de utilizar otras estructuras de datos comprimidas para representar este tipo de problemas.

Para la transformación de párrafos a vectores de palabras, utilizamos la clase ***CountVectorized***[4] de la librería ***sklearn***. A continuación se podrán observar las primeras palabras del diccionario de todas las palabras encontradas en los párrafos de entrenamiento:

{'let': 1328, 'it': 1229, 'alone': 68, 'to': 2470, 'billiards': 229, 'come': 437, 'charmian': 386, 'the': 2406, 'beds': 180, ……}

Esta clase utiliza una **dt-matrix** para la representación de los párrafos, a continuación, los primeros 4 componentes de la representación obtenida de nuestro problema:

(0, 1328) 2: el párrafo "0" contiene 2 palabras asociadas al índice 1328.
(0, 1229) 1: el párrafo "0" contiene 1 palabra asociadas al 1229.
(0, 68) 1: el párrafo "0" contiene 1 palabra asociadas al 68.
(0, 2470) 1: el párrafo "0" contiene 1 palabra asociadas al 2470.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html#

2.1.4 n-gramas y TF-IDF

Los **n-gramas**[5] son secuencias continuas de palabras o símbolos, o tokens en un documento. En términos técnicos, se pueden definir como las secuencias vecinas de elementos en un documento.

Un ejemplo de código usando librería ***nltk***:

```
from nltk import ngrams
sentence = 'Como se forman estos 2-gramas'
n = 2
unigrams = ngrams(sentence.split(), n)
for grams in unigrams:
print(grams)
```

El resultado del ejemplo es el siguiente: { ('Como', 'se') , ('se', 'forman') , ('forman', 'estos') , ('estos', '2-gramas') }

Anteriormente, representamos los párrafos mediante la técnica de conteo ***bag of words***, lo cual tiene un problema: los párrafos más largos tendrán valores de conteo promedio más altos que los párrafos más cortos, aunque puedan hablar sobre los mismos temas. Esto puede generar que párrafos similares sean interpretados como párrafos distintos, por este problema de escala.

Para evitar esto, podemos dividir el número de ocurrencias de cada palabra en un párrafo por el número total de palabras en ese párrafo: este nuevo valor obtenido se denomina **tf** por **Term Frequency**.

Otra mejora que se pude realizar a la matriz de conteo es reducir el peso de las palabras que aparecen en muchos párrafos y, por lo tanto, son menos informativas que las que aparecen en pocos párrafos. Esta reducción de escala se llama **tf-idf (Term Frequency - Inverse Document Frequency)**.

2.1.5 Mapa de Párrafos con Análisis de Componentes Principales (PCA)

La cantidad de párrafos para analizar durante el entrenamiento son 438, mediane la técnica ***PCA*** estos 438 puntos serán proyectos al espacio de R^2 y se interpretará el resultado.

En la Figura 7, se puede observar está reducción dimensional, de la cual no se observa una separación de clases o de personajes, los puntos están dispersos, pero hay mucha superposición entre diferentes clases, lo anterior fue realizado sin eliminar ***Stop Words***, utilizando solo ***1-grama*** y sin utilizar la reducción de peso por ***idf***.

Eliminando las **Stop Words**, utilizando una representación de ***1-grama*** y ***2-gramas*** en simultáneo y usando la reducción ***idf*** comentada en la sección anterior, tampoco se observa una mejora en la visualización de la representación de los datos en 2D, ver Figura 8. Para esta última visualización, se observa que los datos no están tan dispersos, hay más concentración de los datos en la ventana [(-0.1 , 0.1),(0.1 , -0.2)] y mucha superposición entre diferentes clases.

De estas figuras podríamos concluir que no hay evidencia de tener una separación entre clases en alta dimensión, esto último no significa que no exista tal separación, significa que con ***PCA***, si existe, esto no es observable. Se utilizaron otras técnicas de reducción de dimensionalidad de la librería ***Sklearn***: ***TrucatedSVD*** (recomendada para reducir matrices dispersas como nuestro caso) y ***T-SNE***, obteniendo las mismas conclusiones que con ***PCA***.

Para explicar porque no es posible ver con ***PCA*** la separación de clases (suponiendo que existe), podemos observar la varianza explicada de las 2 primeras componentes principales, cuya proporción respecto a la varianza total de los datos es muy baja (1,6%), además si calculamos la varianza explicada acumulada de los 10 primeros componentes principales (ver Figura 9) alcanzamos una varianza explicada acumulada de 6,2%, que también es muy baja, recién utilizando 300 componentes principales (ver Figura 10) podríamos alcanzar un 80% del total de la varianza explicada. Esto último significa que, si pudiéramos "ver" la proyección de los párrafos en R^{300}, la varianza acumulada de los datos proyectados podría alcanzar un valor aceptable para evaluar si hay o no separación de clases.

[5] https://www.analyticsvidhya.com/blog/2021/09/what-are-n-grams-and-how-to-implement-them-in-python/

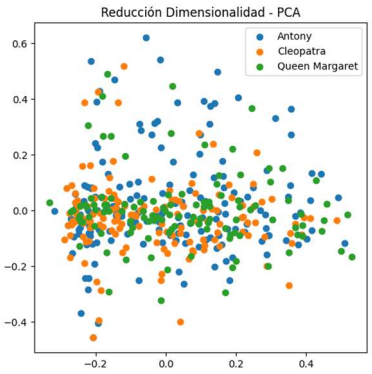


Figura 7: PCA con Stop Word, n_gram_range=(1,1) e idf=False

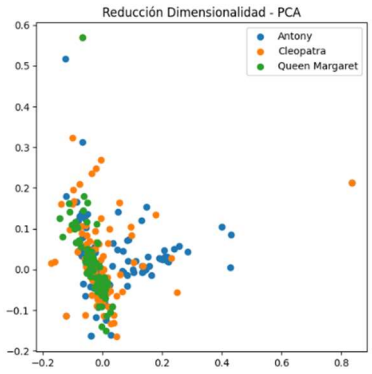


Figura 8: PCA sin Stop Word, n_gram_range=(1,2) e idf=True

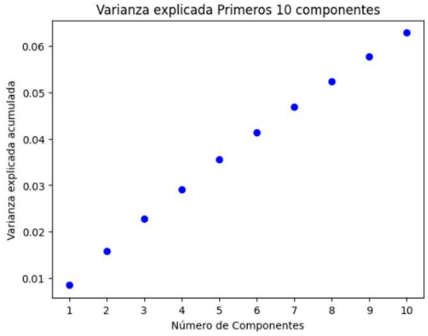


Figura 9: Varianza explicada acumulada, primeros 10 componentes

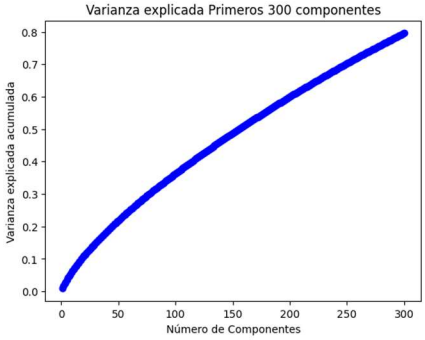


Figura 10: Varianza explicada acumulada, primeros 300 componentes

2.2 Entrenamiento y Evaluación de Modelos

2.2.1 Entrenamiento Modelo Multinomial Naive Bayes, Test, métricas y matriz de confusión.

Se entrenaron los datos del conjunto de entrenamiento utilizando el clasificador ***Multinomial Naive Bayes*** con los siguientes parámetros: ***{"stop_words": 'english', "ngram": (1,1), "idf": False, "alpha":*** 1}

En el ***Anexo A2.4*** se puede observar una descripción más detallada del principio de funcionamiento de este Clasificador.

Con el modelo obtenido se procedió a realizar las predicciones de los párrafos pertenecientes al conjunto de test, para medir el resultado obtenido se utilizaron las siguientes métricas de desempeño:

- ***Accuracy***: Dado un conjunto de muestras etiquetas por el clasificador, ***Accuracy*** es una medida global de la habilidad del clasificador de etiquetar correctamente las muestras.
- ***Precision***: Dado un conjunto de muestras etiquetadas por el clasificador, ***Precision*** es una medida de la habilidad del clasificador de NO etiquetar como positiva una muestra que es negativa.
- ***Recall***: Dado un conjunto de muestras de una clase, ***Recall*** es una media de la habilidad del clasificador de encontrar todas las muestras positivas de esa clase.

Existe un ***trade-off*** entre ***Precision*** y ***Recall***, no pueden mejorarse las dos en simultáneo, la mejora de una va en detrimento de la otra. Además, una métrica puede ser más importante que la otra en algunas tareas, la importancia de cada una está sujeta al tipo de tarea que están midiendo. Por más información sobre estas métricas ir al ***Anexo A2.1.***

El resultado de ***Accuracy*** obtenido fue de **0.596**, en la Tabla 2 se pueden observar ***Precision*** y ***Recall*** obtenidos por clase o personaje, y en la Figura 11 : Matriz de Confusión Resultado Test la Matriz de Confusión.

- Las métricas de la Tabla 2, ***Precision*** y ***Recall***, pueden ser calculadas con la información obtenida de la matriz de Confusión, para esto es necesario definir cuatro contadores: **True Positive** (TP), **False Negative** (FN), **True Negative** (TN) y **False Positive** (FP).
  En la Tabla 3 se pueden observar sus definiciones en función de la clasificación de la predicción y del verdadero valor a predecir, por ejemplo: un párrafo predicho se cataloga como **FP** si el párrafo **NO** pertenecía a la clase **A** y la predicción del clasificador fue que **SI** pertenecía a la clase **A**.

	Precision	Recall	F1-Score
Antony	0.52	0.93	0.67
Cleopatra	0.73	0.36	0.48
Queen Margaret	0.86	0.37	0.52

Tabla 2: Métricas obtenidas de las predicciones del conjunto de párrafos para Test.

	TRUE	PREDICTED
TP: True Positive	A	A
FP: False Positive	NOT(A)	A
FN: False Negative	A	NOT(A)
TN: True Negative	NOT(A)	NOT(A)

Tabla 3: Clasificación de las Predicciones en función de la predicción obtenida y el valor verdadero de la etiqueta.

En función de las definiciones de ***Recall*** y ***Precision***, y la definición de los distintos contadores, podemos definir cuantitativamente las métricas de la siguiente forma:

- ***Precision*** **= TP/(TP+FP)**
- ***Recall*** **= TP/(TP+FN)**

En la Tabla 4 se puede observar cómo se obtienen los valores de TP y FP para el cálculo de ***Precision*** de cada clase. Por ejemplo, si quiero calcular ***Precision*** del personaje ***Antony*** = $TP_A/(\Sigma (TP_A + FP_A)) = 71/(71+37+28)=0,52$.

En la Tabla 5 se puede observar cómo se obtienen los valores de TP y FN para el cálculo de ***Recall*** de cada clase. Por ejemplo, si quiero calcular ***Recall*** del personaje ***Antony*** = $TP_A/(\Sigma (TP_A + FN_A)) = 74/(74+4+1)=0,93$

		$\Sigma (TP_A + FP_A)$	$\Sigma (TP_C + FP_C)$	$\Sigma (TP_Q + FP_Q)$
TRUE LABEL	Antony	TP_A	FP_C	FP_Q
	Cleopatra	FP_A	TP_C	FP_Q
	Queen Margaret	FP_A	FP_C	TP_Q
		Antony	Cleopatra	Queen Margaret
		PREDICTED LABEL		

Tabla 4: Lectura de la matriz de confusión para calcular Precision.

TRUE LABEL	Antony	TP_A	FN_A	FN_A	$\Sigma (TP_A + FN_A)$
	Cleopatra	FN_C	TP_C	FN_C	$\Sigma (TP_C + FN_C)$
	Queen Margaret	FN_Q	FN_Q	TP_Q	$\Sigma (TP_Q + FN_Q)$
		Antony	Cleopatra	Queen Margaret	
		PREDICTED LABEL			

Tabla 5: Lectura de la matriz de confusión para calcular Recall

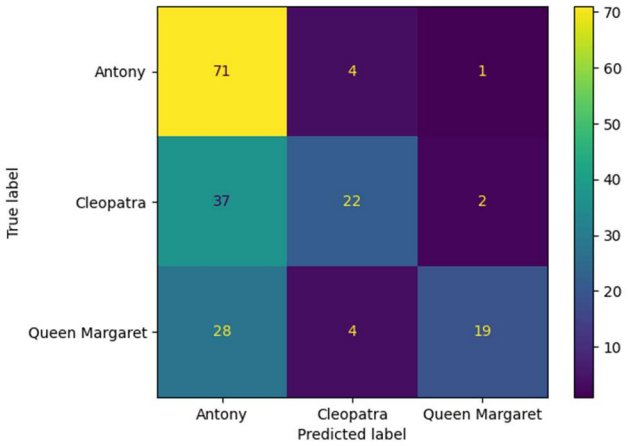


Figura 11 : Matriz de Confusión Resultado Test

A continuación algunos comentarios sobre la Matriz de Confusión:

- Las frases de ***Antony*** se confunden poco (Alto ***Recall***), hay solo cinco párrafos que se confunde (4 se predicen que los dijo ***Cleopatra*** y 1 que lo dijo ***Queen Margaret***). Para alcanzar este alto ***Recall*** se predicen muchos párrafos como ***Antony***, de los cuales muchos son FP, por lo tanto, ***Precision*** es bajo.
- Para ***Cleopatra*** y ***Queen Margaret*** el comportamiento es similar, sus frases se confunden mucho (Alto FN, Bajo ***Recall***), la mayoría de las frases se predicen que son dichas por ***Antony*** (esto genero el bajo ***Precision*** de ***Antony***). Pero cuando el clasificador predice que el párrafo es de algunas de ellas, le erra poco (Bajo FP, Alto ***Precision***)
- Si el objetivo es poder tener un alto valor de ***Recall*** en todas las clases, el clasificador entrenado no es efectivo.

Que problemas puede tener el hecho de mirar solamente el valor de *Accuracy*?

- Esta métrica mide el nivel de aciertos o True Positives de todos los personajes (clases), en forma global.
- Si solo miramos esta métrica, podemos omitir observar que el clasificador puede estar trabajando correctamente para un conjunto de clases y puede estar trabajando con bajo desempeño para otro conjunto de clases.
- Para evaluar el desempeño local de las clases es mejor observar métricas locales, como ***Precision***, ***Recall*** y ***F1***, que dan mejores estimaciones del desempeño del clasificador.

Que problemas podemos tener con un desbalance de datos mayor?

- Un desbalance de datos genera clases mayoritarias y clases minoritarias en el entrenamiento, podríamos esperar un valor de ***Accuracy*** alto, pero pueden existir clases minoritarias que el desempeño de aciertos sea bajo (bajo ***Recall***), porque al ser una clase minoritaria (con baja probabilidad de ocurrencia) el clasificador puede con frecuencia decir que la muestra “NO” pertenece a esta clase minoritaria.
- Se puede observar en la parte 2.5, cuando se cambia al personaje ***Antony*** por ***Falstaff*** (personaje con mayor cantidad de párrafos asociados en todas las obras), un mayor desbalance en la distribución de párrafos, que tiene como consecuencia una baja detección en el test de las clases minoritarias (***Cleopatra*** y ***Queen Margaret***).

2.2.2 Validación Cruzada (Cross Validation)

2.2.2.1 Definición

El objetivo es estimar la tasa de aciertos (***Accuracy***) que tendremos al realizar las predicciones con el conjunto de test, para ello existen una serie de técnicas que pueden ser usadas para estimar este valor usando los datos de entrenamiento.

Método : Conjunto de Validación

Esta técnica consiste en dividir aleatoriamente el conjunto de datos en dos partes: el conjunto de entrenamiento y el conjunto de validación. El modelo se entrena usando los datos de entrenamiento y dicho modelo se usa para predecir las salidas del modelo utilizando como entradas el conjunto de validación. La tasa de acierto resultante es un estimador de la tasa de acierto que tendremos en el test.

Este método es conceptualmente simple y fácil de implementar, pero tiene 2 desventajas o inconvenientes:

1) La tasa de acierto puede ser altamente variable dependiendo de cuáles observaciones se incluyen en el conjunto de entrenamiento y cuáles se incluyen en el conjunto de validación.
2) Ya que los métodos estadísticos tienden a comportarse peor cuando se entrenan con pocas observaciones, esto sugiere que la tasa de acierto del conjunto de validación puede tender a "sobreestimar" esta tasa para el modelo cuando se usa en todos los datos.

A continuación veremos otro método para resolver estas dos desventajas presentadas.

Método : Validación Cruzada k-Fold (k-Grupos).

Este método consiste en dividir aleatoriamente el conjunto de datos total en **k** grupos de aproximadamente igual tamaño, el primer grupo es tratado como el conjunto de validación y el modelo es entrenado con los restantes k-1 grupos. El estimador es obtenido usando el grupo de validación.

Este proceso es repetido k veces; cada vez, un grupo diferente de observaciones es tratada como conjunto de validación. Entonces se obtienen k estimaciones de la tasa de acierto y nuestro estimador seré el promedio de estos resultados. Con k=n, tenemos un caso especial que se conoce como el método ***Leave-One-Out*** (por más detalle ir al ***Anexo A2.2).***

Cuando ejecutamos estos métodos, nuestro objetivo es determinar que tan bien un método estadístico de aprendizaje automático se comportará con datos independiente o con datos "frescos" (datos que no conoce el modelo entrenado); en este caso nos interesa calcular el estimador de la tasa de aciertos del test. En este método hay un compromiso sesgo-varianza asociado con la elección del valor de k en la validación cruzada de k-grupos.

2.2.2.2 Búsqueda de hiperparámetros con Validación Cruzada

Para este problema bajo estudio, se utilizó el método de Validación Cruzada ***Stratified k-fold*** de la librería ***sklern***. el cual utiliza el método ***k-fold*** comentado en la sección anterior con la particularidad que los K grupos formados contiene la misma cantidad de muestras de cada clase. Se utilizó como parámetro principal K=4, con el objetivo de poder tener una relación cercana a 70:30% entre datos de entrenamiento y datos para la validación.

Para los distintos experimentos se utilizaron los siguientes parámetros:

a) Parámetros relacionados a la Transformación de los párrafos a representación numérica:
   1) Uso o no uso de ***Stop Words***
   2) Uso de ***Uni-grama***, ***Bi-grama***, etc.
   3) Uso o no de ***idf***: ***inverse document frecuency***: ponderación inversa de palabras más frecuentes.

b) Parámetros del Clasificador *Multinomial Naive Bayes*:

1) α : ***Laplace Smoothing*** (Laplace smoothing es una técnica que soluciona el problema de probabilidad cero en el clasificador ***Naive Bayes***)

Se utilizaron 12 combinaciones distintas, obteniéndose el mejor ***Accuracy*** promedio de validación en el experimento 12, en la Tabla 12 : Atributos de los Capítulosse pueden observar los resultados de todos los experimentos.

N° Exp.	Stop Word	n-grama	Idf	α	Accuracy Validación
1	None	(2,2)	False	1	0.404
2	None	(4,4)	False	1	0.406
3	None	(1,1)	False	1	0.413
4	None	(3,3)	False	1	0.415
5	None	(1,1)	True	1	0.441
6	English	(1,2)	True	1	0.507
7	English	(1,2)	False	1	0.530
8	English	(1,1)	True	1	0.555
9	English	(1,1)	False	1	0.557
10	English	(1,1)	False	0.75	0.566
11	English	(1,1)	False	0.5	0.582
12	English	(1,1)	False	0.1	0.612

Tabla 6 : Resultados Validación Cruzada.

La Figura 12: Grafico comparativo de resultados de Accuracy en todos los experimentos. es un gráfico de violín con la distribución de los resultados de ***Accuracy*** obtenidos en cada experimento (se repitió K=4 veces), dónde se puede observar la mediana y la varianza. En la Figura 13 y Figura 14, se puede observar la distribución de los resultados de ***Recall*** y ***Precision*** obtenidos en cada experimento para ***Queen Margaret***.

Algunos comentarios sobre el grafico de violín observado en la Figura 12:

- Se observa que el experimento 12 obtiene el mejor resultado, este elimina las Stop Words de los párrafos, usa solo palabras individuales(***uni-grama***), no utiliza la transformación ***idf***(utiliza la matriz de frecuencias), y utiliza un α igual a 0.1.
- Los valores más bajos de ***Accuracy*** (experimentos 1 a 4) resultan cuando no se usa la eliminación de Stop Words.
- El Experimento 5 muestra una mejora con respecto a los anteriores porque usa la transformación ***idf***, quitando relevancia a las Stop-Words más frecuentes del conjunto de párrafos analizados.
- En los experimentos 6 a 9, se observa una mejora al introducir la eliminación de Stop Words.
- Finalmente en los experimentos 10 a 12, se varía el coeficiente α, observándose una mejora en ***Accuracy*** a medida que este parámetro disminuye, generando un óptimo en α =0.1

Algunos comentarios sobre las Figura 13 y Figura 14, ***Recall*** y ***Precision*** del personaje ***Queen Margaret***:

- Se observa una mejora considerable en el ***Recall*** obtenido (cercano a 0.65) relativo al obtenido en el 1er experimento mostrado en la Tabla 2 (0. 37).
- ***Precision*** tiene un comportamiento inverso, ha disminuido (cercano a 0.75) comparado con el valor presentado en la Tabla 2 (0.86).

En el ***Anexo A2.3***, se puede ver cómo cambian las métricas (***Precision*** y ***Recall***), por experimento, para los personajes ***Antony*** y ***Cleopatra***. ***Cleopatra*** se comporta similar a ***Queen Margaret***, mejora el ***Recall*** y disminuye ***Precision***, y ***Antony*** en sentido inverso, decrece ***Recall*** pero mejora ***Precision***.

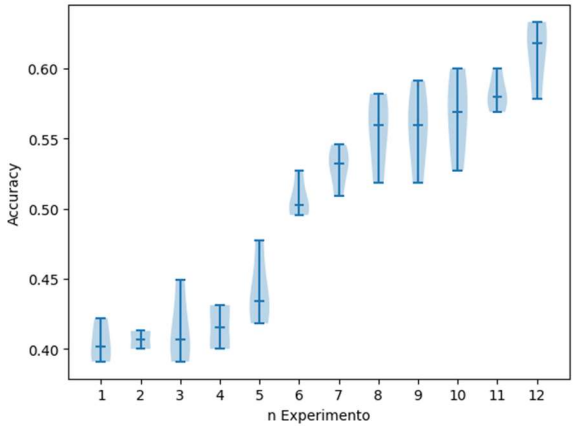


Figura 12: Grafico comparativo de resultados de Accuracy en todos los experimentos.

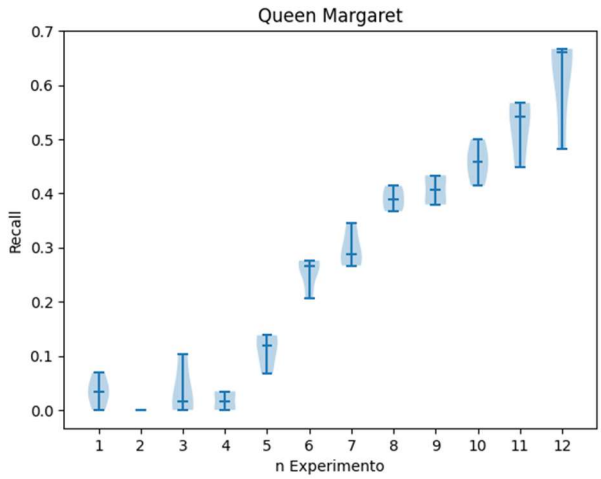


Figura 13 : Distribución Recall en cada experimento - Queen Margaret.

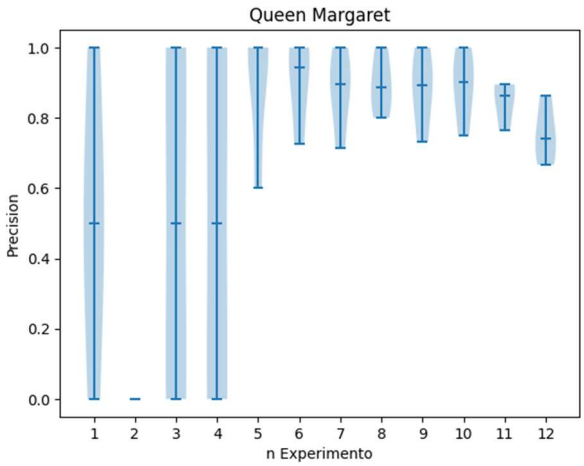


Figura 14: Distribución Precision en cada experimento - Queen Margaret.

2.2.3 2do Entrenamiento con Parámetros Óptimos.

Con los parámetros utilizados en el experimento 12, que se observan a continuación, se entrenó nuevamente el clasificador ***MultinomialNB***: **{"stop_words": 'english', "ngram": (1,1), "idf": False,"alpha": 0.10}**

Con este nuevo modelo, el resultado de ***Accuracy*** obtenido en el test alcanzó el valor de **0.633,** en la Tabla 7 se pueden observar los resultados de ***Precision*** y ***Recall***, y en la Figura 15: Matriz de Confusión Resultado Test, utilizando parámetros obtenidos con Validación Cruzada. se puede observar la matriz de confusión obtenida.

	Precision	Recall	F1-Score
Antony	0.61	0.78	0.68
Cleopatra	0.57	0.49	0.53
Queen Margaret	0.79	0.59	0.67

Tabla 7: Métricas obtenidas de las predicciones del conjunto de párrafos para Test.

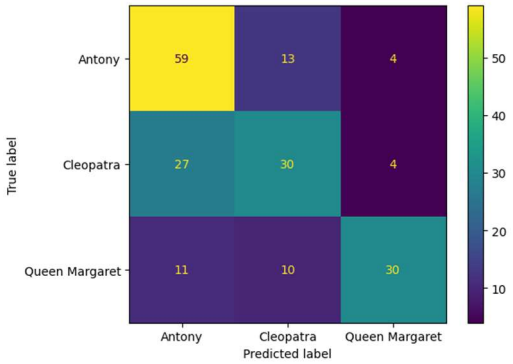


Figura 15: Matriz de Confusión Resultado Test, utilizando parámetros obtenidos con Validación Cruzada.

Algunas observaciones relativas a las métricas obtenidas y a la matriz de confusión:

- El ***Accuracy*** obtenido **(0.633),** con este nuevo conjunto de parámetros, ha mejorado con respecto al valor obtenido con los parámetros iniciales **(0.595).**
- Se observa una mejora en las métricas de las clases ***Cleopatra*** y ***Queen Margaret***, en detrimento del desempeño de ***Antony***, comparando Tabla 2 con Tabla 7.
- Aunque disminuyó, ***Queen Margaret*** sigue teniendo el mejor ***Precision***, idéntico caso con ***Antony***, que aunque bajó sigue manteniendo el mejor ***Recall***.
- La mejora del ***Accuracy*** está relacionada al aumentan significativamente de los TP de los personajes ***Cleopatra*** y ***Queen Margaret*** (comparando las matrices de confusión de las Figura 11 y Figura 15), lo que se ve reflejado en una mejora de ***Recall***. Esta mejora es más significativa que la disminución del ***Recall*** del personaje ***Antony***.

Limitaciones de utilizar un modelo basado en bag-of-words o tf-idf en cuanto al análisis de texto

1. Se pierde la información que es comunicada por el orden de las palabras en los párrafos.
2. Existe la posibilidad de tener dos representaciones numéricas similares cuando los dos párrafos asociados pueden contener diferente significado semántico. Para resolver esto, existen técnicas que si toman en cuenta el significado del párrafo para tener representaciones numéricas diferentes cuando el párrafo es diferente.

2.2.4 Evaluación de otros Clasificadores

Se evaluaron dos modelos más de clasificación utilizando la librería ***Sklern***.

1) Complement Naive Bayes
2) SVC: Support Vector Classification

Todos ellos fueron evaluados usando Validación Cruzada para la búsqueda del óptimo de los hiperparámetros definidos para cada problema. En la Tabla 8, se puede observar un resumen de ***Accuracy*** obtenido en Validación Cruzada y ***Accuracy*** obtenido en el Test.

En la Figura 16: Distribución Accuracy de cada Clasificador evaluado, se muestra un gráfico de violín comparativo entre las soluciones obtenidas de ***Accuracy*** para los distintos clasificadores explorados y en las Figura 17 a Figura 19 se pueden observar un gráfico comparativo del ***Recall*** para cada personaje. Algunos comentarios sobre las Figura 17 a Figura 19:

- No en forma rigurosa, solo mirando el gráfico de violín para ***Accuracy*** y la concentración de los datos de cada distribución, se espera que el clasificador ***ComplementNB*** tenga el mejor desempeño **en promedio** al clasificar “nuevos” párrafos.

De los gráficos de ***Recall*** por personaje, se observa que para la clase ***Antony*** (clase mayoritaria), el clasificador con mejor ***Accuracy*** (***ComplementNB***) genera el peor ***Recall***, pero para las clases minoritarias se observa lo contrario, con ***ComplementNB*** se producen los mayores valores de ***Recall***, y esto último pesa más en el resultado global de este clasificador.

Clasificador	Parámetros óptimos	Accuracy CV	Accuracy Test
MultinombialNB	clf__alpha: 0.10 vect__ngram_range: (1, 1) vect__norm: l2 vect__stop_words: english vect__use_idf: False	0.612	0.633
SVC	clf__C: 1.0 clf__kernel: linear vect__ngram_range: (1, 1) vect__norm: l2 vect__stop_words: english vect__use_idf: True	0.612	0.649
ComplementNB	clf__alpha: 1 vect__ngram_range: (1, 1) vect__norm: l2 vect__stop_words: english vect__use_idf: True	0.630	0.665

Tabla 8 : Resultados Validación Cruzada (Parámetros, Accuracy CV) y Test (Accuracy)

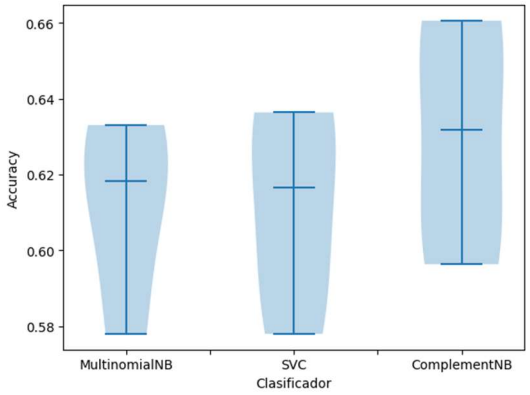


Figura 16: Distribución Accuracy de cada Clasificador evaluado

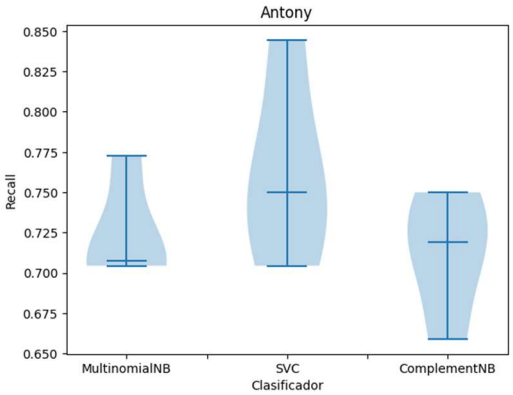


Figura 17: Distribución Recall, personaje Antony, de cada Clasificador evaluado

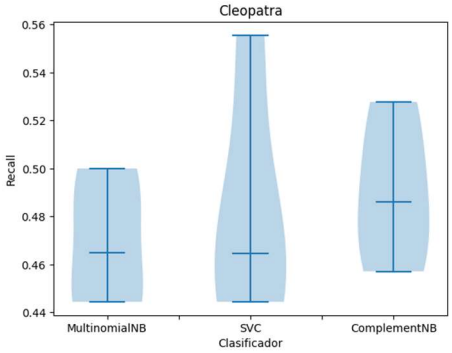


Figura 18: Distribución Recall, personaje Cleopatra, de cada Clasificador evaluado

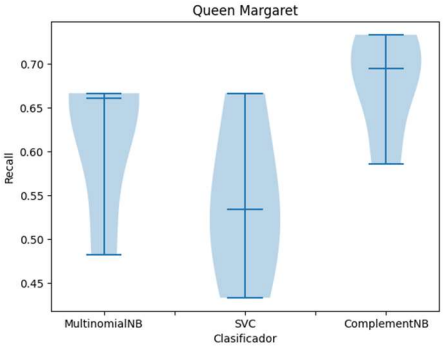


Figura 19 : Distribución Recall, personaje Queen Margaret, de cada Clasificador evaluado

A continuación unos breves comentarios sobre el clasificador ***ComplemntNB***[67] y ***SVC***:

1) ***ComplementNB:*** Es un clasificador que usa como base al clasificador ***MultinomialNB***, pero realiza algunas correcciones para mejorar los problemas que se generan por los supuestos realizados en el clasificador ***MultinomialNB***.
   En el ***Anexo A2.4,*** se puede ver una descripción del clasificador ***MultinomialNB***, el cual tiene varios supuestos:
   1 – Modela la generación de datos con Distribución Multinomial
   2 – Baja Ponderación (P(C_j): probabilidad de clase j) para clases minoritarias, desbalance de datos.
   2 – Independencia entre las palabras (***features***) de una misma frase.
   3 – Independencia en la posición dónde se encuentra la palabra en la frase.

Este clasificador propone:

- Mejorar el modelo asignado a la generación de los datos, proponiendo una distribución Potencial y no Multinomial.
- Para mejorar el desbalance, este clasificador propone agregar una clase complementaria.
- Para mejorar el problema que genera asumir la independencia entre las palabras, este clasificador propone ponderar los pesos de las clases que tiene una dependencia fuerte entre las palabras utilizadas y ponderar menos las que tienen una dependencia más débil.

2) SVC[8]: Support Vector Classification.

- Utiliza como base el algoritmo SVM, ampliamente utilizado en Machine learning para tareas de Regresión o Clasificación.
- Dados los datos de entrenamiento y sus etiquetas, el algoritmo resuelve un problema de optimización lineal para encontrar un hiperplano que maximice el margen o separación entre el hiperplano y las clases a separar.
- Es un clasificador lineal, por lo tanto si el conjunto de datos no es linealmente separable, el algoritmo lo resuelve mapeando los datos a un espacio de mayor orden dimensional, mediante la aplicación de una función Kernel, donde los datos sean linealmente separables.

2.2.5 Analizar desbalance entre Clases modificando Personajes.

2.2.5.1 Cambiando Personaje para análisis: Antony por Falstaff

Falstaff es el personaje con mayor cantidad de párrafos acorde a la Figura 4, para hacer nuevamente el análisis de párrafos se cambió a ***Antony*** por ***Falstaff***, con el objetivo de desbalancear más el dataset y observar las consecuencias. En la Figura 20, se puede observar cómo aumentó el desbalance al utilizar este personaje.

Posteriormente se procedió a realizar los mismos pasos que para el caso anterior:

1) Transformación de párrafos a representación numérica.
2) PCA
3) Entrenamiento con Clasificador ***MultinomialNB***.
4) Predicciones y Evaluación de Métricas: ***Accuracy***, ***Precision***, ***Recall***.
5) Búsqueda de Parámetros Óptimos con Validación Cruzada
6) 2do Entrenamiento con Parámetros Óptimos, Test y evaluación de métricas.

[6] Paper: Tackling the Poor Assumptions of Naïve Bayes Text Classifiers – MIT.
[7] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.ComplementNB.html
[8] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

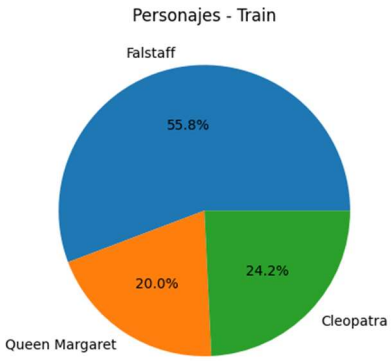


Figura 20 : Distribución de Párrafos de Entrenamiento Falstaff.

A continuación se presentan los resultados más relevantes de las etapas mencionadas anteriormente:

2 – **PCA** : resultados similares a los obtenidos en el caso anterior, la varianza explicada acumulada para dos componentes es de 1.5%. De la visualización en 2D no se puede distinguir una separación de clases o personajes.

3 – **Entrenamiento con MultinomiaNB**: El entrenamiento del clasificador se hizo con los mismos parámetros iniciales que el caso anterior: **{"stop_words": 'english', "ngram": (1,1), "idf": False,"alpha": 1}**

4 - **Predicciones y Evaluación de Métricas: *Accuracy*, *Precision*, *Recall*.**

- La primera predicción con los párrafos de test generó como resultado un ***Accuracy*** de **0.602** y en la Tabla 9 se puede observar las restantes métricas por clase obtenidas del Test. En la Figura 21 se puede observar la Matriz de Confusión de los resultados obtenidos.
- Observando la Tabla 9, los comentarios sobre las métricas ***Recall*** y ***Precision***, son similares al caso anterior, excepto que se acentúa la consecuencia del desbalance entre clases, casi el 96% de las predicciones están asociadas a ***Falstaff***, cuando la cantidad de párrafos dentro del conjunto Test de este personaje es de 55.8%. Por ese motivo, el ***Recall*** de ***Falstaff*** es 1 y el ***Precision*** es bajo, 0.58. Lo contrario pasa con ***Cleopatra*** y ***Queen Margaret*** (clases minoritarias), cuando clasifica un párrafo con alguna de estas clases no le erra(***Precision*** es 1.00) pero hay muchos párrafos de estas clases que son etiquetados como si fueran de ***Falstaff*** (Bajo ***Recall***).

5 - Búsqueda de Parámetros Óptimos con Validación Cruzada:

- Al igual que en el caso anterior, el mejor resultado de ***Accuracy*** lo obtenemos utilizando el siguiente conjunto de parámetros: **{"stop_words": 'english', "ngram": (1,1), "idf": False,"alpha": 0.10}**

6 - 2do Entrenamiento con Parámetros Óptimos, Test y evaluación de métricas.

- ***Accuracy*** obtenido **0.693**, mayor al obtenido con los parámetros iniciales **0.602**.
- Las métricas por clase se pueden observar en la Tabla 9 y la Matriz de Confusión en la Figura 22.
- Se puede observar de los resultados en la Tabla 9, que mejora el ***Recall*** para las clases Minoritarias indicando una mejora en el desempeño del Clasificador.
- Como ultima observación, la mejora del ***Recall*** no es tan significativa si la comparamos con el caso anterior, donde había desbalance entre clases, pero no era tan notorio como en este caso.

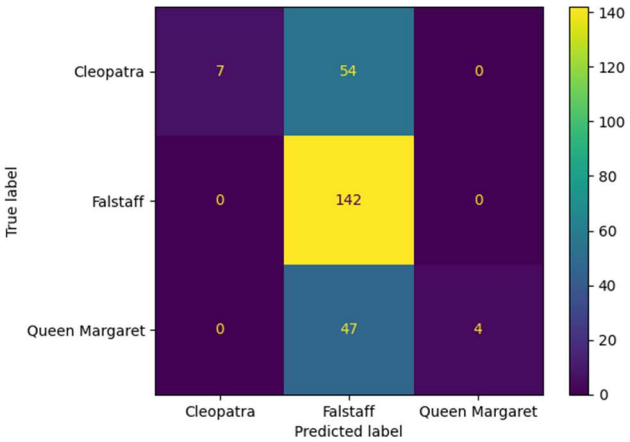


Figura 21: Matriz de Confusión Resultado Test, Parámetros Iniciales.

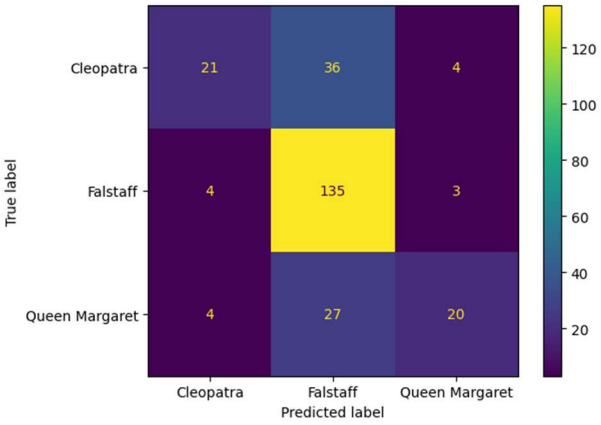


Figura 22: Matriz de Confusión Resultado Test, Parámetros óptimos.

	Parámetros Iniciales			Parámetros óptimos		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Falstaff	0.58	1.00	0.74	0.68	0.95	0.79
Cleopatra	1.00	0.11	0.21	0.72	0.34	0.47
Queen Margaret	1.00	0.08	0.15	0.74	0.39	0.51

- *Tabla 9: Métricas obtenidas de las predicciones del conjunto de párrafos para Test.*

2.2.5.2 Sobremuestreo y Submuestreo

El submuestreo y sobremuestreo[9] son técnicas que intenta reducir el desbalance entre clases. Para nuestro caso, el uso de submuestreo implicaría reducir la cantidad de párrafos de los personajes mayoritarios (por ejemplo: ***Faltstaff***) con el objetivo de equilibra la cantidad de párrafos del conjunto de datos. El uso de sobremuestreo implicaría aumentar los párrafos de los personajes minoritarios (por ejemplo: ***Cleopatra*** y ***Queen Margaret***) con el objetivo de equilibra la cantidad de párrafos del conjunto de datos.

2.6 Técnicas Alternativas para extraer Features de Texto.

Word embeddings es una técnica alternativa para transformar texto en vectores numéricos, un ejemplo de uso de esta técnica es ***Word2vec***[10], que se ocupa del contexto y la coherencia de las palabras, utilizando una red neuronal.

Los dos puntos más importantes, de los cuales se diferencia de ***Bag of Words***, son los siguientes:

1) Los vectores numéricos se crean a partir de palabras individuales, por lo tanto un párrafo no sería un vector numérico (como en ***Bag of Words***), sino que un párrafo sería un conjunto de vectores.
2) Además de almacenar información individual de la palabra (como en ***Bag of Words***), este modelo almacena información contextual considerando las palabras vecinas.

[9] Fuente : https://statologos.com/submuestreo/

[10] Fuente : https://towardsdatascience.com/all-you-need-to-know-about-bag-of-words-and-word2vec-text-feature-extraction-e386d9ed84aa

Al utilizar este modelo, el cual utiliza información de contexto, se esperaría que disminuyan la cantidad de confusiones entre clases o personajes, lo cual impactaría en una mejora de las métricas. En nuestro problema, se esperaría una mejora del ***Accuracy*** y del ***Recall***.

2.7 Modelo Fastext

Se entrenó el modelo con los párrafos de [***Antony***, ***Cleopatra***, ***Queen Margaret***], obteniéndose los siguientes resultados en la predicción de los párrafos del conjunto de Test:

- ***Accuracy*** Test: **0.634**
- ***Precision/Recall***:

	Precision	Recall	F1-Score
Antony	0.61	0.74	0.67
Cleopatra	0.60	0.61	0.60
Queen Margaret	0.76	0.51	0.61

Tabla 10: Métricas obtenidas de las predicciones del conjunto de párrafos para Test.

- Matriz de Confusión:

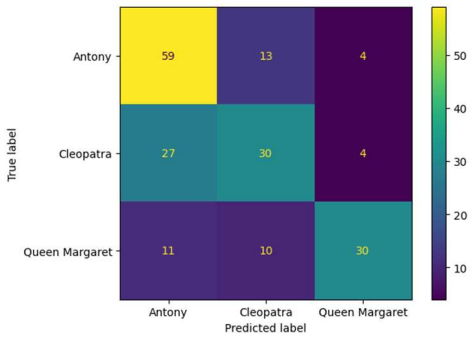


Figura 23: Matriz de Confusión Resultado Test, utilizando fasttext.

Se observa un resultado similar al obtenido en la parte 2.3, ver Tabla 7. La única diferencia sustancial es que mejora el resultado de ***Recall*** de la clase ***Cleopatra***, pero disminuye el resultado de ***Recall*** de la clase ***Queen Margaret***, compensándose para obtener un resultado de ***Accuracy*** similar al obtenido en la parte 2.3.

ANEXO

TAREA 1

A1.1 Descripción Conceptual de la Base de Datos

A continuación se pueden observar tablas con 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 11 : 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 12 : 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 13 : 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 14 : Atributos de los Personajes

A1.2 Información Útil sobre la Construcción de la Base de Datos.

De la página web[11] del proyecto que originó la base de datos, se puede obtener información útil sobre la construcción de esta,

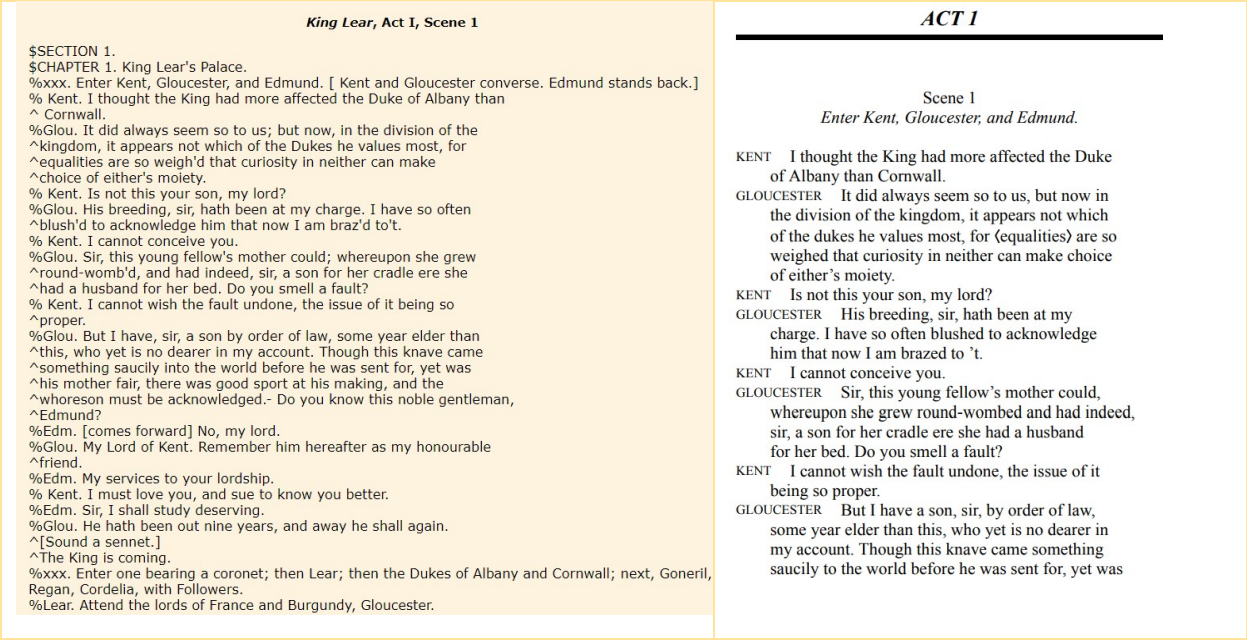
- El texto original utilizado fue preparado para ser interpretado por un analizador de lenguaje, con el objetivo de extraer atributos o información relevante del texto (actos, escenas, personajes, acotaciones de dirección, etc.). Para esto se identificó la información relevante con los siguientes caracteres
    - **$:** 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 24 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.

[11] 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 24 : 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).

- 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 ***Sonnets*** 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.

A1.3 Limpieza y acondicionamiento de Palabras

A1.3.1 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**[12], 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.
****
PAROLLES retire]

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

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.

A1.3.2 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***[13] 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

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

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.

A1.4 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***[14].

Figura 25 : Palabras más frecuentes en todas las obras

Las 10 palabras más frecuentes se pueden observar en la Figura 25. Utilizando las 10 palabras más frecuentes de cada género, 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***[15]), 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:

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

[15]https://www.readwritethink.org/sites/default/files/resources/lesson_images/lesson1031/terms.pdf

- **Thee**: Forma arcaica de *you.*
- **Hath**: Forma arcaica de la conjugación del verbo *have* en tiempo presente, tercera persona del singular.
- **O**[16]: 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[17]. 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.

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.

A1.5 Personajes con más Palabras en las Obras

En la Figura 26, 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 27, 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.

[16] https://www.shakespeareswords.com/Public/Prices.aspx?ReturnUrl=/Public/Glossary.aspx?letter=o

[17] https://aclanthology.org/W18-2502.pdf

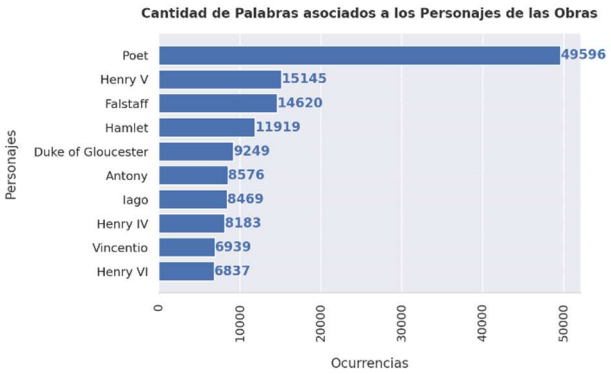


Figura 26 : Cantidad de Palabras por Personajes en todas las Obras

Figura 27 : Cantidad de Palabras por Personajes en todas las Obras de Teatro.

A1.6 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 28, 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 29, 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***[18], con un agrupamiento de 4 años. En la Figura 30, 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).

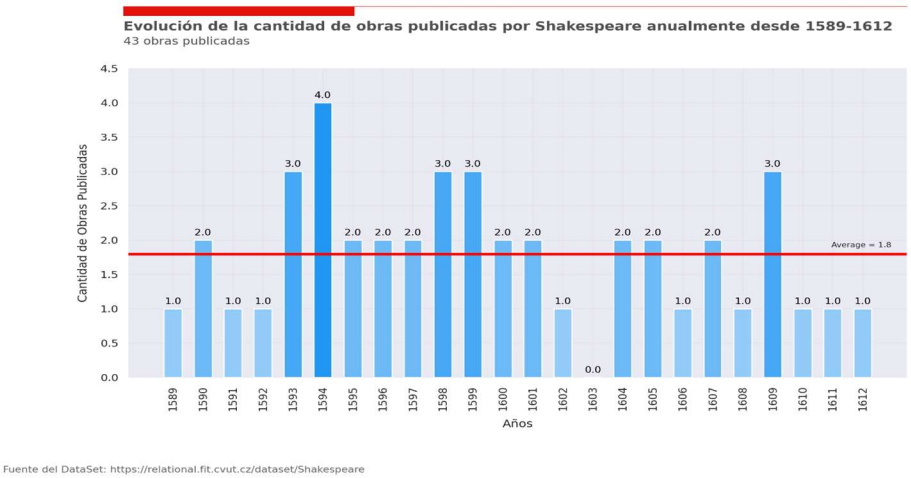


Figura 28 : Evolución de la cantidad de obras publicadas por Shakespeare anualmente desde 1589 a 1612.

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

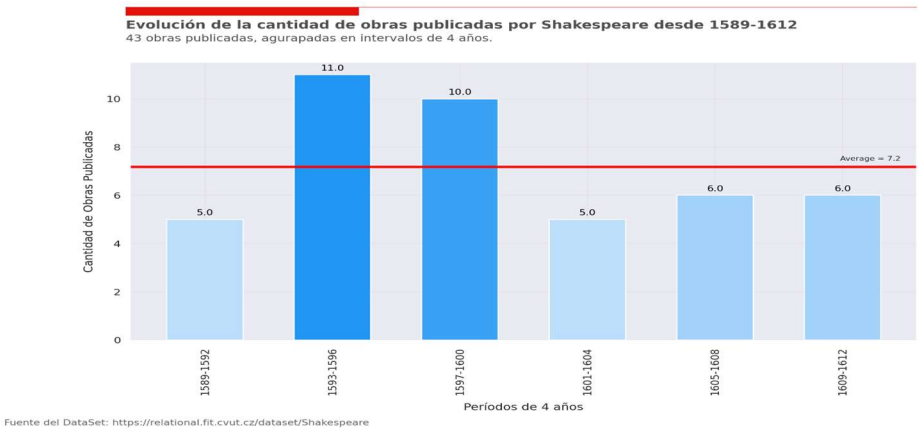


Figura 29 : Evolución de la cantidad de obras publicadas por Shakespeare, agrupadas cada 4 años, desde 1589 a 1612.

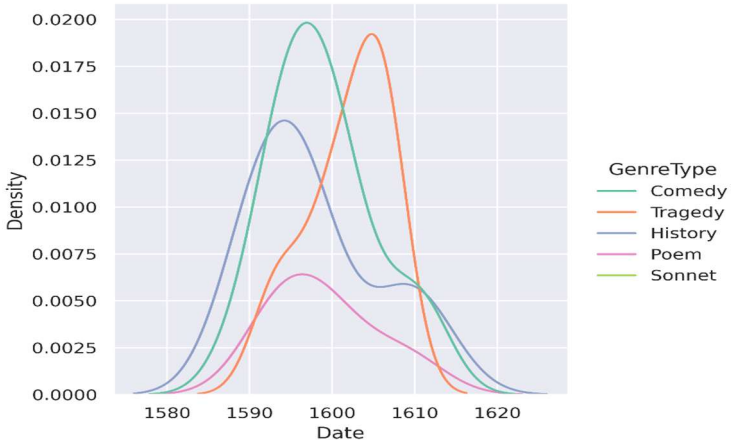


Figura 30 : Estimación de la Densidad de Probabilidad de la cantidad de Publicaciones por Género, agrupación en períodos de 4 años

A1.7 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.

TAREA 2

A2.1 Precision – Recall.

Las métricas ***Precision*** y ***Recall***, también se pueden ver desde el punto de vista de la recuperación de información relevante de los datos, de la siguiente forma:

1) Dado el conjunto de datos para test o información relevante de varias clases. Si miro la información relevante de una clase y la comparo con la información importante recuperada, podría definir esta métrica de comparación como ***Recall***.

   Recall = información relevante recuperada / información relevante.

   Del punto de vista de probabilidad, podríamos verlo como un estimador de una probabilidad condicional: P(Recuperar info/ info relevante)

2) Dado el conjunto de información recuperada, si miro la información relevante de la clase, podría definir esta comparación como ***Precision***.

   Precision = información relevante recuperada / información recuperada.

   Del punto de vista de probabilidad, podríamos verlo como un estimador de una probabilidad condicional: P(Recuperar info relevante / info recuperada)

Puedo definir como prioridad que la tarea me recupere la mayor cantidad de información de una clase X, generalmente esto trae como consecuencia que mucha información no relevante o de otras clases distintas de X también sea recuperadas(etiquetadas como de la clase X: False Positive), mejorando el ***Recall*** (maximizar la recuperación de información relevante) Vs empeora ***Precision*** (alta tasa de False Positive o recuperación de información no relevante).

Lo inverso, solo quiero recuperar información relevante y tener baja tasa de recuperación de información no relevante (baja cuenta de FP), con esto se maximiza el ***Precision***, pero al manejar poca información el ***Recall*** será bajo porque no se recuperará toda la información importante de la clase.

En general uno quiere tener un valor aceptable de ***Recall*** mientras se tolera únicamente un cierto % de Falsos Positivos que generan una baja en el ***Precision***. Una medida de ***trade-off*** que se usa comúnmente es una media harmónica entre estos dos valores: F1 Score = $((1/\alpha)(1/p) +(1/\alpha)(1/r))^{-1}$

A2.2 Validación Cruzada Dejar-Uno-Fuera (Live One Out).

Se construye el modelo entrenando con los n-1 datos y se realiza la predicción para el valor dejado fuera. Entonces se obtiene una pobre estimación de la tasa de error ya que se hizo sobre una sola observación.

Si repetimos el proceso anterior dejando fuera otro dato diferente , volvemos a obtener otra pobre estimación de la tasa de error.

Repetimos esto n veces y hacemos el promedio de todas las tasas de error obtenidas, se obtiene siempre los mismos resultados debido a que no es aleatorio la selección del conjunto de entrenamiento y validación.

A2.3 **Resultados Validación Cruzada – Personajes Cleopatra y Antony.**

A continuación los gráficos de ***Recall*** y ***Precision*** para ***Antony*** y ***Cleopatra***, obtenidos en el proceso de Verificación Cruzada. El experimento que tiene máximo ***Accuracy*** es el experimento 12, se observa que ***Antony***, para estos parámetros, decrece su ***Recall***. Esto está asociado a la mejora en el ***Recall***, tanto en ***Cleopatra*** cómo en ***Queen Margaret***.

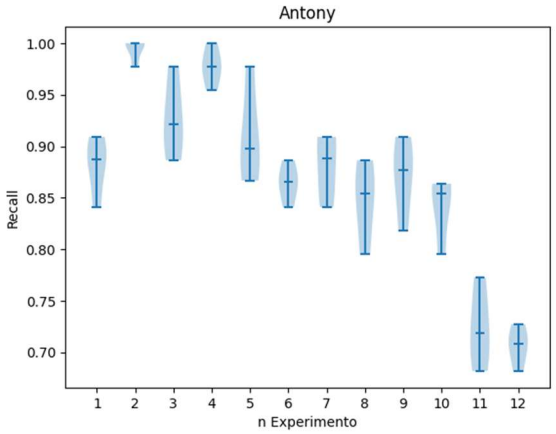


Figura 31 : Recall Antony - Validación Cruzada.

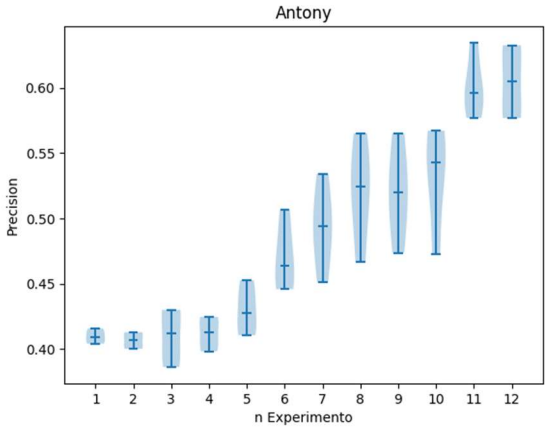


Figura 32 : Precision Antony - Validación Cruzada.

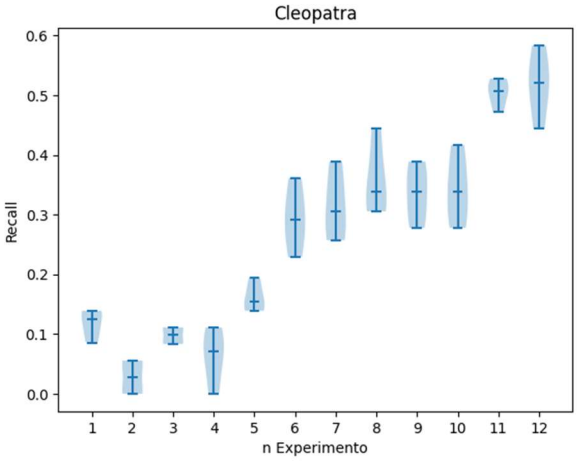


Figura 33 : Recall Cleopatra - Validación Cruzada.

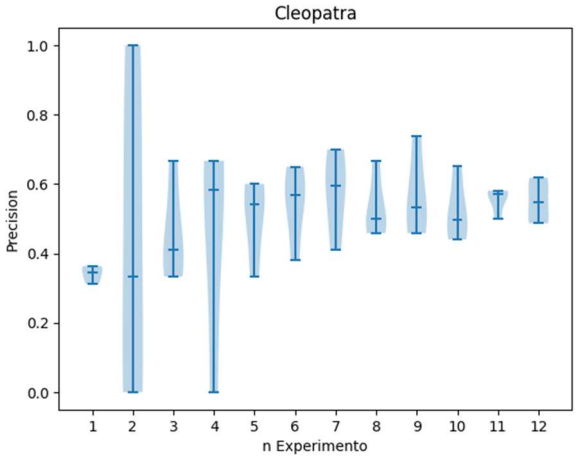


Figura 34 : Precision Cleopatra - Validación Cruzada.

A2.4 Descripción del Clasificador Multinomial Naive Bayes

- El clasificador ***Multinomial Naive Bayes***[19][20] utiliza el teorema de Bayes para clasificar una muestra de entrada ***W***, utilizando la Distribución Multinomial para modelar la generación de los datos. Existirá un modelo por clase o generador de datos.
- Es necesario estimar cuales son los parámetros de la distribución Multinomial, utilizando los datos de entrenamiento, para esto, posterior a tener una realización de ***W***, representada por el vector de datos $[w_1,w_2,....,w_k]$, que en nuestro caso es un párrafo, y considerando el espacio de clases generadoras de datos $Y=[c_1,c_2,....,c_m]$, en nuestro caso ***Y=['Antony', 'Cleopatra', 'Queen Elizabeth']***, para clasificar a ***W*** es necesario evaluar que tan "similar" es el párrafo ***W*** comparado con los párrafos que generan las distintas clases. Esta comparación de similitud se realiza,

  1) Computando las probabilidades condicionales $P(c_1/W), P(c_2/W),.....,P(c_m/W)$: esto se interpreta como: dado un párrafo ***W***, cual es la probabilidad de que haya sido generado por las clases $c_1, c_2,...,c_m$? A esta probabilidad, se le conoce como probabilidad a posteriori, ya que se computa posteriormente a tener la muestra ***W***.

  2) Determinando la clase que computó la máxima probabilidad a posteriori (como una medida de similitud) dado el párrafo ***W***: si a esta clase le llamamos C_p, se puede escribir $C_p = arg\ max\ P(c_j/W)$.
  Para poder computar la probabilidad a posteriori, se utiliza la regla de bayes que relaciona información a priori o conocida con información a posteriori o a determinar.
  La regla de Bayes es la siguiente: $P(C_p/W)=P(C_p,W)/P(W)=P(W/C_p)P(C_p)/P(W)$, por lo tanto podemos reescribir: $C_p = arg\ max\ P(c_j/W) = argmax\ P(W/C_p)P(C_p)/P(W) = argmax\ P(W/C_p)P(C_p) = argmax\ P([w_1,w_2,....,w_k]/Cp)P(Cp)$
  Entonces, debemos conocer a priori la siguiente información:
  - $P([w_1,w_2,....,w_k]/C_p)$: probabilidad que W fue generado por C_p.
  - $P(C_p)$: Probabilidad que el ***W*** esté en C_p, sin conocer nada de ***W***.

  Para estimar $P([w_1,w_2,....,w_k]/C_p)$ es necesario evaluar todas las posibles combinaciones de $[w_1,w_2,....,w_k]$, lo cual requeriría tener mucha información para poder inferir todas las probabilidades, para simplificar el problema el clasificador asume que las variables w_i son independientes (***NAIVE SOLUTION***), por lo tanto $P([w_1,w_2,....,w_k])$ se puede computar como la multiplicación individual de cada $P(w_i)$ y la cantidad de parámetros se reduce radicalmente a ***K*** parámetros, que serían los $P(w_i)$ por cada clase C_j ➔ ***kxm*** parámetros tendrá el sistema. Estos parámetros deben ser estimados del dataset de entrenamiento:
  - $P(c_j)= N_j/N$, donde N_j es la cantidad de párrafos de la clase **j** que hay en todo el dataset y ***N*** es la cantidad de párrafos totales. Esto puede generar algún inconveniente en la predicción de clases minoritarias, ya que la ponderación será baja relativo a clases mayoritarias.
  - Para estimar $P([w_i]/C_j)$: probabilidad de que el modelo C_p genere la palabra w_i, es necesario utilizar un modelo de generación de datos, en este caso se utilizar la Distribución Multinomial para cada clase C_j. Se utiliza el MLE para estimar este parámetro de la distribución, el cual es:
    $P(w_i/C_j) = T_{ji}/\ sum_i(T_{ji})$, donde T_{ji} es la cantidad de veces que la palabra w_i fue generada por la clase C_j, sobre todas las palabras generadas por esa clase. Con este enfoque, si en el dataset de entrenamiento de la clase C_j no aparece la palabra w_i, entonces este parámetro es cero. Si en un párrafo generado por C_j en el test, aparece la palabra w_i, y si el parámetro queda definido como 0, entonces el clasificador computará como probabilidad a posteriori 0 y no predice que esa frase fue generada por la clase C_j. Para evitar esto se agrega un parámetro del clasificador, conocido como ***Laplace smothing =*** α, que suma este valor en el numerador y denominador del estimador: $P(w_i/C_j)$

[19] https://github.com/kingkarlito/machine-learning-books/blob/master/Introduction%20to%20Information%20Retrieval%202009.pdf
[20] https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html

= ($T_{ji} + \alpha$) /($sum_i(T_{ji}) + \alpha$), se puede interpretar como una probabilidad a priori que se le asigna a la ocurrencia de w_i en la clase c_j.

Otra suposición realizada, además de la independencia de las palabras w_i generadas, es la independencia en la posición de generación de la palabra dentro de la frase, por lo tanto el conteo de palabras dentro de un párrafo es independiente de la posición, por lo tanto es posible adoptar el modelo ***bag of Words*** para el conteo de palabras. Esto tiene como desventaja que se pierde la información que es comunicada por el orden de las palabras en los párrafos.

En ocasiones las suposiciones realizadas de independencia condicional de generación de palabras e independencia posicional no son tan verdaderas, ejemplos:

- La palabra ***Hong Kong*** – hay una dependencia fuerte entre estas palabras, casi siempre se encontrarán como ***Hong Kong*** y casi nunca como ***Kong Hong***.

Porque el clasificador tiene un buen desempeño cuando su modelo es una simplificación de la realidad?

- El parámetro utilizado para decidir, la probabilidad condicional, es una multiplicación de muchos factores, lo cual tiende a cero. Y la estimación con respecto al valor esperado es muy mala. Igualmente el resultado obtenido en la estimación da que la clase con mayor probabilidad es la que generó el párrafo, por lo tanto, aunque la estimación de la probabilidad no es buena, si es buena la clasificación, que es lo importante.

- Para el modelo Multinomial, ***W*** es un conjunto de ***K*** variables aleatorias, que se pueden representar como $\{W_1,\ldots\ldots W_k\}$, que cuentan la cantidad de veces que el valor w_i es observado en un ensayo de n pruebas. Para la clasificación de texto, la variable aleatoria W_i es la Palabra ***i***, por lo tanto w_i es la cantidad de veces que esa palabra aparece en un ensayo (párrafo) de ***n*** (cantidad de palabras en el párrafo) pruebas. El modelo también asume que las ***K*** variables aleatorias son independientes.

En nuestro ejemplo $X=[w_1,w_2,\ldots..,w_k]$ es un párrafo, cada componente tiene el conteo de las palabras encontradas en ese párrafo (***k=2821***: cantidad de palabras encontradas en todos los párrafos de entrenamiento), y la suma de esos conteos debería dar n, que sería el tamaño del ensayo. En este caso n es variable, porque la longitud de los párrafos puede ser distinta. Normalizando el conteo, lo que significa dividir cada componente por la suma de todos los conteos, pasar a frecuencia, alcanzaría para que ***n*** fuera fijo e igual a 1. Entonces, Podemos decir que para nuestro modelo ***bag-of-word*** , un párrafo se puede modelar con una distribución multinomial de parámetros ***k=2821*** y ***n=1***. Cada componente xi o palabra debe tener asociada una probabilidad de ocurrencia p_i.