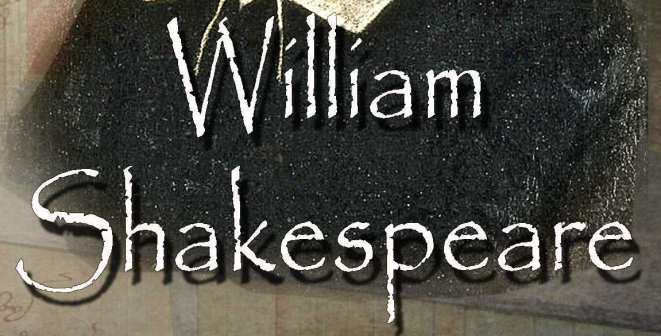




Tareas 1 y 2 SHAKESPEARE



William
Shakespeare

Curso de Introducción a la Ciencia de Datos 2023

Lic.Ec.Federico Bentos.

C.I.: 3.991.400-0

Grupo: 20



RESUMEN

Este trabajo se centra en el Análisis de la Obra de Shakespeare bajo una mirada de ciencia de datos y procesamiento del lenguaje natural, realizando limpieza de datos, análisis exploratorios, análisis descriptivos, análisis multivariados de reducción de dimensionalidad, y entrenamiento de diferentes modelos de clasificación con el fin de conocer y poder predecir a partir de un texto a que personaje de la Obra del Autor pertenecen.

ENTREGABLE TAREA 1

Parte 1: Cargado y limpieza de datos

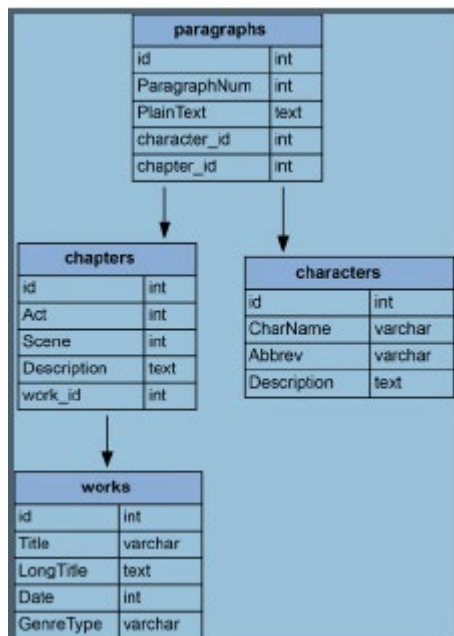
Se probó las primeras celdas del Notebook entregado y funcionaron correctamente.

Se cargaron todas las tablas (4) con la función: `load_table(table_name, engine)`.

La Base de Datos Relacional de la Obra de Shakespeare, brindada en el curso se compone de 4 tablas:

1. Tabla Works: Con los registros de cada Obra realizada por Shakespeare.
`df_works = load_table("works", engine)`
2. Tabla Chapters: Con los Capítulos de cada una de la Obras de Shakespeare, cada Obra se compone de N Capítulos, y cada Capítulo pertenece a una sola Obra.
`df_chapters = load_table("chapters", engine)`
3. Tabla Paragraphs: Con todos los Párrafos de la Obra de Shakespeare, cada Capítulo contiene N Párrafos, y cada Párrafo pertenece a un solo Capítulo.
`df_paragraphs = load_table("paragraphs", engine)`
4. Tabla Characters: Con todos los Personajes de la Obra de Shakespeare, cada Párrafo esta asociado a un Único Personaje y cada Personaje puede tener N Párrafos, en particular estar incluidos en diferentes Obras.

`df_characters = load_table("characters", engine)`



Se realizó Join entre las diferentes Tablas, para extraer la información relevante y se generó un DataFrame de pandas con registros por Palabras (Words) de cada Párrafo, de cada Capítulo, de cada Obra, para el Personaje Asociado.

Calidad de Datos

Respecto a Datos Faltantes, se evaluaron las 4 tablas Descriptas con la librería pandas a través de los DataFrames contruidos:

Los únicos valores faltantes o nulos se registran en el campo descripción de la tabla de Personajes.

1. Tabla de Faltantes:

Tabla de Obras: Faltantes	Tabla de Capítulos: Faltantes	Tabla de Párrafos: Faltantes	Tabla de Personajes: Faltantes	Tabla de Palabras: Faltantes																																																																																												
<table><tr><td></td><td>0</td></tr><tr><td>id</td><td>0</td></tr><tr><td>Title</td><td>0</td></tr><tr><td>LongTitle</td><td>0</td></tr><tr><td>Date</td><td>0</td></tr><tr><td>GenreType</td><td>0</td></tr></table>		0	id	0	Title	0	LongTitle	0	Date	0	GenreType	0	<table><tr><td></td><td>0</td></tr><tr><td>id</td><td>0</td></tr><tr><td>Act</td><td>0</td></tr><tr><td>Scene</td><td>0</td></tr><tr><td>Description</td><td>0</td></tr><tr><td>work_id</td><td>0</td></tr></table>		0	id	0	Act	0	Scene	0	Description	0	work_id	0	<table><tr><td></td><td>0</td></tr><tr><td>id_paragraphs</td><td>0</td></tr><tr><td>ParagraphNum</td><td>0</td></tr><tr><td>PlainText</td><td>0</td></tr><tr><td>character_id</td><td>0</td></tr><tr><td>chapter_id</td><td>0</td></tr><tr><td>text_expanded</td><td>0</td></tr><tr><td>CleanText</td><td>0</td></tr><tr><td>WordList</td><td>0</td></tr><tr><td>id_characters</td><td>0</td></tr><tr><td>CharName</td><td>0</td></tr></table>		0	id_paragraphs	0	ParagraphNum	0	PlainText	0	character_id	0	chapter_id	0	text_expanded	0	CleanText	0	WordList	0	id_characters	0	CharName	0	<table><tr><td></td><td>0</td></tr><tr><td>id</td><td>0</td></tr><tr><td>CharName</td><td>0</td></tr><tr><td>Abbrev</td><td>5</td></tr><tr><td>Description</td><td>646</td></tr></table>		0	id	0	CharName	0	Abbrev	5	Description	646	<table><tr><td></td><td>0</td></tr><tr><td>id_words</td><td>0</td></tr><tr><td>ParagraphNum</td><td>0</td></tr><tr><td>character_id</td><td>0</td></tr><tr><td>chapter_id</td><td>0</td></tr><tr><td>word</td><td>0</td></tr><tr><td>id_characters</td><td>0</td></tr><tr><td>CharName</td><td>0</td></tr><tr><td>id_chapters</td><td>0</td></tr><tr><td>Act</td><td>0</td></tr><tr><td>Scene</td><td>0</td></tr><tr><td>Description</td><td>0</td></tr><tr><td>work_id</td><td>0</td></tr><tr><td>Title</td><td>0</td></tr><tr><td>LongTitle</td><td>0</td></tr><tr><td>Date</td><td>0</td></tr><tr><td>GenreType</td><td>0</td></tr><tr><td>unos</td><td>0</td></tr></table>		0	id_words	0	ParagraphNum	0	character_id	0	chapter_id	0	word	0	id_characters	0	CharName	0	id_chapters	0	Act	0	Scene	0	Description	0	work_id	0	Title	0	LongTitle	0	Date	0	GenreType	0	unos	0
	0																																																																																															
id	0																																																																																															
Title	0																																																																																															
LongTitle	0																																																																																															
Date	0																																																																																															
GenreType	0																																																																																															
	0																																																																																															
id	0																																																																																															
Act	0																																																																																															
Scene	0																																																																																															
Description	0																																																																																															
work_id	0																																																																																															
	0																																																																																															
id_paragraphs	0																																																																																															
ParagraphNum	0																																																																																															
PlainText	0																																																																																															
character_id	0																																																																																															
chapter_id	0																																																																																															
text_expanded	0																																																																																															
CleanText	0																																																																																															
WordList	0																																																																																															
id_characters	0																																																																																															
CharName	0																																																																																															
	0																																																																																															
id	0																																																																																															
CharName	0																																																																																															
Abbrev	5																																																																																															
Description	646																																																																																															
	0																																																																																															
id_words	0																																																																																															
ParagraphNum	0																																																																																															
character_id	0																																																																																															
chapter_id	0																																																																																															
word	0																																																																																															
id_characters	0																																																																																															
CharName	0																																																																																															
id_chapters	0																																																																																															
Act	0																																																																																															
Scene	0																																																																																															
Description	0																																																																																															
work_id	0																																																																																															
Title	0																																																																																															
LongTitle	0																																																																																															
Date	0																																																																																															
GenreType	0																																																																																															
unos	0																																																																																															

2. Tabla de Nulos:

Tabla de Obras: Nulos	Tabla de Capítulos: Nulos	Tabla de Párrafos: Nulos	Tabla de Personajes: Nulos	Tabla de Palabras: Nulos																																																																																												
<table><tr><td></td><td>0</td></tr><tr><td>id</td><td>0</td></tr><tr><td>Title</td><td>0</td></tr><tr><td>LongTitle</td><td>0</td></tr><tr><td>Date</td><td>0</td></tr><tr><td>GenreType</td><td>0</td></tr></table>		0	id	0	Title	0	LongTitle	0	Date	0	GenreType	0	<table><tr><td></td><td>0</td></tr><tr><td>id</td><td>0</td></tr><tr><td>Act</td><td>0</td></tr><tr><td>Scene</td><td>0</td></tr><tr><td>Description</td><td>0</td></tr><tr><td>work_id</td><td>0</td></tr></table>		0	id	0	Act	0	Scene	0	Description	0	work_id	0	<table><tr><td></td><td>0</td></tr><tr><td>id_paragraphs</td><td>0</td></tr><tr><td>ParagraphNum</td><td>0</td></tr><tr><td>PlainText</td><td>0</td></tr><tr><td>character_id</td><td>0</td></tr><tr><td>chapter_id</td><td>0</td></tr><tr><td>text_expanded</td><td>0</td></tr><tr><td>CleanText</td><td>0</td></tr><tr><td>WordList</td><td>0</td></tr><tr><td>id_characters</td><td>0</td></tr><tr><td>CharName</td><td>0</td></tr></table>		0	id_paragraphs	0	ParagraphNum	0	PlainText	0	character_id	0	chapter_id	0	text_expanded	0	CleanText	0	WordList	0	id_characters	0	CharName	0	<table><tr><td></td><td>0</td></tr><tr><td>id</td><td>0</td></tr><tr><td>CharName</td><td>0</td></tr><tr><td>Abbrev</td><td>5</td></tr><tr><td>Description</td><td>646</td></tr></table>		0	id	0	CharName	0	Abbrev	5	Description	646	<table><tr><td></td><td>0</td></tr><tr><td>id_words</td><td>0</td></tr><tr><td>ParagraphNum</td><td>0</td></tr><tr><td>character_id</td><td>0</td></tr><tr><td>chapter_id</td><td>0</td></tr><tr><td>word</td><td>0</td></tr><tr><td>id_characters</td><td>0</td></tr><tr><td>CharName</td><td>0</td></tr><tr><td>id_chapters</td><td>0</td></tr><tr><td>Act</td><td>0</td></tr><tr><td>Scene</td><td>0</td></tr><tr><td>Description</td><td>0</td></tr><tr><td>work_id</td><td>0</td></tr><tr><td>Title</td><td>0</td></tr><tr><td>LongTitle</td><td>0</td></tr><tr><td>Date</td><td>0</td></tr><tr><td>GenreType</td><td>0</td></tr><tr><td>unos</td><td>0</td></tr></table>		0	id_words	0	ParagraphNum	0	character_id	0	chapter_id	0	word	0	id_characters	0	CharName	0	id_chapters	0	Act	0	Scene	0	Description	0	work_id	0	Title	0	LongTitle	0	Date	0	GenreType	0	unos	0
	0																																																																																															
id	0																																																																																															
Title	0																																																																																															
LongTitle	0																																																																																															
Date	0																																																																																															
GenreType	0																																																																																															
	0																																																																																															
id	0																																																																																															
Act	0																																																																																															
Scene	0																																																																																															
Description	0																																																																																															
work_id	0																																																																																															
	0																																																																																															
id_paragraphs	0																																																																																															
ParagraphNum	0																																																																																															
PlainText	0																																																																																															
character_id	0																																																																																															
chapter_id	0																																																																																															
text_expanded	0																																																																																															
CleanText	0																																																																																															
WordList	0																																																																																															
id_characters	0																																																																																															
CharName	0																																																																																															
	0																																																																																															
id	0																																																																																															
CharName	0																																																																																															
Abbrev	5																																																																																															
Description	646																																																																																															
	0																																																																																															
id_words	0																																																																																															
ParagraphNum	0																																																																																															
character_id	0																																																																																															
chapter_id	0																																																																																															
word	0																																																																																															
id_characters	0																																																																																															
CharName	0																																																																																															
id_chapters	0																																																																																															
Act	0																																																																																															
Scene	0																																																																																															
Description	0																																																																																															
work_id	0																																																																																															
Title	0																																																																																															
LongTitle	0																																																																																															
Date	0																																																																																															
GenreType	0																																																																																															
unos	0																																																																																															

3. Tabla de Duplicados:

Tabla de Obras: Duplicados	Tabla de Capítulos: Duplicados	Tabla de Párrafos: Duplicados	Tabla de Personajes: Duplicados	Tabla de Palabras: Duplicados
0	0	0	0	175270

No se encuentran tampoco registros duplicados en las Tablas originales de la base de datos, y solo se encuentran duplicados en la tabla de Palabras ("words"), pero no son un problema de calidad de datos.

Solamente se trabajó en limpieza y transformación de Datos, para la construcción del DataFrame de Palabras, debido a la cantidad de caracteres incluidos entre las palabras, como signos de puntuación, pregunta, exclamación, separación, etc. Dicho problema fue atacado con la función `clean_text()` la cual se amplió a los caracteres faltantes. Además de los problemas de puntuación, nos encontramos con las contracciones en Ingles. Para este último problema se resolvió utilizar la librería **pycontractions** creándose la función: **expand_contractions()** a partir de aplicar el método "expand_contractions" de dicha librería. La expansión no es tan simple ya que requiere conocimiento contextual para elegir las palabras correctas de reemplazo en el Párrafo, esto implica bastante tiempo de cómputo.

La librería usa un enfoque de tres pasos. Primero, las contracciones simples con una sola regla se reemplazan. En el segundo paso, si hay contracciones con múltiples reglas, se procede a reemplazar todas las combinaciones de reglas para producir todos los textos posibles. Cada texto se pasa luego por un corrector gramatical y se calcula la distancia del movimiento de palabras (WMD) entre él y el texto original. Las hipótesis se ordenan por menor número

de errores gramaticales y menor distancia del texto original y se devuelve la hipótesis superior como la forma expandida. El conteo de errores gramaticales elimina las peores opciones, pero hay muchos casos que no tienen o tienen el mismo número de errores gramaticales. En estos casos, el WMD funciona como el desempate. El WMD es el costo acumulativo ponderado mínimo requerido para mover todas las palabras del texto original a cada hipótesis. Esto aprovecha el modelo vectorial semántico subyacente elegido, como Word2Vec, GloVe o FastText. Como la diferencia entre cada hipótesis es solo el reemplazo de una contracción con su expansión, la hipótesis “más cercana” al texto original será la que tenga la mínima distancia euclidiana entre el par de palabras de contracción y expansión en el espacio de incrustación. - (AI) Bing-CHAT.

Luego de la primer presentación el grupo de corrección noto que también existían instrucciones del “stage directions” dentro de los propios párrafos de los personajes, por lo cual se decidió crear una nueva limpieza, realizando la sustracción de estas instrucciones del stage directions dentro de los párrafos de los personajes a través de expresiones regulares, quitando “[. * ? \]” con el módulo de Python “re”.

En resumen, se expandió las contracciones con **pycontractions** en cada Párrafo, se convirtió a minúsculas y se reemplazaron los signos de puntuación por el espacio, y se eliminaron las instrucciones del stage directions en los párrafos de los personajes para entrenar posteriormente los modelos de aprendizaje, tal cual fue sugerido en el curso.

Personajes con más Párrafos:

Respecto a los Personajes con más párrafos, aparecen primero “**Stage Directions**” y “**Poet**” que no son personajes propiamente, el primer caso son las instrucciones de dirección en la mayoría de las Obras que no son Poemas o Sonetos, y el segundo que aparece es justamente quien se asocia a estos 2 últimos géneros mencionados.

	Count
CharName	
(stage directions)	3751
Poet	766
Falstaff	471
Henry V	377
Hamlet	358
Duke of Gloucester	285
Othello	274
Iago	272
Antony	253
Richard III	246

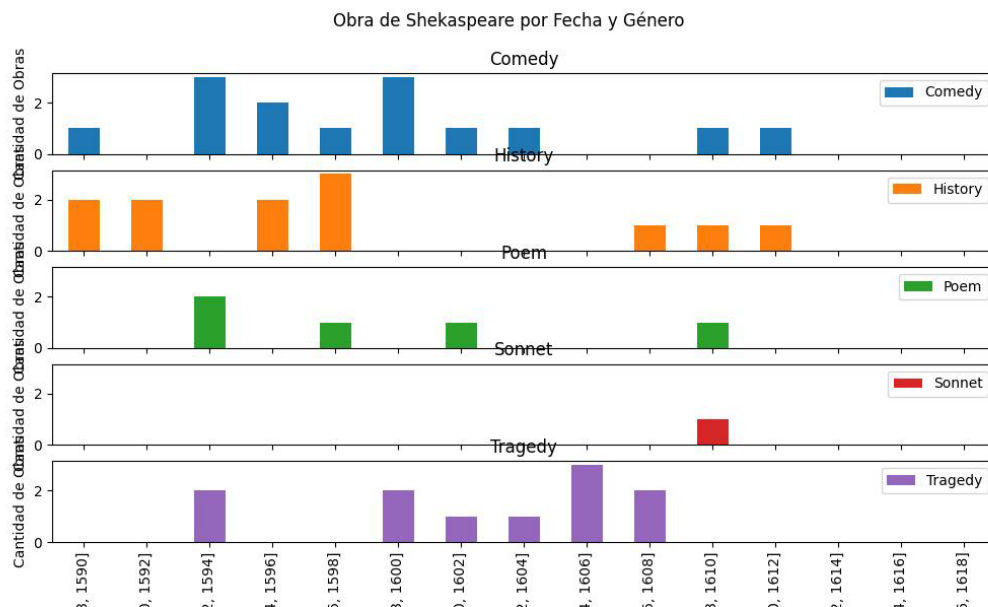
El personaje con más Párrafos, teniendo en cuenta las anteriores aclaraciones es **Falstaff**. Según Wikipedia, Sir John Falstaff es un personaje de ficción creado por el dramaturgo inglés William Shakespeare. Aparece en tres obras de Shakespeare y recibe un elogio en una cuarta. Su importancia como personaje plenamente desarrollado se formó principalmente en las obras Enrique IV, 1ª parte y 2ª parte, donde es compañero del príncipe Hal, el futuro rey Enrique V de Inglaterra. Un notable elogio de Falstaff se presenta en el Acto II, Escena III de Enrique V, donde Falstaff no aparece como personaje en escena, sino que su muerte es narrada por la señora Quickly en términos que algunos estudiosos han atribuido a la descripción que hiciese Platón de la muerte de Sócrates tras beber cicuta. En comparación, Falstaff es presentado como el bufonesco pretendiente de dos mujeres casadas en Las alegres comadres de Windsor.

Luego en un siguiente grupo lo siguen “**Enrique V**” y “**Hamlet**”, luego en un tercer grupo aparecen “**Duke of Gloucester**”, “**Othello**”, “**Iago**”, “**Antony**” y “**Richard III**”.

Obra de Shakespeare a lo largo del tiempo en cuanto a Géneros:

En el comienzo de la Obra de Shakespeare se puede encontrar 2 ciclos de creación de Obras en el Género de **Historia** alternado con **Comedia**. Luego en el final de su Obra se observa un Ciclo de creación de Obras en el Género de la **Tragedia**.

Hay un Ciclo intermedio en el que su Obra disminuye en cantidades. Si bien su creación de **Poemas** y **Sonetos** no fue tan cuantiosa, se observa distribuida a lo largo de su Obra.



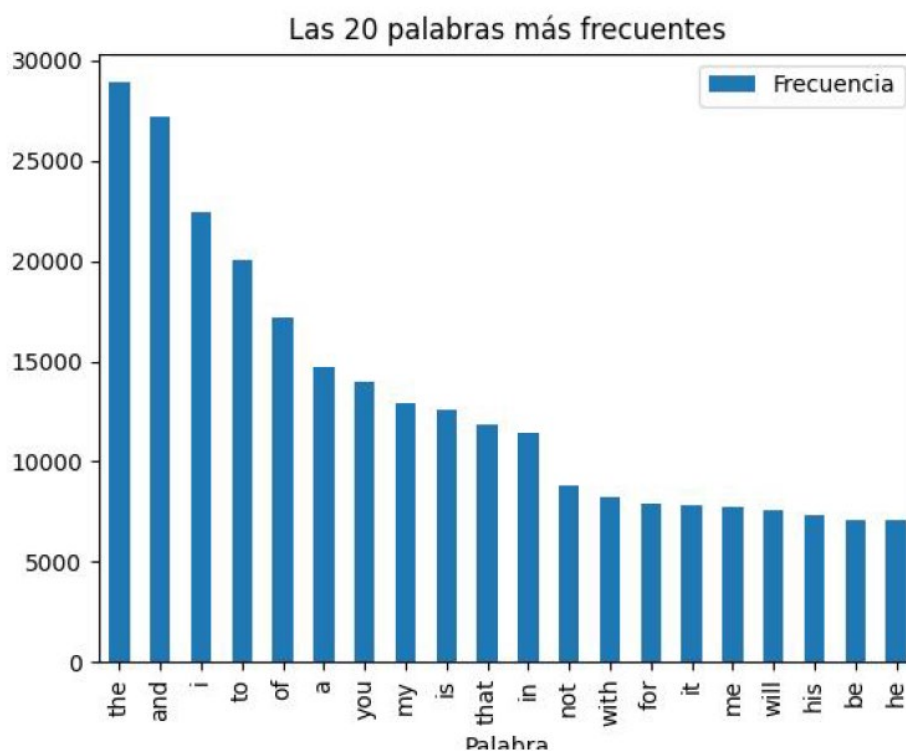
Conteo de Palabras (df_words):

Para la creación de la Tabla de Palabras se partió de la tabla de Párrafos, se expandieron las contracciones del inglés con la librería **pycontractions** como se explicó anteriormente, luego se pasaron a minúsculas todos los caracteres, luego se corrigieron los signos de puntuación con la función `clean_text()`. Luego se realizó un `Split` sobre el campo "CleanText" asignándolo a la columna "WordList" dividiendo cada elemento de la lista. Por último se hizo un `explode()` sobre ese campo para crear un registro para cada palabra y en el DataFrame de Palabras se eliminaron los campos; "CleanText", "PlainText", "text_expanded" y se renombró la columna "WordList" por "word".

Parte 2: Conteo de palabras y visualizaciones

Palabras más frecuentes considerando toda la obra:

Si analizamos las palabras más utilizadas en la Obra de Shakespeare, encontramos que obviamente la Frecuencia de las 20 palabras más utilizadas refieren a artículos y palabras de conjunción.



Comedy History



Podrían también plantearse visualizaciones dinámicas utilizando librerías como **pivottablejs** para explorar en mayor profundidad los datos con *Tablas Dinámicas* o sugerir utilizar Nubes de Palabras de modo tridimensional.

Personajes con mayor cantidad de Palabras:

Cantidad de Palabras por Personaje
Palabras por Personaje Palabras Únicas por Personaje

("", 'Personajes')	("", 'Cantidades de Palabras')	("", 'Personajes')	("", 'Cantidades de Palabras Únicas')
Poet	49495	Poet	7537
(stage directions)	16181	Henry V	3214
Henry V	15159	Falstaff	2907
Falstaff	14614	Hamlet	2799
Hamlet	12033	Duke of Gloucester	2308
Duke of Gloucester	9342	Henry IV	2201
Antony	8667	Antony	2104
Iago	8516	Iago	2040
Henry IV	8230	Queen Margaret	1922
Vincenzio	6998	(stage directions)	1917

En este caso podemos observar como en los Párrafos que los Personajes con más palabras son **“Poet”** y **“Stage Directions”**. Que no son Personajes propiamente dichos y refieren a Obras asociadas a los *Poemas* y *Sonetos* en el primer caso y en el resto de las Obras al segundo. Si consideramos la mayor cantidad de palabras únicas por cada Personaje vemos que **“Stage Directions”** cae en el ranking debido a que se repiten con mucha frecuencia los mismos tipos de órdenes y textos, para el caso de **“Poet”** no se verifica esto debido a que se trata de la voz de los *Poemas* y *Sonetos*.

Teniendo en cuenta estas consideraciones se podrían quitar de los análisis estos pseudo Personajes y encontraríamos en dichos rankings a Personajes como **“Enrique V”**, **“Falstaff”**, etc.

Preguntas y análisis posibles a partir de estos datos:

Podrían a partir de estos datos textuales, a partir de palabras y párrafos, realizar análisis de sentimientos, predicción de Personajes en base a posibles textos expresados o Párrafos de la Obra de Shakespeare, caracterizar a los Personajes y Clasificarlos en base a nuevas categorías a partir de análisis de clúster.

Podrían responderse preguntas como:

1. ¿Qué obras tienen más personajes y qué relación hay entre el número de personajes y la longitud de la obra?
2. ¿Qué palabras son las más frecuentes en cada obra y qué temas o emociones reflejan?
3. ¿Qué personajes tienen más líneas y qué papel desempeñan en la trama?
4. ¿Qué obras tienen más escenas y cómo se distribuyen las escenas entre los actos?
5. ¿Qué obras tienen más variación lingüística y qué factores pueden influir en ello?

Hay muchas más posibilidades de explorar los datos con técnicas de análisis cuantitativo o cualitativo. Por ejemplo, se podría usar el análisis de frecuencias, el análisis de contenido, el análisis de redes o el análisis de sentimientos.

ENTREGABLE TAREA 2

Parte 1: Dataset y representación numérica de texto

Se crea un Dataset reducido de sólo 3 personajes; "Antony", "Cleopatra", "Queen Margaret".

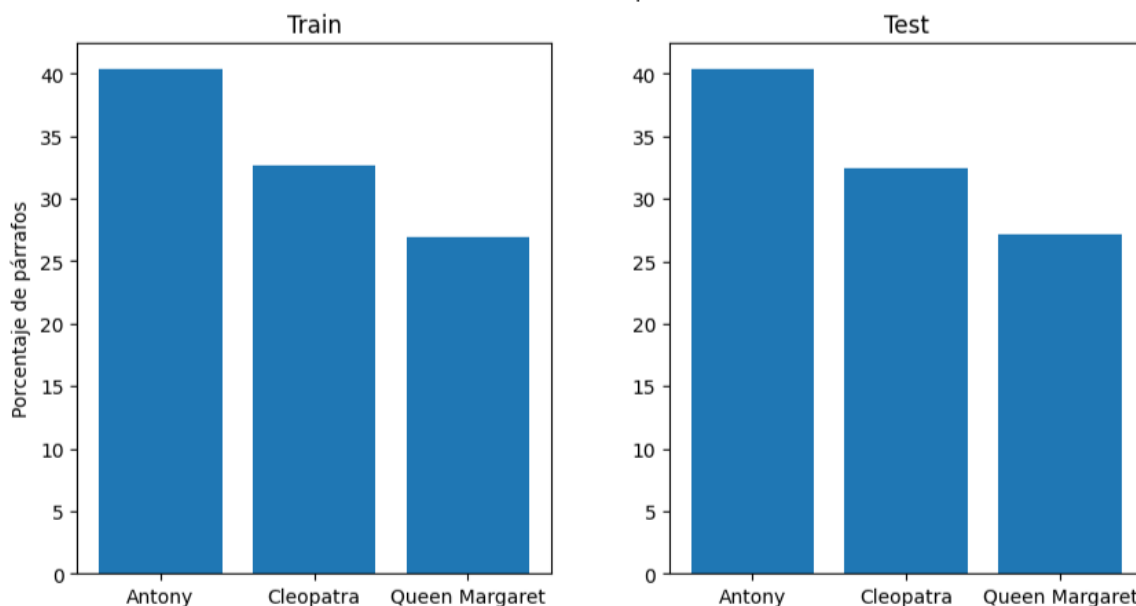
Se crea un conjunto de test del 30% del total, utilizando muestreo estratificado:

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=0, stratify=y)
```

Tamaños de Train/Test: 438/188

El balance de párrafos de cada personaje es similar en train y test:

Distribución de párrafos



Bag of Words:

La técnica de conteo de palabras o bag of words es una forma de representar textos como vectores numéricos, donde cada elemento del vector corresponde a la frecuencia de una palabra en el texto. Esta técnica es útil para extraer características de los textos que se pueden usar en modelos de aprendizaje automático, como la clasificación de documentos.

La técnica funciona de la siguiente manera:

1. Se toma una colección de documentos (por ejemplo, un conjunto de entrenamiento) y se extraen todas las palabras distintas que aparecen en ellos. Estas palabras forman el vocabulario del modelo.
2. Se asigna un índice a cada palabra del vocabulario, de forma que se pueda identificar cada palabra con un número.
3. Se toma cada documento y se cuenta el número de veces que aparece cada palabra del vocabulario en él. Estos conteos se almacenan en un vector, donde cada posición corresponde al índice de una palabra del vocabulario. Así, se obtiene una representación numérica del documento basada en la ocurrencia de las palabras.

La matriz resultante tiene tantas filas como documentos y tantas columnas como palabras en el vocabulario. Esta matriz es una **sparse matrix** o matriz dispersa porque tiene muchos valores cero. Esto se debe a que normalmente el vocabulario es muy grande y cada documento contiene solo una pequeña fracción de las palabras posibles. Por lo tanto, se suelen usar técnicas de compresión o reducción de dimensionalidad para almacenar o procesar estas matrices de forma más eficiente.

N-Grama:

Un n-grama es una secuencia contigua de n elementos de un texto o un discurso. Los elementos pueden ser caracteres, palabras o frases, y n puede ser cualquier número entero. Por ejemplo, cuando n es 2, se llama bigrama

a la secuencia de dos elementos. De forma similar, una secuencia de tres elementos se llama trígrama, y así sucesivamente. Los n-gramas se usan para modelar el lenguaje y extraer características de los textos.

TF-IDF:

La representación numérica Term Frequency - Inverse Document Frequency (TF-IDF) es una forma de ponderar las palabras de un texto según su importancia relativa. Esta transformación se basa en dos conceptos:

1. Term Frequency (TF): Es la frecuencia de una palabra en un documento. Se calcula como el número de veces que aparece la palabra en el documento dividido por el número total de palabras en el documento. Esto refleja cuán relevante es una palabra para un documento en un corpus.
2. Inverse Document Frequency (IDF): Es la medida inversa de la frecuencia de una palabra en un corpus. Se calcula como el logaritmo del número total de documentos en el corpus dividido por el número de documentos que contienen la palabra. Esto refleja cuán común o rara es una palabra en todo el corpus.

La puntuación TF-IDF se obtiene multiplicando los valores de TF e IDF para cada palabra. Esto da como resultado un peso mayor para las palabras que son más importantes para un documento específico, pero que no aparecen con frecuencia en el corpus, lo que reduce el efecto de las palabras comunes como artículos o preposiciones.

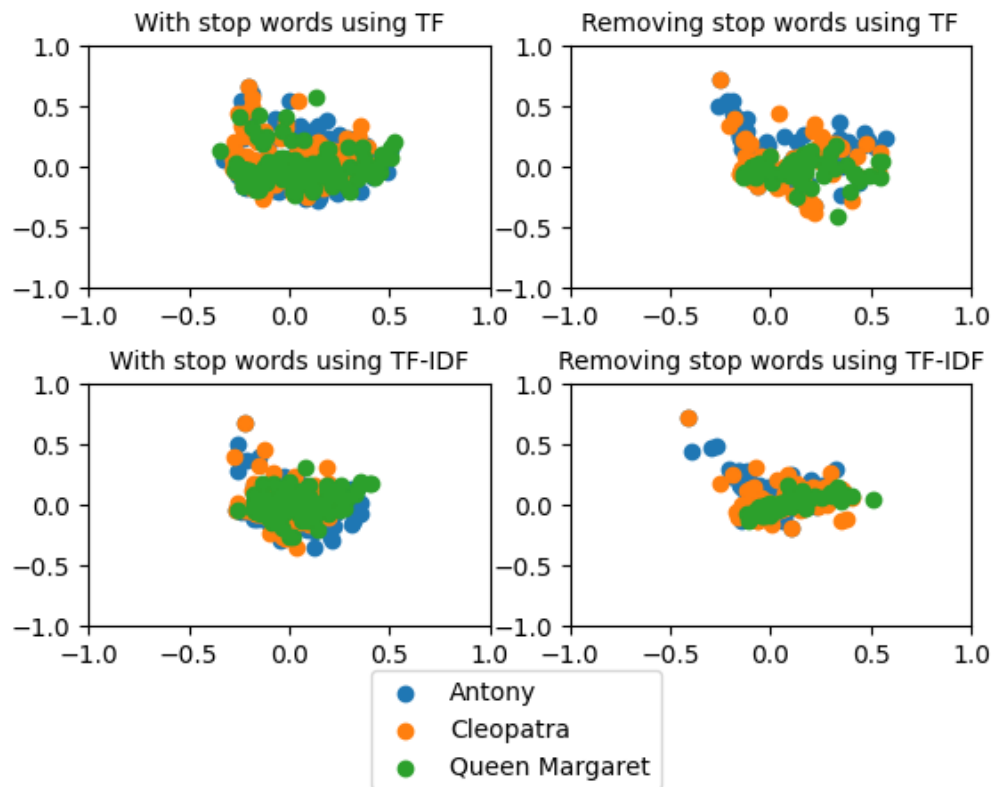
PCA (Análisis de Componentes Principales):

La técnica PCA (Principal Component Analysis) es una técnica estadística para reducir la dimensionalidad de un conjunto de datos. Consiste en transformar linealmente los datos a un nuevo sistema de coordenadas donde la mayor parte de la variación de los datos se puede describir con menos dimensiones que los datos iniciales. Esto permite simplificar y visualizar los datos, así como eliminar el ruido y la redundancia.

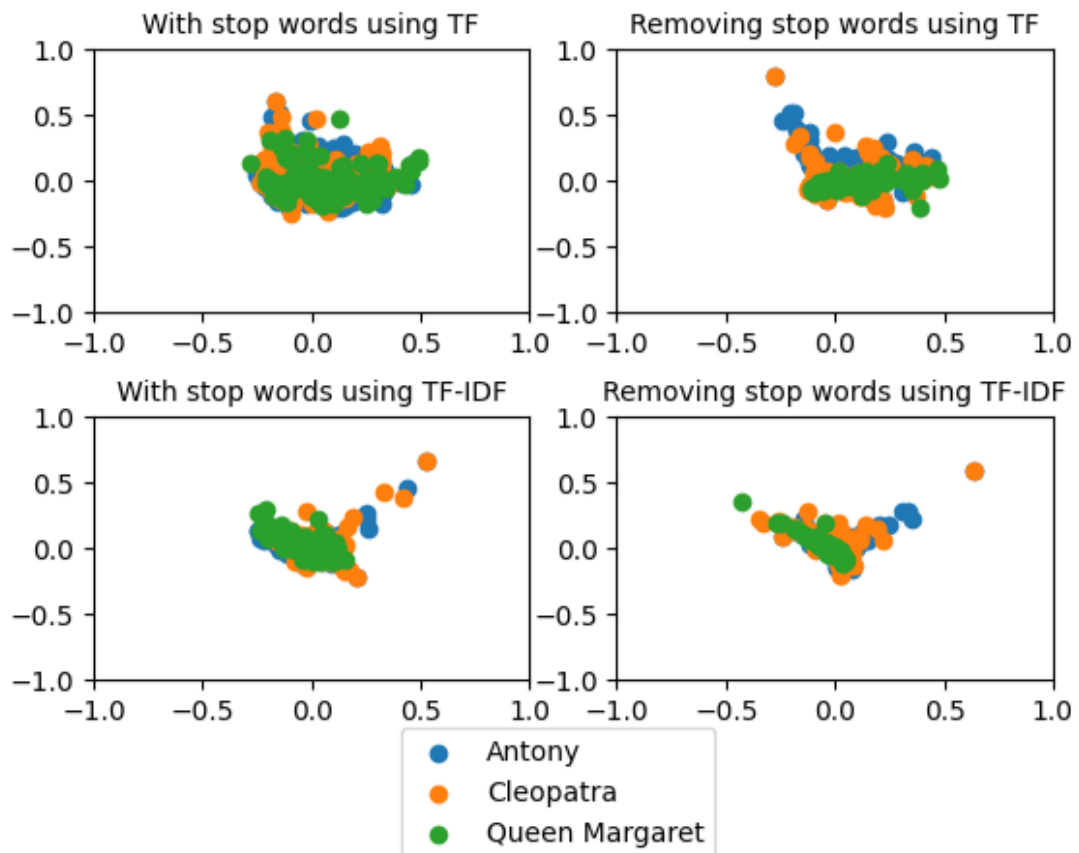
La técnica PCA funciona de la siguiente manera:

1. Se calcula la matriz de covarianza de los datos, que mide las relaciones lineales entre las variables originales.
2. Se calculan los vectores y valores propios de la matriz de covarianza, que representan las direcciones y magnitudes de la máxima variación de los datos, respectivamente.
3. Se ordenan los vectores y valores propios de mayor a menor según su importancia, y se seleccionan los primeros k vectores propios, donde k es el número de dimensiones deseadas. Estos vectores propios se llaman componentes principales (PCs).
4. Se proyectan los datos originales sobre el espacio generado por los componentes principales, obteniendo así una nueva representación de los datos con menos dimensiones.
5. Los Componentes Principales son combinaciones lineales de las variables originales, donde por su construcción son ortogonales entre sí, siendo LI.

PCA por Personaje tomando unigramas en el conjunto de entrenamiento



PCA por Personaje tomando bigramas en el conjunto de entrenamiento



Analizando PCA tanto en Unigramas, como en Bigramas, el dejar las “**stop words**” dificulta la discriminación en el plano factorial de los 3 personajes escogidos. Los Datos parecen seguir una estructura mas clara de representación en el plano factorial cuando se remueven las “**stop words**”, esto sin duda mejora aún mas al utilizar la técnica TF-IDF que al no incluir idf (=false). Esto se verifica tanto para Unigramas como Bigramas, aunque con esto último parece ser mas claro la diferenciación dentro del plano factorial.

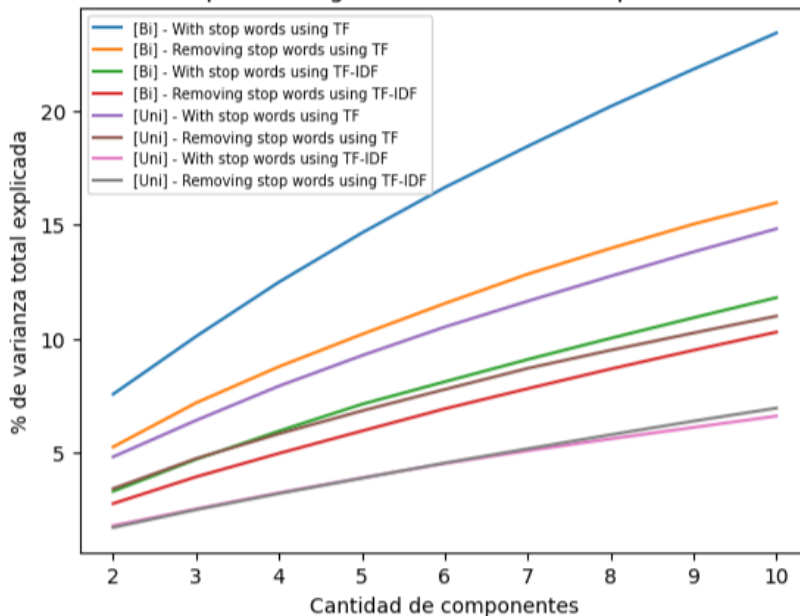
De todos modos, la varianza capturada por estos PCA individualmente considerados es muy baja en todos los casos, esto se ejemplifica en los cálculos de las siguientes tablas para Unigramas y Bigramas mostrados en los gráficos anteriores:

Varianza capturada por PCA Nº1 in unigramas - 7.57%
 Varianza capturada por PCA Nº2 in unigramas - 5.25%
 Varianza capturada por PCA Nº3 in unigramas - 3.3%
 Varianza capturada por PCA Nº4 in unigramas - 2.77%

Varianza capturada por PCA Nº1 in bigramas - 4.82%
 Varianza capturada por PCA Nº2 in bigramas - 3.42%
 Varianza capturada por PCA Nº3 in bigramas - 1.79%
 Varianza capturada por PCA Nº4 in bigramas - 1.71%

Considerando como aumenta la varianza capturada por los primeros 10 componentes principales encontramos que en ninguno de los casos se logra capturar mucho más del 25% de la varianza total:

Varianza explicada según la cantidad de componentes de PCA



Por todo esto resulta muy difícil separar los personajes utilizando solamente 2 componentes principales, debido a la gran dimensionalidad del problema y la utilización de técnicas basadas en matrix sparse que son poco eficientes y utilizan como se explicó, niveles de dimensionalidad muy alta.

Parte 2: Entrenamiento y Evaluación de Modelos

Entrene el modelo Multinomial Naive Bayes:

	precision	recall	f1-score	support
Antony	0.429	0.961	0.593	76
Cleopatra	0.667	0.131	0.219	61
Queen Margaret	0.833	0.098	0.175	51
accuracy			0.457	188
macro avg	0.643	0.397	0.329	188
weighted avg	0.616	0.457	0.359	188

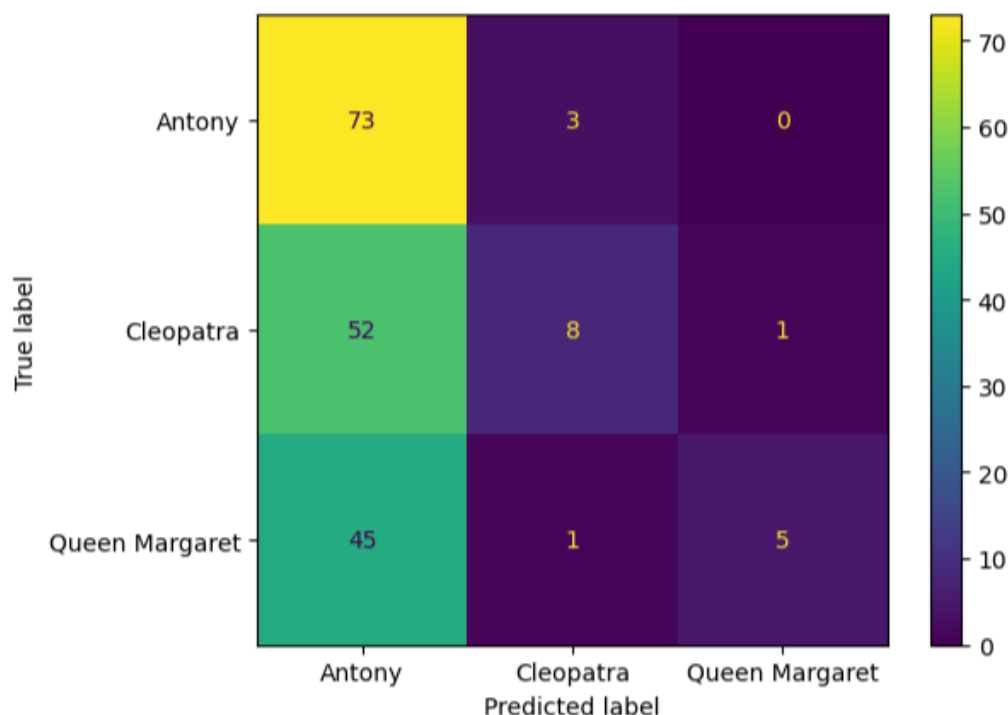
El valor de accuracy (0.46) es una medida que resume el rendimiento de un modelo de clasificación como el número de predicciones correctas dividido por el número total de predicciones. Es fácil de calcular e intuitivo de entender, por lo que es la medida más común para evaluar los modelos clasificadores.

Sin embargo, el valor de accuracy puede ser engañoso cuando el conjunto de datos tiene un desbalance de clases, es decir, cuando la distribución de las clases es muy desigual. Por ejemplo, si tenemos un conjunto de datos con

un soporte de 188 donde 76 pertenecen a la clase “Antony” y muchos menos a las clases “Cleopatra” y “Queen Margaret”, y el modelo predice siempre la clase “Antony”, tendrá un accuracy de más del 40%, pero no será capaz de clasificar correctamente ningún ejemplo de las clases minoritarias. Esto implica que el valor de accuracy no refleja adecuadamente la capacidad del modelo para discriminar entre las diferentes clases, especialmente cuando la clase minoritaria es la más relevante o interesante.

Si el desbalance de clases fuera aún mayor, el valor de accuracy sería aún más inadecuado para medir el rendimiento del modelo.

Por estas razones, se recomienda usar otras medidas que tengan en cuenta el desbalance de clases, como la precisión, el recall, el F1-score o el AUC-ROC. Estas medidas se basan en la matriz de confusión, que muestra el número de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN) que produce el modelo.



Esto significa que:

Para Antony, la precisión es 0.4294, el recall es 0.9605, el F-score es 0.5934 y el soporte es 76. Para Cleopatra, la precisión es 0.66666667, el recall es 0.1311, el F-score es 0.2191 y el soporte es 61. Para Queen Margaret, la precisión es 0.8333, el recall es 0.0980, el F-score es 0.1754 y el soporte es 51.

La precisión y el recall se relacionan con la matriz de confusión de la siguiente manera:

1. La precisión de una clase es el número de verdaderos positivos (TP) dividido por el número de verdaderos positivos más falsos positivos (TP + FP). Es decir, la precisión mide cuántos de los párrafos que se predijeron como pertenecientes a una clase realmente lo son.
2. El recall de una clase es el número de verdaderos positivos (TP) dividido por el número de verdaderos positivos más falsos negativos (TP + FN). Es decir, el recall mide cuántos de los párrafos que realmente pertenecen a una clase se predijeron correctamente.
3. El F-score es la media armónica de la precisión y el recall, lo que significa que da un valor único que representa tanto la precisión como el recall en una sola métrica. El F-score más alto posible es 1.0, lo que indica una precisión y un recall perfectos, y el más bajo posible es 0, si la precisión o el recall son cero.

Esto muestra que el Modelo tiende a predecir casi siempre la clase mayoritaria “Antony”, por eso los recall para las clases minoritarias son tan bajos.

Cross-Validation:

La técnica de validación cruzada o cross-validation es una técnica que se usa para evaluar el rendimiento de un modelo de aprendizaje automático en datos no vistos. Consiste en dividir los datos disponibles en varios pliegues o subconjuntos, usando uno de estos pliegues como conjunto de validación y entrenando el modelo en los pliegues restantes. Este proceso se repite varias veces, cambiando el pliegue de validación en cada iteración, y se calcula una medida de rendimiento promedio sobre todas las iteraciones.

La validación cruzada permite estimar cómo se comportará el modelo en la práctica, evitando el sobreajuste y el sesgo que pueden ocurrir al usar un solo conjunto de entrenamiento y prueba. También permite comparar y seleccionar diferentes modelos o hiperparámetros, eligiendo aquellos que tengan un mejor rendimiento promedio en la validación cruzada.

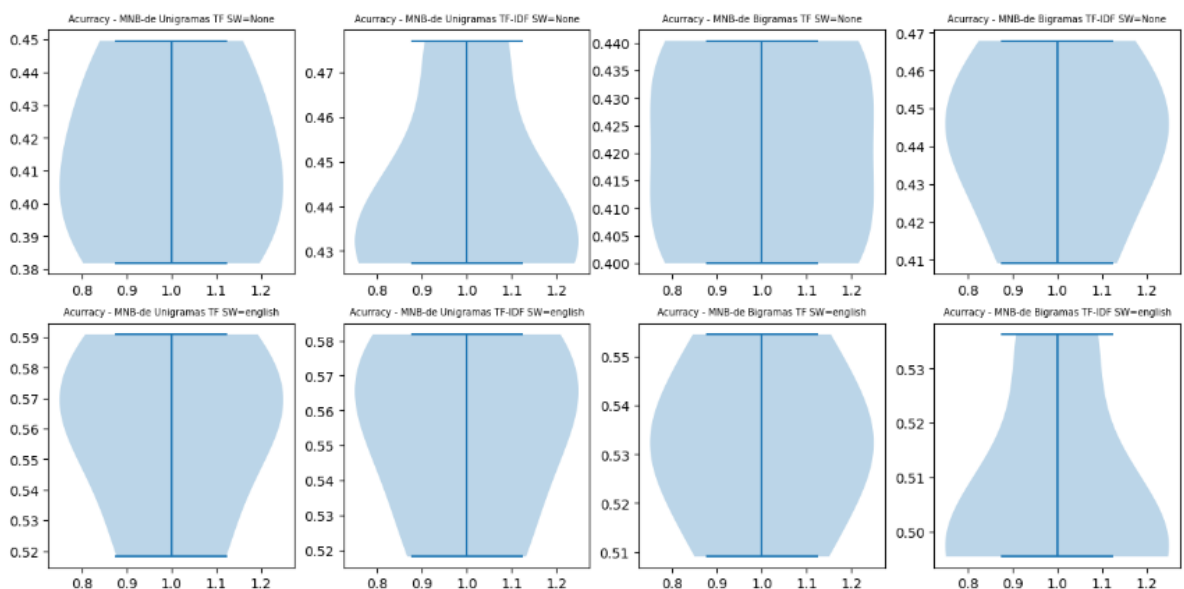
La forma más común de validación cruzada es la k-fold cross-validation, donde k es el número de pliegues o subconjuntos en los que se divide el conjunto de datos. Por ejemplo, si k es 4, los datos se dividen en 4 partes iguales, y se usa una de ellas como validación y las otras cuatro como entrenamiento. Se repite este proceso 4 veces, usando cada parte como validación una vez, y se obtiene una medida de rendimiento (por ejemplo, accuracy o f1-score) para cada iteración. Finalmente, se calcula el promedio y la desviación estándar de las medidas obtenidas.

Se estimaron 8 diferentes variantes de modelos o hiperparámetros, obteniéndose las siguientes métricas:

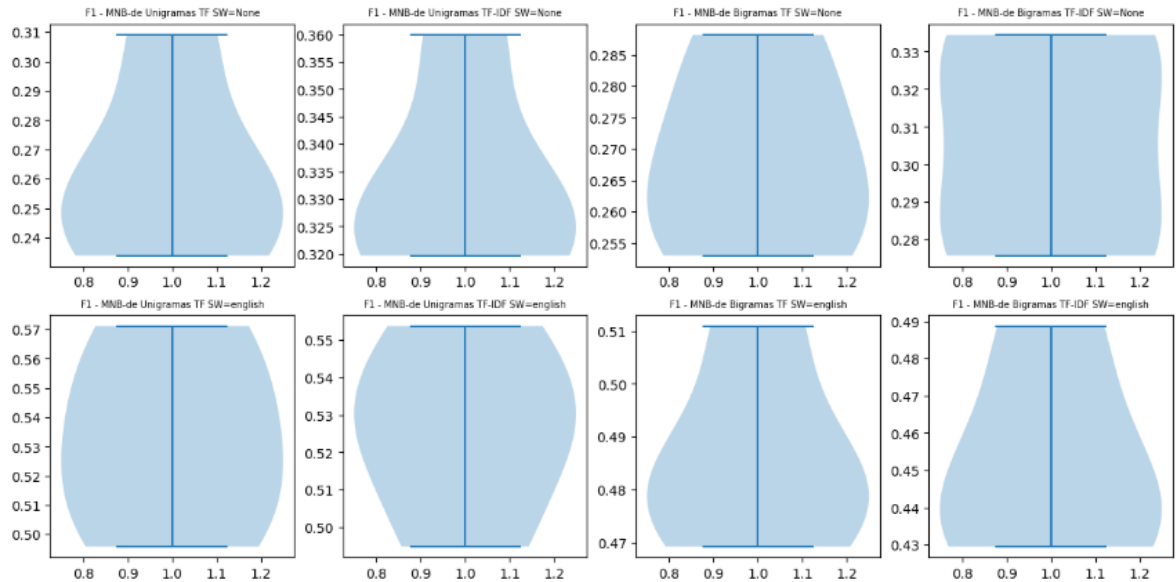
```
MNB-de Unigramas TF SW=None mean_acc=0.4133 std_acc=0.0253 mean_f1=0.2623
MNB-de Unigramas TF-IDF SW=None mean_acc=0.4430 std_acc=0.0199 mean_f1=0.3332
MNB-de Bigramas TF SW=None mean_acc=0.4202 std_acc=0.0163 mean_f1=0.2683
MNB-de Bigramas TF-IDF SW=None mean_acc=0.4407 std_acc=0.0210 mean_f1=0.3052
MNB-de Unigramas TF SW=english mean_acc=0.5594 std_acc=0.0264 mean_f1=0.5319
MNB-de Unigramas TF-IDF SW=english mean_acc=0.5548 std_acc=0.0239 mean_f1=0.5259
MNB-de Bigramas TF SW=english mean_acc=0.5320 std_acc=0.0161 mean_f1=0.4858
MNB-de Bigramas TF-IDF SW=english mean_acc=0.5068 std_acc=0.0172 mean_f1=0.4517
```

El mejor modelo estimado de acuerdo a el accuracy y el f1-score resulta ser: MNB-de Unigramas TF SW=english mean_acc=0.5594 std_acc=0.0264 mean_f1=0.5319. Si vemos la distribución de las medidas de bondad del ajuste de los 8 modelos a través de violín plot observamos esto:

Para el Accuracy:



Para el f1-score:

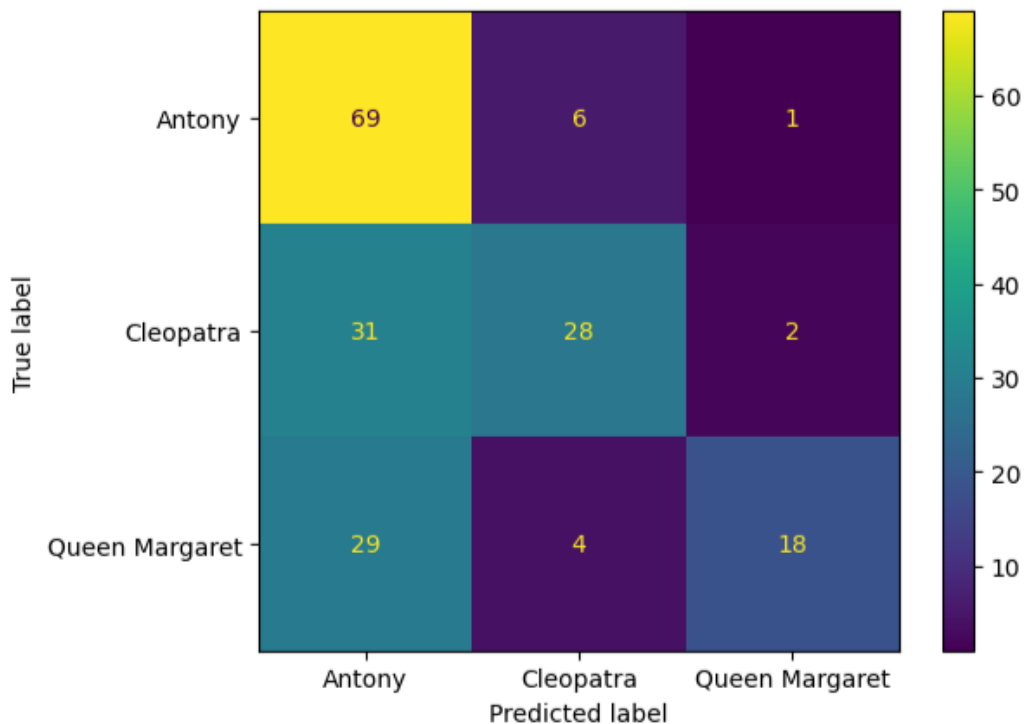


El mejor modelo es el Multinomial Naive Bayes con representación de Unigramas, stop words en english, con transformación TF, aunque con IDF apenas se aprecian diferencias.

Entrenamiento con Mejor Modelo:

	precision	recall	f1-score	support
Antony	0.535	0.908	0.673	76
Cleopatra	0.737	0.459	0.566	61
Queen Margaret	0.857	0.353	0.500	51
accuracy			0.612	188
macro avg	0.710	0.573	0.580	188
weighted avg	0.688	0.612	0.591	188

El modelo mejora los recall tan bajos encontrados anteriormente para las clases minoritarias, y mejora el accuracy al 61,2%.



Si miramos la matriz de confusión la diagonal principal creció de manera significativa respecto al primer entrenamiento fuera de considerar los hiperparámetros.

Los modelos basados en bag-of-words o tf-idf tienen algunas limitaciones en cuanto al análisis de texto, tales como:

1. No capturan el orden ni la estructura de las palabras en el texto, lo que puede afectar al significado y la semántica. Por ejemplo, las frases “el perro mordió al hombre” y “el hombre mordió al perro” tendrían la misma representación con bag-of-words o tf-idf, pero tienen significados muy diferentes.
2. No capturan la similitud ni la relación entre las palabras, lo que puede dificultar la generalización y la comprensión del contexto. Por ejemplo, las palabras “gato” y “felino” son sinónimos, pero tendrían vectores distintos con bag-of-words o tf-idf, y no se podría medir su cercanía semántica.
3. Pueden generar vectores de alta dimensionalidad y dispersos, lo que puede aumentar la complejidad computacional y el riesgo de sobreajuste. Esto ocurre especialmente cuando el vocabulario es muy grande y cada documento contiene solo una pequeña fracción de las palabras posibles.

Evaluación con un Modelo Adicional: SVC

Se realizó un nuevo ajuste por hiperparámetros para el modelo SVC.

La estimación con hiperparámetros por modelo SVC utilizando gridsearchCV consiste en buscar los mejores valores de los parámetros que afectan al rendimiento del modelo de clasificación Support Vector Classifier (SVC), usando la técnica de validación cruzada o cross-validation.

Los parámetros que se pueden ajustar en un modelo SVC son, entre otros:

1. C: Es el parámetro de regularización que controla el equilibrio entre el margen y el error de clasificación. Un valor alto de C implica una mayor penalización por los errores de clasificación, lo que puede llevar a un sobreajuste. Un valor bajo de C implica una mayor tolerancia a los errores de clasificación, lo que puede llevar a un subajuste.
2. kernel: Es el tipo de función que se usa para transformar los datos a un espacio de mayor dimensión donde sean linealmente separables. Los tipos de kernel más comunes son ‘linear’, ‘poly’, ‘rbf’ y ‘sigmoid’.
3. gamma: Es el parámetro que controla la complejidad del kernel. Un valor alto de gamma implica una mayor influencia de cada punto de datos, lo que puede llevar a un sobreajuste. Un valor bajo de gamma implica una menor influencia de cada punto de datos, lo que puede llevar a un subajuste.
4. degree: Es el grado del polinomio cuando se usa el kernel ‘poly’. Un valor alto de degree implica una mayor complejidad del kernel, lo que puede llevar a un sobreajuste. Un valor bajo de degree implica una menor complejidad del kernel, lo que puede llevar a un subajuste.
5. La técnica de gridsearchCV consiste en definir una rejilla o malla de posibles valores para cada uno de estos parámetros, y probar todas las combinaciones posibles usando validación cruzada para medir el rendimiento del modelo con cada combinación. La validación cruzada implica dividir los datos en varios pliegues o subconjuntos, y usar uno de ellos como conjunto de prueba y los demás como conjunto de entrenamiento. Este proceso se repite varias veces, cambiando el pliegue de prueba en cada iteración, y se calcula una medida de rendimiento promedio (por ejemplo, accuracy) para cada combinación de parámetros.

La implementación de gridsearchCV en scikit-learn es la clase GridSearchCV, que recibe como argumentos:

1. estimator: El modelo que se quiere ajustar, en este caso un objeto SVC.
2. param_grid: Un diccionario o una lista de diccionarios con los nombres y los valores de los parámetros que se quieren explorar.
3. scoring: La medida de rendimiento que se quiere optimizar, por ejemplo ‘accuracy’.
4. cv: El número de pliegues o subconjuntos para la validación cruzada, por ejemplo 5.

La clase GridSearchCV implementa la interfaz habitual de un estimador: cuando se llama al método fit con los datos, se prueban todas las combinaciones posibles de parámetros y se guarda la mejor combinación. Luego se puede acceder a los atributos y métodos del mejor estimador, como score, predict o best_params_.

Se definieron los siguientes hiperparámetros en esta implementación:

```
param_grid = {'C': [0.01, 0.1, 0.5, 0.9, 1, 1.35, 1.5, 1.7, 2, 10], 'kernel': ['linear', 'rbf', 'poly', 'sigmoide'], 'degree': [2], 'gamma': [0.001, 0.01, 0.1, 1, 1.5, 2, 10]}
grid_search = GridSearchCV(SVC(), param_grid, cv=5, scoring='accuracy', verbose=2)
```

Esto origino 280 modelos posibles con 1400 ajustes:

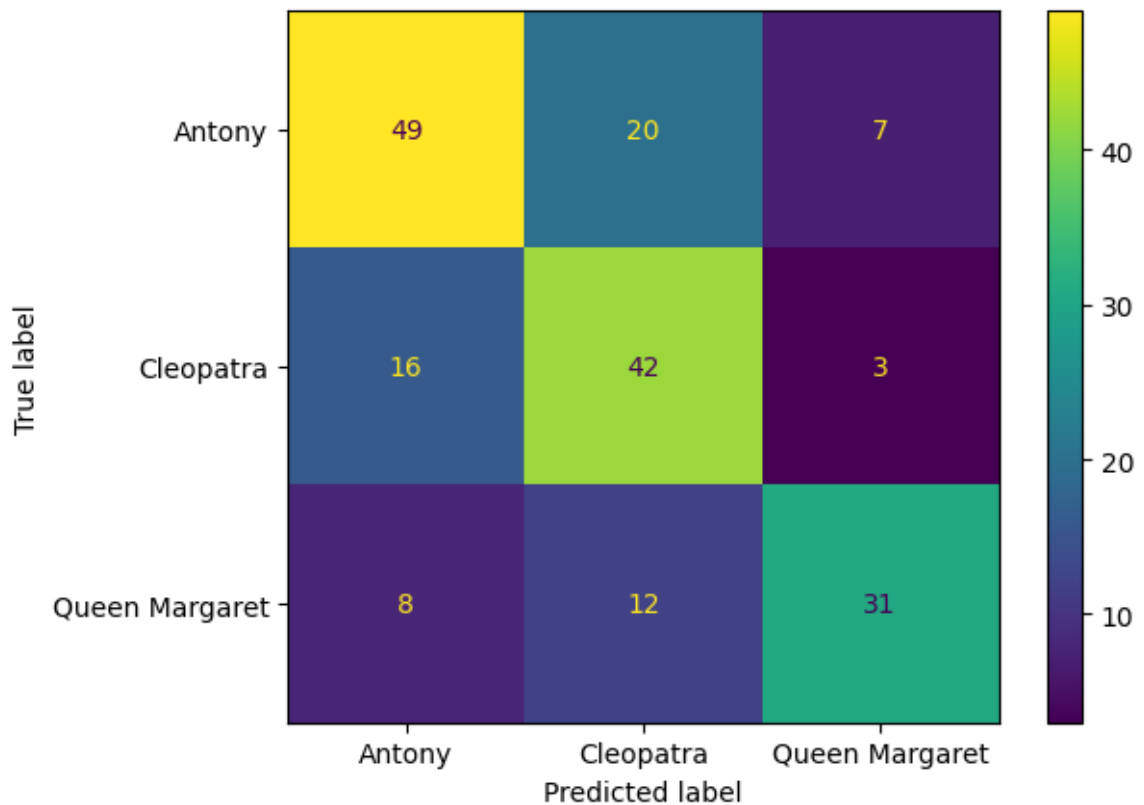
Fitting 5 folds for each of 280 candidates, totalling 1400 fits
El mejor ajuste se dio para:

```
{'C': 10, 'degree': 2, 'gamma': 0.1, 'kernel': 'rbf'}
```

Este modelo de clasificación presento las siguientes métricas:

	precision	recall	f1-score	support
Antony	0.671	0.645	0.658	76
Cleopatra	0.568	0.689	0.622	61
Queen Margaret	0.756	0.608	0.674	51
accuracy			0.649	188
macro avg	0.665	0.647	0.651	188
weighted avg	0.661	0.649	0.651	188

Presentando la siguiente Matriz de Confusión:



Sin duda este modelo con más hiperparámetros genero una leve mejora en cuanto a las métricas, aumentando un poco más los recall respecto a los modelos anteriores, mejorando el dejar de clasificar sistemáticamente en la clase mayoritaria y dando un accuracy más alto. Lo que se perdió en precisión de la clase mayoritaria se compenso a través de mirar los f1-score para todas las clases.

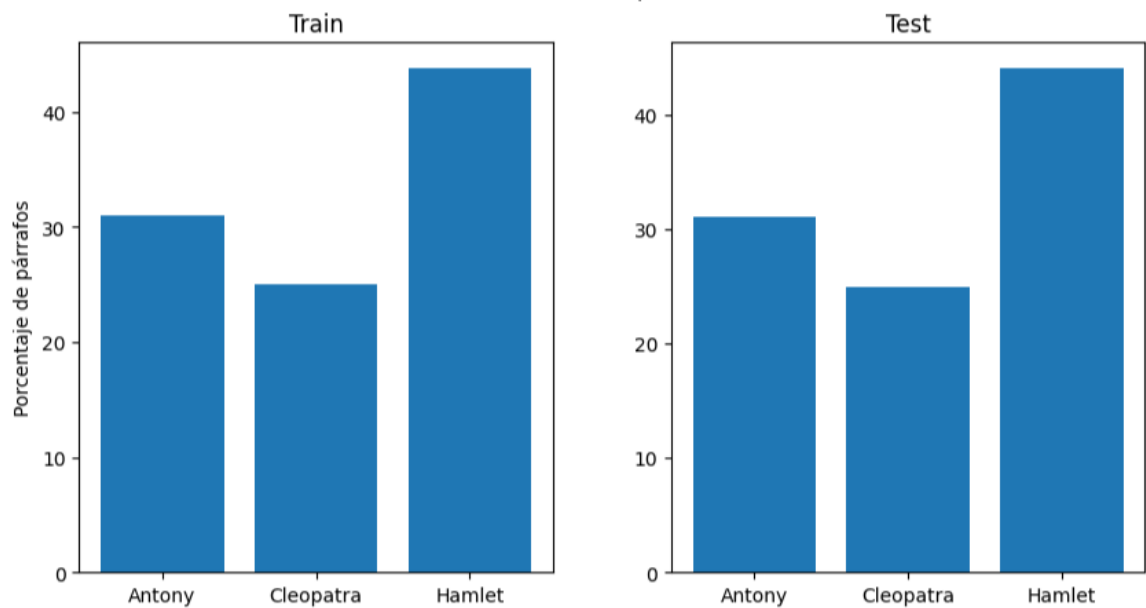
Evalúe el problema cambiando al menos un personaje:

Usaremos estos personajes

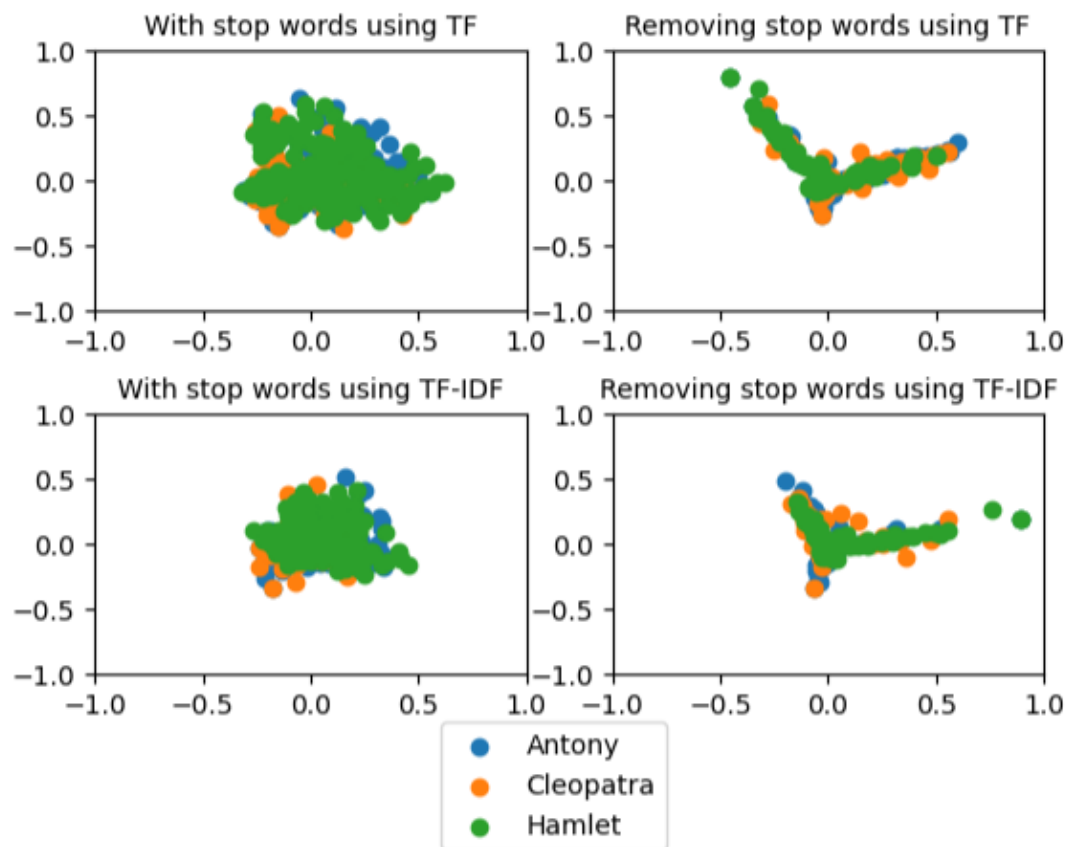
characters2 = ["Antony", "Cleopatra", "Hamlet"]

Ahora la clase mayoritaria pasa a ser "**Hamlet**".

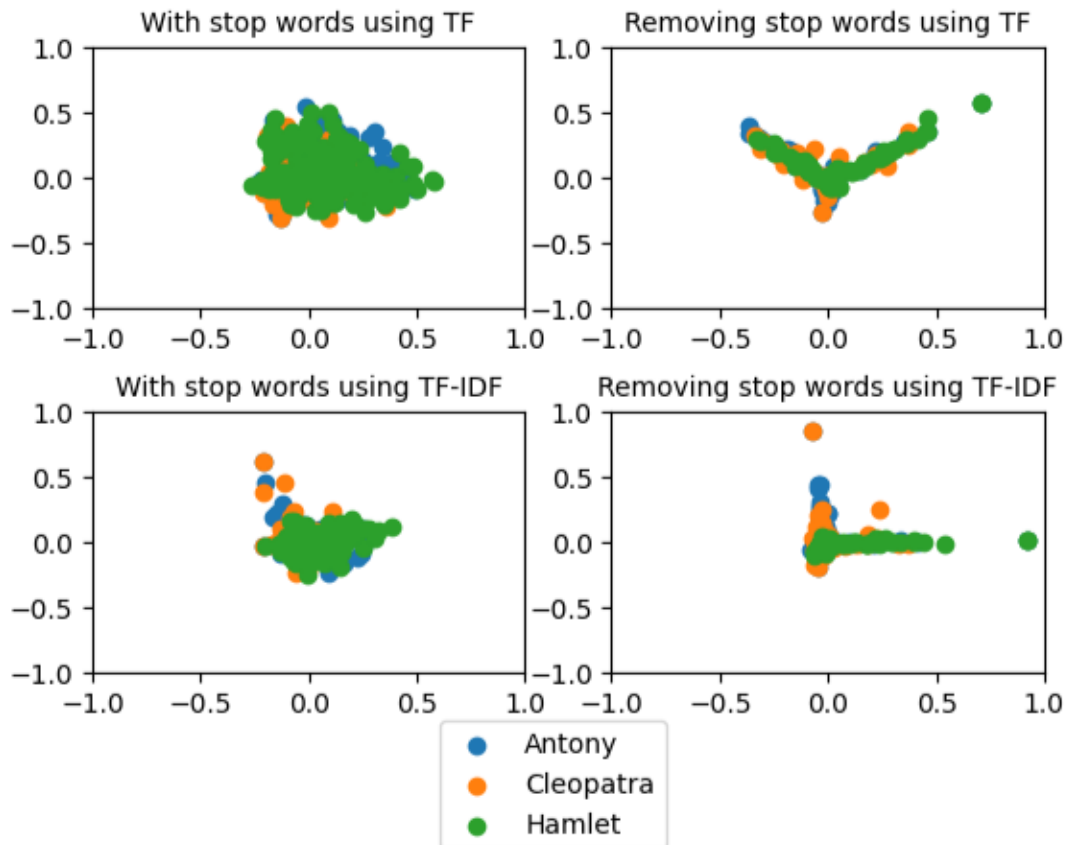
Distribución de párrafos



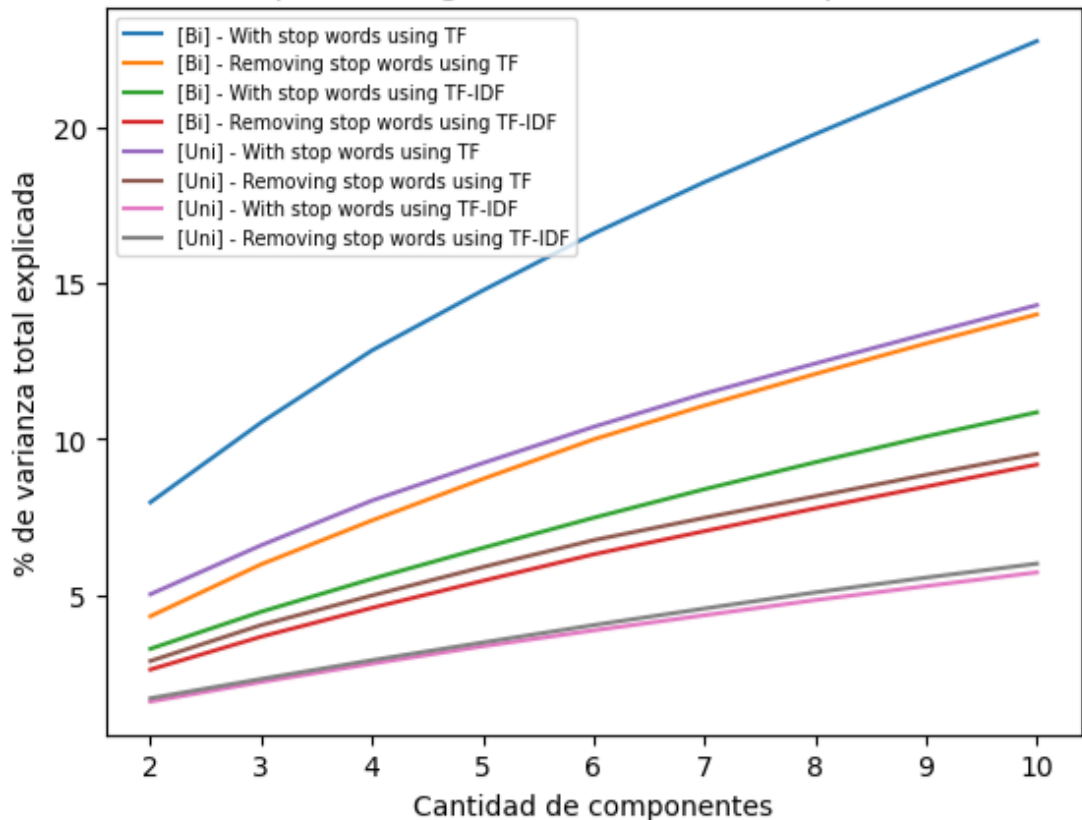
PCA por Personaje tomando unigramas en el conjunto de entrenamiento



PCA por Personaje tomando bigramas en el conjunto de entrenamiento



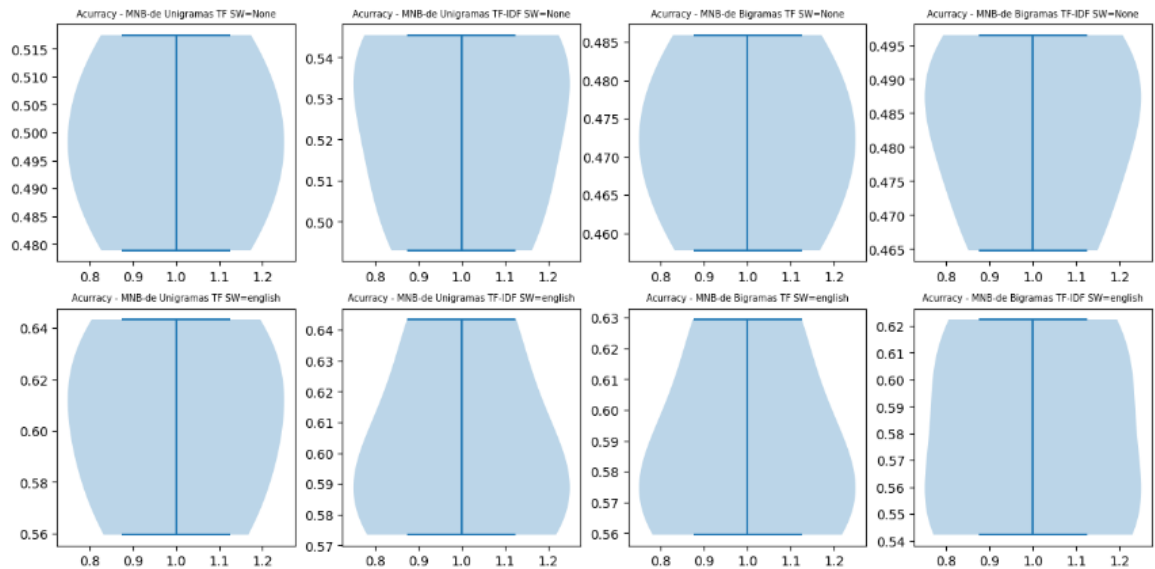
Varianza explicada según la cantidad de componentes de PCA



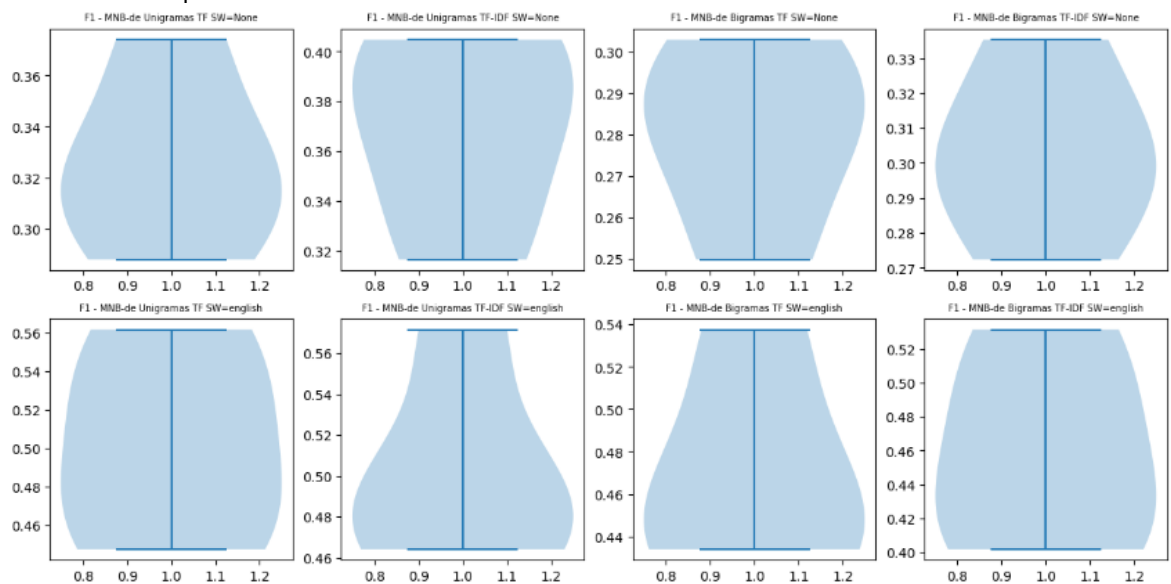
Las conclusiones son similares al punto estudiado con los personajes Anteriores, la varianza explicada en los primeros 10 componentes principales no alcanza el 25% en ninguno de los casos. Obviamente no se pueden diferenciar correctamente los personajes en el plano factorial (primeros 2 componentes). La búsqueda de hiperparámetros arroja los mismos resultados que el análisis anterior para el accuracy, mientras que visto por f1-score mejora con la inclusión de idf.

MNB-de Unigramas TF SW=None mean_acc=0.4982 std_acc=0.0141 mean_f1=0.3258
 MNB-de Unigramas TF-IDF SW=None mean_acc=0.5227 std_acc=0.0208 mean_f1=0.3681
 MNB-de Bigramas TF SW=None mean_acc=0.4719 std_acc=0.0103 mean_f1=0.2797
 MNB-de Bigramas TF-IDF SW=None mean_acc=0.4824 std_acc=0.0120 mean_f1=0.3025
 MNB-de Unigramas TF SW=english mean_acc=0.6035 std_acc=0.0314 mean_f1=0.5011
 MNB-de Unigramas TF-IDF SW=english mean_acc=0.6017 std_acc=0.0267 mean_f1=0.5020
 MNB-de Bigramas TF SW=english mean_acc=0.5877 std_acc=0.0267 mean_f1=0.4703
 MNB-de Bigramas TF-IDF SW=english mean_acc=0.5789 std_acc=0.0329 mean_f1=0.4590

Gráficos de Violín para el Accuracy:



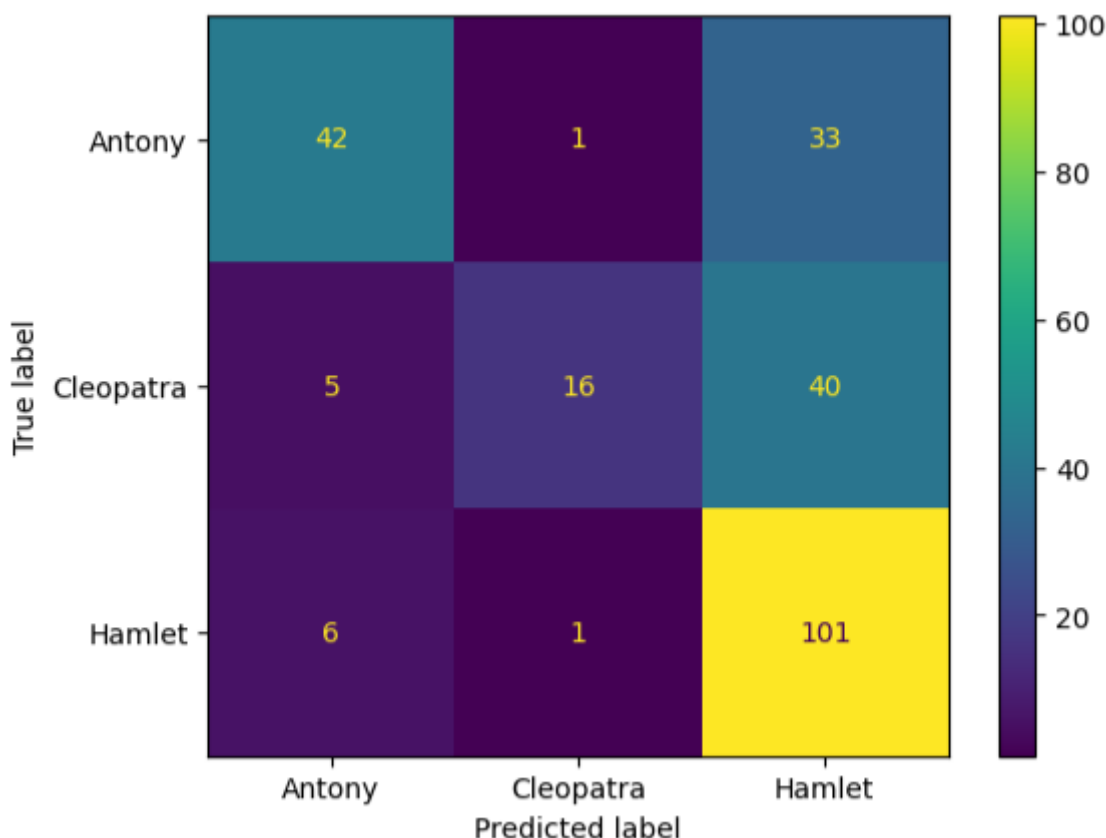
Gráficos de Violín para el f1-score:



El mejor modelo estimado de acuerdo a el accuracy resulta ser: MNB-de Unigramas TF SW=english mean_acc=0.6035 std_acc=0.0314 mean_f1=0.5011

	precision	recall	f1-score	support
Antony	0.792	0.553	0.651	76
Cleopatra	0.889	0.262	0.405	61
Hamlet	0.580	0.935	0.716	108
accuracy			0.649	245
macro avg	0.754	0.583	0.591	245
weighted avg	0.723	0.649	0.619	245

La matriz de confusión muestra que el modelo tiende a predecir con el desbalance de clases, la clase mayoritaria Hamlet.



Técnicas de Sobre muestreo y submuestreo:

Para mitigar el problema de clases desbalanceadas existen diferentes técnicas:

1. Sobremuestreo aleatorio: Consiste en replicar las muestras de la clase minoritaria hasta equilibrar el conjunto de datos. Es la técnica más simple, pero puede causar sobreajuste.
2. SMOTE (Synthetic Minority Oversampling Technique): Consiste en crear nuevas muestras sintéticas de la clase minoritaria usando un algoritmo de vecinos k-más cercanos. Evita la pérdida de información, pero puede crear puentes entre clases.
3. Borderline SMOTE: Consiste en sobremuestrear solo los ejemplos minoritarios cercanos al límite con la clase mayoritaria, evitando los puntos de ruido o atípicos.
4. KMeans SMOTE: Consiste en aplicar el algoritmo KMeans para agrupar las muestras minoritarias y luego generar nuevas muestras sintéticas dentro de cada grupo.
5. SVM SMOTE: Consiste en usar una máquina de vectores de soporte (SVM) para encontrar el hiperplano que separa las clases y luego generar nuevas muestras sintéticas cerca del hiperplano.
6. ADASYN (Adaptive Synthetic Sampling): Consiste en generar más muestras sintéticas para las clases minoritarias que son más difíciles de clasificar, usando una medida de densidad basada en los vecinos k-más cercanos.
7. SMOTE-NC (SMOTE for Nominal and Continuous features): Consiste en una variante de SMOTE que puede manejar features de texto que son nominales o categóricos, usando una medida de distancia adaptada.

Técnicas Alternativas para extraer features de Texto:

El word embedding es una técnica que consiste en representar las palabras con vectores de números reales, de manera que se capture el significado, el contexto y la similitud de las palabras.

Las ventajas de usar word embedding respecto a BoW y TF-IDF son las siguientes:

1. El word embedding permite capturar el significado y la semántica de las palabras, mientras que BoW y TF-IDF solo se basan en la frecuencia de las palabras. Por ejemplo, el word embedding puede reconocer que "gato" y "felino" son sinónimos, mientras que BoW y TF-IDF los tratarían como palabras distintas.

2. El word embedding permite medir la similitud entre las palabras mediante la distancia o el ángulo entre sus vectores. Por ejemplo, el word embedding puede calcular que “gato” y “perro” son más similares entre sí que “gato” y “jirafa”, mientras que BoW y TF-IDF no podrían hacerlo.
3. El word embedding permite reducir la dimensionalidad de los datos, al representar las palabras con vectores de tamaño fijo. Por ejemplo, el word embedding puede usar vectores de 300 dimensiones para representar miles de palabras, mientras que BoW y TF-IDF tendrían una dimensión por cada palabra del vocabulario.

Modelo FastText:

El modelo de FastText es una herramienta de aprendizaje automático para la clasificación de texto y la generación de word embeddings. Este modelo se basa en el uso de redes neuronales que aprenden a representar las palabras y los documentos como vectores numéricos, y que pueden predecir la categoría de un texto a partir de sus características.

Al entrenar con el Modelo de FastText propuesto se obtuvieron los siguientes resultados:

	precision	recall	f1-score	support
Antony	0.615	0.737	0.671	76
Cleopatra	0.644	0.623	0.633	61
Queen_Margaret	0.737	0.549	0.629	51
accuracy			0.649	188
macro avg	0.665	0.636	0.644	188
weighted avg	0.658	0.649	0.647	188

Algunas ventajas y desventajas de utilizar este modelo son:

Ventajas:

1. Es rápido y fácil de entrenar, ya que usa una arquitectura simple y eficiente.
2. Es capaz de manejar textos en varios idiomas y alfabetos.
3. Es capaz de capturar el contexto y la semántica de las palabras y los documentos.
4. Es capaz de generar word embeddings que pueden ser usados para otras tareas o aplicaciones.

Desventajas:

1. No es capaz de manejar textos muy largos o complejos, ya que usa una ventana fija para analizar las palabras.
2. No es capaz de capturar la ambigüedad o la ironía de los textos, ya que usa una representación vectorial que no considera estos aspectos.
3. No es capaz de incorporar conocimiento externo o previo al modelo, ya que solo se basa en los datos disponibles.

Los resultados arribados con este modelo son similares al mejor modelo encontrado con SVC:

{ 'C': 10, 'degree': 2, 'gamma': 0.1, 'kernel': 'rbf' }				
	precision	recall	f1-score	support
Antony	0.671	0.645	0.658	76
Cleopatra	0.568	0.689	0.622	61
Queen_Margaret	0.756	0.608	0.674	51
accuracy			0.649	188
macro avg	0.665	0.647	0.651	188
weighted avg	0.661	0.649	0.651	188

En accuracy coinciden para el modelo general y las métricas de precisión y recall para algunos personajes son levemente mejores en uno que en el otro y viceversa, pero realmente muy similares.