

2 de julio de 2024

Introducción a la Ciencia de Datos

Tarea 2 - Grupo 5

Agustín Ghazarian
Sofía Gervaz

Tarea 2

En esta etapa se entrenará un modelo de lenguaje natural utilizando la base de datos presentada anteriormente. Se utilizará una versión reducida de la misma, considerando los personajes “Antony”, “Cleopatra” y “Queen Margaret”. En la Tabla 1 se presenta la cantidad de párrafos para cada uno de estos personajes.

Para estos personajes se dividieron los párrafos en *train* y *test*, siendo los mismos el 70% y 30% respectivamente. El tamaño de *train* alcanza los 433 párrafos mientras que para *test* es de 188. La proporción de párrafos de cada personaje en *train* y *test* se observa en la Figura 1.

Tabla 1 - Cantidad de párrafos por personaje.

Personaje	Total
Antony	253
Cleopatra	204
Queen Margaret	169

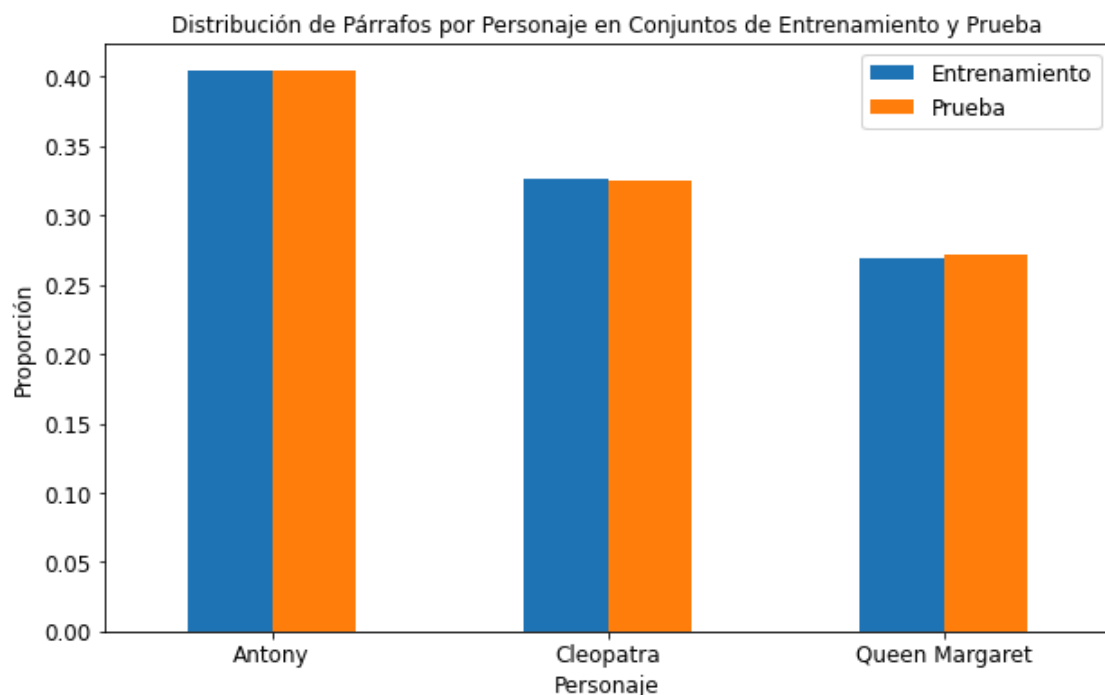


Figura 1 - Balance de párrafos por personaje en train y test.

Se procede a transformar el texto de entrenamiento a representación numérica utilizando *bag of words*. Este método asigna una fila a cada párrafo y una columna a cada palabra que se encuentra. Cada posición i,j de la matriz es completada con la cantidad de veces que aparece la palabra j en el párrafo i .

A modo de ejemplo, se presenta el siguiente texto formado por dos párrafos:

Párrafo 1:

El gato negro duerme en la silla. La casa está en silencio.

Párrafo 2:

La luna brilla en el cielo. El perro ladra en el jardín.

Con estos párrafos, la matriz *bag of words* es la que se presenta en la Tabla 2.

Tabla 2 - Matriz *bag of words* para caso de ejemplo.

	brilla	casa	cielo	duerme	el	en	está	gato	jardín	la	ladra	luna	negro	perro	silencio	silla
Párrafo 1	0	1	0	1	1	2	1	1	0	2	0	0	1	0	1	1
Párrafo 2	1	0	1	0	3	2	0	0	1	1	1	1	0	1	0	0

Como los textos suelen tener muchos párrafos, lo habitual es que cada párrafo tenga pocas de las palabras que aparecen en todo el conjunto de entrenamiento. Por lo tanto, la matriz *bag of words* resultante es mayoritariamente ceros. La representación de matriz dispersa aprovecha esta característica, guardando únicamente aquellos valores no nulos y su posición. Con este mecanismo se ahorra un gran uso de memoria que se destinaría a almacenar valores nulos.

De igual modo al método *bag of words*, el método *Term Frequency-Inverse Document Frequency* (TF-IDF) es una representación numérica de un determinado texto, pero donde cada término recibe un peso basado en su frecuencia de ocurrencia en el párrafo y su rareza en el conjunto de párrafos. Esta transformación tiene en cuenta tanto la importancia local (frecuencia del término en el párrafo) como la importancia global (inversa de la frecuencia del término en todos los párrafos). De esta forma, las palabras que se repiten mucho en cada párrafo pero además se repiten en el conjunto de párrafos (por ejemplo *stop words*), pierden relevancia. Dando entonces más peso a aquellas palabras “claves” al momento de diferenciar personajes.

Un N-grama es una secuencia contigua de N elementos (palabras o caracteres) de un texto dado. Los N-gramas se utilizan comúnmente en procesamiento de lenguaje natural para modelar el contexto local dentro del texto.

El método de componentes principales (PCA) sirve para reducir la dimensionalidad de problemas complejos. En este caso, se tiene un conjunto de 438 párrafos formados por 2,807 palabras. Se trata de encontrar los ejes que expliquen de mejor modo la nube de puntos existente. En las Figuras 2 y 3 se presentan las gráficas resultantes al reducir la dimensionalidad del problema a 2 componentes; en la Figura 2 para el conjunto de entrenamiento original y en la Figura 3 para el conjunto de entrenamiento donde se filtraron las *stop words*, se usó IDF y se consideraron bigramas.

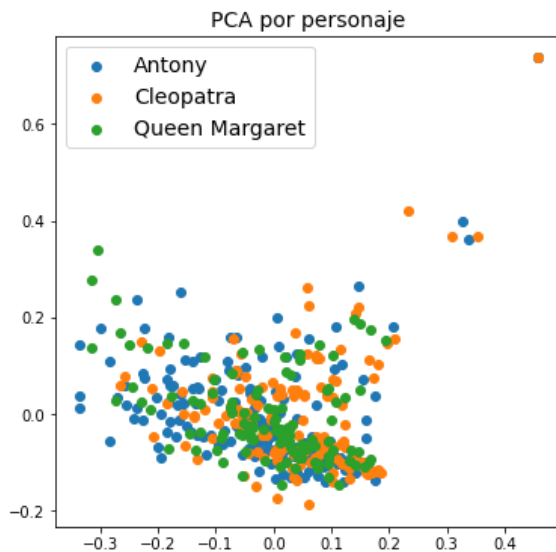


Figura 2 - PCA. dos componentes

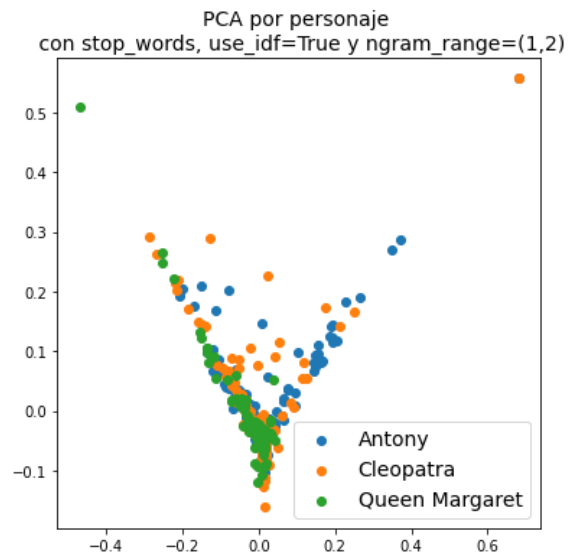


Figura 3 - PCA dos componentes,
stop_words=english, use_idf=True, ngram_range(1,2)

Debido a que la reducción de dimensionalidad es muy elevada, pasando de un problema de dimensión 2,807 a un problema de dimensión 2, se pierde gran parte de la información por lo que la varianza explicada por el nuevo modelo es baja (ver Figura 4).

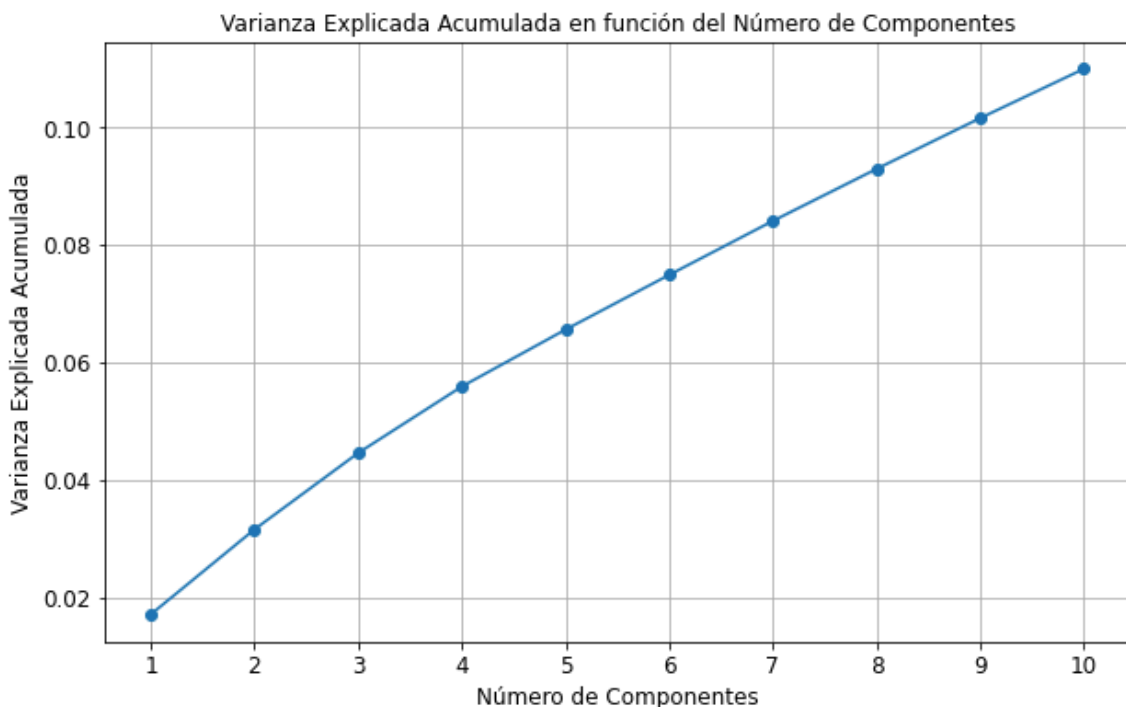


Figura 4 - Varianza explicada vs Número de componentes de PCA utilizados.

Al aplicar el filtrado de *stop words*, el uso de IDF y considerar bigramas, la representación TF-IDF se vuelve más útil para discriminar entre personajes, con una mayor separación en los puntos entre personajes y un menor solapamiento (ver Figs. 2 y 3).

Se entrenó el modelo Multinomial Naive Bayes y se evaluarán los resultados en términos de *accuracy*, matriz de confusión, precisión y *recall*.

Accuracy: El valor de *accuracy* mide la proporción de predicciones correctas sobre el total de predicciones. Sin embargo, no evalúa los resultados para cada personaje ni evalúa si los mismos se encuentran balanceados.

Matriz de Confusión: La matriz de confusión muestra las verdaderas etiquetas frente a las etiquetas predichas, proporcionando una visión detallada de los errores de clasificación para cada clase.

Precision y *Recall*:

- Precisión: Mide la proporción de predicciones correctas sobre el total de predicciones realizadas para un personaje específico.
- Recall: Mide la proporción de predicciones correctas sobre el total de párrafos realmente asignados al personaje.

El valor de *accuracy* hallado es 0.5 mientras que la siguiente tabla muestra tanto el parámetro precisión como *recall* para cada personaje y la matriz de confusión correspondiente.

Observar solamente el valor de *accuracy* (variable que considera la totalidad de las predicciones, no cada personaje en particular) puede llevar a error en el nivel de confiabilidad en la predicción en cada personaje. Por ejemplo, si la muestra fuera extremadamente desbalanceada, el modelo puede tender a asignar todos los párrafos al personaje con mayor peso, y eso daría como resultado un alto valor de *accuracy*, aunque el modelo falla en predecir realmente los párrafos correspondientes a otros personajes.

Tabla 3 - Parámetros de precisión y recall obtenidos.

	Precisión	<i>Recall</i>
Antony	0.91	0.20
Cleopatra	0.63	0.20
Queen Margaret	0.64	0.95

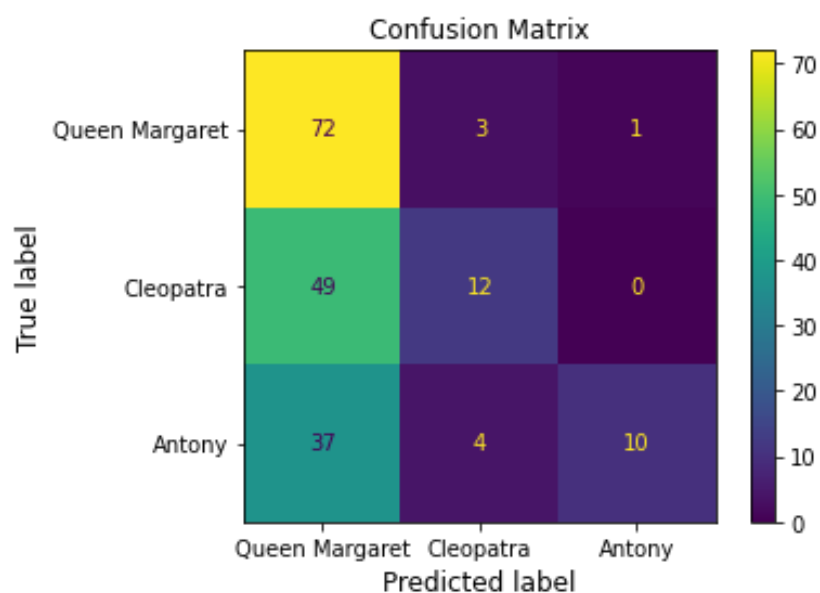


Figura 5 - Matriz de confusión.

La técnica de validación cruzada es un método para evaluar el rendimiento de un modelo. Consiste en dividir el conjunto de datos en múltiples subconjuntos o *folds*. El modelo se entrena en una combinación de estos *folds* y se evalúa en el *fold* restante. Este proceso se repite varias veces, y los resultados se promedian para obtener una estimación más confiable del rendimiento del modelo.

Del análisis anterior se desprende que para nuestro caso el mejor modelo resulta con los parámetros ($\text{ngram_range}=(1,1)$, $\text{use_idf}=\text{False}$, $\alpha=0.01$) (ver Figura 6). Utilizando este modelo se alcanza un valor de *accuracy* de 0.63, y los siguientes valores de precisión y *recall* (ver Tabla 3).

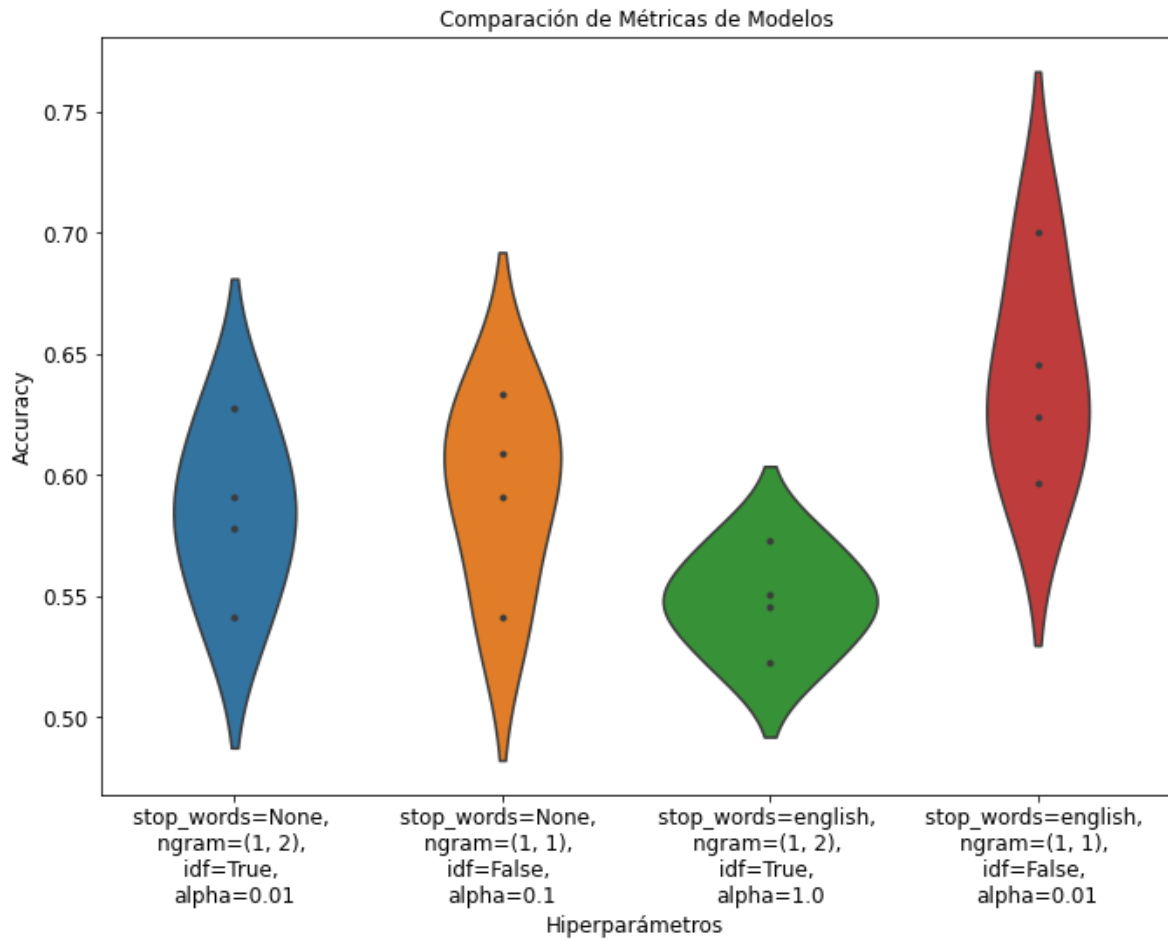


Figura 6 - Gráfico de violín para distintos parámetros de entrenamiento de modelo.

Tabla 4 - Parámetros de precisión y *recall* obtenidos.

Personaje	Precisión	<i>Recall</i>
Antony	0.75	0.71
Cleopatra	0.61	0.46
Queen Margaret	0.59	0.72

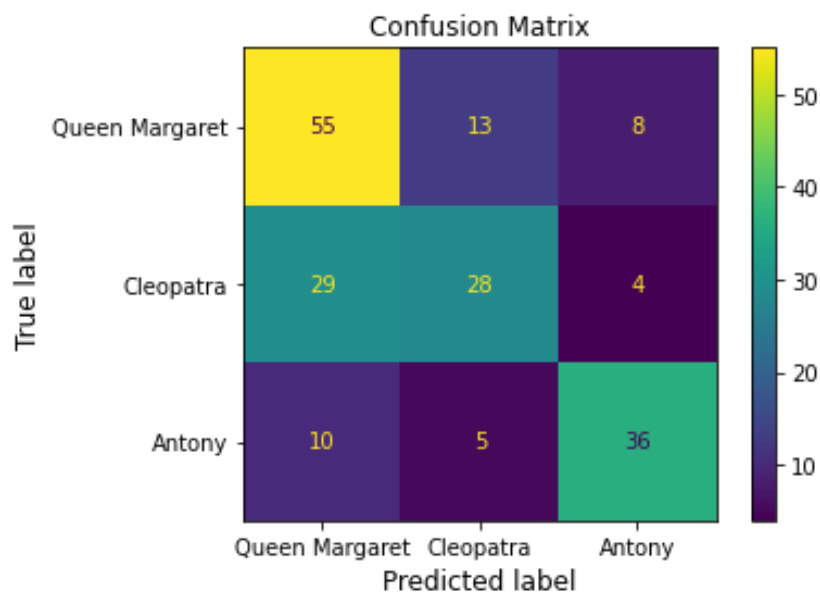


Figura 7 - Matriz de confusión.

Una de las limitaciones asociadas al uso de un modelo basado en *bag of words* es la abordada en el método TF-IDF y es que con *bag of words* no se considera la importancia relativa de las palabras, resultando que no se tenga en cuenta que hay palabras más informativas que otras. Además, el modelo no tiene en cuenta el orden de las palabras en el texto, haciendo que se pierda contexto y que frases con distinto significado puedan tener igual representación. Por el contrario, al no capturar relaciones semánticas, palabras que son en realidad sinónimos van a representarse por dimensiones distintas. Además, el modelo suele llevar a representaciones de dimensionalidad alta y podría ser difícil de manejar cuando se trabaja con grandes volúmenes de datos.

Al igual que el *bag of words*, TF-IDF no considera el orden de las palabras, por lo que se pierde contexto y sintaxis. Tampoco captura relaciones semánticas, haciendo que palabras que son sinónimos se representen como dimensiones separadas. Esta técnica también suele llevar a representaciones de dimensionalidad alta que pueden ser difíciles de manejar. Por otro lado, a diferencia de *bag of words*, al considerar la importancia relativa de las palabras, hace que la técnica pueda ser sensible a la longitud del documento a estudiar.

Para evaluar otro modelo además del Multinomial Naive Bayes, vamos a utilizar el clasificador de Máquinas de Soporte Vectorial (*Support Vector Machine*, SVM). Este algoritmo de clasificación busca encontrar el hiperplano que mejor separa las diferentes clases en el espacio de características. El objetivo del SVM es maximizar el margen, es decir, la distancia entre el hiperplano separador y los puntos de datos más cercanos de cada clase, que se llaman vectores de soporte.

El *accuracy* resultante es de 0.62 y los valores de precisión y *recall* son los que se presentan en la Tabla 4.

Tabla 5 - Parámetros de precisión y *recall* obtenidos.

Personaje	Precisión	Recall
Antony	0.85	0.65
Cleopatra	0.54	0.43
Queen Margaret	0.57	0.76

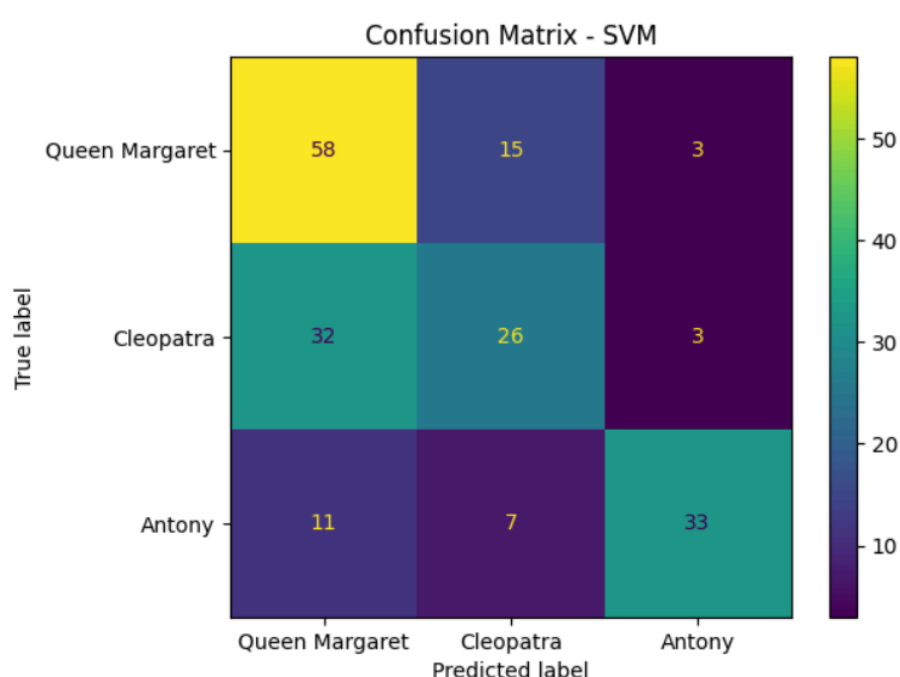


Figura 8 - Matriz de confusión SVM.

Se procede a realizar el análisis cambiando el personaje “Antony” por “Henry V”. Este personaje tiene más participación que “Cleopatra” y “Queen Margaret”. Este desbalance puede llevar a que el modelo tienda a predecir más veces el resultado “Henry V”, dado que esto resultaría en un alto nivel de *accuracy*.

Tabla 6 - Cantidad de párrafos por personaje.

Personaje	Total	Train	Test
Henry V	377	264	113
Cleopatra	204	143	61
Queen Margaret	169	118	51

Al entrenar el modelo y hacer la predicción, se obtiene un *accuracy* de 0.51 pero se observa que el modelo adjudica los párrafos casi exclusivamente a “Henry V”. Esto da como resultado valores malos de *recall* y precisión para cada personaje.

En el caso de “Cleopatra” la precisión es 1 ya que sólo una vez predijo este resultado y fue acertado, pero el *recall* es de 0.02 ya que falló en predecir 60 párrafos, los que fueron asignados a “Henry V”. Para “Queen Margaret” no realizó predicciones, por lo que tanto el *recall* como la precisión fueron nulos. Por otra parte, para “Henry V” el modelo tuvo una precisión de 0.5 y un *recall* de 1.

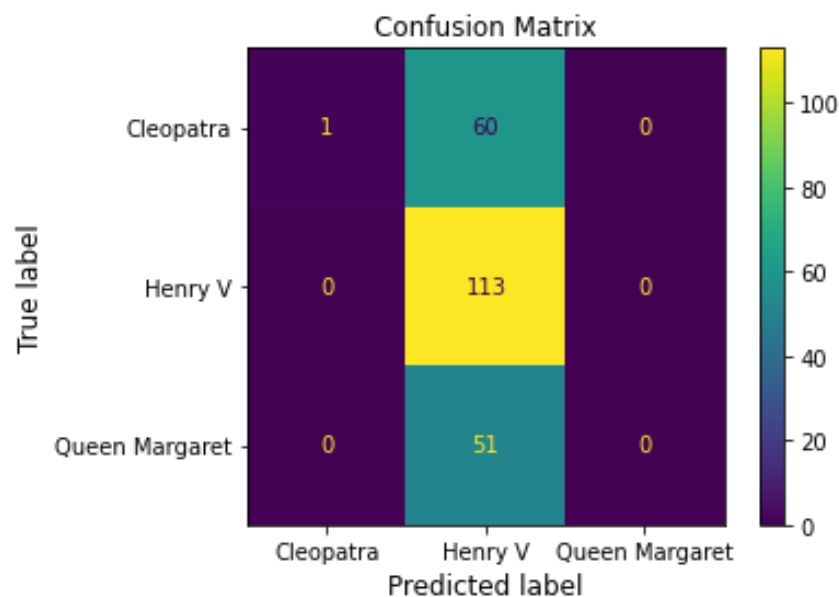


Figura 9 - Matriz de confusión variando un personaje.

Existen técnicas más avanzadas que abordan las limitaciones mencionadas para *bag of words* y TF-IDF. Un ejemplo es el modelo de *word embedding Word2Vec*, una técnica de *deep learning* desarrollada por Google que se basa en generar representaciones vectoriales (o *embeddings*) de palabras. Así, a diferencia de los modelos trabajados en la tarea, que representan palabras de manera independiente y sin contexto, *Word2Vec* crea *embeddings* que capturan relaciones semánticas entre palabras.

Para hacer esto, la técnica utiliza dos arquitecturas principales:

1. *Continuous bag of words*: se predice una palabra dado su contexto (ventana de palabras).
2. *Skip-gram*: se predicen las palabras circundantes (dentro de la ventana definida) dada una palabra central.

Estos modelos aprenden los *embeddings* de las palabras usando una red neuronal simple, que busca predecir los vecinos de una palabra.

Para usar la técnica, se limpian los datos igual que para los modelos trabajados en la tarea, se define una ventana de palabras que determina la extensión del contexto y se entrena la red neuronal para aprender los *embeddings* de las palabras. En *continuous bag of words* la entrada es el contexto y la salida la palabra objetivo, en *skip-gram* es el inverso. La red ajusta los pesos de la capa oculta y estos pesos ajustados son los *embeddings*.

Como esta técnica capta las relaciones semánticas entre las palabras, palabras con significados similares estarán cercanas en el espacio vectorial. Lo que entendemos que sería beneficioso para adjudicar párrafos a personajes ya que facilita la identificación del estilo y vocabulario específico de cada personaje. El hecho de que considere el contexto también puede ayudar a distinguir entre personajes, sobretodo si son de obras diferentes.