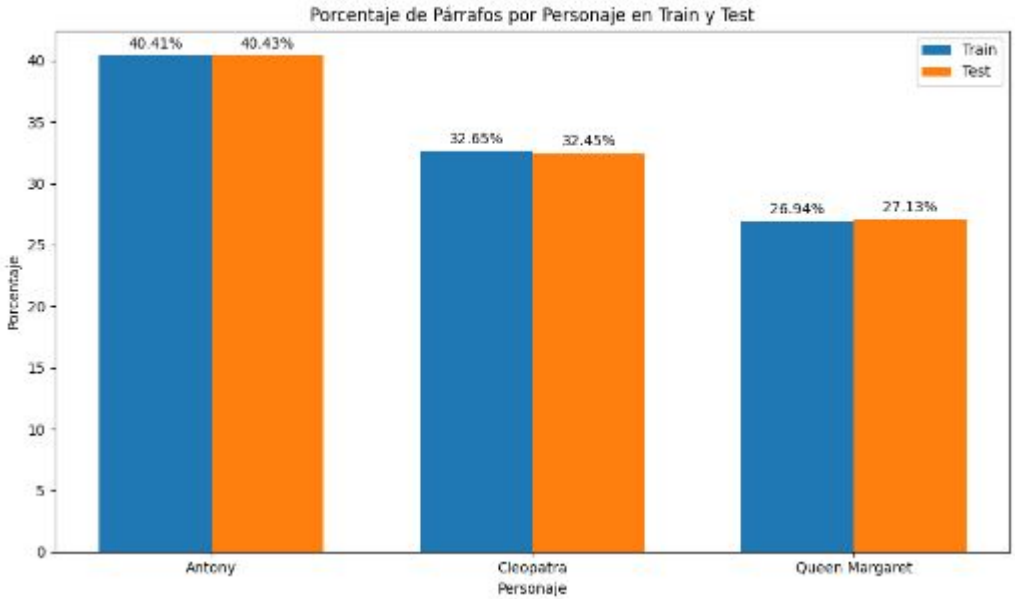
Introducción a la ciencia de datos - Tarea 2

Grupo 20 Auriello - Molina
02/07/2024

Reducción de Dataset y train/test

- Antony, Cleopatra y Queen Margaret.
- Conjuntos de entrenamiento y prueba. Muestreo estratificado, asegurando que el 30% de los datos se destinara al conjunto de prueba y que las proporciones de los personajes se mantuvieran consistentes en ambos conjuntos.
- Hay un balance en train y test de la cantidad de párrafos para los tres personajes seleccionados.



Procesamiento Lenguaje Natural

Bag of Words

La técnica Bag of Words convierte el texto en una representación numérica. Se construye un vocabulario de palabras conocidas de todo el corpus (conjunto de documentos) y luego se cuenta cuántas veces aparece cada palabra en cada documento.

Vocabulario

Índice	Palabra
0	amigos
1	casa
2	el
3	en
4	gato
5	ladra
6	la
7	perro
8	se
9	sienta
10	son
11	ventana
12	y

La matriz resultante de aplicar la técnica Bag of Words a un corpus es conocida como una matriz dispersa (sparse matrix).

- Texto 1: "El gato se sienta en la ventana"
- Texto 2: "El perro ladra en la casa"
- Texto 3: "El gato y el perro son amigos"

Matriz Resultante

	amigos	casa	el	en	gato	ladra	la	perro	se	sienta	son	ventana	y
Texto 1	0	0	1	1	1	0	1	0	1	1	0	1	0
Texto 2	0	1	1	1	0	1	1	1	0	0	0	0	0
Texto 3	1	0	2	0	1	0	0	1	0	0	1	0	1

Procesamiento Lenguaje Natural

TF-IDF

n-grama es una secuencia continua de n elementos (palabras, caracteres, etc.) en un texto. En el procesamiento de lenguaje natural, los n-gramas se utilizan para capturar la relación entre palabras y la estructura de un texto.

- **Transformación TF-IDF (Term Frequency - Inverse Document Frequency)** es una técnica que pondera la frecuencia de las palabras en un documento considerando también la frecuencia inversa en el corpus total.
- **Term Frequency (TF)**: La frecuencia de una palabra en un documento.
- **Inverse Document Frequency (IDF)**: Una medida de cuánto de común o rara es una palabra en todos los documentos del corpus.

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log\left(\frac{N}{\text{DF}(t)}\right)$$

Matriz TF-IDF:

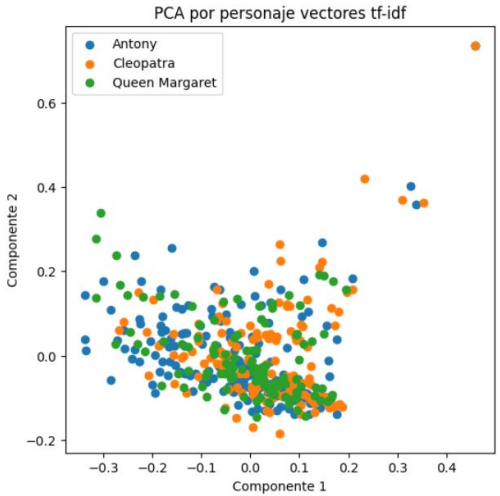
	amigos	casa	el	el gato	el perro	en	en la	gato	gato el	gato se	...	se	se sienta	sienta	sienta en	son	son amigos	ventana
Texto 1	0	0	0.18456	0.237655	0	0.237655	0.237655	0.237655	0	0.312487	...	0.312487	0.312487	0.312487	0.312487	0	0	0.312487
Texto 2	0	0.348349	0.205741	0	0.264928	0.264928	0.264928	0	0	0	...	0	0	0	0	0	0	0
Texto 3	0.338858	0	0.40027	0.25771	0.25771	0	0	0.25771	0.338858	0	...	0	0	0	0	0.338858	0.338858	0

Las palabras que son comunes en muchos documentos reciben una puntuación más baja, mientras que las palabras que son más distintivas para un documento reciben una puntuación más alta.

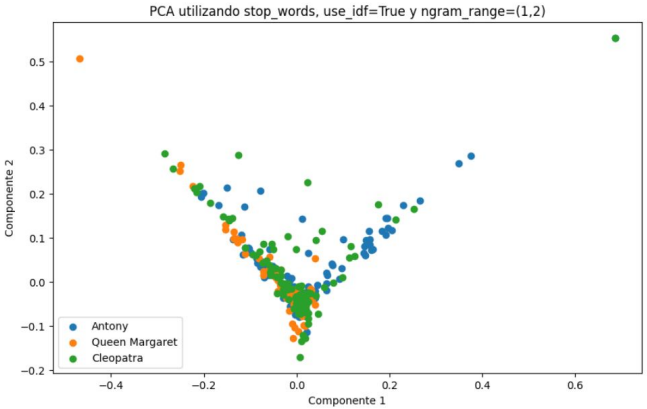
Procesamiento Lenguaje Natural

PCA

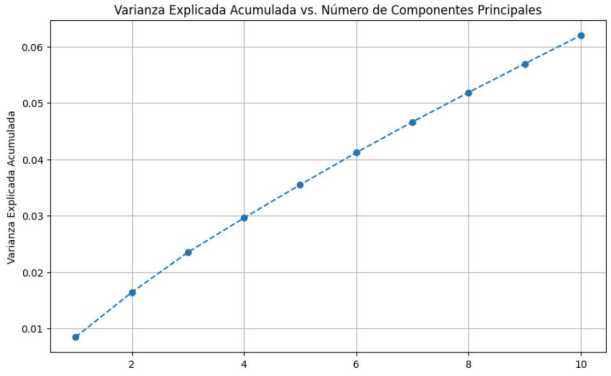
Aplicación de PCA sobre los vectores TF-IDF



PCA sobre el conjunto de entrenamiento considerando stop words en inglés y ngram=(1,2)



Variación de la Varianza Explicada



- Utilizando solo dos componentes principales, no es posible distinguir completamente entre los textos de Antony, Cleopatra y Queen Margaret.
- Observamos que las primeras dos componentes principales explican aproximadamente el 2% de la varianza total en los datos. Incluso al considerar hasta 10 componentes principales, la varianza explicada acumulada no supera el 6%.

Entrenamiento y Evaluación de Modelos

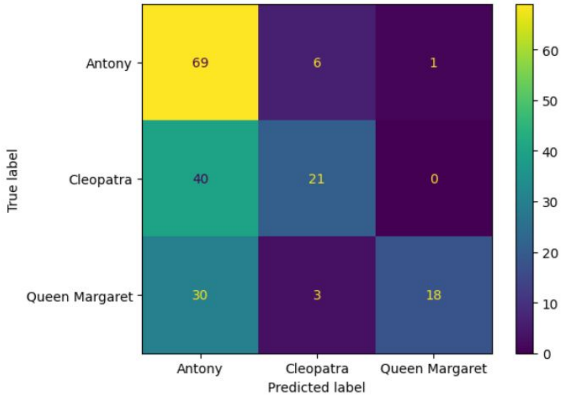
Multinomial Naive Bayes

La principal ventaja de este modelo es que estima distintas distribuciones para cada característica.

Accuracy: 0.574468085106383

Precisión y recall

	precision	recall	f1-score	support
Antony	0.50	0.91	0.64	76
Cleopatra	0.70	0.34	0.46	61
Queen Margaret	0.95	0.35	0.51	51
accuracy			0.57	188
macro avg	0.71	0.54	0.54	188
weighted avg	0.68	0.57	0.55	188



Entrenamiento y Evaluación de Modelos

Cross-Validation

La validación cruzada es una técnica utilizada para evaluar la capacidad de generalización de un modelo de aprendizaje automático. Consiste en dividir el conjunto de datos en varias partes o "pliegues". En cada iteración, una de las partes se utiliza como conjunto de validación y las restantes como conjunto de entrenamiento.

Rendimiento Global:

- Los modelos que utilizan stop words en inglés y bigramas tienden a tener accuracy más altas en general, con menores valores mínimos y máximos, indicando que estos parámetros podrían ser más efectivos.
- Por otro lado, los modelos sin stop words muestran una mayor dispersión y en algunos casos menor accuracy, lo que sugiere que la eliminación de stopwords podría mejorar el rendimiento.

Entrenamiento y Evaluación de Modelos

Regresión Logística

Es lineal en el espacio de los parámetros y puede manejar relaciones no lineales al combinar múltiples características. Este modelo es adecuado para problemas de clasificación binaria y multiclase, y puede ser regularizado para prevenir el sobreajuste.

Seleccionamos la regresión logística para este ejemplo debido a su simplicidad, eficiencia y buen rendimiento en problemas de clasificación de texto.

Entrenamiento y Evaluación de Modelos

FastText

La innovación de fastText fue plantear las palabras como una “bolsa” de n-gramas. Este proceso divide la palabra en pequeños grupos de caracteres, los agrupa cada tantos caracteres como se les haya indicado, luego se crea un único vector como la suma o promedio de todos los caracteres del n-grama. Este proceso de subdivisión de las palabras le da una ventaja con respecto a word2vec cuando se tiene que enfrentar con variaciones de palabras o palabras que están fuera del vocabulario con el que fue entrenado.

Final Classification Report with fastText:

	precision	recall	f1-score	support
Coriolanus	0.54	0.75	0.63	57
Lear	0.72	0.52	0.60	56
Queen_Margaret	0.75	0.65	0.69	51
accuracy			0.64	164
macro avg	0.67	0.64	0.64	164
weighted avg	0.67	0.64	0.64	164