Tarea a ser resuelta

La tarea abordada en este código se refiere a la clasificación de documentos de texto en categorías específicas. El conjunto de datos utilizado para esta tarea es "20 Newsgroups", que consiste en mensajes de grupos de noticias clasificados en varias categorías temáticas.

Proceso de normalización

Para llevar a cabo la tarea de clasificación, se aplicó un proceso de normalización de texto en el que se realizaron las siguientes etapas:

1. **Tokenización**: El texto original se dividió en tokens (palabras) utilizando la biblioteca spaCy.
2. **Limpieza de Texto**: Se realizó una limpieza del texto eliminando caracteres no alfabéticos y otras irregularidades del texto.
3. **Eliminación de Stop Words**: Se eliminaron las palabras comunes conocidas como "stop words" que generalmente no aportan información relevante para la clasificación.
4. **Lematización**: Se lematizaron las palabras para reducirlas a su forma base, lo que ayuda a unificar las diferentes formas de una palabra.

Representación de texto

Se crearon tres representaciones diferentes de texto a partir de los documentos normalizados:

1. **Binarizada**: Se utilizó una representación binaria que codifica la presencia o ausencia de palabras en un documento.
2. **Frecuencia de Términos**: Se creó una representación que cuenta la frecuencia de cada palabra en un documento.
3. **TF-IDF (Term Frequency-Inverse Document Frequency)**: Se utilizó una representación ponderada que considera la frecuencia de las palabras en un documento y la rareza de las palabras en el conjunto de documentos.

Métodos de Machine learning

Para la tarea de clasificación de texto, se emplearon los siguientes métodos de aprendizaje automático:

1. **Regresión Logística**: Se utilizó para clasificar documentos en categorías basándose en las representaciones de texto.

2. **Naive Bayes Multinomial**: Se empleó para llevar a cabo la clasificación utilizando el clasificador Naive Bayes Multinomial.

3. **Multilayer Perceptron (MLP)**: Se incorporó un clasificador MLP para la tarea de clasificación de texto. El MLP es un tipo de red neuronal que puede aprender relaciones más complejas en los datos.

Experimento	Normalización de texto	Representación de texto	Método de machine learning	f-score promedio
1	Tokenización + stopwords + text_cleaning + lematización	Binarizado	Logistic regression	0.79
2	Tokenización + stopwords + text_cleaning + lematización	Binarizado	MultinomialNB	0.80
3	Tokenización + stopwords + text_cleaning + lematización	Binarizado	MLP	0.83
4	Tokenización + stopwords + text_cleaning + lematización	Frecuencia	Logistic regression	0.79
5	Tokenización + stopwords + text_cleaning + lematización	Frecuencia	MultinomialNB	0.82
6	Tokenización + stopwords + text_cleaning + lematización	Frecuencia	MLP	0.82
7	Tokenización + stopwords + text_cleaning + lematización	TF-IDF	Logistic regression	0.82
8	Tokenización + stopwords + text_cleaning + lematización	TF-IDF	MultinomialNB	0.81
9	Tokenización + stopwords + text_cleaning + lematización	TF-IDF	MLP	0.84