TECNOLÓGICO NACIONAL DE MÉXICO

INTELIGENCIA ARTIFICIAL

UNIDAD 2 – TAREA 3

NOMBRE:

SARABIA GUZMÁN JESÚS ALDHAIR

ZAMUDIO LIZÁRRAGA BRYAN MARTÍN

NUMERO DE CONTROL:

22170824

22170855

FECHA:

30/03/2025

MAESTRO:

ZURIEL DATHAN MORA FÉLIX

Descripción del algoritmo:

Este código implementa un clasificador de correos electrónicos utilizando el algoritmo de Naive Bayes, basado en las probabilidades condicionadas de las características extraídas de los correos (palabras en este caso). El código preprocesa los datos de texto, realiza la extracción de características usando TF-IDF y luego clasifica los correos como "Spam" o "No Spam" según la probabilidad calculada.

Dependencias:

- **NumPy**: Para el manejo de arrays y cálculos matemáticos.
- **Pandas**: Para la manipulación de datos.
- **scikit-learn**: Para la vectorización del texto (TF-IDF).
- **nltk**: Para la normalización del texto, eliminación de palabras vacías (stopwords) y lematización.

Funciones:

- **preProcesar(texto):**
    - Recibe un texto y lo preprocesa eliminando caracteres especiales, convirtiendo a minúsculas, lematizando y filtrando palabras vacías.
    - Devuelve el texto procesado.
- **clasificar(email):**
    - Recibe un correo electrónico, lo preprocesa, lo vectoriza usando TF-IDF, y calcula la probabilidad de que el correo sea spam o no spam utilizando la fórmula de Bayes.
    - Devuelve "Spam" si la probabilidad de spam es mayor a 0.3, de lo contrario, devuelve "No Spam".

Pasos del Código:

1. Importación de Librerías:

- Se importan las bibliotecas necesarias para la manipulación de datos y el procesamiento de texto, como numpy, pandas, nltk, y sklearn.

2. Carga del Dataset:

- Se carga el dataset spam_assassin.csv que contiene correos etiquetados como "spam" o "no spam" y sus respectivos contenidos de texto.

3. Preprocesamiento del texto:

- Se define la función preprocesar() que realiza el preprocesamiento sobre el texto de cada correo en el dataset: convierte el texto a minúsculas, elimina caracteres especiales, lematiza las palabras y filtra las stopwords en inglés.

4. Extracción de características con TFIDF

- Se utiliza TfidVectorizer para convertir el texto en valores numéricos basado en TFIDF

5. Cálculo de Probabilidades:

- Se calcula la probabilidad de que un correo sea spam (P_spam) y no spam (P_no_spam).
- Se calculan las frecuencias de las características (palabras) en correos etiquetados como spam y no spam.
- Se calcula la probabilidad de cada característica dado que el correo es spam (P(características|Spam)) y dado que el correo no es spam (P(características|No Spam)).

6. Clasificación de Correos:

- Se define la función clasificar(email) que toma un correo, lo preprocesa y lo vectoriza usando el modelo TF-IDF entrenado.
- Se calcula la probabilidad de que el correo sea spam usando la fórmula de Bayes

- Si la probabilidad de que el correo sea spam es mayor a 0.3, se clasifica como "Spam", de lo contrario, como "No Spam".

8. Evaluación del Modelo:

- Se evalúa el desempeño del clasificador comparando las predicciones con las etiquetas reales del dataset.
- Se calcula la precisión (porcentaje de clasificaciones correctas) y la recuperación (porcentaje de correos spam correctamente clasificados).

9. Impresión de Resultados:

- Se imprime la precisión y la recuperación obtenida en el modelo.

Prueba de ejecución