PRUEBA DE CONOCIMIENTOS DATOS NO ESTRUCTURADOS I

Hebert Gomez

Clasificación de Documentos Basada en Contenido

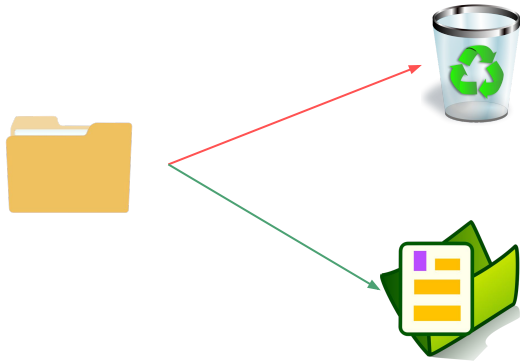
Se requiere implementar un sistema que clasifique imágenes de documentos en dos categorías: con contenido relevante y sin contenido relevante.

1. Filtrado de imágenes sin datos de entrenamiento.

Se clasifican documentos basados en su contenido utilizando técnicas de procesamiento de imágenes y aprendizaje **no supervisado** para mejorar la organización y accesibilidad de los documentos.

Definición de Rutas:

- **Rutas de Entrada y Salida:**
  - **Entrada:** Carpeta **images** con todas las imágenes de documentos.
  - **Salidas:**
    - **Con contenido:** Carpeta para documentos con contenido relevante.
    - **Sin contenido:** Carpeta para documentos sin contenido relevante.



Preprocesamiento y Extracción de Características

Mediante la implementación de **OpenCV** se realiza**:**

- La conversión de las imágenes a escala de grises para una mejor consistencia en el análisis.
- Redimensionamiento de imágenes a un tamaño uniforme de 128x128 píxeles.
- Normalización de imágenes para tener valores entre 0 y 1, lo que facilita el procesamiento.

Con la tecnica **HOG (Histogram of Oriented Gradients):**

- Extracción de características que capturan la estructura y gradiente de las imágenes.
- HOG es eficaz para detectar patrones y objetos dentro de las imágenes, lo cual es crucial para identificar contenido relevante en documentos.

Reducción de Dimensionalidad

Para mejorar el rendimiento y optimizar recursos, se emplea la reducción de la dimensionalidad con **PCA** con un parámetro de **n_components=50**, lo que ofrece un buen equilibrio entre retener suficiente información y reducir significativamente la dimensionalidad de los datos originales**.** Con esto se quiere lograr:

- Transformación de características en un conjunto de componentes principales no correlacionadas.

- Reducción de la cantidad de datos para acelerar el proceso de clustering y mejorar la eficiencia computacional.

Clustering

Se implementa el método K-means Clustering, el cual es un algoritmo que agrupa datos en K clústeres, en este caso, dos clústeres o categorías: documentos con contenido relevante y documentos sin contenido relevante.

K-means minimiza iterativamente la suma de las distancias al cuadrado entre los datos y los centroides de los clústeres, este enfoque asegura que los datos dentro de cada clúster sean lo más similares posible entre sí, optimizando así el agrupamiento.

Clasificación y Movimiento de Imágenes:

- **Asignación de Clústeres:**
  - Cada imagen se asigna a un clúster basado en las etiquetas generadas por K-means.
- **Reubicación de Imágenes:**
  - Las imágenes se mueven a las carpetas correspondientes utilizando `os.rename`.

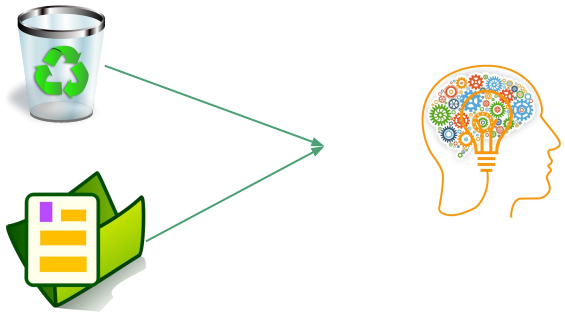
Con estos procesos se genera la clasificación de las imágenes en los directorios correspondientes, teniendo en cuenta su contenido:

Filtrado de imágenes con etiquetas

El método actual mejora el proceso de clasificación de documentos mediante la combinación de técnicas avanzadas de procesamiento de imágenes y aprendizaje supervisado. Este enfoque proporciona una clasificación más precisa y eficiente, facilitando una mejor organización y accesibilidad de los documentos.

Definición de Rutas:

- **Carpeta de entrada:** trained_images/
  - folders = ['Blanco', 'Documentos']
- **Carpetas de salida:** docs_con_contenido/ y docs_sin_contenido/



Técnicas Utilizadas

OpenCV y HOG

- Carga y preprocesamiento de imágenes
- Extracción de características estructurales

PCA

- Reducción de dimensionalidad para conservar características relevantes y mejorar eficiencia computacional

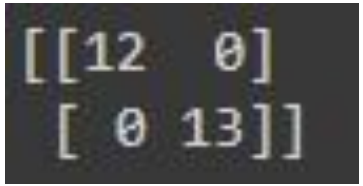
K-means Clustering

- Algoritmo de agrupamiento para separar imágenes en dos categorías (**n_clusters=2**).

Modelo Secuencial de Keras

- Modelo de red neuronal utilizado para clasificar las imágenes basándose en las características extraídas.
- Se define un modelo secuencial de Keras con capas **Dense** para clasificar las imágenes en las dos clases (documentos con y sin contenido relevante.

Matriz de Confusión



La matriz de confusión es una herramienta fundamental en la evaluación del rendimiento de los modelos de clasificación. Proporciona una imagen clara de cómo el modelo está categorizando las muestras y permite identificar rápidamente cualquier desequilibrio o problema en las predicciones.

Interpretación del Resultado:

- **Precisión del Modelo:** El modelo tiene una precisión perfecta para ambas clases, ya que todos los ejemplos se clasificaron correctamente.
  - Para la clase 0, se predijeron correctamente las 12 muestras y ninguna se clasificó incorrectamente.
  - Para la clase 1, se predijeron correctamente las 13 muestras y ninguna se clasificó incorrectamente.

En este caso, la matriz de confusión muestra que el modelo tiene un rendimiento excelente en la tarea de clasificación de documentos con y sin contenido relevante.

Aplicación para sistema de filtrado de imágenes

Para la “apificación” del desarrollo, se sugiere tener un sistema compuesto por: