|  |  |  |
| --- | --- | --- |
| T3 | **Minería de Datos**  Ana Karen Huerta Hernández #13  Isabel Nava Martínez #21 | 8”C” |

**Introducción OCR**

(Reconocimiento Óptico de Caracteres), o OCR, es una tecnología que le permite convertir diferentes tipos de documentos, tales como documentos en papel escaneados, PDF archivos o imágenes captadas por una cámara digital en datos con opción de búsqueda y funcionalidad de editar.

**Requerimientos**

* Equipo de cómputo
* Spyder instalado
* Imágenes Binarias

**Clasificación**

El método que se utilizó para clasificar es KNN para clasificación en donde la entrada es introducir el número de vecinos que queremos encontrar y nos devuelve esos vecinos y sus características además de clase y el número de instancia.

**Conjunto de imágenes**

Las imágenes binarias tienen un tamaño de 55x88, Las clases son del 1 hasta el número 9 y alfabéticamente de la letra A *hasta la* Z.

**Imágenes**

Una imagen binaria es una imagen digital que tiene únicamente dos valores posibles para cada pixel Normalmente, los colores utilizados para su representación son negro y blanco aunque puede usarse cualquier pareja de colores. Uno de los colores se emplea como fondo y el otro para los objetos que aparecen en la imagen.

En el programa fue desarrollado en Python  es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. En éste código se toman las carpetas que contienen carpetas con imágenes del 0 hasta la letra Z que con estas imágenes el programa.

**¿Qué es un DataSet?**

El DataSet es una representación de datos residente en memoria que proporciona un modelo de programación relacional coherente independientemente del origen de datos que contiene.

Un DataSet representa un conjunto completo de datos, incluyendo las tablas que contienen, ordenan y restringen los datos, así como las relaciones entre las tablas.

**Generar DataSet**

* Primer paso las Imágenes segmentadas se basa en leer carpetas.
* Segundo paso, dentro de cada una de las carpetas se lee cada imagen.
* Tercer paso por cada una de las imágenes se va obteniendo las características.

**Características:**

Característica 1



Saca la razón= columnas/filas.

**Característica 2**



Ejemplo si la imagen es de 5\*4 su área será 20

Y recorre filas y columas cuando encuentra unos(1) los va contabilizando

Al final divide el área entre los (1’s) encontrados

En esta característica se saca el área de la imagen y se cuentan los unos(1) que existen en la imagen

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 |  |  | 1 |  |

**Característica 3, 4,5 (Columnas)**



Cuenta los números de (1’s) y los divide entre el valor del vector. Este procedimiento se realiza 3 veces ya que como lo indica la imagen se hace a la mitad, 3/4y ¼ de la antes mencionada.

**Característica 6, 7,8 (Filas)**

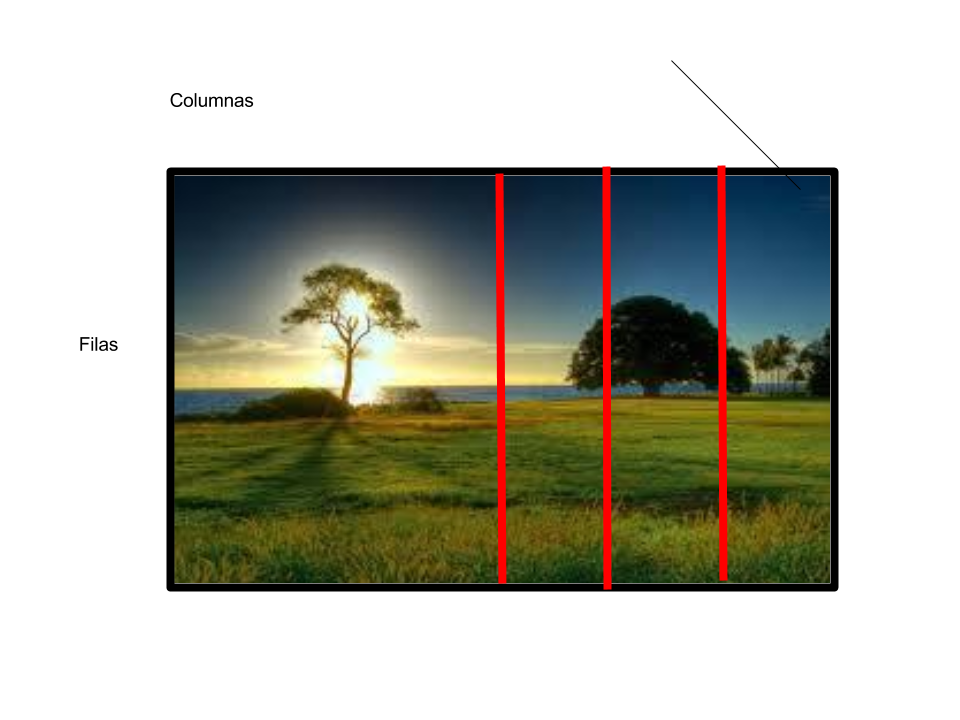
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1 |  |  | 1 |  |



Cuenta los números de (1’s) y los divide entre el valor del vector. Este procedimiento se realiza 3 veces ya que como lo indica la imagen se hace a la mitad, 3/4y ¼ de la antes mencionada.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 |  | 1 |  | 1 |

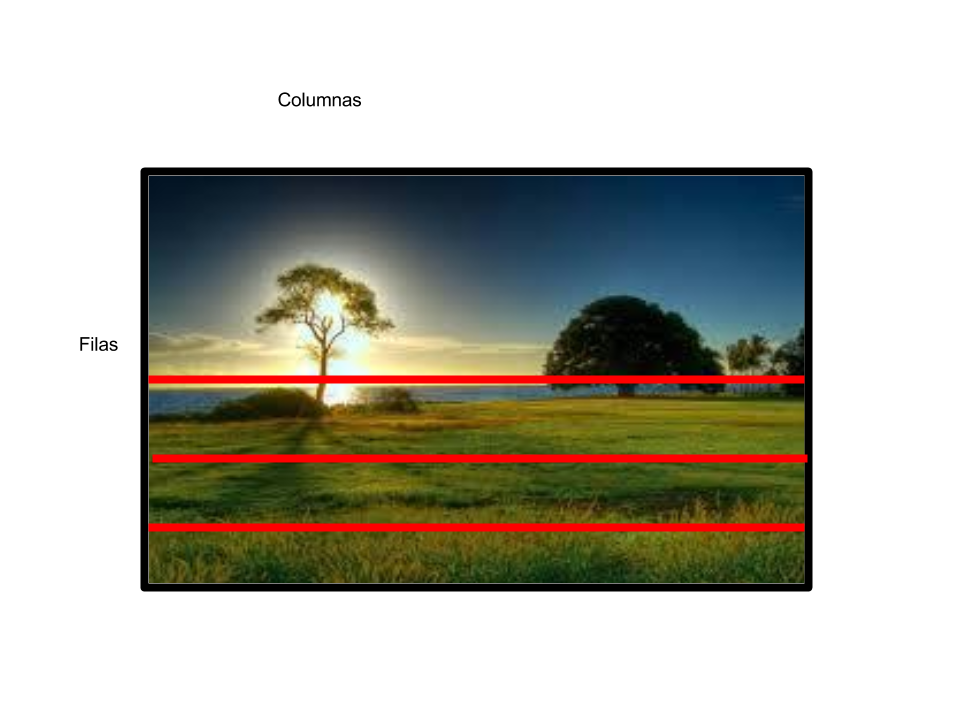
**Característica 9, 10,11**



Cambio de 0 a 1

En estas características se cuentan los cambios de 1 a 0 en las columnas de ½,3/4 y ¼ y después divide los cambios entre 2 para determinar los cortes en ese vector.

**Característica 12, 13,14**



|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 |  | 1 |  | 1 |

Cambio de 0 a 1

En estas características se cuentan los cambios de 1 a 0 en las filas de ½,3/4 y ¼ y después divide los cambios entre 2 para determinar los cortes en ese vector.

Knn

Menú

**Funcionamiento del programa**

Cargar Dataset

Obtener vecinos más cercanos

Clasificar imagen

Recorrido de directorios

Crear DataSet

Extracción de Características

Generar Dataet

Principal

Opciones

Métodos

Salida

**Código disponible**

**https://github.com/anakarenhuhe/codeOCR.git**