T V1 3	**Minería de Datos** #03 Iván Briones Salazar #10 Donaldo Flores Molina	8° B	E#

Requerimientos:

***Computadora con Zorin Linux

*Sublime Text3

*Carpeta con imágenes del 0 al 9 segmentadas y binarizadas.

OCR

OCR (Optical Character Recognition) es el reconocimiento óptico de caracteres, es una tecnología que le permite convertir diferentes tipos de documentos, como escaneados, archivos PDF, imágenes captadas por una cámara digital en datos con opción de búsqueda y funcionalidad de editar.

Para reconocer los caracteres, el software inspecciona la imagen, buscando formas que coincidan con los rasgos de los caracteres. Éste buscará coincidencias con los caracteres y fuentes disponibles en el programa, identificando los caracteres a través del análisis de sus características.

Conjunto de Imágenes

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

Conjunto: Se tiene un total de 2380 imágenes, distribuidas en 10 carpetas con 238 imágenes cada una.

Las siguientes imágenes muestran un ejemplo de las dimensiones de cada imagen contenida en las carpetas:

Carpeta 0:

Imagen 1.- Dimensiones estándar de la imagen: 56x88 píxeles.

Carpeta 1:

Imagen 2.-Dimensiones estándar de la imagen: 32x86 píxeles.

DataSet

Un DataSet es un conjunto de datos, representa un conjunto completo de datos.

El Dataset utilizado en esta práctica se obtiene a partir de las siguientes características:

Característica 1

Se obtienen la razón de filas y columnas de toda la imagen:

Razón total = filas/columnas

Imagen 3.- Imagen binaria que indica filas y columnas de la imagen.

Característica 2

Calcular el número de unos que existen en toda la imagen y este resultado se divide sobre el tamaño total de la imagen, es decir: #1's / (#Columnas * #filas)

Imagen 4.- Imagen binaria en donde los unos corresponden a l área que ocupa la imagen.

Característica 3, 4 y 5

Calcula cuantos unos existen en la fila que se encuentra a la mitad, a un cuarto y a tres cuartos de la imagen, y este resultado se divide sobre el número total de columnas.

Característica 6, 7,8

Calcula cuantos unos existen en la columna que se encuentra a la mitad, a un cuarto y a tres cuartos de la imagen, este resultado se divide sobre el número total de filas.

Imagen 5.- Imagen binaria que indica las posiciones en las que se calculan el número de unos existentes en la imagen, características 3 a 8.

Característica 9, 10,11

Calcula el número de cortes, es decir: el número de cambios de cero a uno y uno a cero que existen en la imagen, en la mitad, un cuarto y tres cuartos de las filas de la misma.

Característica 12, 13,14

Calcula el número de cortes, es decir: el número de cambios de cero a uno y uno a cero que existen en la imagen, en la mitad, un cuarto y tres cuartos de las columnas de la misma.

Cortes →

Imagen 6.- Imagen binaria que indica los cortes en las posiciones de la imagen 6.

*Nota: La nueva instancia es una imagen segmentada y binarizada.

Clasificación

Se utilizó el método KNN para clasificar, que consiste en introducir el número de vecinos que se encuentren más cercanos a la nueva instancia, a partir de ellos se realiza una votación para saber que clase se repite más entre estos. La clase con mayor repetición será a la que pertenece la nueva instancia.

Procedimiento:

Generación del dataset:

Fase 1: Leer las carpetas

Fase 2: Por carpeta leer imágenes

Fase 3: Extraer las 14 características de cada una de las imágenes y se guardan en un archivo .csv con el nombre de: dataset.csv

Clasificación:

Fase 4: Pedir al usuario que seleccione una imagen que será la nueva instancia.

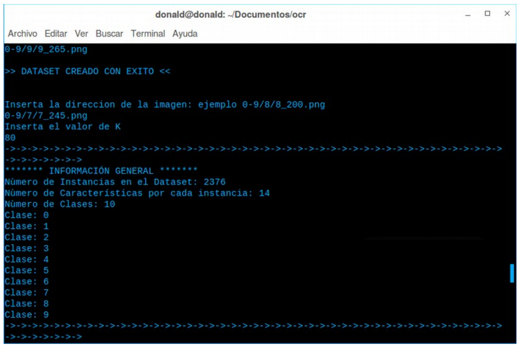
Fase 5: Se obtienen las 14 características de la nueva instancia.

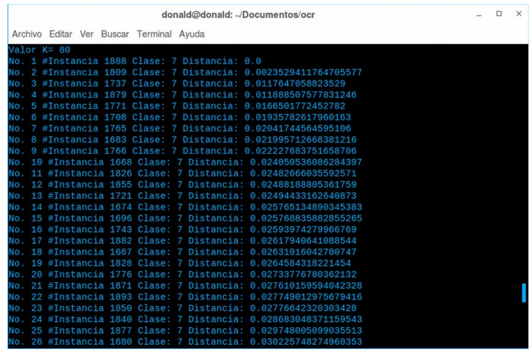
Fase 6: Se pide al usuario el valor de K.

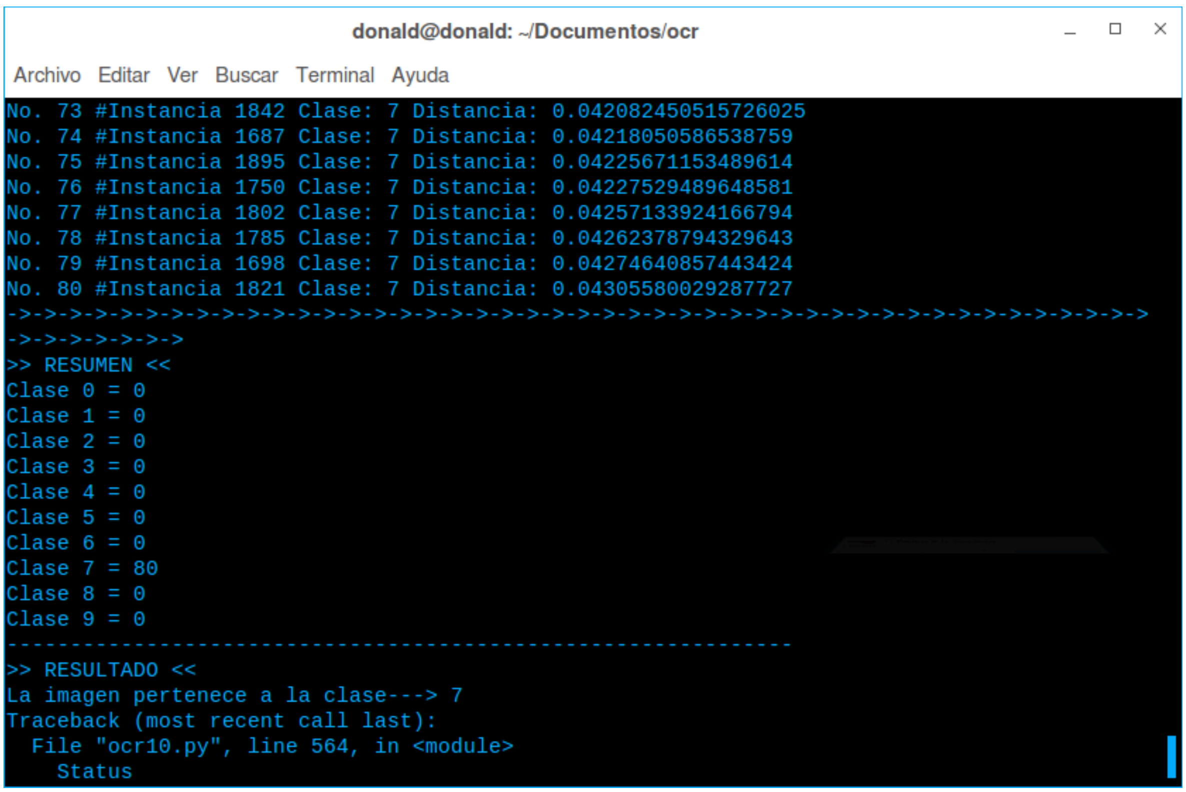
Fase 7: Se aplica KNN entre el dataset y la nueva instancia.

Fase 8: Se muestra en pantalla a que clase pertenece la nueva instancia.

Resultados:







Salida del programa

Código Github

http://github/osogoloso/OCR/