T3	**Minería de Datos** #15 Narda Teresa Pérez Arellano #24 Paul Torres Rivera	8°B	E1

Requerimientos:

- Equipo de cómputo con Spyder instalado
- Carpeta de Imágenes binarias

Introducción

OCR (Reconocimiento Óptico de Caracteres), es una tecnología que permite convertir diferentes tipos de documentos escaneados, como PDF archivos o imágenes captadas por una cámara digital, en datos con opción de búsqueda y funcionalidad de editar. Este nos permite conocer características únicas de un número o letra. Así va comparando cada carácter para ver a que pertenece, si numero o letra.

Conceptos

DataSet

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.

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.

Clasificación

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

Conjunto de imágenes

Para poder generar el DataSet, se utilizara un conjunto de imágenes. Estas imágenes son binarias y están categorizadas. Estas imágenes van del 0 al 9 y de la A a la Z (imagen 1).

Imagen 1: Conjunto de imágenes de distintas categorías.

Las imágenes tienen un tamaño promedio de 55x88, son aproximadamente de 250 a 300 imágenes por clase. Las clases que hay son: 1,2,3,4,5,6,7,8,9,A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y *y* Z.

En total son: 15,378 imágenes binarias

Creación del DataSet

Para generar el DataSet se necesitó realizar los siguientes pasos:

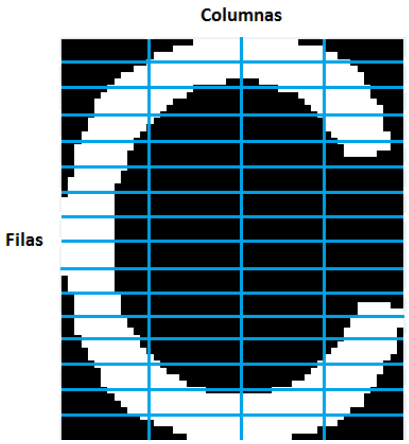
Paso 1: Leer carpetas con imágenes segmentadas.

Paso 2: Por cada carpeta se lee cada imagen.

Paso 3: Por cada imagen se obtienen las siguientes características y se escribe en un documento CSV:

Característica 1:

Es la razón de filas y columnas de la imagen. Razón= columnas/filas (imagen 2).



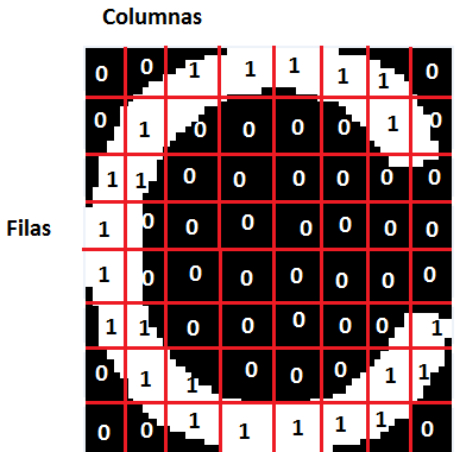
Ejemplo, si la imagen es de 4 columnas y 16 filas se debe realizar una división.

Razón 4/16 = 0.25

Imagen 2: Imagen binaria C, con 4 columnas y 16 filas

Característica 2:

Se calcula el numero de 1’s que existen en la imagen y se divide entre la razón (imagen 3).



Ejemplo si la imagen es de 8*8 su área será 64

Se debe de recorrer las filas y columnas, cuando encuentra un uno, los va contabilizando. Al final divide el número de 1’s entre el área de la imagen.

Imagen 3: imagen binaria C de 8 filas y 8 columnas, con 1’s y 0’s.

Característica 3, 4, 5, 6, 7, 8:

Se cuenta el número de 1’s y los divide entre el valor de cada vector (imagen 4). Razón = 1s/tamVec

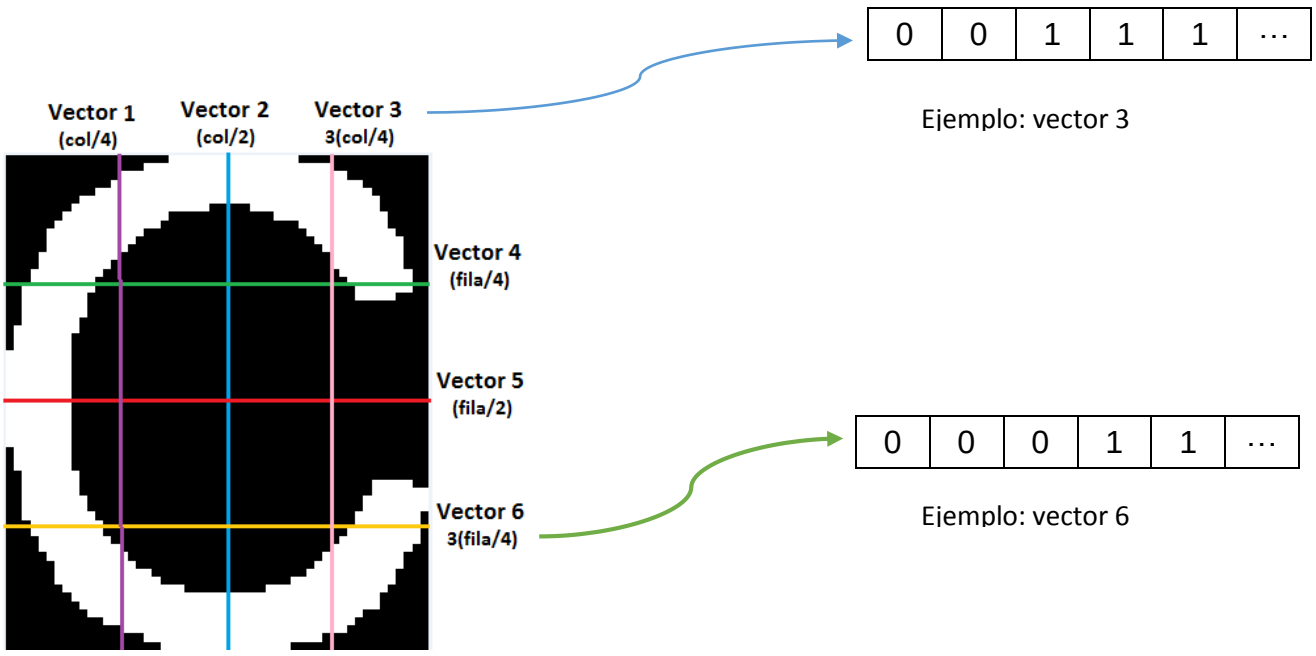


Imagen 4: imagen binaria C con 6 vectores que dividen esta imagen.

Característica 9, 10, 11, 12, 13, 14:

Se cuentan los cambios de 0 a 1 y viceversa en los vectores (imagen 5). La característica es el número de cambios (cortes) que hay en la imagen.

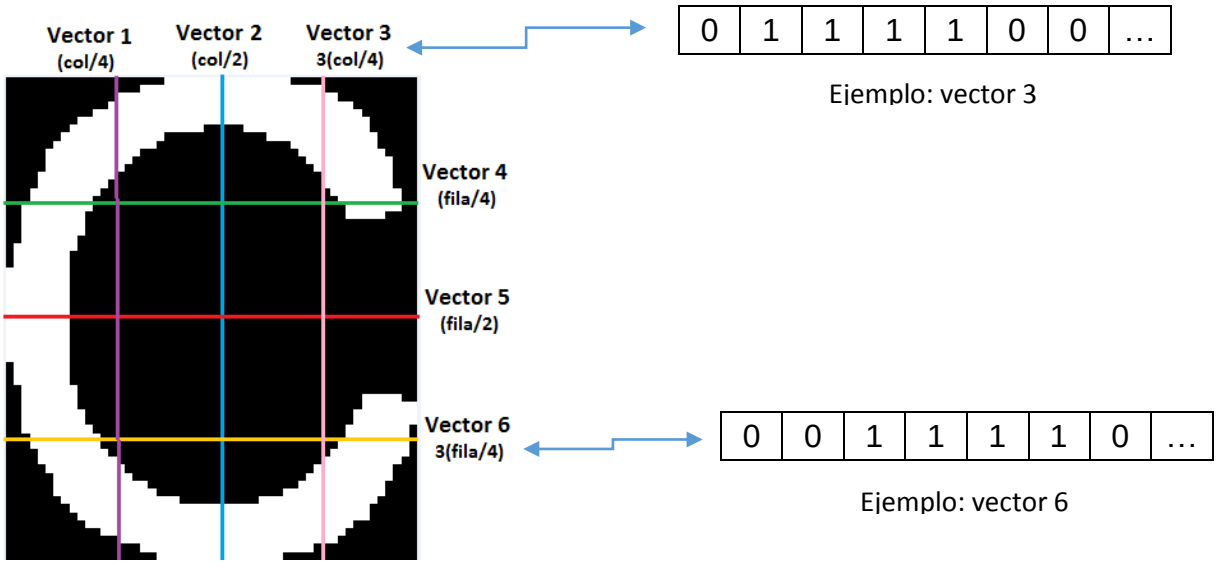
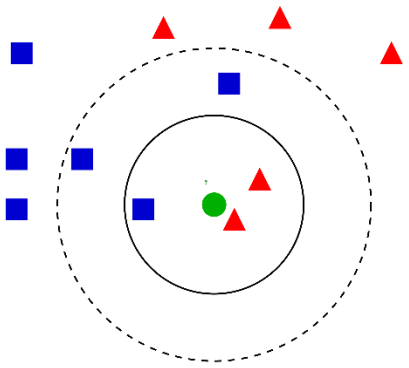


Imagen 5: imagen binaria C con 6 vectores que dividen esta imagen.

Paso 4: Una vez generado el DataSet se aplicara un método de clasificación. En este caso se utilizó KNN pero se puede aplicar cualquier otro método de clasificación. KNN clasifica dependiendo de las características obtenidas de cualquier instancia. En este método se ingresa una nueva instancia con características propias. Se debe de determinar un numero k de instancias a comparar. Este número k, debe ser impar ya que si se compara con un número par de instancias jamás podremos decir a que grupo de características pertenece. Una vez comparada la nueva instancia con el conjunto de características (DataSet), se determinara los "vecinos cercanos" con ello sabremos a que grupo pertenece.



Ejemplo: Tenemos una nueva instancia representada por el circulo verde. Queremos obtener k=3 vecinos cercanos. Este método nos devuelve los 3 vecinos más cercanos. Con esta información podemos decir que de esos tres vecinos el círculo es más cercano al grupo de triángulos rojos.

Imagen 6: imagen ejemplo de clasificación knn

Corrida

Capturas de pantalla del funcionamiento del programa.

Al correr el programa, se imprime un menú (imagen 7).

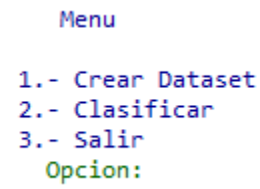


Imagen 7: Menú de OCR.

Si se elige la opción uno, se debe de contar con el conjunto de imágenes para poder generar el DataSet. Se mostrará un mensaje de como obtiene las características (imagen 8).

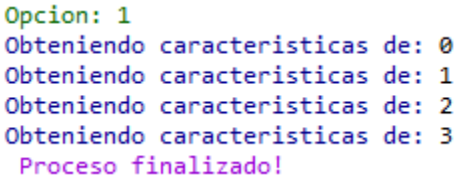


Imagen 8: Opción 1, obtención de características de cada clase.

Si se elige la opción dos, se debe de contar con el DataSet generado en el paso anterior. Se mostrará la información general del DataSet y se pedirá continuar (imagen 9).

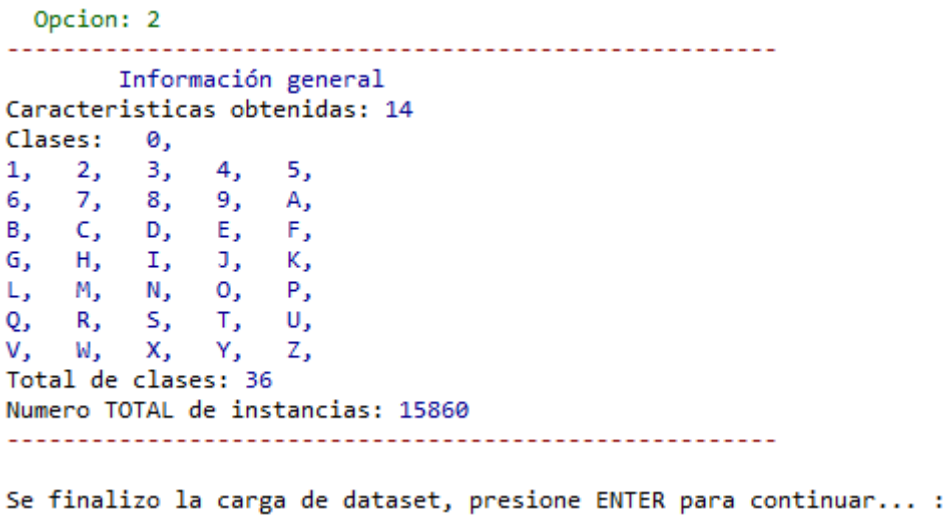


Imagen 9: Opción 2, Información general del DataSet.

Después se pide ingresar una imagen binaria para determinar sus características y se pedirá el número k de vecinos para comparar las características de la imagen binaria con las del conjunto de imágenes (DataSet). Se mostrará la información de los k vecinos más cercanos. Se imprime un resumen de KNN, mostrando el número de instancias y la votación. Al final se mostrara un mensaje determinando a que pertenece la imagen (imagen 10).

```
Ingresa el nombre de la imagen: 7.png

Ingresa el numero K: 7
-----------------------------------------------------
K vecino   Linea(Instancia)  Clase:  Distancia:

1              1668           7       0.00000

2              1705           7       0.01000

3              1858           7       0.01000

4              1857           7       0.01000

5              1747           7       0.02000

6              1838           7       0.02000

7              1693           7       0.02000
-----------------------------------------------------

     Resumen KNN

Instancias de la clase   Votación
       7                7

     LA IMAGEN ES UN: 7  <-----
```

Imagen 10: Opción 2, Ingreso de imagen, numero de k e impresión de k vecinos, resumen KNN y determinación de imagen.

Si se elige la opción tres, se imprime un mensaje de adiós y termina el programa (imagen 11).

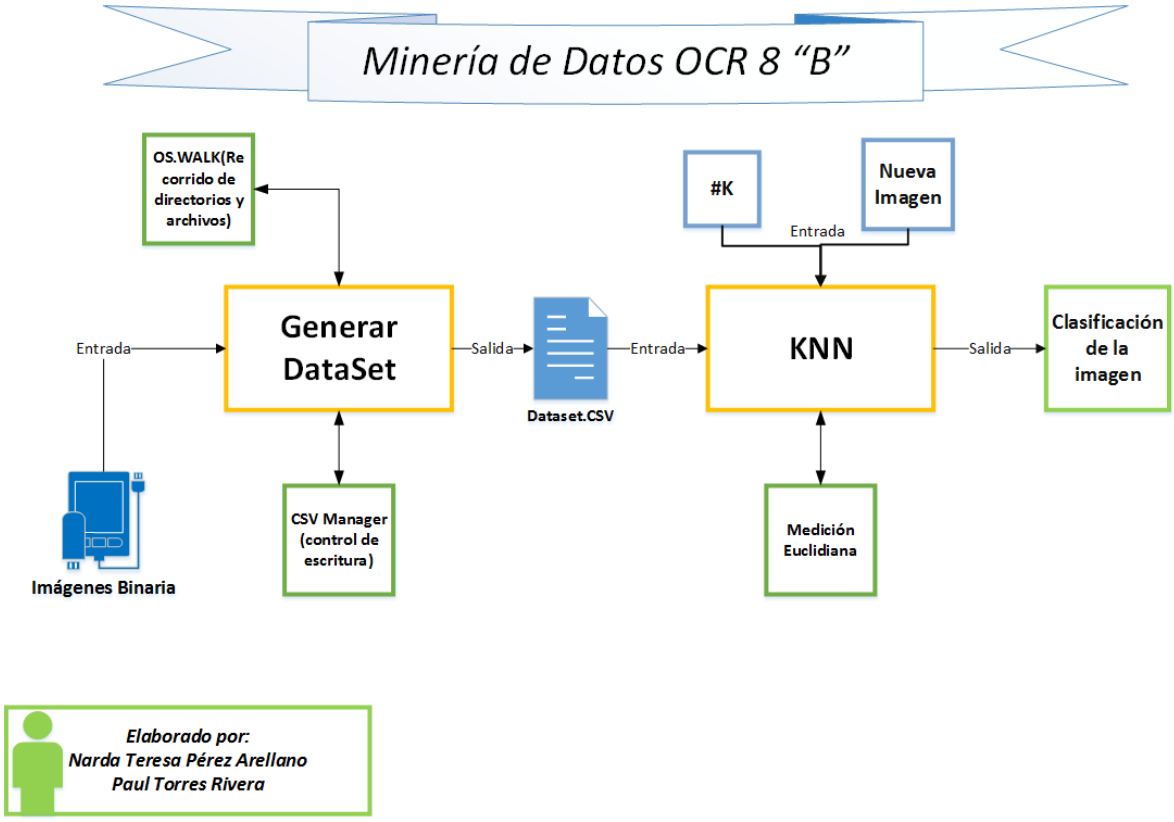
```
          Menu

    1.- Crear Dataset
    2.- Clasificar
    3.- Salir
      Opcion: 3

            Adios...!
```

Imagen 10: Opción 2, Salida del programa.

Diagrama de bloques



Enlace de Github:

http://github.com/poltorres7/Clasificacion-OCR/tree/master/ocrP