T3	Minería de datos 24 Pérez León Gabriela 30 Tecuapacho López Jorge Gontran	8C

Introducción

SIGLAS OCR: OCR es la sigla de Optical Character Recognition, una expresión en lengua inglesa que puede traducirse como Reconocimiento Óptico de Caracteres.

Para que sirve OCR

Es una tecnología que permite transformar el contenido de una imagen en texto plano. Normalmente, el contenido de una imagen que suele transformarse es aquél asociado a cadenas de texto, si bien algunas aplicaciones para OCR permiten transformar otro tipo de objetos gráficos contenidos en una imagen, como pueden ser, por ejemplo, códigos de barras.

En donde se aplica OCR:

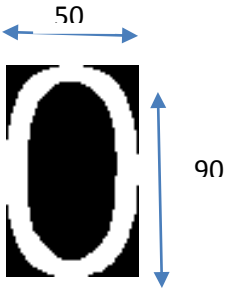
- Reconocimiento de texto manuscrito
- Reconocimiento de matrículas
- Indexación en bases de datos
- Reconocimiento de datos estructurados con ROC Zonal

Requerimientos

1. Sistema operativo Windows 8.1
2. Anaconda 2.4.1 para 64 bits
3. Python 3.5.1
4. Microsoft Excel 2013

Conjunto de imágenes

El conjunto de imágenes usadas para la creación del DataSet está compuesto por 2,380 imágenes binarias divididas en 10 carpetas, son imágenes binarizadas las cuales anteriormente pasan por un proceso de segmentación donde se separan las imágenes, después se eliminan las imágenes de ruido y quedan solo las imágenes binarizadas.



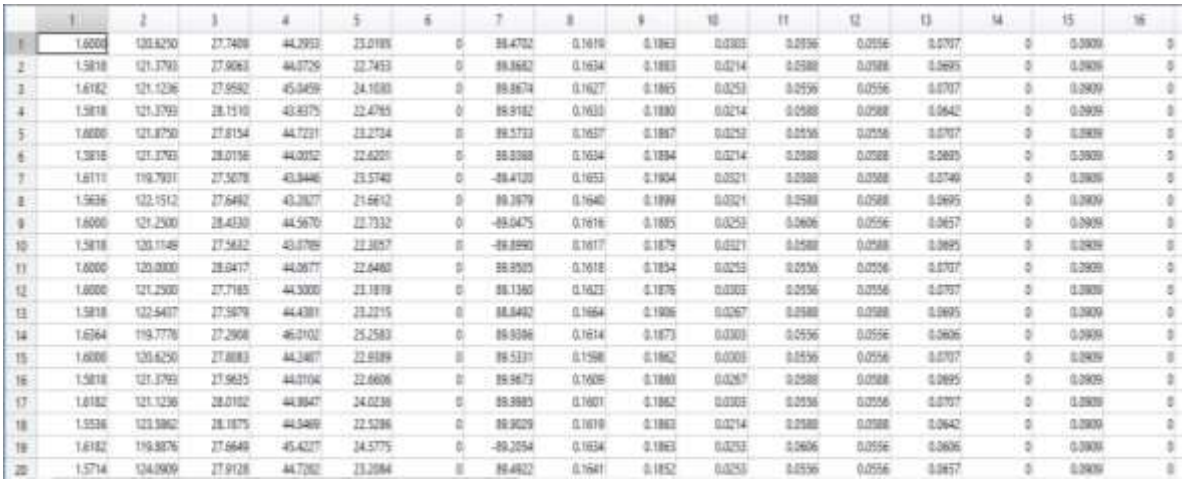
Total números 0-9 = 2380

Tamaño de las imágenes

Las imágenes oscilan en un tamaño entre 50x90 píxeles

Dataset

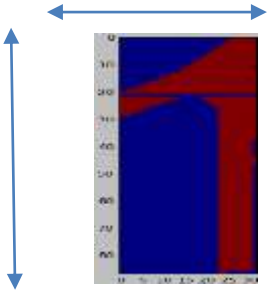
Representación de datos residente en memoria que proporciona una modelo de programación relacional coherente independientemente del origen de datos que contiene. El DataSet contiene en sí, un conjunto de datos que han sido volcados desde el proveedor de datos.Dataset con 2380 instancias y cada una está compuesta por 14 características



Como se genera el dataset

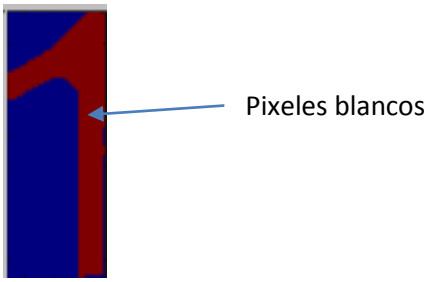
El dataset se genera mediante 14 características que son:

1. Razón Filas columnas



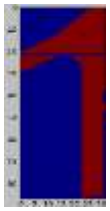
La primera característica es el resultado de obtener la relación que existe entre el alto y el ancho de la imagen en la cual se dividen las filas entre las columnas.

2. Razón Pixeles Blancos



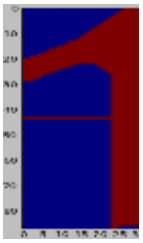
Segunda característica tomamos en consideración la cantidad de pixeles blancos de la imagen, y son sumados.

3. Cambios Primera Línea Horizontal



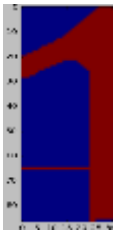
Tercera característica detectara los cambios de color de la primera línea horizontal que se ubica a un cuarto de la imagen.

4. Cambios Segunda Línea Horizontal



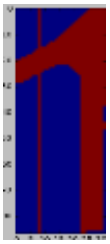
Cuarta característica detectara los cambios de color de la segunda línea horizontal que se ubica a la mitad de la imagen.

5. Cambios Tercera Línea Horizontal



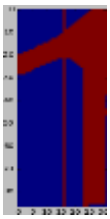
Quinta característica detectara los cambios de color de la tercera línea horizontal que se ubica a tres cuarto de la imagen.

6. Cambio Primera Línea Vertical



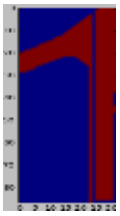
Sexta característica detectara los cambios de color de la primera línea vertical que se ubica a un cuarto de la imagen.

7. Cambio Segunda Línea Vertical



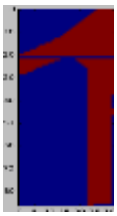
Séptima característica detectara los cambios de color de la segunda línea vertical que se ubica a la mitad de la imagen.

8. Cambio Tercera Línea Vertical



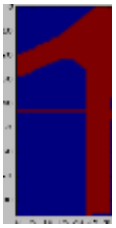
Octava característica detectara los cambios de color de la tercera línea vertical que se ubica a tres cuartos de la imagen.

9. Contar Pixeles Primera Línea Horizontal



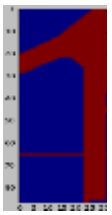
Novena característica cuenta los pixeles blancos que se encuentran en la primera línea horizontal que se ubica a un cuarto de la imagen.

10. Contar Pixeles Segunda Línea Horizontal



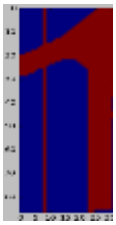
Decima característica cuenta los pixeles blancos que se encuentran en la segunda línea horizontal que se ubica a la mitad de la imagen.

11. Contar Pixeles Tercera Línea Horizontal



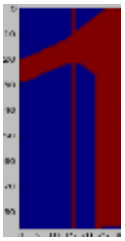
Onceava característica cuenta los pixeles blancos que se encuentran en la tercera línea horizontal que se ubica a tres cuartos de la imagen.

12. Contar Pixeles Primera Línea Vertical



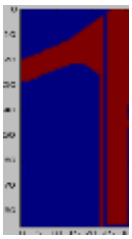
Doceava característica cuenta los pixeles blancos que se encuentran en la primera línea vertical que se ubica a un cuarto de la imagen.

13. Contar Pixeles Segunda Línea Vertical



Treceava característica cuenta los pixeles blancos que se encuentran en la segunda línea vertical que se ubica a la mitad de la imagen.

14. Contar Pixeles Tercera Línea Vertical



Catorceava característica cuenta los pixeles blancos que se encuentran en la tercera línea vertical que se ubica a tres cuartos de la imagen.

Que es K-nn para clasificación

Consiste en que dado una colección de registros cada registro contiene un conjunto de variables denominado “x” con una variable adicional “y” el objetivo es predecir la clase a la que pertenece cada registro. Para ello será necesario un conjunto de prueba la tabla de testing se utiliza para determinar la precisión del modelo.

Practica OCR

1. Correr el programa *Menu.py*
2. Elegir una de las tres opciones
3. Si es seleccionada la opción 1 generara el dataset con las 10 clases y 2380 instancias en un archivo *csv*
4. Si es seleccionada la opción 2 se pide el nombre de la imagen con extensión *.png* y el número de k-vecinos, se clasificara la imagen y se mostrara el resultado de los vecinos más cercanos y la clase a la que pertenece.
5. Si es seleccionada la opción 3 el programa termina su ejecución

```
No. Total de instancias:  2380
Instancia del K vecino mas cercano:  239

K vecinos mas cercanos:
Instancia: 239          Distancia: 0.0000     Clase 1
Instancia: 347          Distancia: 2.2361     Clase 1
Instancia: 294          Distancia: 2.2369     Clase 1

Número de Instancias por clase:
 3    Instancias de la clase: 1
 0    Instancias de la clase: 9
 0    Instancias de la clase: 8
 0    Instancias de la clase: 7
 0    Instancias de la clase: 6
 0    Instancias de la clase: 5
 0    Instancias de la clase: 4
 0    Instancias de la clase: 3
 0    Instancias de la clase: 2
 0    Instancias de la clase: 0


La imagen es un : 1
```

Diagrama del funcionamiento del programa *Menu.py*.