

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:

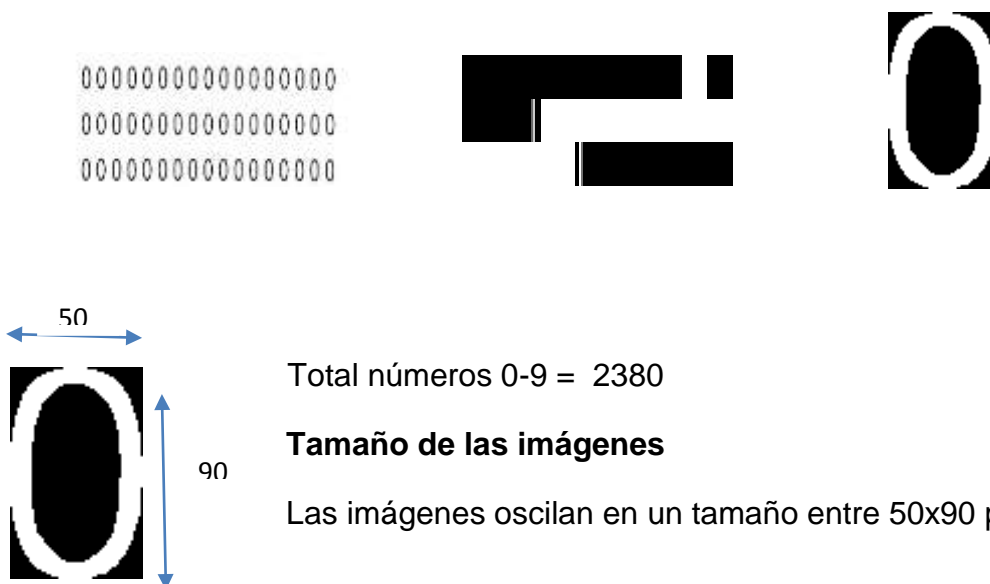
- Reconocimiento de texto manuscrito
- Reconocimiento de matrículas
- Indexación en bases de datos
- Reconocimiento de datos estructurados con ROC Zonal

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.



Dataset

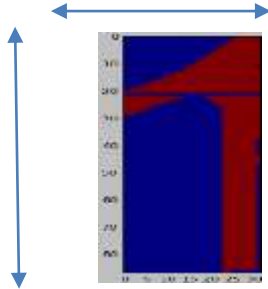
Representación de datos residente en memoria que proporciona un modelo de programación relacional coherente independientemente del origen de datos que contiene. El DataSet contiene en sí, un conjunto de datos que han sido volcados desde el proveedor de datos. Dataset con 2380 instancias y cada una está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

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	1.6000	120.6250	27.7400	44.2953	23.0100	0	88.4702	0.1610	0.1863	0.0303	0.0556	0.0556	0.0707	0	0.0600	0
2	1.5818	121.3793	27.9063	44.0729	22.7453	0	88.8682	0.1634	0.1883	0.0214	0.0588	0.0588	0.0695	0	0.0600	0
3	1.6182	121.1236	27.9582	45.0459	24.1083	0	88.8674	0.1627	0.1865	0.0253	0.0556	0.0556	0.0707	0	0.0600	0
4	1.5818	121.3793	28.1510	43.8375	22.4785	0	88.9182	0.1633	0.1880	0.0214	0.0588	0.0588	0.0642	0	0.0600	0
5	1.6000	121.0750	27.8154	44.7231	23.2734	0	88.5733	0.1637	0.1867	0.0253	0.0556	0.0556	0.0707	0	0.0600	0
6	1.5818	121.3793	28.0196	44.0052	22.6201	0	88.9388	0.1634	0.1894	0.0214	0.0588	0.0588	0.0695	0	0.0600	0
7	1.6110	119.7931	27.5078	43.8446	23.5742	0	88.4130	0.1653	0.1904	0.0321	0.0588	0.0588	0.0740	0	0.0600	0
8	1.5636	122.1512	27.6482	43.3877	21.5612	0	88.3979	0.1640	0.1899	0.0321	0.0588	0.0588	0.0695	0	0.0600	0
9	1.6000	121.2500	28.4330	44.5670	22.7332	0	88.0475	0.1616	0.1885	0.0253	0.0606	0.0556	0.0657	0	0.0600	0
10	1.5818	120.1148	27.5832	43.0789	22.3657	0	88.8890	0.1617	0.1879	0.0321	0.0588	0.0588	0.0695	0	0.0600	0
11	1.6000	120.0000	28.6417	44.0677	22.6480	0	88.8905	0.1618	0.1854	0.0253	0.0556	0.0556	0.0707	0	0.0600	0
12	1.6000	121.2500	27.7185	44.3000	23.1819	0	88.1360	0.1623	0.1876	0.0303	0.0556	0.0556	0.0707	0	0.0600	0
13	1.5818	122.5407	27.5978	44.4381	23.2215	0	88.8482	0.1664	0.1906	0.0267	0.0588	0.0588	0.0695	0	0.0600	0
14	1.6264	119.7778	27.2908	46.0102	25.2583	0	88.9386	0.1614	0.1873	0.0303	0.0556	0.0556	0.0606	0	0.0600	0
15	1.6000	120.6250	27.8883	44.2407	22.8189	0	88.5331	0.1598	0.1862	0.0303	0.0556	0.0556	0.0707	0	0.0600	0
16	1.5818	121.3793	27.9635	44.0104	22.6008	0	88.9673	0.1609	0.1880	0.0267	0.0588	0.0588	0.0695	0	0.0600	0
17	1.6182	121.1236	28.0182	44.8847	24.0236	0	88.9883	0.1601	0.1862	0.0303	0.0556	0.0556	0.0707	0	0.0600	0
18	1.5536	123.3862	28.1875	44.5489	22.5286	0	88.9029	0.1619	0.1883	0.0214	0.0588	0.0588	0.0642	0	0.0600	0
19	1.6182	119.8876	27.6648	45.4227	24.5775	0	88.2054	0.1634	0.1883	0.0253	0.0606	0.0556	0.0606	0	0.0600	0
20	1.5714	124.0909	27.9128	44.7262	23.2884	0	88.4922	0.1641	0.1852	0.0253	0.0556	0.0556	0.0657	0	0.0600	0

Como se genera el dataset

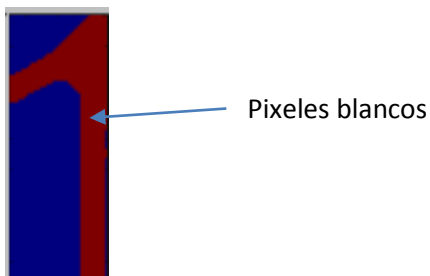
El dataset se genera mediante 14 características que son:

1. Razón Filas columnas



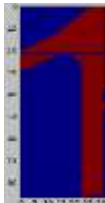
La primera característica es el resultado de obtener la relación que existe entre el alto y el ancho de la imagen en la cual se dividen las filas entre las columnas.

2. Razón Pixeles Blancos



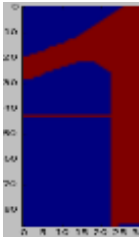
Segunda característica tomamos en consideración la cantidad de pixeles blancos de la imagen, y son sumados.

3. Cambios Primera Línea Horizontal



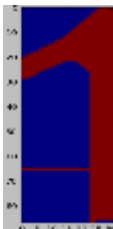
Tercera característica detectara los cambios de color de la primera línea horizontal que se ubica a un cuarto de la imagen.

4. Cambios Segunda Línea Horizontal



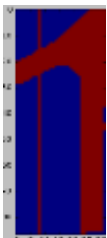
Cuarta característica detectara los cambios de color de la segunda línea horizontal que se ubica a la mitad de la imagen.

5. Cambios Tercera Línea Horizontal



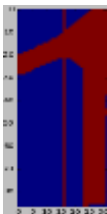
Quinta característica detectara los cambios de color de la tercera línea horizontal que se ubica a tres cuarto de la imagen.

6. Cambio Primera Línea Vertical



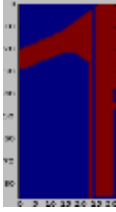
Sexta característica detectara los cambios de color de la primera línea vertical que se ubica a un cuarto de la imagen.

7. Cambio Segunda Línea Vertical



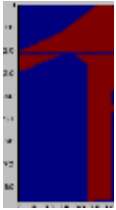
Séptima característica detectara los cambios de color de la segunda línea vertical que se ubica a la mitad de la imagen.

8. Cambio Tercera Línea Vertical



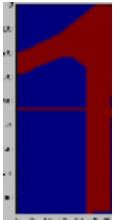
Octava característica detectara los cambios de color de la tercera línea vertical que se ubica a tres cuartos de la imagen.

9. Contar Pixeles Primera Línea Horizontal



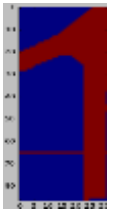
Novena característica cuenta los pixeles blancos que se encuentran en la primera línea horizontal que se ubica a un cuarto de la imagen.

10. Contar Pixeles Segunda Línea Horizontal



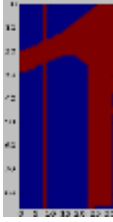
Decima característica cuenta los pixeles blancos que se encuentran en la segunda línea horizontal que se ubica a la mitad de la imagen.

11. Contar Pixeles Tercera Línea Horizontal



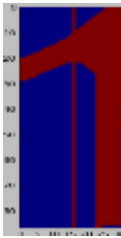
Onceava característica cuenta los pixeles blancos que se encuentran en la tercera línea horizontal que se ubica a tres cuartos de la imagen.

12. Contar Pixeles Primera Línea Vertical



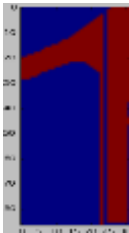
Doceava característica cuenta los pixeles blancos que se encuentran en la primera línea vertical que se ubica a un cuarto de la imagen.

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.

