|  |  |  |  |
| --- | --- | --- | --- |
| **T3** | **Minería de Datos**  #26 Yllescas Lamas Sandro Ivan  #30 Vásquez Conde Erick Ivan | **8B** | **#** |

**Requerimientos:**

* Python 3.5
* Una computadora
* Carpeta con imágenes segmentadas
* Método: Clasificación “KNN”

**Introducción**

**¿Qué es OCR?**

El OCR (Optical Character Recognition) es una tecnología que trata de emular la capacidad del ojo humano para reconocer objetos. Concretamente es un software que permite el reconocimiento óptico de los caracteres contenidos en una imagen (documento escaneado o fotografía), de forma que estos se vuelven comprensibles o reconocibles para un ordenador, obteniendo como resultado final un archivo en un formato de texto editable. El formato del archivo de salida (txt. pdf. etc.) dependerá de las posibilidades que ofrezca el software.

**¿Cómo funciona OCR?**

Para reconocer los caracteres, el software inspecciona la imagen pixel a pixel, buscando formas que coincidan con los rasgos de los caracteres.

En función del nivel de complejidad o grado de desarrollo del software, éste buscará coincidencias con los caracteres y fuentes disponibles en el programa, o tratará de identificar los caracteres a través del análisis de sus características, de forma que el reconocimiento de los mismos no se limite exclusivamente a un determinado número de fuentes.

El OCR puede analizar los elementos del documento (bloques de texto, imágenes, tablas…), examinando los espacios en blanco y descomponiendo el texto en líneas, palabras y caracteres, de forma que el programa puede formular distintas hipótesis y cotejarlas con los diccionarios contenidos por el mismo (actualmente los programas contienen diccionarios en distintos idiomas), para formar palabras y textos completos.

**¿Qué es knn?**

Es un método de clasificación supervisada (Aprendizaje, estimación basada en un conjunto de entrenamiento y prototipos) que sirve para estimar la función de densidad. Este es un método de clasificación no paramétrico, que estima el valor de la función de densidad de probabilidad o directamente la probabilidad a posteriori de que un elemento X pertenezca a la clase C, a partir de la información proporcionada por el conjunto de prototipos. En el proceso de aprendizaje no se hace ninguna suposición acerca de la distribución de las variables predictoras.

**IMÁGENES**

Utilizamos un total de 6 imágenes que están ubicadas en una carpeta nombrada

C:\Users\Erik Ivan\AppData\Local\Microsoft\Windows\INetCache\Content.Word\2.pngE:\simpleOCR-master\trainingSet\arialSegmented\6\6_010.pngC:\Users\Erik Ivan\AppData\Local\Microsoft\Windows\INetCache\Content.Word\8.png

“imágenes segmentadas”

Imagen 1. Imagen segmentada utilizada

**¿Qué es dataset?**

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

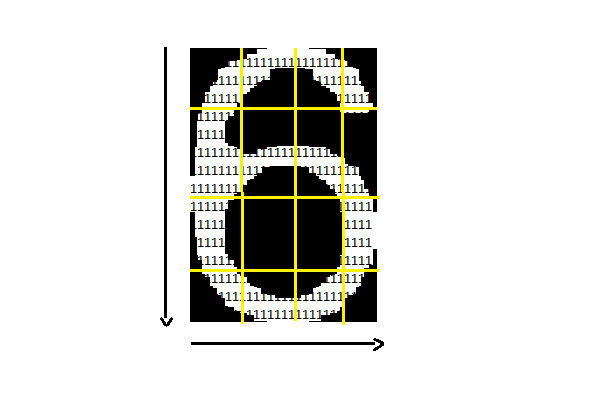
El **DataSet** contiene en sí, un conjunto de datos que han sido volcados desde el proveedor de datos.

**Creación de dataset**

Las 14 características de cada una de las imágenes se ingresaron en un archivo en Excel con todos sus datos.

**Características**

Característica 1: Se obtiene el número de columnas entre el número de filas, regresa la relación entre 1´s con respecto al tamaño de la imagen (total de 1´s/tamaño total de la imagen).



Característica 2: El número de 1´s entre el tamaño de la imagen, regresa la relación entre el ancho sobre el alto.



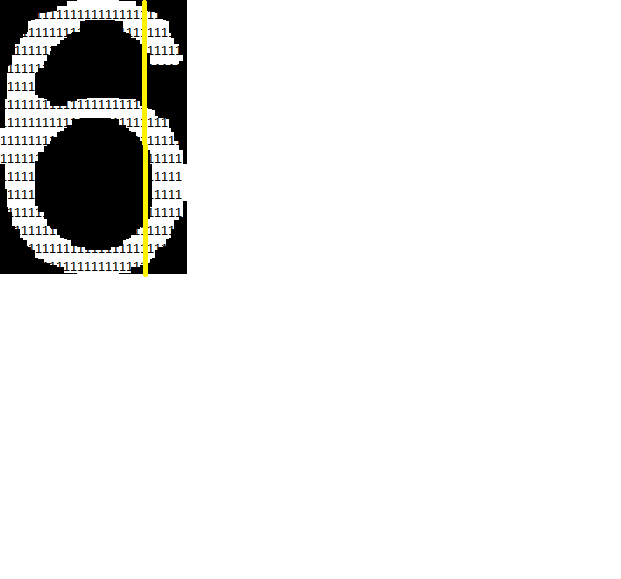
Característica 3: El número de 1´sentre filas, número columnas entre 2, regresa la relación de 1´s que en encuentran a la mitad de la imagen de forma vertical con respecto al tamaño total de la imagen.



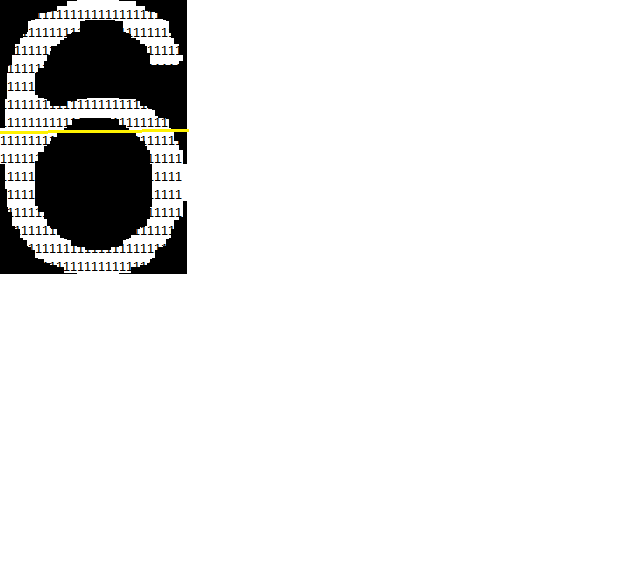
Característica 4: El número de 1´sentre filas, número columnas entre 4, regresa la relación de 1´s que en encuentran a 1/4 de la imagen de forma vertical con respecto al tamaño total de la imagen.



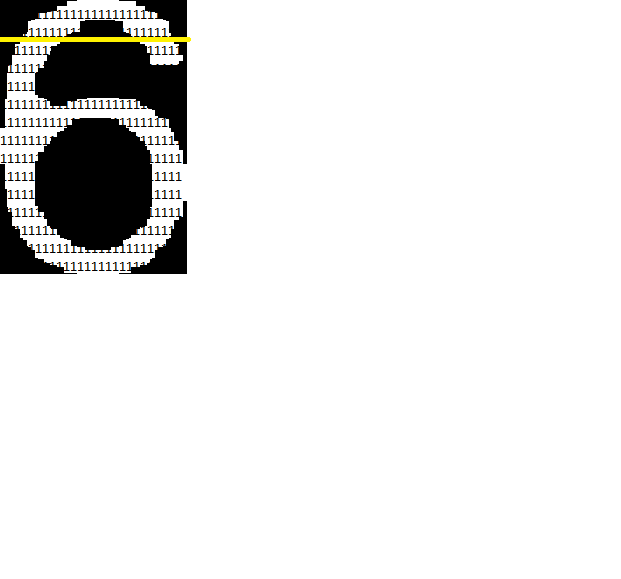
Característica 5: El número de 1´sentre filas, 3 (número columnas entre 4), regresa la relación de 1´s que en encuentran a 3/4 de la imagen de forma vertical con respecto al tamaño total de la imagen.



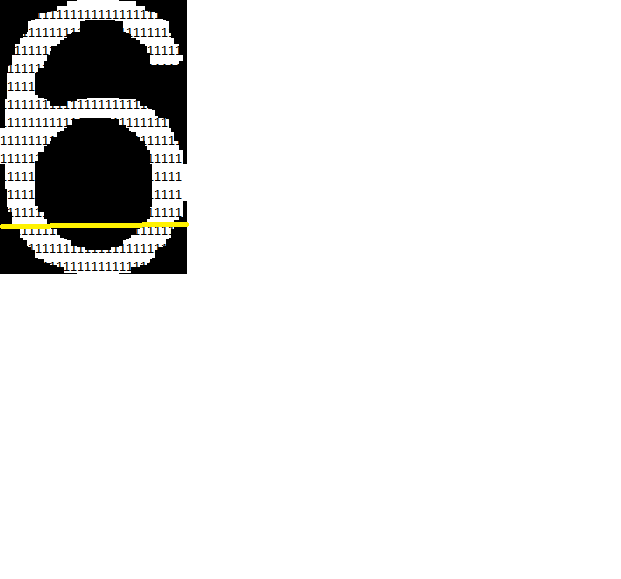
Característica 6: El número de 1´sentre filas, filas entre 2, regresa la relación de 1´s que en encuentran a 1/2 de la imagen de forma horizontal con respecto al tamaño total de la imagen.



Característica 7: El número de 1´sentre filas, filas entre 4, regresa la relación de 1´s que en encuentran a 1/4 de la imagen de forma horizontal con respecto al tamaño total de la imagen.



Característica 8: El número de 1´sentre filas, filas 3 (número de columnas 4), regresa la relación de 1´s que en encuentran a 3/4 de la imagen de forma horizontal con respecto al tamaño total de la imagen.



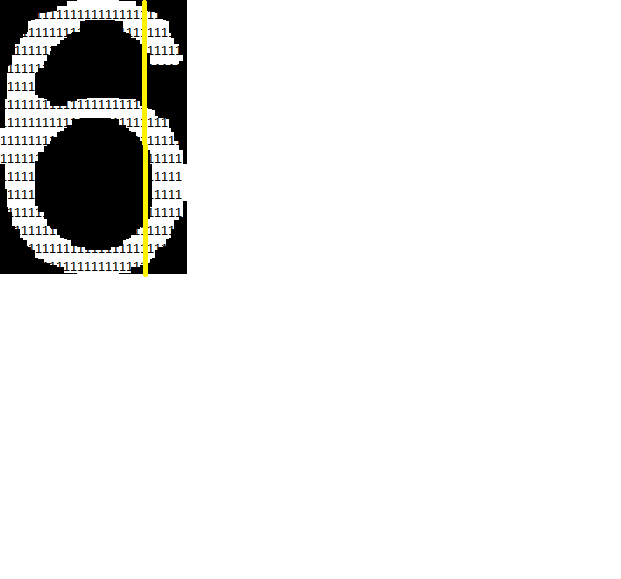
Característica 9: Números de cortes 0´s entre 1´s, regresa la relación de cortes que en encuentran a 1/2 de la imagen de forma vertical (los cambios que hace entre 1´s y 0´s con respecto al tamaño total de la imagen).



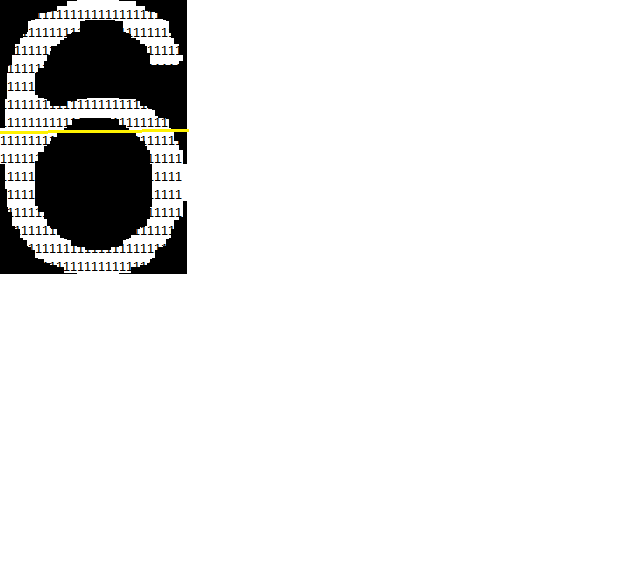
Característica 10: Números de cortes 0´s entre 1´s, regresa la relación de cortes que en encuentran a 1/4 de la imagen de forma vertical (los cambios que hace entre 1´s y 0´s con respecto al tamaño total de la imagen).



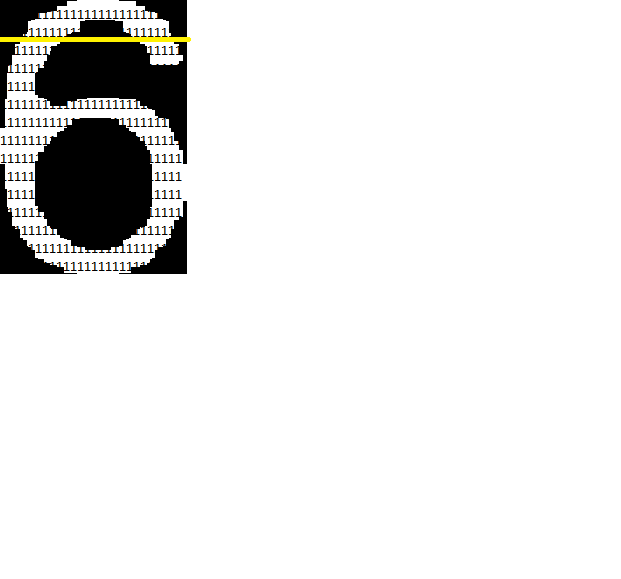
Característica 11: Números de cortes 0´s entre 1´s, regresa la relación de cortes que en encuentran a 3/4 de la imagen de forma vertical (los cambios que hace entre 1´s y 0´s con respecto al tamaño total de la imagen).



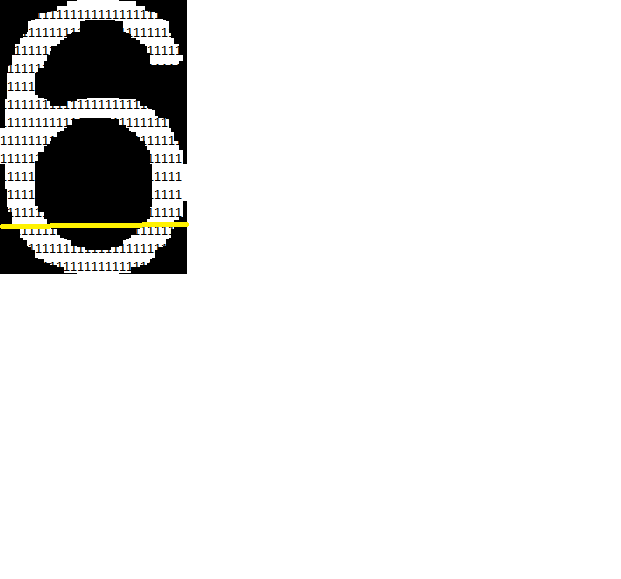
Característica 12: Números de cortes 0´s entre 1´s, regresa la relación de cortes que en encuentran a 1/2 de la imagen de forma horizontal (los cambios que hace entre 1´s y 0´s con respecto al tamaño total de la imagen).



C13: Números de cortes 0´s entre 1´s, regresa la relación de cortes que en encuentran a 1/4 de la imagen de forma horizontal (los cambios que hace entre 1´s y 0´s con respecto al tamaño total de la imagen).



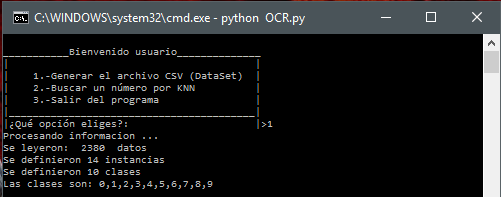
C14: Números de cortes 0´s entre 1´s, regresa la relación de cortes que en encuentran a 3/4 de la imagen de forma horizontal (los cambios que hace entre 1´s y 0´s con respecto al tamaño total de la imagen).



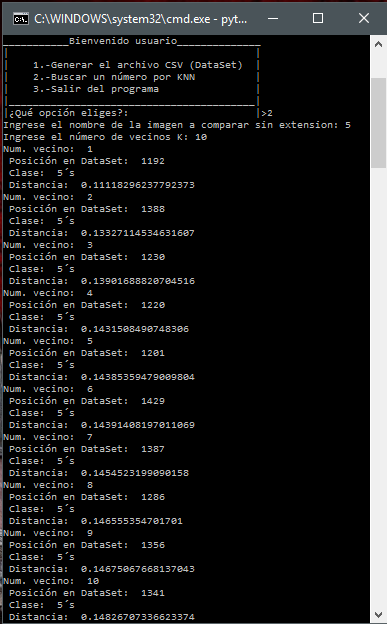
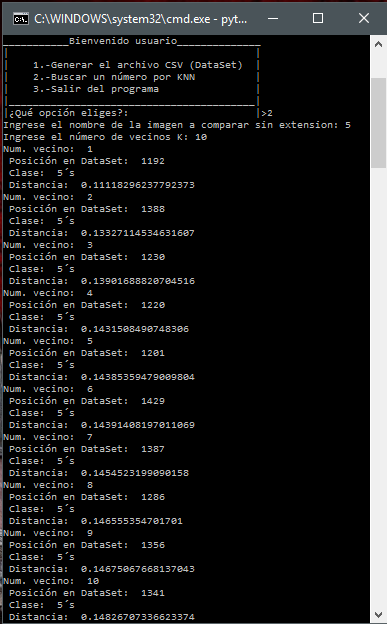
* **Método de Clasificación KNN**

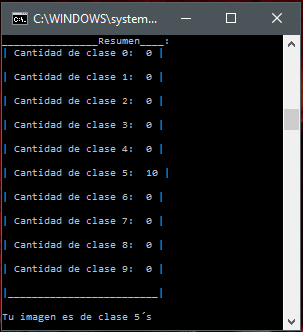
Recibe la imagen de la nueva instancia y el número de vecinos a obtener, con este método se aplica el reconocimiento de los numero si son 1´s o 0´s

* **Resultados**



Muestra menú y genera CSV (DATASET)





Búsqueda y clasificación

* **URL**

Ubicación del OCR completo

https://github.com/Yllescas/OCR.git