Informe pruebas de extracción de texto tesseract ocr

Para la primera etapa de pruebas se evalúa la capacidad de extracción de la herramienta tesseract analizando sus resultados de extracción de texto para distintas imágenes que divergen en formato y complejidad visual.

Prueba 1

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

La imagen es un archivo de formato png, posee fondo blanco, el texto contiene caracteres en mayuscula tambien en minuscula, ademas hay palabras de distinto color y al final de la imagen se presenta un logo que contiene la frase tesseract ocr.

Resultado de extracion de texto

Como resultado de la extraccion la herramienta consiguio extraer las frases principales respetando las muyusculas y minusculas, tambien conserva la dispocision de las frases, pero no extrajo el texto adjunto como parte del logo.

Captura de pantalla de computadora

Descripción generada automáticamente

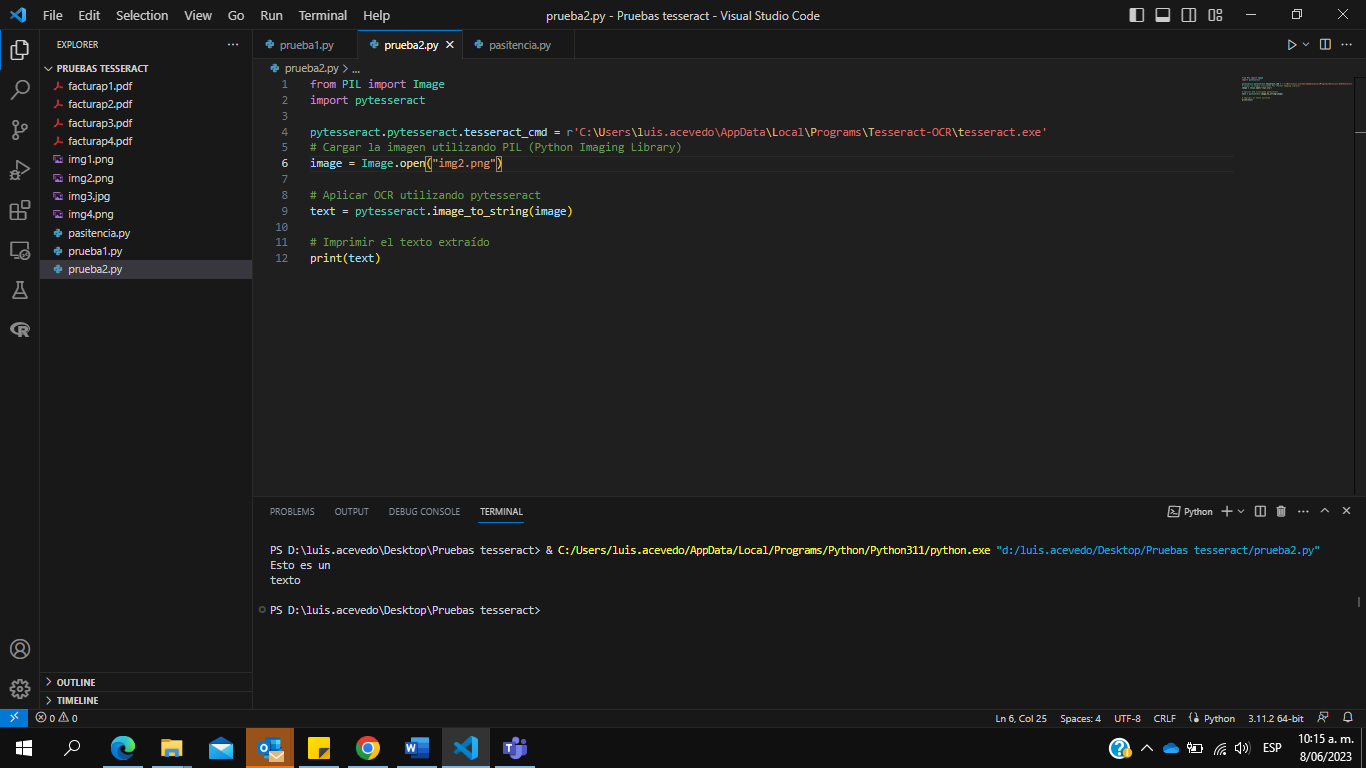
Prueba2

La imagen es un archivo png con un texto simple y fondo blanco.



Resultado de extracion de texto

Como resultado de la extraccion la herramienta consiguio extraer la frase principal respetando las muyusculas y minusculas, tambien conserva la dispocision de las frase, detalles como el acento en la vocales no fue extraido a texto



Prueba 3

La imagen es un archivo jpg con una sola palabra con estilo de fuente cursivo, el fondo es un paisaje ruidoso en colores y contrastes ademas se puede apreciar caracteres en otro idioma como parte del paisaje de fondo



Como resultado de la extraccion la herramienta no logro extraer extraer ningun tipo de informacion

Prueba 4

La imagen es un archivo png, compuesto por un fondo blanco y una frase dispuesta en 3 lineas con una caligrafia artistica o cursiva

Texto

Descripción generada automáticamente

Resultado de extracion de texto

Como resultado de la extraccion la herramienta no logro obtener ningun tipo de informacion

Prueba 5

La imagen es un archivo jpg, compuesto por un fondo blanco y una sola palabra en caligrafia cursiva

Texto, Pizarra

Descripción generada automáticamente

Resultado de extracion de texto

Como resultado de la extraccion la herramienta no logro extraer extraer ningun tipo de informacion

Prueba 6

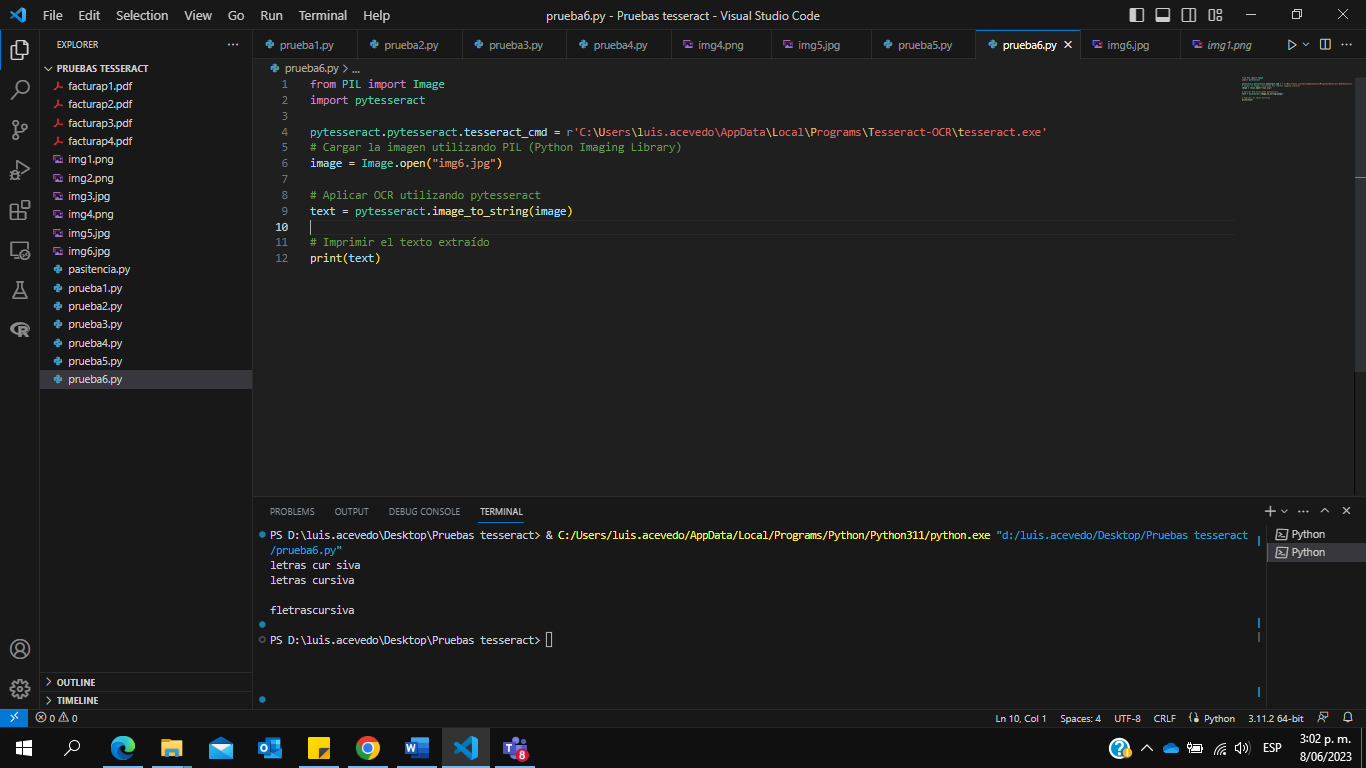
La imagen es un archivo jpg, el fondo es claro y plano con una lineas rojas que enmarcan los laterales derecho e izquierdo, el texto a extraer son 3 lineas con la misma frase pero con distinto tipo de fuente

Texto

Descripción generada automáticamente

Resultado de extracion de texto

La herramienta logro realizar extracion de texto de las 3 lineas, se realizo una extracion inprecisa de las lineas 1 y 3 la cuales en la imagen se presentaba con una fuente cursiva mas prominente



Conclusiones

La herramienta logra distinguir entre mayusculas y minusculas

La herramienta busca coservar la dispociosion del texto extraido

Las imágenes deben ser saneadas para lograr la extracion del texto, se debe eliminar el ruido de la imagen, buscando fondos planos.

La herramienta no esta en capacidad de leer fuentes de letras artisticas y cursivas

Etapa 2 de pruebas

Para la segunda etapa de pruebas se evalúa la capacidad de extracción de la herramienta tesseract analizando sus resultados de extracción de texto para distintas facturas que divergen en formato y complejidad visual, para la realización del proceso de extracción de texto desde un archivo pdf tesseract requiere acompañamiento de librerías diseñadas para la manipulación de este tipo de archivos, estas librerías se encargan de cambiar el formato del archivo de pdf a un formato de imagen valido como parámetro de entrada para tesseract, de lo anterior se puede inferir que la extracción de texto desde un pdf no es más que una extracción de texto de una imagen mas compleja

Prueba 7

Resultado de extracción de texto

El archivo que corresponde a una hoja blanco con contenido de una factura fue extraído totalmente correcto más sin embargo una línea de texto se extrajo en un lugar orden equivocado

El texto en negrilla corresponde al texto extraído

**#1024**

**BILLED TO: Really Great Company**

**PAY TO: Avery Davis**

**123 Anywhere St., Any City**

**123-456-7890**

**Bank Really Great Bank**

**Account Name John Smith**

**BSB 000-000**

**Account Number 0000 0000**

**DESCRIPTION RATE HOURS AMOUNT**

**Content Plan $50/hr 4 $200.00**

**Copy Writing $50/hr 2 $100.00**

**Website Design $50/hr 5 $250.00**

**Website Development $100/hr 5 $500.00**

**SEO $50/hr 4 $200.00**

**Sub-Total**

**Package Discount (30%) $375.00**

**TOTAL $875.00**

**Payment is required within 14 business days of invoice date.**

**remittance to hello@reallygreatsite.com.**

**factura de prueba numero uno**

**Please send**

Tabla

Descripción generada automáticamente

Prueba 8

**RECIBO DE COMPRA**

**Vendido a: Rimberio y asociados Fecha: 31 de enero del 2023**

**Calle Cualquiera 123, No. de recibo: 001**

**Cualquier Lugar, CP: 12345**

**Cantidad Concepto Precio Total**

**1 Concepto $10.00 $10.00**

**1 Concepto $10.00 $10.00**

**1 Concepto $10.00 $10.00**

**1 Concepto $10.00 $10.00**

**SUBTOTAL: $40.00**

**Método de pago: Transferencia bancaria IMPUESTOS: 0.00%**

**Informacién de Pago:**

**Rimberio y asociados TOTAL: $40.00**

**Banco Cualquiera**

**Cuenta: 0123456789**

**Gracias por su compra.**