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

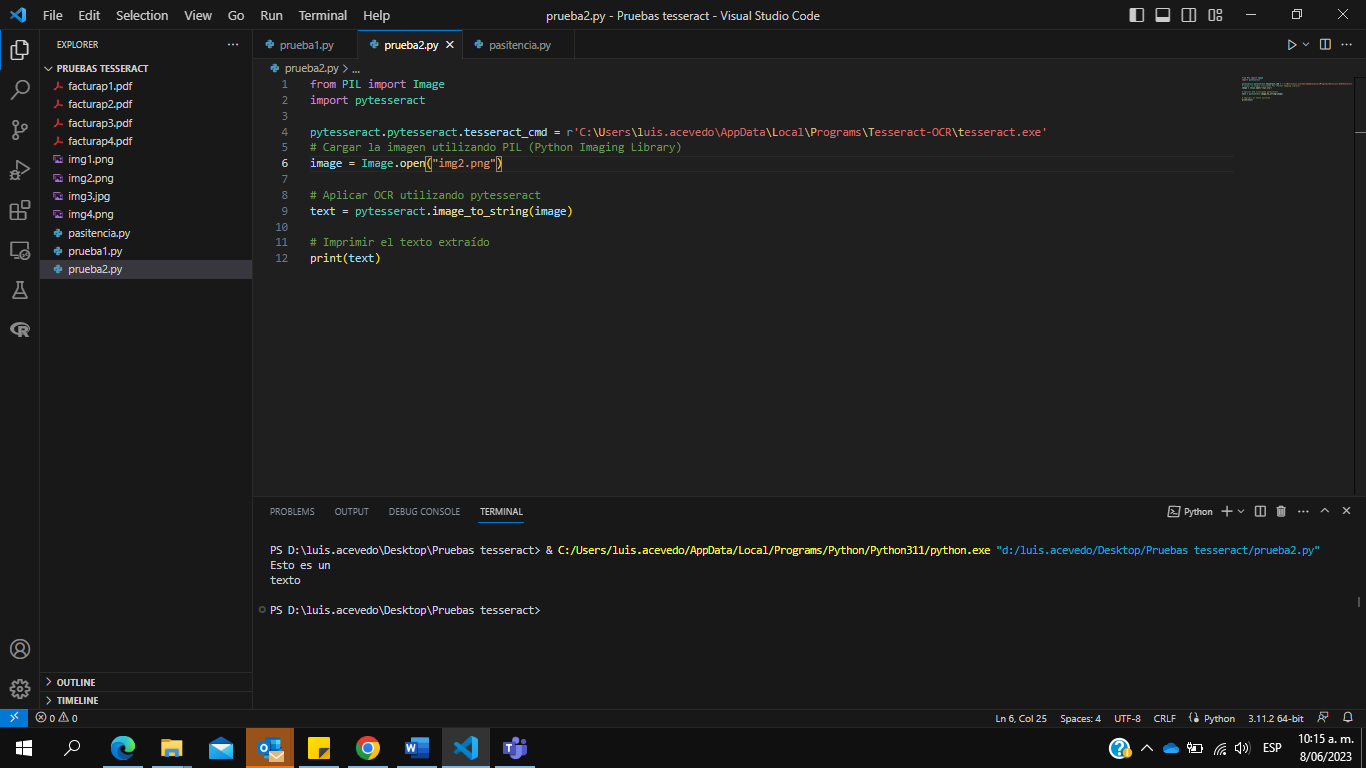
Prueba2

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

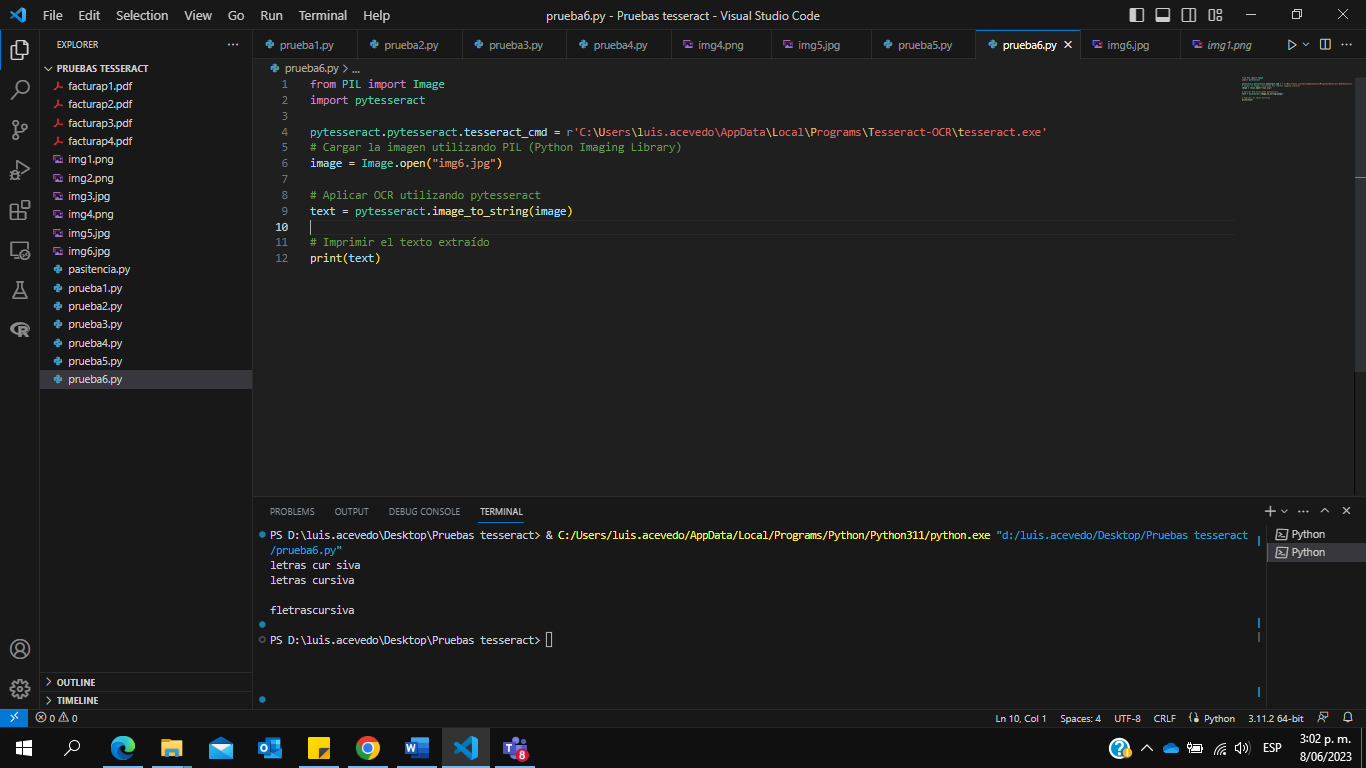
La imagen es un archivo jpg, el fondo es claro y plano con una lineas rojas que enmarcan los laterales derecho e izquierdo, el texto a extraer son 3 lineas con la misma frase pero con distinto tipo de fuente

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

Resultado de extracción de texto

La factura digital con logotipo y bastante ruido fue extraída con éxito en su mayoría, la herramienta presenta dificultades para diferenciar algunas vocales minúsculas

**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.**

**Tabla

Descripción generada automáticamente**

**Etapa 3 de pruebas**

En esta etapa se evalua la capacidad de extracion de la herramienta mediante pruebas con un archivo escaneado, se toman varios frentes de analis sobre este documento buscando establecer parametros mas concisos, los frentes de analisis son :

Presicion en un documento de varias paginas y precision en una unica pagina (analisis sobre los resultados sobre el mismo sector de archivo), modificaciones que mejoran la calidad de extracion, comparacion entre la extracion del archivo escaneado y el archivo en version digital, fuentes con mejor conpatibilidad.

Para el analisis de la calidad de extracion se realizan diversas comparacion respecto a la primera extraccion y se resaltan las novedades descubiertas, se usa rojo para nuevos errores, amarillo para cambios o variociones y verde para correciones tales que la extracion es igual al archivo original

Presicion en un documento de varias paginas y precision en una unica pagina (analisis sobre los resultados sobre el mismo sector de archivo)

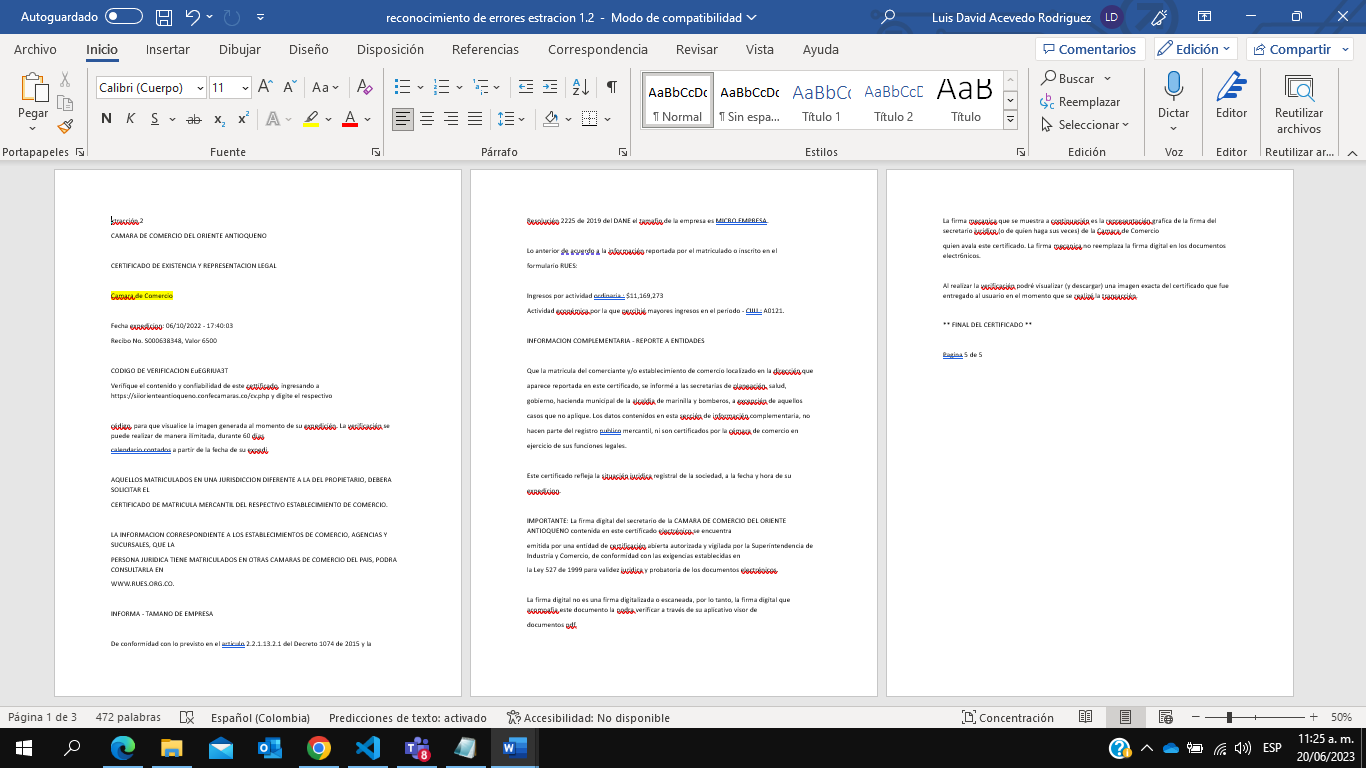
**Presicion en un documento de varias paginas**: en esta prueba se debio realizar la extracion 2 veces, dado a que la primera extracion presento demasiadas incosistencia

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

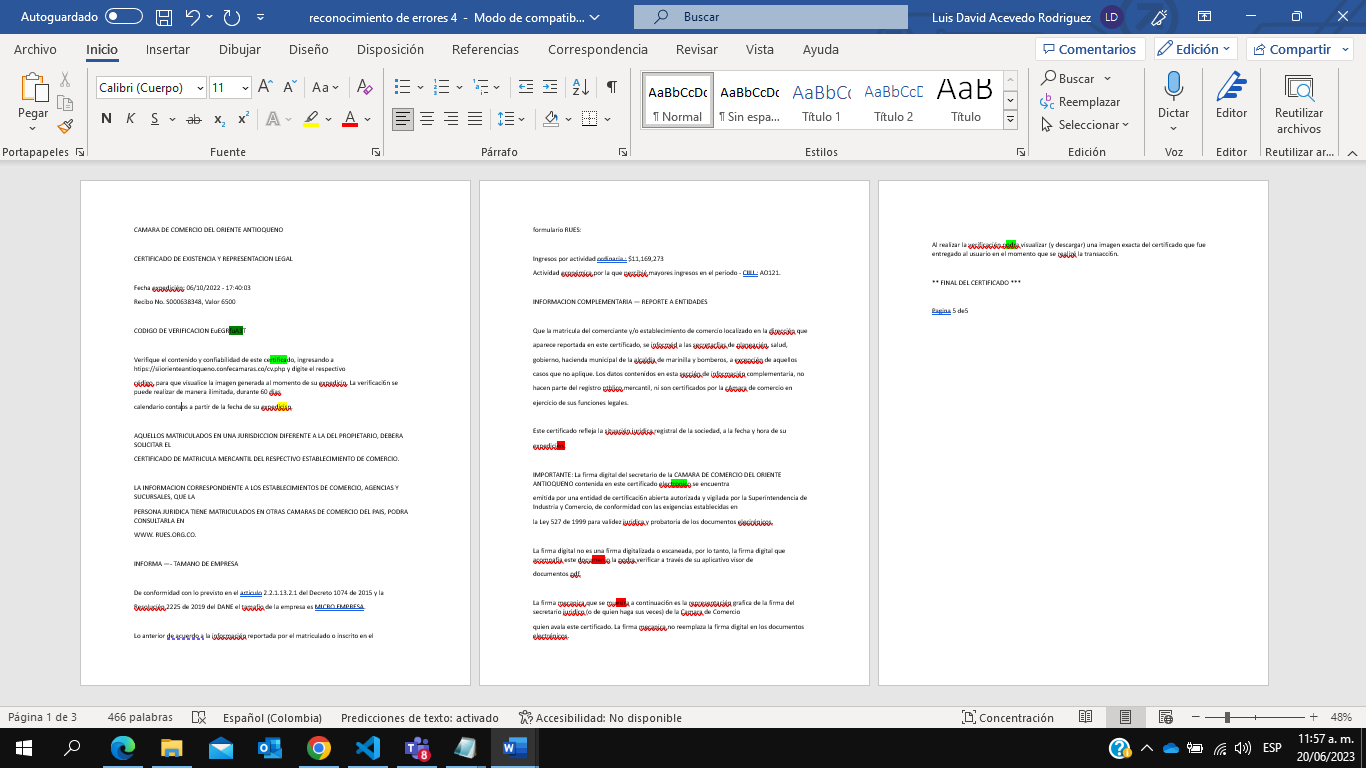
extracción 1 documento completo

las inconsistencias más relevantes que motivaron la segunda extracción se trataron de palabras que no fueron extraídas, para la segunda extracción y posteriores a esta la herramienta si extrajo todas las palabras del archivo o las interpretaciones de estas.



extracción 2 documento completo

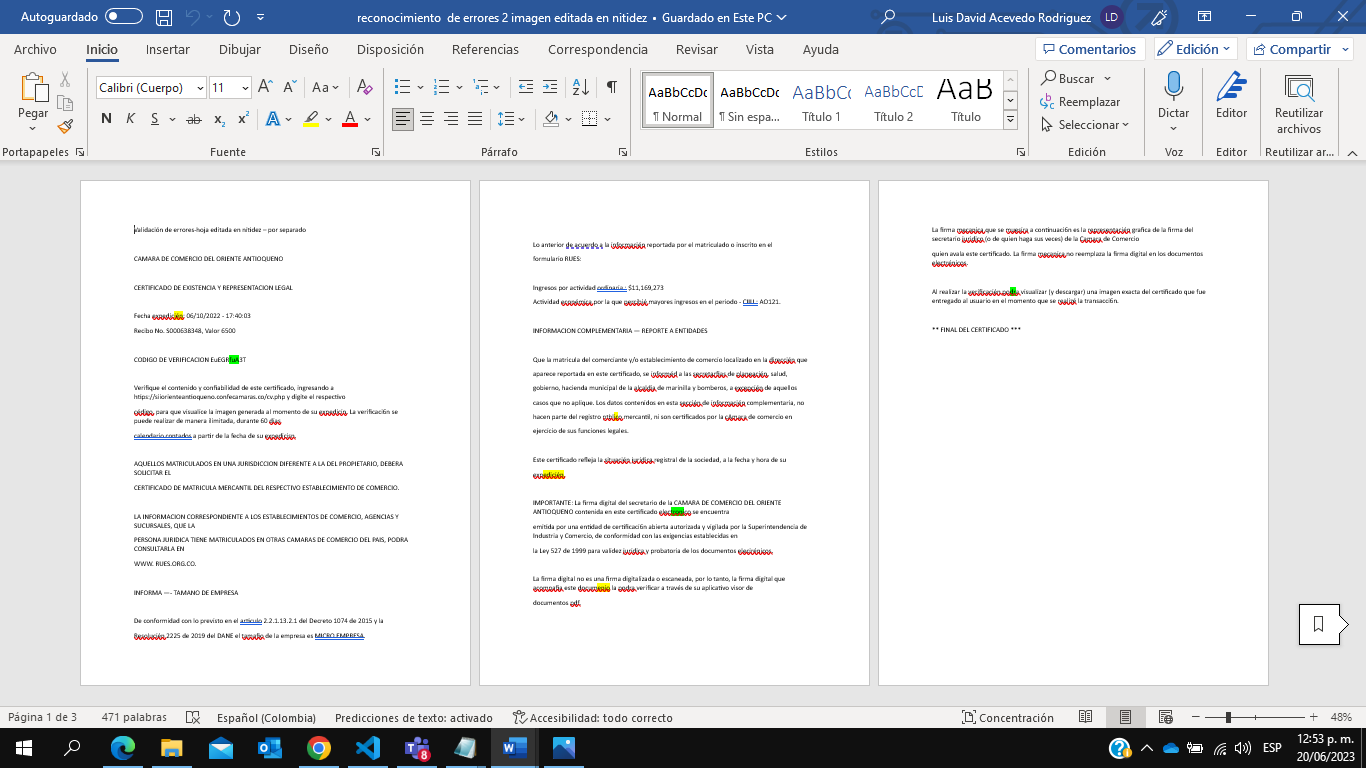
**Precision en una unica pagina :** esta prueba se llevo acabo varias veces se presentaron correciones y variaciones



extracción 3 página individual

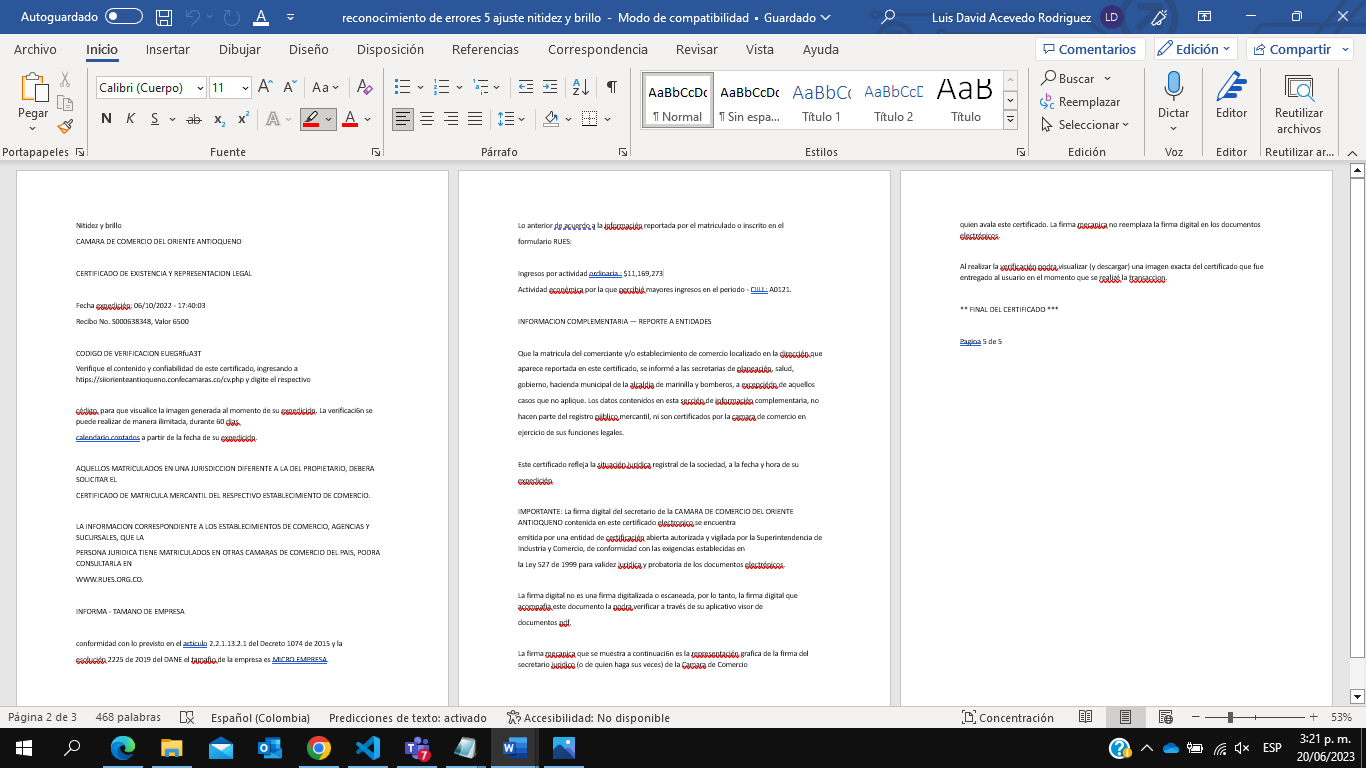
**Modificaciones que mejoran la calidad de extracion**

**Mejoramiento de nitidez:** sometiendo el archivo a un proceso de nitidez mediante open cv se logra que la herramienta realice una extracion perfecta sobre un codigo de verificacion y este avance se mantiene constante en pruebas posteriores



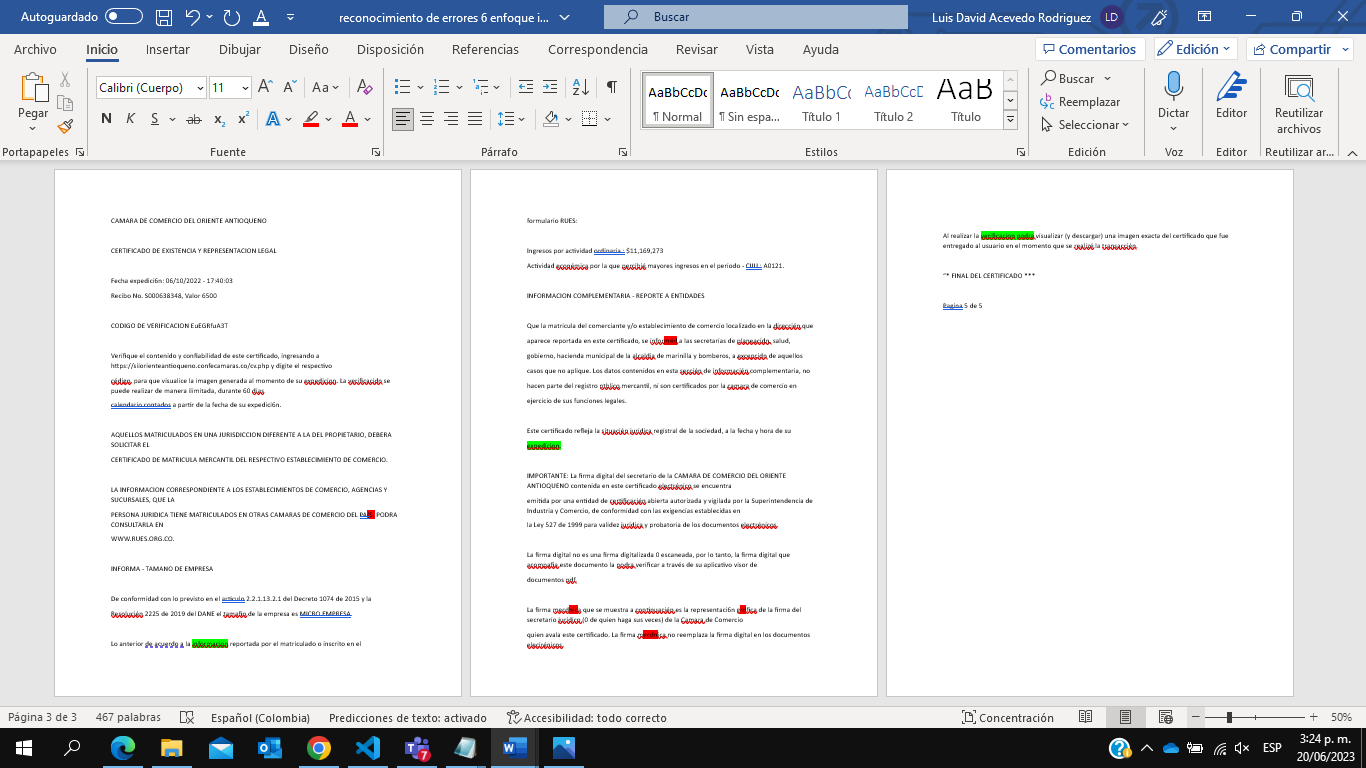
extracción 4 página individual ajuste de nitidez

**Mejoramiento de nitidez y brillo:** en esta prueba no hubo cambios respecto a la prueba anterior



extracción 5 página individual ajuste de nitidez y brillo

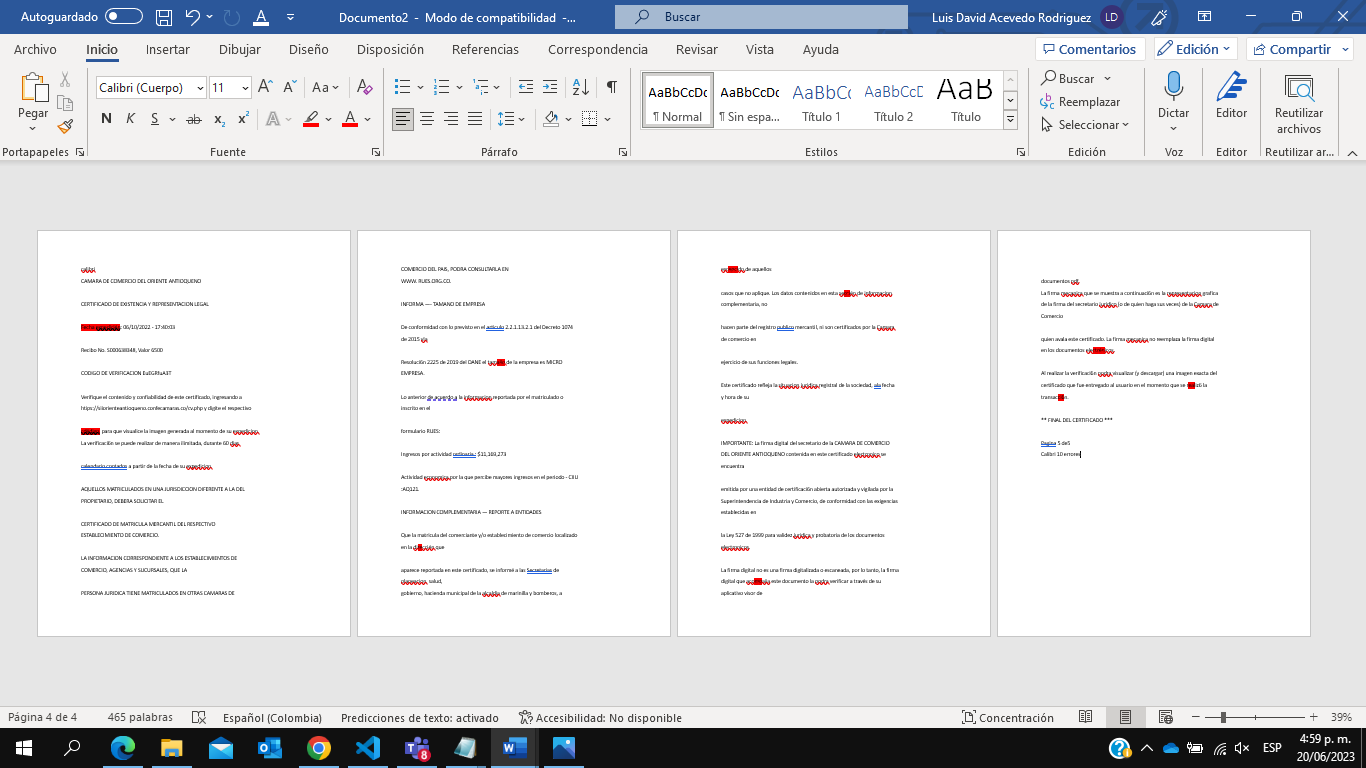
**Mejoramiento mediante proceso de enfoque inverso :**  mediante open cv se realiza un procedimento de mejoramiento de imagen para realsar la calidad del texto, donde se presentaron correciones, pero tamien suergieron nuevas equivocaciones



extracción 6 página individual proceso de enfoque inverso

**Archivo en version digital :** se realiza la digitalizacion del archivo para comparar los resultados de la extracion con versiones escaneadas, buscando descartar un margen de mejora en el tratamiento de las versiones escanedas.

**Fuentes con mejor compatibilidad:** en conjunto con la digitalizacion de los archivos se realizan algunas versiones del documento digital variando el tipo de fuente, basado en indagaciones que exponen a las fuentes tipo serif las cuales son aquellas que tienen pequeños detalles decorativos en los extremos de las letras tipo Times New Roman, Georgia, Courier New y Cambria y las fuentes sans-serif las fuentes las cuales no tienen los remates decorativos en los extremos de las letras son fuentes más limpias tipo Arial, Verdana, Calibri y Helvetica; como las fuentes que están dentro de la capacidad de la herramienta para lograr buenos resultados de extracción

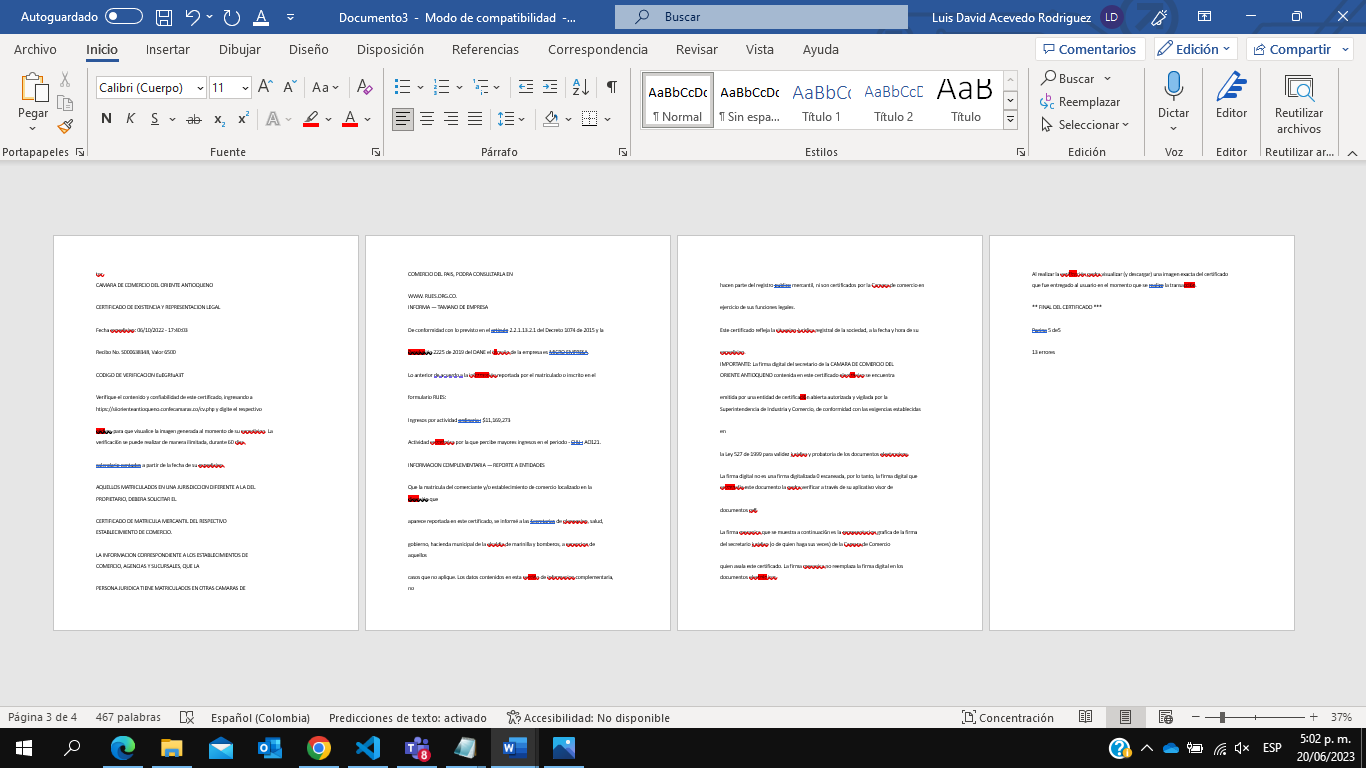


calibri

Una captura de pantalla de una computadora

Descripción generada automáticamente

arial



times new roman

La digitalización del documento ayuda significativamente y entre las fuentes a recomendar validado con la prueba las de tipo sans-serif son las que aportan mayor claridad a la extracción y con la que los resultados presentan menos errores

**Conclusiones**

* Conforme mas extenso sea el archivo o más información contenga la herramienta presenta mayor imprecisión
* El documento ideal buscar para una extracción de texto exitosa es un archivo digitalizado, con una fuente sans-serif o serif, con un contraste fuerte entre fondo y el contenido a extraer, preferiblemente la fuente deber ser mayor a o igual a 12 puntos
* La herramienta tiene mayor precisión con los caracteres mayúsculos
* El mejoramiento de imagen potencia la capacidad de extracción
* La herramienta tiene dificultades para diferencias caracteres minúsculos muy específicos, entonces se puede sistematizar una solución por que las imprecisiones en la extracción no son arbitrarias
* Las oportunidades de mejora de la herramienta se presentan en la diferenciación de las vocales o, e y con menor frecuencia la a, con las consonantes la letra ñ y en ocasiones la f, las vocales e y o se confunden entre si y en ocasiones la o es confundida con el numero 6, la a es ocasionalmente confundida con la d.

**Etapa 4 de pruebas**

En esta etapa se busca perfeccionar la extracción de texto de la herramienta mediante la configuración de parámetros y métodos de la misma, esta etapa es totalmente experimental dado a que se busca determinar que ajustes benefician mas el modelo para el tipo de requerimientos que está basado en los formatos de documentos de los cuales se desea extraer información

**Parámetro de lenguaje:** mediante indagación se descubrió que la herramienta esta predeterminada en inglés, posterior a esto se realizo el ajusto para indicar en que idioma estaba el texto a extraer, dando una ejecución perfecta como resultado de la extracción con dicho parámetro modificado.

Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente

extracción perfecta, pagina individual, mod: parámetro lenguaje

realizando la extracción de un documento compuesto de varias páginas, aun se presentan incongruencias en la extracción, siendo el análisis tomado de una muestra se asume que en el resto de las páginas también hay errores.

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

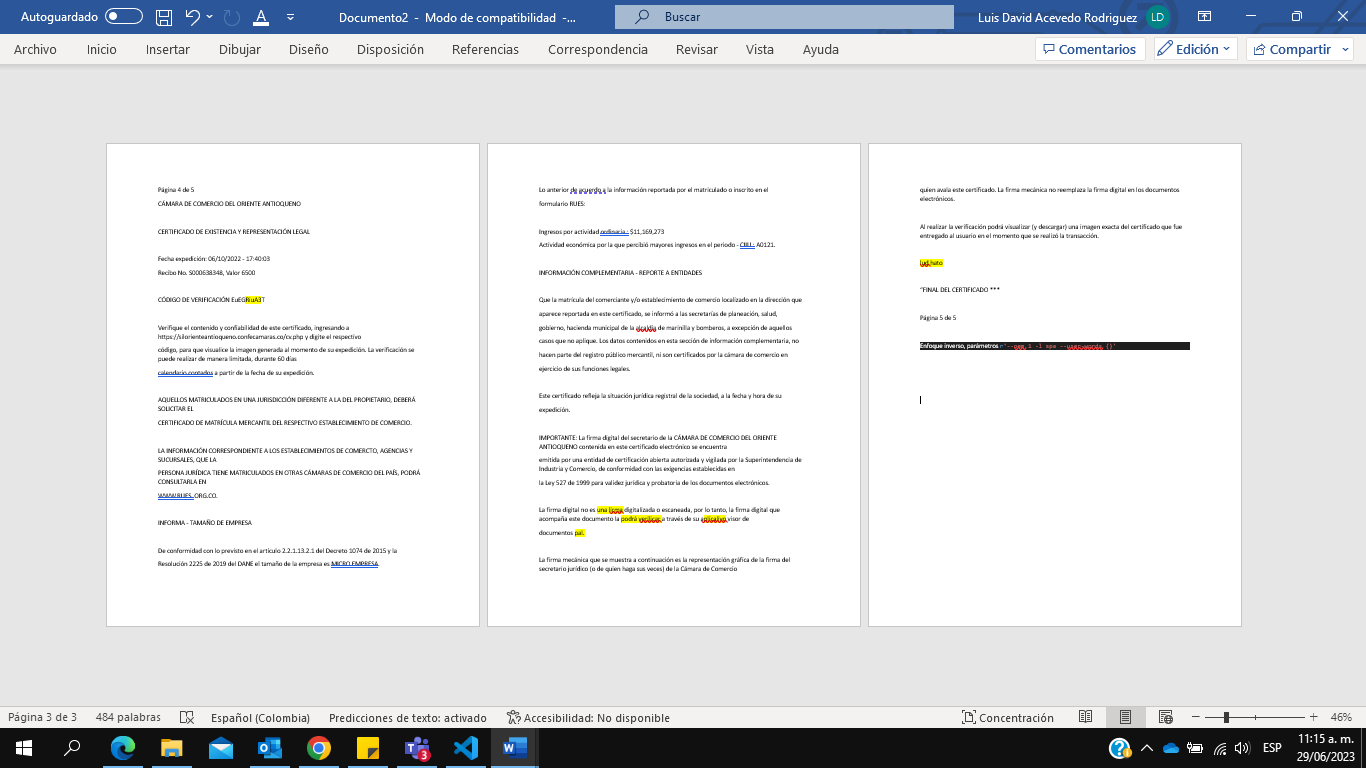
documento original completo, mod: parámetro lenguaje

mediante indagación y ensayos de prueba y error, se empiezan a ajustar los mejores parámetros de extracción, en esta parte más enfocados a las configuraciones de la herramienta que al procesamiento del archivo

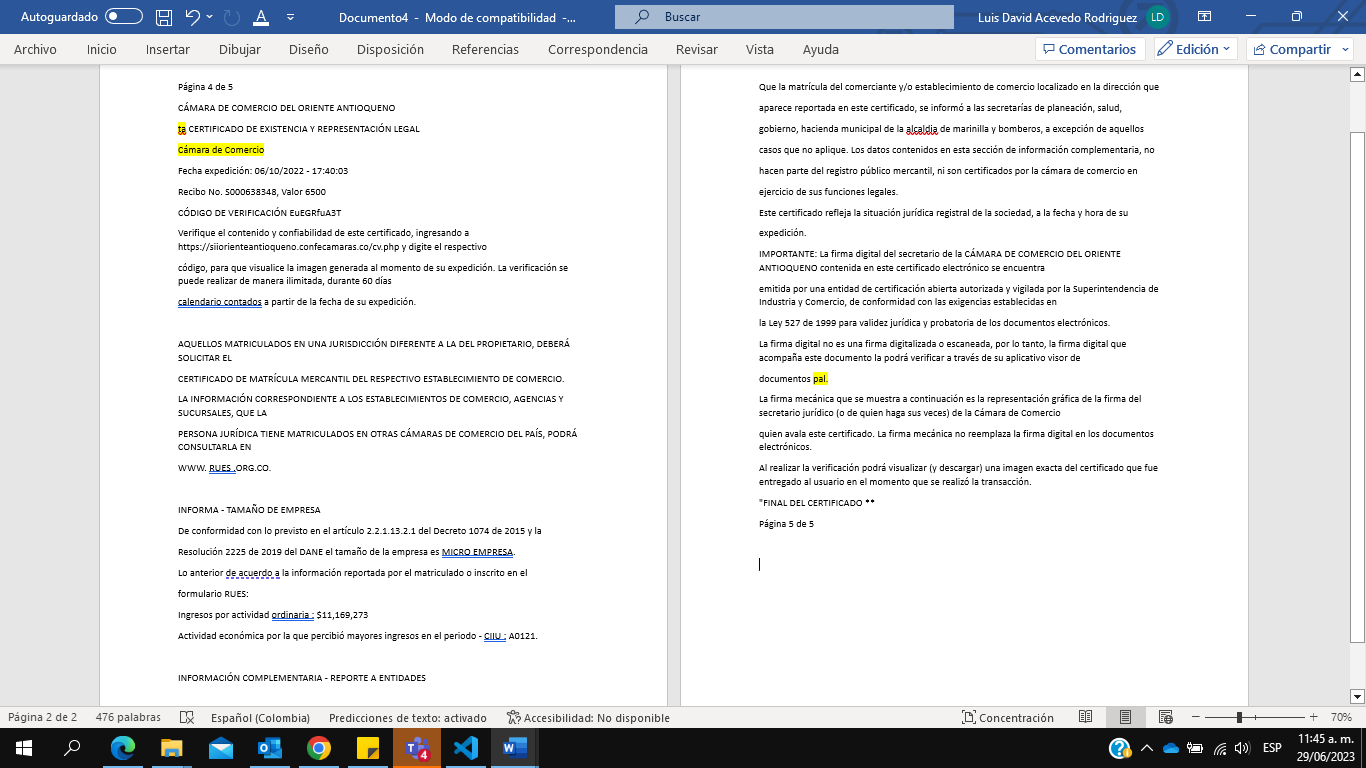
Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

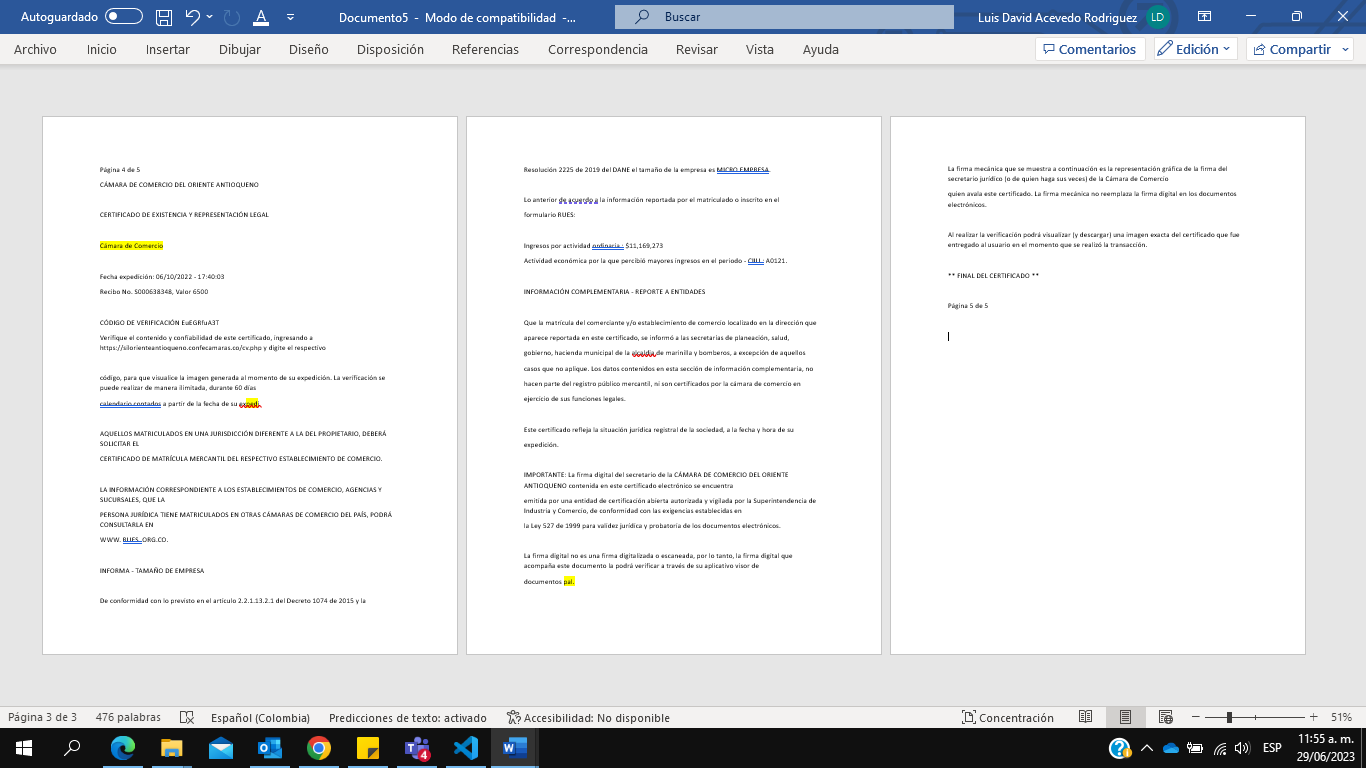
enfoque inverso, parámetro leg,seg,mt, precisión de 98,57



Enfoque inverso, parámetros r'--oem 1 -l spa --user-words {}' precisión 98.51



Original Parámetros todos r'--oem 1 --psm 6 -l spa --user-words {}' precisión 98.93

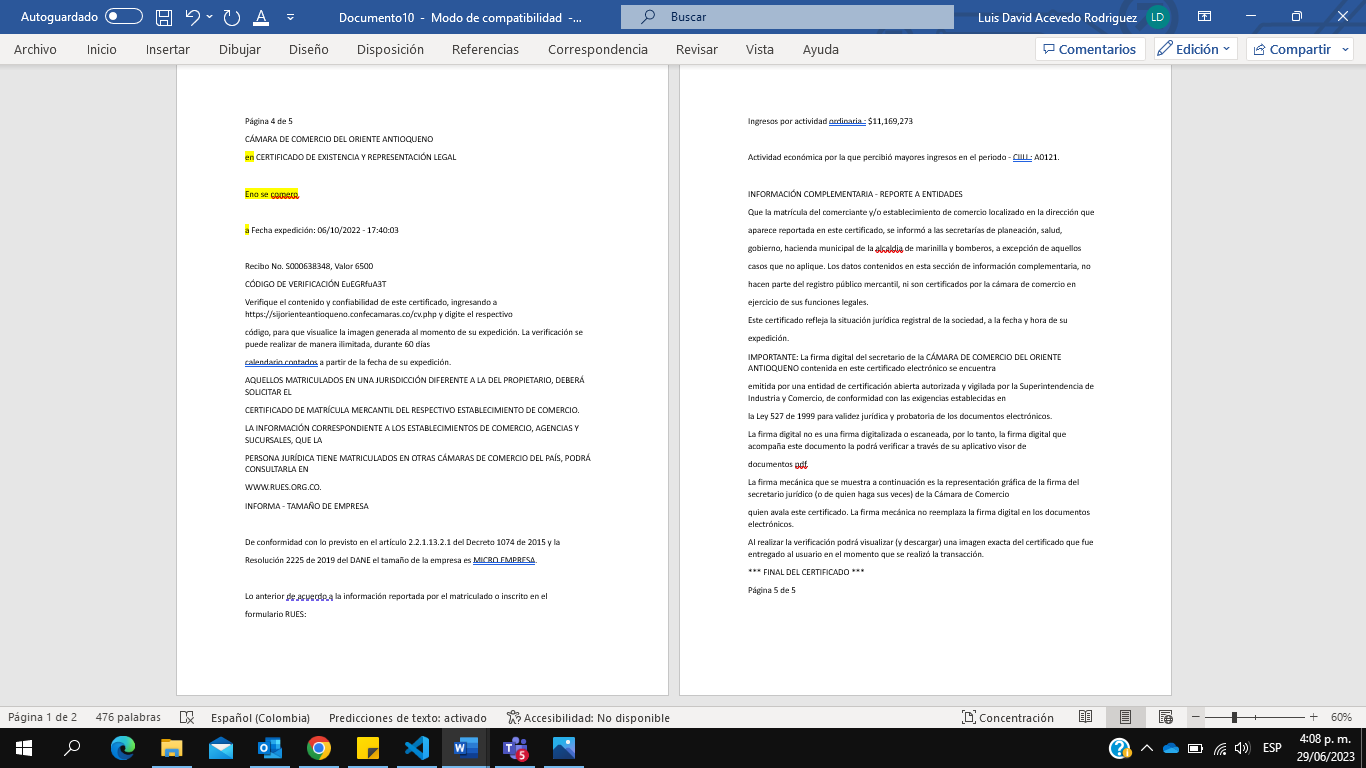


Original Parámetros --oem 1 -l spa --user-words {} precisión 98.30

Interfaz de usuario gráfica, Aplicación, Word

Descripción generada automáticamente

Todos los parámetros '--oem 1 --psm 6 -l spa --user-words {}', " imagen nivel de brillo 38 , contraste 1.0 , nitidez, ecualizada" documento formato imagen, precisión 99.1



Enfoque inverso Todos los parámetros '--oem 1 --psm 6 -l spa --user-words {}', documento formato imagen, precisión 98.9

En este punto se ha alcanzado la combinación mas óptima para la extracción del texto en este formato, se logró una precisión suficiente sobre la extracción, hay que contemplar que el modelo de evolución para determinar dicha precisión sobre la muestra consiste en :

* Contemplar la cantidad de palabras del documento
* La cantidad de palabras del documento van a ser el parámetro de referencia para el estadístico de muestra, la cantidad de palabras va ser igual al numero de eventos totales.
* Sobre la cantidad contemplada, se verifica y de define que:
* 1 palabra faltante es igual a un error
* 1 palabra de sobra (repetida o parte del ruido del archivo) es igual a un error
* 1 palabra incompleta es igual a un error
* 1 carácter erróneo en una palabra es igual a un error
* El número de errores totales se divide sobre el numero de casos posibles, se multiplica por 100 y se obtiene un valor que representa el porcentaje de error, este porcentaje se le es restado al valor del 100 % y se obtiene el valor de precisión.

Este porcentaje es representativo, mas no es un indicador que pueda definir que combinación es la mas apropiada, dado a la forma en la que está definido el método evaluativo, si embargo el indicador referencia una ruta para un posterior análisis.

Basados en la prueba mas optima, la cual extrajo todo el texto del archivo se procede realizar el análisis del archivo completo, se recalca que en esta extracción no se obtuvo la máxima precisión alcanzada en las pruebas dado a que este indicador bajo por información de ruido que también fue extraída, mas sin embargo ya se trabaja en la eliminación del ruido vía open cv

Interfaz de usuario gráfica, Texto, Aplicación, Word

Descripción generada automáticamente

Precisión de extracción 99 %

21 errores

16 de ruido-76 %

23.8 % de extracción

Se espera luego de la eliminación de ruido restante una precisión del 99.78 %