Facultad de Ciencias Empresariales

Implementación de un prototipo para extraer información de documentos por medio de plantillas a través de técnicas de reconocimiento de patrones y OCR.

Proyecto de título para optar al título de Ingeniero Civil en informática

Autores: Miguel Csori – Matías Martínez

Profesor Guía: Rodrigo Torres Avilés

Carola Andrea Figueroa Flores

Agradecimientos

Luego de varios años de esfuerzos y sacrificios, finalmente me encuentro en la recta final. Me gustaría agradecer a toda mi familia por mantener su fe en mí desde el principio.

A mi abuelo por siempre animarme con sus chistes y siempre repetirme que soy su nieto favorito. Ahora podré por fin decir que soy otro Ingeniero en la familia.

A mi papá que siempre me apoyó a financiar mis estudios para que terminara, sin importar que tanto me demorara. Siento la tardanza.

A mi abuela quien siempre me ha regaloneado desde pequeño y siempre intenta hacerme el gusto. Iremos a comer para celebrar.

Al Francisco por ayudarme a lo largo de la carrera, con quien siempre pude debatir mano a mano sobre como poder mejorar. Al Max por alegrar los días monótonos con tu particular personalidad. Al Hans, Juan, Luis, Miguel y Freddy por ser muy importantes amigos durante estos últimos años, no sería lo mismo sin ustedes.

Finalmente, un gran agradecimiento a la mayor fuente de inspiración en mi vida, mi querida madre quien siempre me ha dado las fuerzas para seguir adelante, ayudándome a mejorar mi futuro y sacrificándose constantemente para salir adelante. Gracias por todos estos años de vida.

Muchas gracias a todos los demás que no alcancé a nombrar, cada uno fue un importante hito en mi historia para llegar a este momento.

Miguel Csori

En esta etapa de la vida donde ya se cierra una etapa y se comienza otra, me es inevitable mirar hacia atrás y ver que el camino que en un inicio se veía largo y una cuesta que prometía ser la más grande de mi vida y ya está llegando a su fin.

De la misma forma es inevitable ver los rostros de todos aquellos que de forma directa o indirecta hicieron más llevadera esta travesía. En especial a mis padres Alonso Martínez y Jovita Rodríguez quienes con su ejemplo de lucha y superación, me dieron los valores y herramientas que me definen como persona y marcaran como seré de profesional, ellos que con sus ánimos y fe en mi me impulsaron a alcanzar la esta meta, todo el esfuerzo plasmado en estos años es para ustedes. A mi hermana Jazmín Martinez que pasear del distanciamiento que las distintas circunstancias hemos tenido, sé que siempre podre contar con tigo solo espero tener la paciencia para aguantarte jajaja, pero igual quiero que sepas que siempre has sido un ejemplo a seguir y que en el fondo de mi corazón te quiero.

También quiero mencionar de forma muy especial a Virginia Jaña, mi amada polola, ella quien me acompaño en cada paso que di en esta última etapa dándome todo su apoyo, fuerza, amor y comprensión, los cuales fueron pártete esencial para llegar a esta meta.

Extiendo este agradecimiento a todos aquellos que no he mencionado pero que sin ellos no estaría acá, a la familia, a todos los amigos y compañeros, profesores y maestros, mascotas, `Select * from seresQueridos;` para ustedes mi eterna gratitud.

Quería dejarlos con una reflexión con todo mi cariño para quien este leyendo este documento:

"Dar un buen ejemplo, no basta con imitar buenos ejemplos. Porque a veces he creído que este asunto del buen ejemplo se aplica en mí sólo en cuanto a que debo imitar a los buenos, las buenas prácticas y conductas, no sería una mala idea, ¿cierto? Pero esto no es efectivo, se vuelve un cambio forzado, artificial y con el paso del tiempo se agota, por el contrario, cuando me he esforzado en ser la mejor persona que puedo ser, entonces y sólo entonces, mis ojos miran lo bello y bueno a mí alrededor con nitidez. Y este esfuerzo por vivir la mejor versión de mí mismo quizá sea ejemplo para alguien más, por lo pronto es un recorrido alegre para uno mismo."

Matias Martínez

Resumen

El ser humano siempre ha tenido la necesidad de registrar la información que han aprendido a lo largo de toda su existencia, desde las talladuras en la roca hasta los registros en papel. Gracias a la tecnología que tenemos actualmente, hemos podido almacenar grandes cantidades de información en un espacio muy reducido, incluso pudiendo subir archivos a la nube sin la obligación de tener un documento físico que acredite la información. En este punto surgió la necesidad de traspasar los datos en papel a digital en una forma más rápida y eficiente que tener que realizarlo manualmente uno por uno.

Los OCR son una tecnología que, a pesar de haber nacido con el objetivo de facilitar la vida a personas discapacitadas, terminó por convertirse en una herramienta capaz de ayudar en un futuro a la digitalización de la información, pudiendo facilitar enormemente esta tarea. Las técnicas utilizadas en este sistema presentan limitaciones, tal como es la concentración de errores causados por imperfecciones en el documento, que pueden ser omitidas o limitadas por medio de plantillas.

Por esto es que en el presente trabajo se investigó el funcionamiento de los OCR y la creación de plantillas para la delimitación del contenido innecesario en los documentos, de esta forma acelerando el procesamiento del programa.

.

Abstract

The human being always had the necessity of register the information that has learned through all of his existence, since the carving in the stones to the register in papers, Thanks to the technology that we have right now, we can allocate great amounts of information in a much reduced space, even can upload things to the cloud without the obligation of having a physical document that credits the information. In this point came the necessity of transfer the data on paper to digital in a way more fast and efficient rather than manually one by one.

The OCR are a technology that, though it born off with the objective of ease the life of disabled people, finally it turns into a tool able to help in the future of the digitalization of the information, being able to make a huge improve in this task. The techniques used in this system present limitations, like the concentration of errors caused by the imperfections in the document, this can be corrected or limited by the use of templates.

Because of that, this work focus on the research of the performance of the OCR and the creation of templates for the delimitation of the unnecessary content in the documents, in this way will speed up the processing of the program.

Índice general

Índice de Figuras

Capítulo I

Problemas de Investigación

En este primer capítulo se abordarán los principios básicos que serán tratados a lo largo de este documento. Se presentarán todos los temas requeridos antes del comienzo del desarrollo del software propiamente tal.

Se comenzará por el planteamiento del problema que encontramos, incluyendo las tecnologías actuales ligadas al tema propuesto y cómo actúan estas conforme al problema. Luego se procederá a la formulación del problema, destacando las falencias que existen actualmente y que podrían ser solucionadas para mejorar el rendimiento del trabajo de las personas. Después se presentarán los objetivos respecto a la problemática, tanto generales como específicos. Finalmente, se plantearán las bases que certifican que esta investigación posee fundamentos sólidos y que pueden ser respaldados por fuentes confiables.

1.1 Planteamiento del problema

La gestión documental es el conjunto de normas técnicas y prácticas usadas para administrar el flujo de documentos de todo tipo en una organización, tales como permitir la recuperación de información desde ellos, determinar el tiempo que los documentos deben guardarse, eliminar los que ya no sirven y asegurar la conservación indefinida de los documentos más valiosos, aplicando principios de racionalización y economía.

En otras palabras, es una actividad casi tan antigua como la escritura, que nació debido a la necesidad de "documentar" o fijar actos administrativos y transacciones legales y comerciales por escrito para dar fe de los hechos. Este tipo de documentos se plasmaron sucesivamente en tablillas de arcilla, hojas de papiro, pergaminos y papel, cuya gestión se fue volviendo cada vez más compleja a medida que crecía el tamaño de los fondos documentales.

En palabras del Reglamento general de archivos, Colombia, en su artículo 1 “Los documentos que conforman los archivos son importantes para la administración y la cultura porque son imprescindibles para la toma de decisiones basadas en antecedentes y porque pasada su vigencia se convierten en fuentes de la historia y componentes valiosos del patrimonio cultural y de la identidad nacional” (Melorose, Perroy, & Careas, 2015). Los documentos son parte esencial de las organizaciones y un instrumento básico para la modernización como para optar por un mejor escenario para el futuro, resultando ser una tarea de crucial importancia.

Actualmente, aún se acostumbra utilizar sistemas de gestión e inventario de documentos físicos, los cuales se almacenan en bodegas (Figura 1), lo que significa un costo de almacenamiento no menor. Estos documentos pueden ser contables (boletas, facturas, guías de despacho, etc), historiales (historiales médicos, inventarios, etc) o empresariales (contratos, convenios, acuerdos, reglamento interno, procedimientos, manuales de maquinarias, etc).

Figura 1: Ej. Bodegas de Archivos y Documentos

Algunas empresas han optado por digitalizar sus documentos para poder solventar lo anteriormente expuesto, ya sea implementando un sistema tecnológico que permite la gestión documental o digitalizando los documentos por medio del uso de un escáner.

Si bien la opción del escáner rectifica el problema del almacenamiento en las bodegas de estos documentos, deja el siguiente problema sin tratar: La información alojada en los documentos digitalizados, al igual que en los documentos físicos, no se puede aprovechar a menos que una persona busque el documento que necesita, para de esta forma, traspasar la información a un sistema ad hoc de esta manera pudiendo aprovechar lo que en estos documentos se guarda.

1.2 Formulación del problema

Como observamos en el capítulo anterior, en las empresas, organizaciones o cualquier organismo público o privado, a diario se archivan grandes cantidades de documentos, ya sea en bodegas abarrotadas de carpetas con documentos importantes o en bases de datos con sus respectivas versiones digitales.

Desde ahí nacen las siguientes interrogantes:

- ¿Existe la posibilidad de utilizar alguna técnica o tecnología que nos permita aprovechar la información presente en dichos documentos?
- ¿Es posible, a través de algún sistema computarizado que utilice un sistema de reconocimiento de patrones, recolectar toda la información contenida en los documentos?
- ¿Puede ser la información contenida en los documentos de utilidad en la toma de decisiones?
- ¿Es posible obtener sólo la información relevante (ser selectivo)?

En este proyecto de título se exploran las tecnologías que permiten la obtención de la información de los documentos y se plantea solucionar la falencia de la lectura que se da con los programas de reconocimiento óptico de caracteres (desde ahora OCR) actuales, al no poder procesar todos los diferentes elementos que no pueden ser reconocidos por una máquina, tales como firmas, timbres, etc. Se plantea la utilización de plantillas que delimiten las áreas de interés a escanear, donde se podrán reconocer los diferentes tipos de documentos y así aprovechar de mejor forma la información presente en estos.

1.3 Objetivos de la investigación

Objetivo general:

- Creación de un prototipo de reconocimiento de patrones, usando tecnología OCR, de manera de digitalizar información alfanumérica en áreas delimitadas del documento.

Objetivos específicos

- Implementación de un OCR para poder identificar y obtener información relevante de facturas digitalizadas en formato de imagen.
- Implementar la creación de distintas plantillas, las cuales serán creadas a la medida de cada usuario.
- Aplicar ROI (áreas de interés) para evitar un esfuerzo extra del OCR.
- Aplicar los datos obtenidos en un caso práctico de utilidad para el caso de una empresa hipotética.

1.4 Justificación e importancia

El proyecto a desarrollar nace en el contexto de la experiencia obtenida por uno de los estudiantes en su práctica profesional en la empresa EXE Ingeniería & Software Ltda., en donde se propone desarrollar e implementar una herramienta que permita la obtención de información relevante presente en documentos escaneados utilizando plantillas, en las cuales se definirán las zonas de los documentos en que se encuentra la información de interés. Los datos que se pretenden obtener se procesarán bajo distintos criterios, con lo cual se podrán realizar actividades tales como gestión documental, traductores, entre otras.

En el contexto de esta actividad de titulación se buscará realizar un prototipo, el cual logre mostrar el funcionamiento de un OCR, con la utilización de ROI para el procesamiento de la información obtenida de estas imágenes.

El tipo de documento elegido para trabajar en esta propuesta son las facturas, debido a su gran relevancia en el área contable, ya que poseen un formato estándar y la ubicación de sus componentes las hacen un documento idóneo para nuestro proyecto.

De esta forma, el prototipo creado podrá concentrarse en zonas en las cuales ya se sabe que existe información relevante, y así evitar un esfuerzo adicional por parte del OCR al intentar la lectura y reconocimiento de todo lo presente en la imagen, que puede contener datos imposibles de analizar tales como los son los timbres, firmas, manchas, etc.

Se considerará el uso de distintas técnicas de reconocimiento de caracteres como el algoritmo de Hough (Conocido también como la transformación de Hough). El algoritmo principalmente consiste en procesar una imagen determinando sus márgenes, transformarla a binario y luego implementar una serie de ecuaciones diferenciales para la detección de puntos de interés en esta imagen.

Con esto se pretende analizar la alternativa de implementar completamente el OCR y añadir un valor agregado como parte de un sistema mayor de gestión documental, siendo de especial utilidad en el análisis de factibilidad para ser aplicados en distintos sistemas específicamente en repositorios documentales, en donde se trabaja principalmente con documentos digitalizados.

1.5 Hipótesis

Si la información registrada en papel no es práctica de manejar y almacenar, entonces la tecnología OCR es capaz de digitalizar los documentos y hacerlos más accesibles.

Capítulo II

Fundamentación Teórica

En este segundo capítulo se observaran los fundamentos teóricos que respaldan el proceso de la presente investigación.

Primero, se explicará en detalle la tecnología OCR desde sus inicios, quién fue el fundador, con qué propósito fue creado y cómo se implementa actualmente esta tecnología. Luego, se presentarán varios ejemplos de distintos trabajos que utilizan variables de OCR, explicando detalladamente en los puntos que se enfocan y diferencian de nuestro proyecto.

En la segunda parte de este capítulo, se exponen las bases teóricas que respaldan los fundamentos que se presentan en este documento. Se especifican las diferentes variantes de los OCR y, finalmente, se explicará extensivamente el funcionamiento paso a paso del principal algoritmo utilizado en todo el proceso del OCR.

2.1 Antecedentes

El concepto de OCR comenzó como una idea en donde se utilizaban tecnologías que involucraban la telegrafía y la capacidad de crear aparatos para que las personas ciegas lograran leer. En 1914, Emmanuel Goldberg desarrolló una máquina que leía los caracteres y los convertía en código telegráfico estándar. En esas mismas fechas, Edmund Fournier d'Albe desarrolló el optófono, un escáner de mano que cuando se movía a través de una página con texto era capaz de producir tonos que correspondían a específicas letras o caracteres. A finales del año 1920 hasta principios de 1930 Emanuel Goldberg desarrolló lo que él llamo una “Máquina estadística” para buscar archivos microfilm usando un sistema de reconocimiento óptico. En 1931 se le dio una patente en Estados Unidos por su invento, la cual fue adquirida más tarde por IBM.

En 1974, Ray Kurzweil fundó la compañía Kurzweil Computer y continuó desarrollando el omni-font OCR, el cual podía reconocer texto impreso en cualquier fuente. Kurzweil decidió que la mayor aplicación para esta tecnología sería crear una máquina que ayudara a leer a las personas ciegas, las cuales conllevaría a que ellas tuvieran una computadora que les leyera el texto. Este dispositivo requería la invención de dos tecnologías que lo habilitaran, el escáner plano CCD y el sintetizador de texto a habla. El 13 de enero de 1976 se reveló un producto final exitoso durante una conferencia dirigida por Kurzweil y los líderes de la federación nacional de ciegos. En 1978, los productos de las computadoras Kurzweil comenzaron a vender versiones comerciales de los programas de reconocimiento óptico de caracteres. LexisNexis fue uno de los primeros clientes, y compró el programa para subir papeles legales y documentos de noticias a sus bases de datos recientemente en línea. 2 años después, Kurzweil vendió su compañía a Xerox, la cual tenía un interés en comercializar la conversión de texto desde papel a computador.

En el año 2000 los OCR estuvieron disponibles online como un servicio (WebOCR), en un ambiente de computación en la nube y en aplicaciones móviles con traducción en tiempo real a otros idiomas.

Varios Sistemas OCR comerciales y opensource están disponibles en varios sistemas de escrituras, como latín, chino, japonés, árabe, coreano, etc.

Se han realizado varias investigaciones respecto a la metodología consistente al OCR en orden a analizar sus funcionalidades para su posterior implementación. Se presenta a continuación una lista de diversos trabajos realizados anteriormente por otros autores, que comparten parte de la tecnología que será utilizada en este trabajo.

- *Conversión de texto manuscrito a formato digital utilizando máquinas de soporte vectorial.* "En este trabajo se aborda el reconocimiento y clasificación de caracteres manuscritos aislados el cual se compone de 3 etapas. La etapa de preprocesamiento utiliza técnicas de procesamiento de imágenes y consiste en la conversión a niveles de gris, detección de bordes, umbralado, dilatación, relleno y etiquetado. Prosigue la etapa de extracción de características, donde se utilizan combinaciones de técnicas de extracción de rasgos compuesta por las tres invariantes de flusser, excentricidad, elongación, factor de compacidad y tamaño. Finalmente, la etapa de clasificación utiliza las máquinas de soporte vectorial multiclase para determinar la clase de la que proviene el objeto o carácter que se está procesando." (Valenzuela & Báez, 2007)

- *Reconocimiento de patrones.* "Una propuesta de corrección automática de test de selección múltiple: En este trabajo se formula una propuesta de diseño y construcción de un sistema que realice la corrección automática de exámenes de selección múltiple y que sirva de modelo alternativo al sistema de corrección manual. Se realizan principalmente los procesos de adquisición y reproceso de datos, extracción de características y toma de decisiones o agrupamiento."(Cisneros, 2007)

- A*pp móvil para reconocer texto en imágenes.* "Este proyecto cuenta con la idea principal de utilizar un lector OCR e integrarlo a una aplicación Android que permita a cualquier usuario, a través de una imagen obtenida a partir de la cámara o del almacenamiento interno del dispositivo, que pueda obtener el texto de la imagen. El procesamiento del núcleo principal del OCR en C++ se divide en dos partes, la primera parte es procesada dentro del mismo dispositivo Android y la segunda es procesada a través de un servidor Apache que una vez procesado el resultado, será devuelto al dispositivo. También se agrega una función de traductor que puede ser utilizada por el usuario una vez obtenida la palabra por la aplicación." (Montes, 2014)
- *Sistema de gestión integral de documentos de archivo para empresas de la construcción del territorio de Camugüey.* "Este trabajo responde al desarrollo de la temática gestión documental como línea de investigación, implícita en el proyecto nacional de innovación y desarrollo "Gerencia de los recursos de información de las organizaciones". La investigación se basa en ofrecer una metodología para el diseño e implementación de un sistema de gestión de documentos, dividida por etapas consecutivas que demuestran resultados sobre la valoración de este proceso en las empresas objeto de análisis." (Campillo, 2010)

2.2 Bases teóricas

Como se ha explicado anteriormente, la tecnología OCR se ha utilizado desde hace varios años como una manera de facilitar la interacción de las personas con discapacidad visual, en el presente forma parte de la vida en la facilitación del trabajo a diario. Esta tecnología puede ser implementada en varios aspectos, como:

- Entrada de datos para documentos.
- Reconocimiento automático de patentes.
- Extraer información de negocios a una lista de contactos.
- Traspasar sistemas anti-bot Captcha.
- Tecnología para asistir personas ciegas.

Con el tiempo se han desarrollado variantes de esta tecnología.

- Optical Character Recognition (OCR): Orientado a texto escrito a máquina, un caracter a la vez.

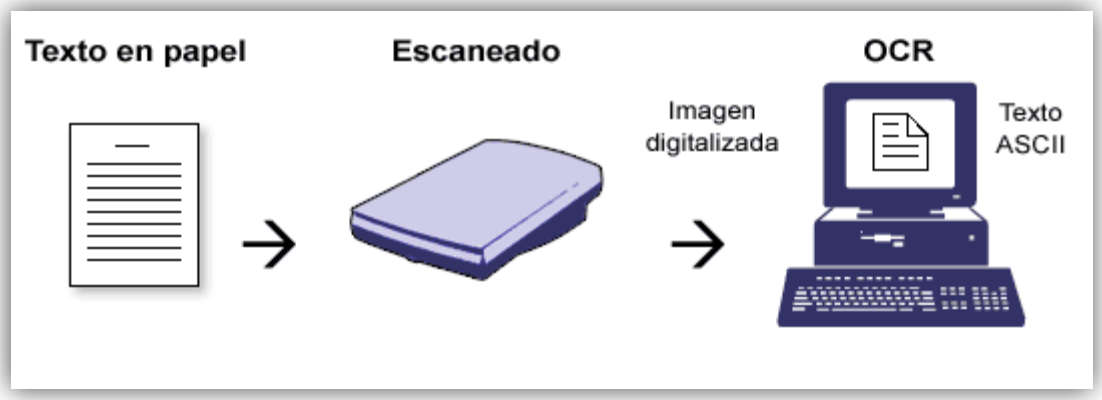


Figura 2: Esquema OCR

- Optical Word Recognition: Orientado a texto escrito a máquina, una palabra a la vez usando un separador de palabras para los idiomas. También es llamado OCR.
- Intelligent Character Recognition (ICR): Orientado principalmente a texto manuscrito o un caracter cursivo a la vez, generalmente incluyendo aprendizaje.

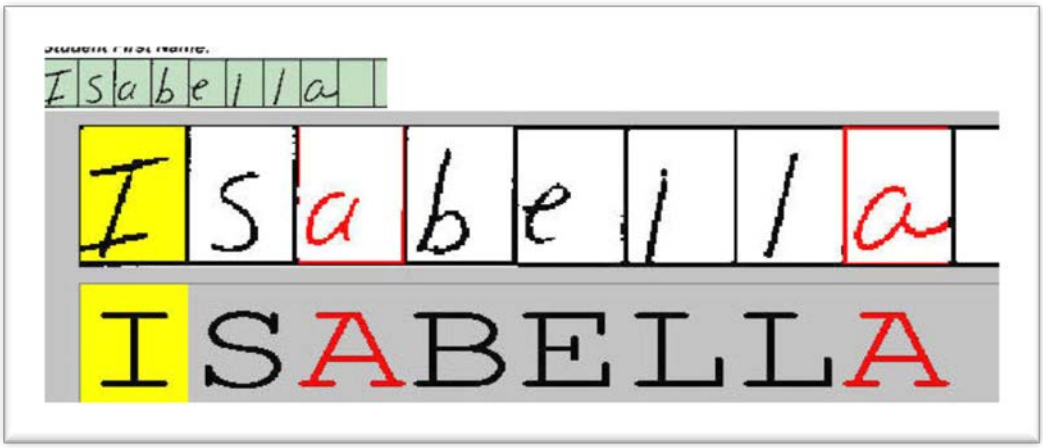


Figura 3: Ejemplo ICR

- Intelligent Word Recognition (IWR): También orientado a texto manuscrito o texto cursivo, una palabra a la vez. Esto es especialmente útil en idiomas donde los caracteres no están separados por texto en cursiva.

En los procesos necesarios para la extracción del texto en un OCR, generalmente aparece el uso de la transformada de Hough (también conocido como el algoritmo de Hough). Este consiste en una técnica de extracción utilizada en el análisis de imágenes, visión por computadora y el procesamiento de imágenes digitales.

2.2.1 Distancia de Levenshtein

La distancia de Levenshtein es una medida de la similitud entre dos palabras, una considerada el origen (Para el siguiente ejemplo la llamaremos *x*) y una objetivo (Para el siguiente ejemplo le llamaremos *y*). Esta distancia es el número de eliminaciones, inserciones o sustituciones requeridas para transformar una palabra en otra. Por ejemplo:

- Suponga dos palabras *x*=”casa” e *y*=”casa”. Si analizamos la distancia de Levenshtein entre *x* e *y*, el resultado final será 0, ya que ambas palabras son iguales.
- Suponga dos palabras *x*=”casa” e *y*=”capa”. Si analizamos la distancia de Levenshtein entre *x* e *y*, el resultado final será 1, porque existe una sustitución en la letra “s” de *x*, por la letra “p” de *y*.

Entre más grande sea la distancia de Levenshtein entre dos palabras, más diferentes serán una de otra. En algunos casos también se conoce como distancia de edición.

El origen del nombre de este algoritmo se debe al matemático ruso Vladimir Levenshtein, que lo desarrolló en el año 1965. Se ha utilizado en diversas categorías, como correctores gramaticales, análisis de ADN, detección de plagios y en OCR.

El algoritmo en pseudocódigo es el siguiente:

```
funcion Levenshtein(s1, s2: cadena):entero
   -sea n1 la longitud de s1
   -sea n2 la longitud de s2
   -si n2 vale 0, devolver n1 y terminar
   -si n1 vale 0, devolver n2 y terminar
   -sea m una matriz de enteros m[n1+1][n2+1]
   //inicializar la primera fila con los valores 0,1,...,n2 y
   //la primera columna con los valores 0,1,...,n1
   -para i=0 hasta n1
      -m[i][0]←i
   -fin para
   -para i=0 hasta n2
      -m[0][i]←i
   -fin para
   //comparar cada caracter de s1 con cada caracter de s2
   //tomando nota de su posición en cada cadena
   //y asignar a cada elemento de m el minimo de:
   // * El elemento de la fila superior más uno
   // * El elemento de la izquierda más uno
   // * El elemento anterior de la diagonal más el coste

   -para cada caracter en la posición i1 de s1 (empezando a contar por 1)
      -para cada caracter i2 de s2 (empezando a contar por 1)
         -si s1[i1] es igual a s2[i2] entonces  //aquí se calcula el coste
            -sea coste←0
         -si no
            -sea coste←1
         -fin si
         -m[i1,i2]←minimo( m[i1-1][i2]+1, m[i1][i2-1]+1, m[i1-1][i2-1]+1 )
      -fin para
   -fin para
   -devolver m[n1][n2]
```

2.2.2 Redes Bayesianas

"Una red Bayesiana es un modelo gráfico que codifica las relaciones probabilísticas sobre unas variables de interés. Por ejemplo, una red Bayesiana podría representar las relaciones entre unas enfermedades y sus síntomas. Basado en los síntomas, la red puede ser usada para computar la probabilidad de la presencia de varias enfermedades. Formalmente, las redes Bayesianas son grafos acíclicos dirigidos (DAG por sus siglas en inglés), cuyos nodos representan variables aleatorias en el significado Bayesiano, o sea, ellos pueden tener cantidades observables, variables latentes, parámetros desconocidos o hipótesis." (Heckerman, 1995)

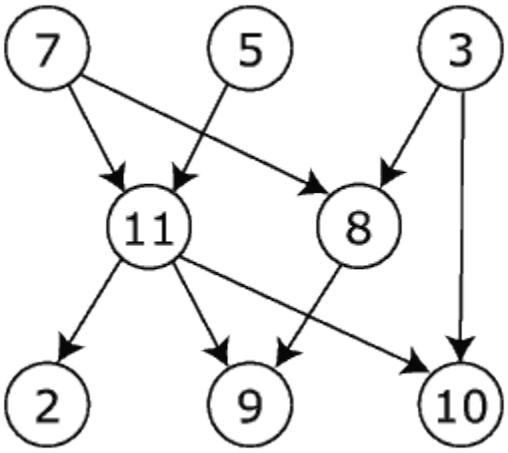


Figura 4: DAG Simple

Las puntas representan dependencias condicionales; los nodos que no están conectados, o sea, no existe ningún camino a ninguna de las variables a la otra en la red Bayesiana, representan variables que son condicionalmente independientes una de otra. Cada nodo está asociado a una función probabilística que toma como ingreso un grupo particular de valores para las variables de los nodos padres, y da como resultado la probabilidad de una variable representada por el nodo, o la probabilidad de distribución si es aplicable.

Por ejemplo, si los nodos padres representan variables booleanas, entonces la probabilidad de la función podría ser representada por una tabla de entrada, una por cada una de las posibles combinaciones de sus padres de ser verdad o falso.

Existen algoritmos eficientes que realizan interferencias y aprenden de las redes Bayesianas. Estas redes Bayesianas que modelan las secuencias de las variables son llamadas redes Bayesianas dinámicas. Las generalizaciones de las redes Bayesianas que pueden representar y resolver problemas de decisión bajo incertidumbre se llaman diagramas de influencia.

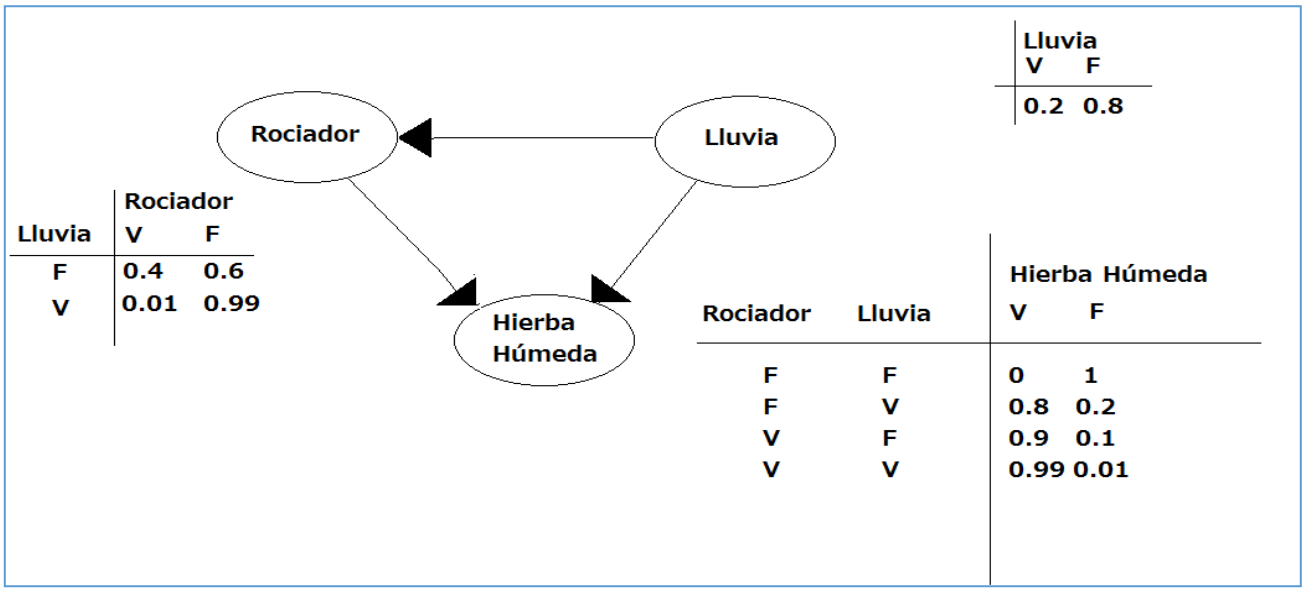


Figura 5: Red Bayesiana Simple

Cuando es usada en conjunción con técnicas estadísticas, el modelo gráfico tiene varias ventajas para el análisis de la información:

1. Una red Bayesiana puede ser utilizada para aprender relaciones causales, y por ende puede ser utilizada para ganar conocimiento sobre el dominio de un problema y para predecir las consecuencias de una intervención.
2. Como el modelo tiene semánticas causales y probabilísticas, es una representación ideal de combinar previamente los conocimientos (Los cuales vienen usualmente en una forma causal) y la información.
3. Los métodos estadísticos Bayesianos en conjunto con las redes Bayesianas ofrecen una eficiente y ética aproximación a evitar el sobreajuste de la información.

2.2.3 Redes Neuronales

El sistema de visión humana es una de las maravillas del mundo. Considerando una secuencia de dígitos escritos en manuscrito como los indicados en la *Figura 21*, se puede apreciar que nosotros tenemos la capacidad de distinguir los caracteres como una secuencia de números. Estos números dan como resultado el número 504192.

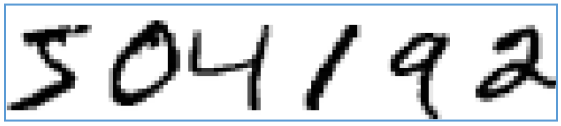


Figura 6: Ejemplo de numeración manuscrita

"En cada hemisferio de nuestro cerebro, los humanos tenemos una corteza primaria visual, también conocida como *V1*, que contiene 140 millones de neuronas con 10 billones de conexiones entre ellas. Y aun así, la visión humana involucra no solo al *V1*, sino a una serie de cortezas visuales, desde la *V2* a la *V5*. Nosotros tenemos en nuestra cabeza un supercomputador, afinado por la evolución de cientos de millones de años y adaptado ampliamente para entender el mundo visual."(Nielsen, 2016)

Reconocer los caracteres en manuscrita no es fácil, sin embargo nosotros los seres humanos somos increíblemente buenos para hacer sentido de lo que nuestros ojos nos muestran. Pero, casi todo el trabajo hecho es inconsciente, y es por eso que no apreciamos realmente que tan difícil es resolver el problema de los sistemas de visualización. La dificultad del reconocimiento de patrones visuales se muestra aparente si intentas que un programa de computadora reconozca los dígitos escritos en manuscrita como los de la *Figura 6.* Lo que parece una tarea fácil para nosotros, repentinamente se torna muy complicado de resolver. La simple intuición nos hace reconocer formas, tomemos por ejemplo el número *9.* Si lo analizamos, nos daremos cuenta que tiene un círculo en la parte superior y una línea vertical en la parte derecha. Pero esto no es simple de explicar por medio de un algoritmo, ya que al momento de especificar las reglas precisas, rápidamente se pierde el punto al intentar explicar casos especiales y puntuales.

Las redes neuronales se aproximan al problema de una manera distinta. La idea es tomar un gran número de caracteres manuscritos, conocidos como “Ejercicios de ejemplo”, y desarrollar un sistema del cual la red pueda aprender de estos. En otras palabras, la red neuronal aprende de los patrones definidos para crear reglas que reconozcan los caracteres manuscritos, y entonces mejorar su precisión.

Figura 7: Ejercicios de ejemplo

Al incrementar el número de ejercicios de ejemplo, la red neuronal puede aprender más sobre la letra manuscrita. Una red de este nivel es capaz de reconocer dígitos con una exactitud del 96% sin intervención humana.

Un tipo de red neuronal muy popular, y el que será utilizada en este proyecto, es el multicapa. Estos están formados por varias capas de neuronas, las cuales están ordenadas desde la entrada hasta la salida y cada neurona puede alimentarse de neuronas de una capa anterior, y alimentar neuronas de la capa superior. Ese tipo de conexiones se denominan conexiones feedforward o hacia adelante. Este tipo de redes solo contienen conexiones entre capas hacia delante. Esto implica que una capa no puede tener conexiones a una que reciba la señal antes que ella en la dinámica de la computación.

Por el contrario, existen algunas redes en donde las capas aparte del orden normal están también unidas desde la salida hasta la entrada en el orden inverso en que viajan las señales de información. Las conexiones de este tipo se llaman conexiones hacia atrás, feedback o retroalimentadas. Este tipo de redes se diferencia en las anteriores en que sí pueden existir conexiones de capas hacia atrás y, por tanto, la información puede regresar a capas anteriores en la dinámica de la red.

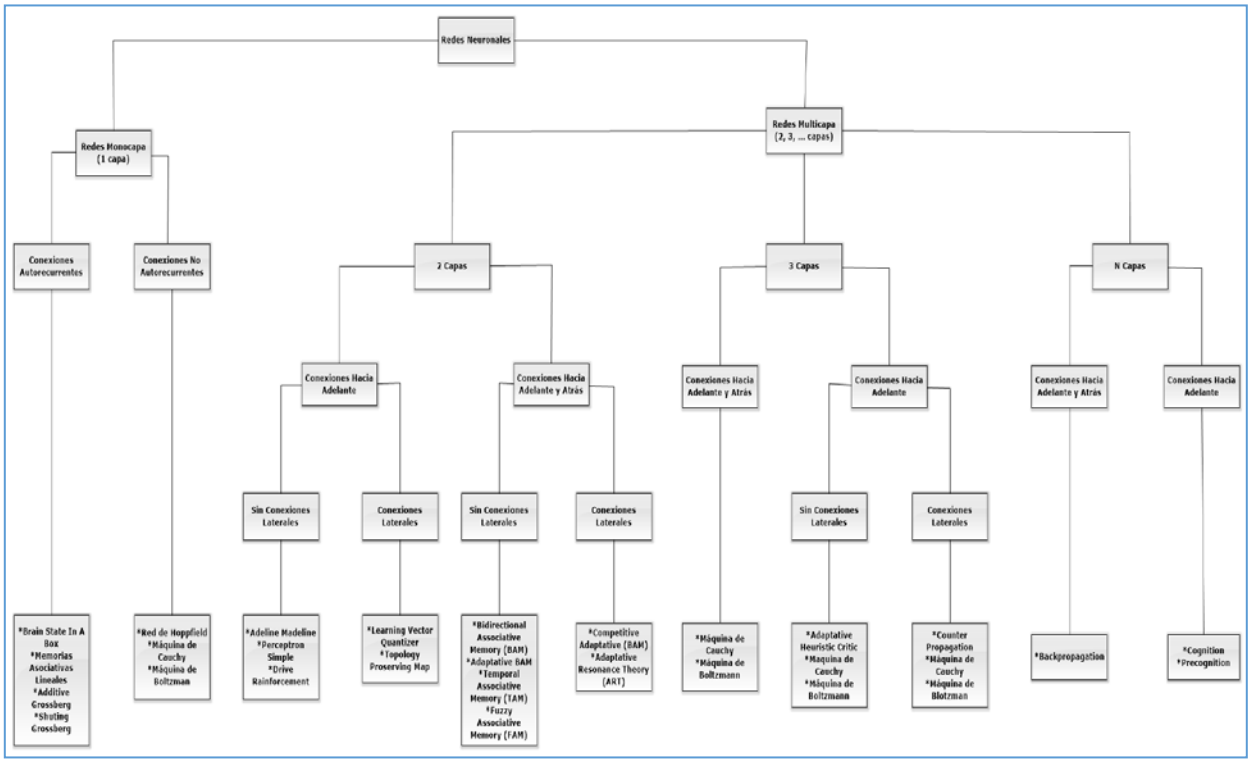


Figura 8: Tipos de redes neuronales y ejemplos

Basándonos en estas formas de aprendizaje, la respuesta que dará una red será basada en los ejercicios de ejemplo que se le fueron otorgados, por lo tanto, es una respuesta conocida. Este tipo de aprendizaje requiere a un profesor, el cual conoce la clasificación correcta para el ingreso de los patrones en el ejercicio de entrenamiento. El objetivo es, típicamente, el generalizar desde estos ejercicios a otros nuevos que no hayan sido observados previamente, de esta forma se darán más o menos respuestas correctas sin la intervención de alguien más. Esto representará si al final del experimento existe un progreso en el aprendizaje de la red.

Por otra parte, el aprendizaje sin supervisión tiene un enfoque diferente. Aquí el objetivo es, en términos simples, encontrar la estructura natural inherente en la información de ingreso. Hay un número de técnicas no supervisadas para aprender esquemas, incluyendo el aprendizaje competitivo, la teoría de la resonancia adaptativa y los mapas autoorganizados (SOM por sus siglas en inglés). Uno de los SOM más conocidos es la red neuronal artificial creada en 1980 por el profesor finlandés Teuvo Kohonen. Las redes o mapas de Kohonen son modelos de red neuronal con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro, inspirado en los modelos biológicos neuronales de los años 70 y modelos de morfogénesis creados por Alan Turing en los años 50.

El objetivo de una red de Kohonen es el mapear los vectores de entrada o patrones, de una dimensión arbitraria N a un mapa discreto con 1 o 2 dimensiones. Los patrones que están cerca uno de otro en el espacio de entrada también deberían estar cerca en este nuevo mapa, deberían estar ordenados topológicamente. Una red de Kohonen está compuesta por una cuadrícula de salidas y N unidades de entrada.

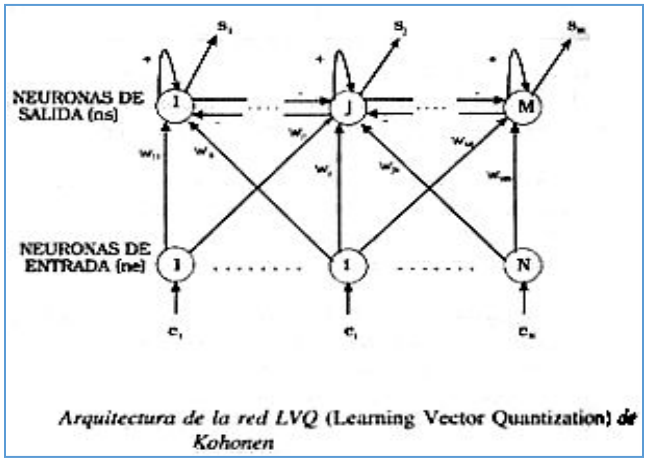


Figura 9: Ejemplo Red de Kohonen

El parámetro de neuronas de entrada está alimentando a cada unidad de salida, como puede verse reflejado en la Figura 9. Las líneas de entrada a cada salida tienen un valor determinado llamado peso. Estos pesos son inicializados inicialmente por números pequeños generados de manera aleatoria.

El proceso de aprendizaje en las redes de Kohonen es el siguiente:

```
Inicializar los pesos para cada unidad de salida
While (Los cambios de pesos sean despreciables)
       For (cada patrón de ingreso)
                Presenta el patrón de ingreso
                Encuentra la unidad de salida ganadora
                Encuentra todas las unidades en el vecindario del ganador
                Actualizar el peso de los vectores a todas esas unidades
       Disminuir cantidad de vecindario de ser necesario
```

La salida ganadora es simplemente la unidad con el peso de vector que tiene la distancia Euclidiana más pequeña en relación al patrón de ingreso. El vecindario de una unidad está definida como todas las unidades dentro de la distancia de la unidad del mapa (No en espacio de peso). Si el tamaño del vecindario es 1, entonces todas las unidades de no más de 1 ya sea horizontal o vertical desde cualquier unidad caerán dentro de su vecindario.

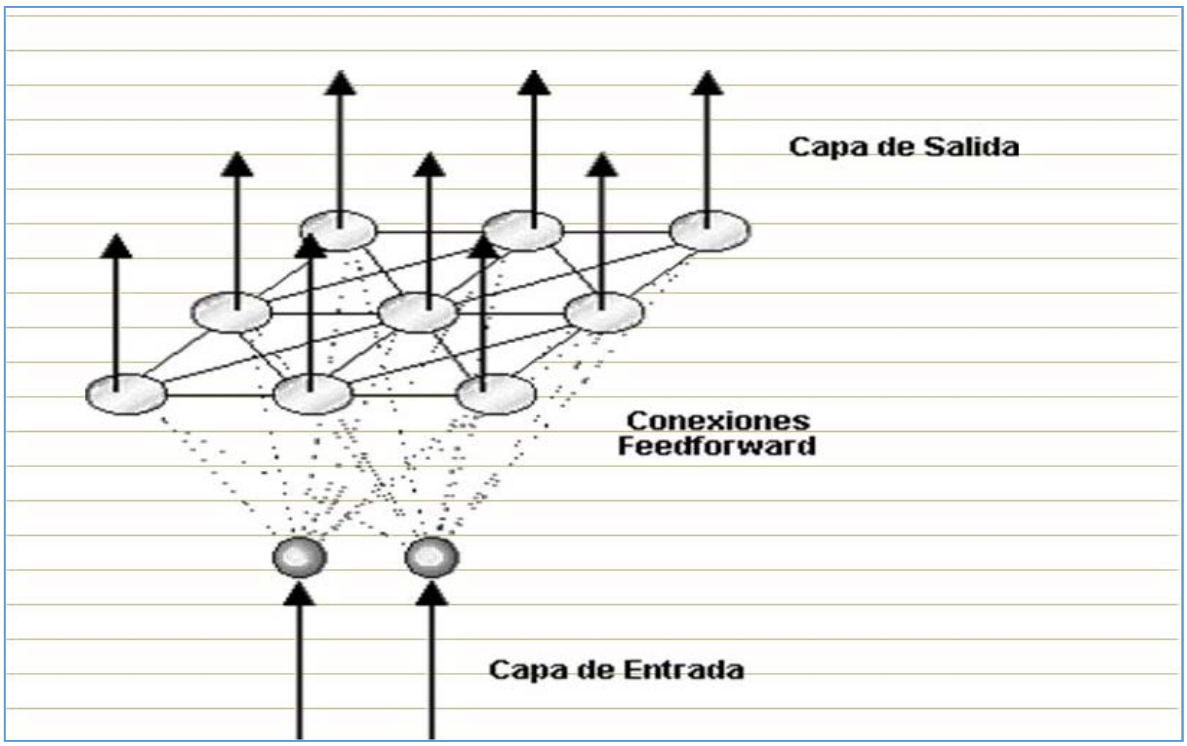


Figura 10: Ejemplo de Red de Kohonen

Anteriormente se explicó que las redes multicapas son capaces de computar un rango mayor de funciones booleanas que las redes de una capa. Sin embargo, el esfuerzo computacional necesario para encontrar la combinación correcta de pesos incrementa sustancialmente cuando más parámetros y topologías más complicadas son considerados. Un método popular que puede manejar tales problemas de aprendizaje a gran escala es el Backpropagation.

Conocida como la abreviación a “Propagación de errores hacia atrás”, es un método de entrenamiento de redes neuronales artificiales que es usado comúnmente en conjunto con un método de optimización como el gradiente de descenso. Este método calcula el gradiente de la “función de pérdida” con respecto a todos los pesos de la red. El gradiente es alimentado con el método de optimización el cual es usado para actualizar los pesos en un intento de minimizar la función de pérdida.

El objetivo de cualquier algoritmo de aprendizaje supervisado es el encontrar una función que mejor logre mapear un grupo de entradas hacia su salida correcta. Un ejemplo simple podría ser el de una tarea de clasificación, donde la entrada es la imagen del animal, y la salida correcta sería el nombre del animal.

El objetivo y motivación para desarrollar el algoritmo de backpropagation es el encontrar una manera de entrenar una red neuronal de multicapas, la cual pueda aprender representaciones internas apropiadas que le permitan aprender cualquier mapeado arbitrario de ingreso a salida. El algoritmo de backpropagation puede ser dividido en dos fases:

1) Propagación: Cada propagación consiste en los siguientes pasos:
    a) Propagación hacia delante de la entrada de un patrón de entrenamiento a través de redes neuronales, en orden para generar las propagaciones de activaciones de las salidas.
    b) Propagación hacia atrás de las propagaciones de activaciones de las salidas a través de las redes neuronales, usando el patrón de entrenamiento objetivo en orden para generar los deltas (Diferencias entre los valores objetivos y los valores de salidas actuales) de todas las salidas y las neuronas ocultas.

2) Actualización del peso: Para cada sinapsis de peso se siguen los siguientes pasos:
    a) Multiplicar el delta de salida y la entrada de activación por el gradiente de peso.
    b) Restar el porcentaje del gradiente del peso.

Este porcentaje influye en la velocidad y calidad del aprendizaje, es también llamado el "Radio de aprendizaje". Entre más grande sea el radio, más rápido se entrenan las neuronas, entre más lento sea el radio, más certero será el aprendizaje. El signo de la gradiente de un peso indica dónde está incrementándose el error, esto es porque el peso debe ser actualizado en la dirección opuesta.

Finalmente, se deben repetir las fases 1 y 2 hasta que el desempeño de la red sea satisfactorio.

2.2.4 ROI

Antiguamente todas las acciones, eventos y registros que contenían algún grado de información eran almacenados en papel, lo cual es incómodo de revisar para los estándares de hoy teniendo en consideración la gran cantidad de datos que se pueden tener en un solo día; eso sin contar los datos antiguos que se quieran revisar para hacer estimaciones estadísticas, recuentos de inventario, ventas del mes, etc. Además, hay varios otros problemas que resaltan a la hora de archivar los documentos en papel. Al utilizar un espacio físico en donde almacenar los datos, es sólo cuestión de tiempo para que la gran recopilación de información ocupe un lugar considerable en cualquier institución. Si este papel se daña, la información contenida en él se perderá ya que es poco probable que se tenga un respaldo.

Eventualmente llegó la era de la información, en donde se tiene una amplia cantidad de contenido multimedia que logra capturar la realidad de manera casi tal cual como nosotros la percibimos, y así almacenarla en pequeños dispositivos que caben perfectamente en la palma de la mano. De esta manera, se pensó en una forma para que se lograra digitalizar los documentos físicos y poder tenerlos en una forma digital para facilitar el respaldo de estos y acceder a ellos de una manera mucho más sencilla.

Así surgieron los OCR, que como se explica en el capítulo anterior, son capaces de extraer el texto contenido en una imagen y traspasarlo a un formato digital. Sin embargo, este procedimiento no es perfecto y tiene inconvenientes conocidos. Uno de ellos es la gran cantidad de contenido innecesario en el documento. Cuando se desee extraer el texto de una página, el OCR debe escanear toda la página del documento para poder determinar donde se encuentra el texto en ellas, pero generalmente se pueden encontrar otros elementos como timbres, firmas, manchas, y otra serie de imperfecciones que dificultan la tarea de poder extraer el texto.

Los ROI se construyeron para darle una solución a esta problemática, de manera que exista una forma de delimitar los alcances del OCR a sólo el área deseada por el usuario, dejando de lado el contenido sobrante. De esta manera se reduce considerablemente el tiempo de procesamiento al tener un margen menor, se reducen los errores al no estar considerados en el área que se desea escanear y se limita la información extraída a sólo lo necesario para el usuario.

Capítulo III

Desarrollo de la propuesta de solución

En este capítulo se va a explicar detalladamente los procedimientos que se utilizaron en el desarrollo de esta solución planteada.

Primero se comenzará por presentar el flujo de trabajo y explicar brevemente cada pasó de este proceso. Se detallan secuencialmente las etapas que comienzan desde la utilización de las plantillas, la extracción del texto en un formato editable, y cómo terminan en el reconocimiento de texto del OCR.

Luego, se mencionan todos los datos administrativos de este proyecto, los requisitos que son necesarios en el programa especificando cada uno de ellos, el diagrama de casos de uso y la explicación de cada caso de uso.

3.1 Metodología de trabajo

En este apartado se van a mencionar y analizar las distintas tareas que son necesarias para llevar a cabo este proyecto de título, desde la creación de las plantillas y su nivel de importancia hasta la utilización y funcionamiento del OCR, para el reconocimiento de los patrones y la posterior lectura de los documentos.

Los procesos de trabajo tienen dos partes fundamentales, que corresponden a la "Creación de plantillas" y el "Reconocimiento de Caracteres".

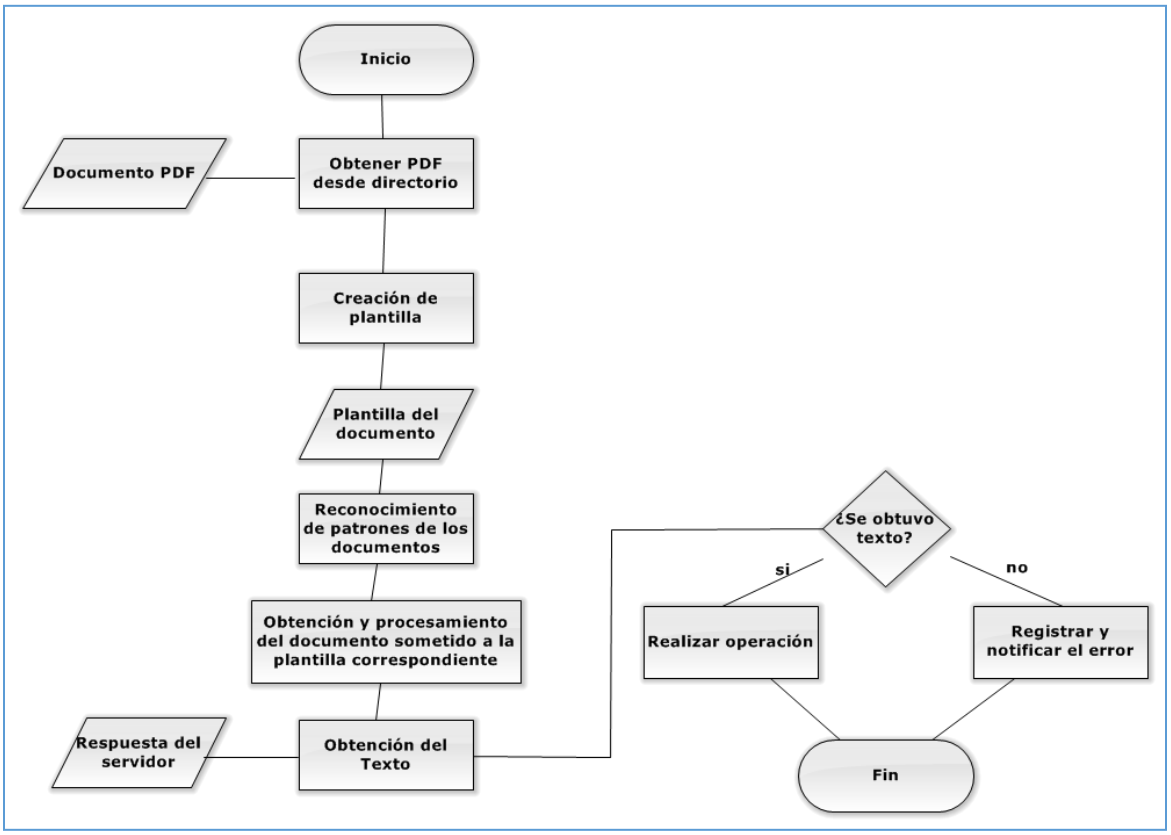


Figura 11: Metodología de Trabajo del OCR

El procedimiento del flujo de trabajo es el siguiente: Se comienza por cargar un documento de formato PDF en el programa. Luego, se crea una plantilla que marcará los cuadros de texto que el usuario crea necesarios, para luego pasar por el reconocimiento de patrones de los documentos proporcionado por el OCR.

Una vez que el documento haya pasado por el procesamiento del texto, se confirma si se ha logrado la extracción de texto, en el caso que no se haya logrado, se mostrará un mensaje de error y se registrará para su posterior análisis, en el caso contrario, se realizará la operación correspondiente.

A continuación, se describirán de forma detallada los dos principales procesos asociados a la creación de plantillas y reconocimiento de caracteres.

1. Creación de plantillas

La creación de las plantillas es un proceso fundamental, ya que en él se busca reducir el área en la cual el OCR trabaja leyendo y reconociendo los distintos archivos que le son entregados. Al tener áreas específicas que escanear y no tener que procesar el documento en su totalidad, se reduce considerablemente el tiempo de procesamiento y, por ende, el tiempo de respuesta que el usuario estará esperando. Además, se tiene en consideración los distintos tipos de etiquetas que determinarán las diferencias entre un tipo de documento u otro al analizar los patrones estructurales.

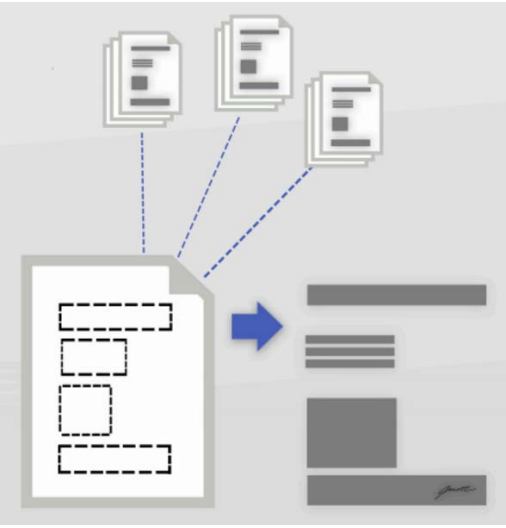


Figura 12: Estructura de plantillas para el reconocimiento de patrones

Para la creación de las plantillas se deben cumplir con las siguientes actividades.

- Seleccionar archivo: En este punto se debe buscar el documento a analizar, el cual debe cumplir con las formas y proporciones ideales para ese tipo de documento en particular.

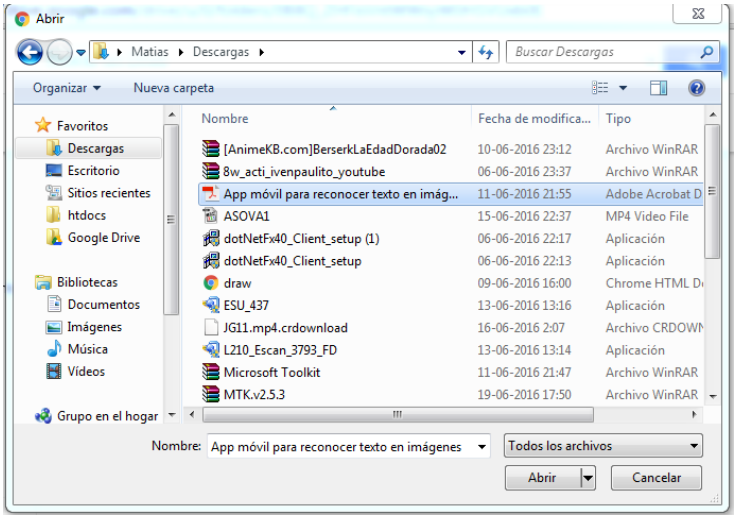


Figura 13: Imagen de selección de documento

- Dibujar las zonas de interés: En esta etapa se “dibuja” o selecciona los párrafos y su ubicación en el documento, estos serán posteriormente sometidos al reconocimiento. Las zonas demarcadas deberán tener un margen extra de selección de unos 10 pixeles mínimo con respecto al área que originalmente se pensaba escanear. Esto es conocido como “margen de error” y es utilizado para evitar problemas de selección de texto. En esta etapa se deben agregar algunos atributos propios de cada párrafo tales como “tipo de selección”, ya que puede ser un título, sub-título, párrafo principal, entre otros.

- Clasificar tipos de párrafos: Se clasifican las distintas marcas dibujadas con el fin de definir cuál es la información importante en esta zona. La clasificación de los párrafos permiten dar prioridades de lectura y reconocimiento otorgando un grado de preferencia sobre otros, dependiendo de estos datos se puede obviar el resto de escaneo innecesario. Ej: Se analiza la fecha de vencimiento de un documento y este resulta vencida, lo que lleva a la finalización del escaneo.

- Reconocer el tipo de documento: Esta actividad consiste en reconocer los patrones de cada plantilla según la estructura y el formato que estos tengan. El formato es crucial para determinar si un documento es un contrato, o si es una factura de cotización. Se tiene en cuenta una cantidad determinada de plantillas, de esta manera se podrá utilizar un reconocimiento de patrones que logrará determinar el origen de procedencia del documento, o sea, si este documento pertenece a una empresa específica.

- Descargar y guardar XML: Proceso en el cual se descarga la plantilla creada para mantener consistencia y así no tener que dibujar la plantilla cada vez que se analiza un mismo tipo de documento.

2. Reconocimiento

Proceso que consiste en someter el documento al OCR después de haber sido clasificado según su patrón y las plantillas que se encuentran en el sistema.

El reconocimiento se realiza en 2 etapas principalmente: Primero, el reconocimiento de la morfología del documento, es decir, el análisis de la distribución y forma de un documento al compararlo con las plantillas en el sistema. Segundo, el análisis del reconocimiento propio de los patrones de caracteres alfanuméricos.

En primera instancia el OCR facilitado por la empresa realiza los siguientes procesos.

- Clasificar los documentos en distintos directorios, verificando la dirección de donde proviene para así agregar el documento. En el caso que no exista, se creará un nuevo directorio y se agregarán los ROI y los documentos correspondientes.

- Implementar las técnicas de visión artificial las cuales ameriten recortar la imagen, utilizando las coordenadas del ROI.xml, con esto se desechan las zonas que sean distintas a las seleccionadas.

- Aplicar la transformada de Hough, para detectar las líneas de los párrafos. Como se trabajará con documentos escaneados, no necesariamente se espera que los documentos estén perfectamente alineados, los párrafos pueden estar rotados, por lo cual el OCR ajusta la rotación de la imagen para capturar las líneas del párrafo. Una vez corregida la rotación, el OCR detecta todas las líneas de la zona demarcada por la plantilla.

- El OCR selecciona y va reduciendo el área, detectando las palabras presentes en cada línea. Luego reduce aún más el área hasta llegar al nivel de las letras.

- Una vez delimitado el caracter, este pasa por una red neuronal la cual entrega como salida la letra a la que corresponde el carácter, para luego de detectada la letra o número correspondiente al abecedario se reconstruyen los párrafos usando el mismo procedimiento pero en retroceso, es decir, se reconstruyen las palabras, luego las oraciones hasta reconstruir el párrafo que el OCR está transformado a datos con los cuales se podrá trabajar más adelante.

- Para evitar problemas de reconocimiento de alguno de los caracteres, la palabra será sometida a una red bayesiana con la cual se comparará la palabra obtenida con un diccionario.

3.2 Especificación de requerimientos del software

El sistema se encargará principalmente de la administración de las diferentes plantillas, las funciones correspondientes a estos procesos, así como lo referente a la extracción del texto presente en las zonas de interés y su posterior reconocimiento por parte del OCR.

Los usuarios del sistema podrán acceder a la página en donde se encuentra localizado el programa y, de esta manera, realizar todas las funciones propuestas. Este software será agregado como un plugin de un programa mayor.

Considerando las diferentes necesidades que se deben realizar, se ha elaborado una lista con los requisitos que este proyecto debe tener en cuenta.

Código	Requisito Funcional	Especificación de Requisito
RQ-01	Cargar archivos	El programa deberá ser capaz de cargar archivos pdf.
RQ-02	Crear plantillas	El programa deberá ser capaz de crear plantillas.
RQ-03	Editar plantillas	El programa deberá ser capaz de editar plantillas.
RQ-04	Eliminar plantillas	El programa deberá ser capaz de eliminar plantillas.
RQ-05	Implementar un historial de plantillas	El programa contará con un historial de las plantillas usadas recientemente.
RQ-06	Implementado en servidor web	El programa deberá estar implementado en un servidor web.
RQ-07	Proveer guía de uso	El programa contará con una guía rápida de uso que facilitará el acceso a los nuevos usuarios.
RQ-08	Contar con una pre-visualización	El programa contará con una pre-visualización de la información que se va a procesar.
RQ-09	Remarcar la selección del ROI	El programa deberá ser capaz de remarcar la selección del ROI y poder diferenciar las distintas áreas delimitadas.
RQ-10	ROI con atributos	Los ROI deberán tener asociados atributos (Nombre, empresa, fecha, tipo de documento, etc).
RQ-11	Enviar múltiples archivos	El programa implementará la posibilidad de enviar múltiples archivos del mismo tipo o carpetas con archivos.
RQ-12	Agregar un buscador de plantillas	El programa deberá implementar un buscador de plantillas.
RQ-13	Visualizar al paginar	El programa presentará una visualización de los elementos que han sido seleccionados al momento de paginar.
RQ-14	Conservar las selecciones al cambiar de página	El programa deberá mantener las áreas seleccionadas de una página al momento de cambiarla, y volver a marcarlas cuando se seleccione nuevamente.
RQ-15	Agregar prioridades	El programa deberá poder determinar los campos que tengan prioridad por sobre otros. Ej: Determinar si se sobrepasa la fecha de caducidad, en ese caso el documento queda automáticamente desechado.
RQ-16	Resultado de la lectura	El programa deberá emitir de salida un xml con el mismo formato de la plantilla, pero con la información capturada dentro.
RQ-17	Documento ideal	El programa trabará en base a un documento ideal, véase, que esté bien alineado.
RQ-18	Margen error	El programa contará con un área grande de escaneo como margen de error.

Figura 12: Requerimientos Funcionales

Código	Requisito no Funcional	Especificación de Requisito
RN-01	Tiempo de respuesta	El programa contará con un corto tiempo de respuesta de menos de 5 segundos.
RN-02	Interfaz	El programa contará con una interfaz amigable y moderna para el usuario.
RN-03	Fácil de usar	El programa deberá ser intuitivo y no tan complejo de utilizar.
RN-04	Seguridad	El sistema deberá autorizar el acceso del cliente, administrador.
RN-05	Notificaciones	El programa deberá emitir notificaciones si se sobrepasa el peso máximo permitido.
RN-06	Servidor	El programa hará uso del servidor interno de la empresa.
RN-07	Base de datos	El programa utilizará una base de datos relacional para el almacenamiento de los datos.
RN-08	Compatibilidad	El sistema deberá visualizarse en cualquier navegador web.
RN-09	Portabilidad	El sistema deberá ser capaz de funcionar remotamente en cualquier dispositivo.

Figura 13: Requerimientos no Funcionales

3.3 Diagrama de casos de uso

A continuación se mostrará el diagrama de casos de uso, seguido por las especificaciones correspondientes.

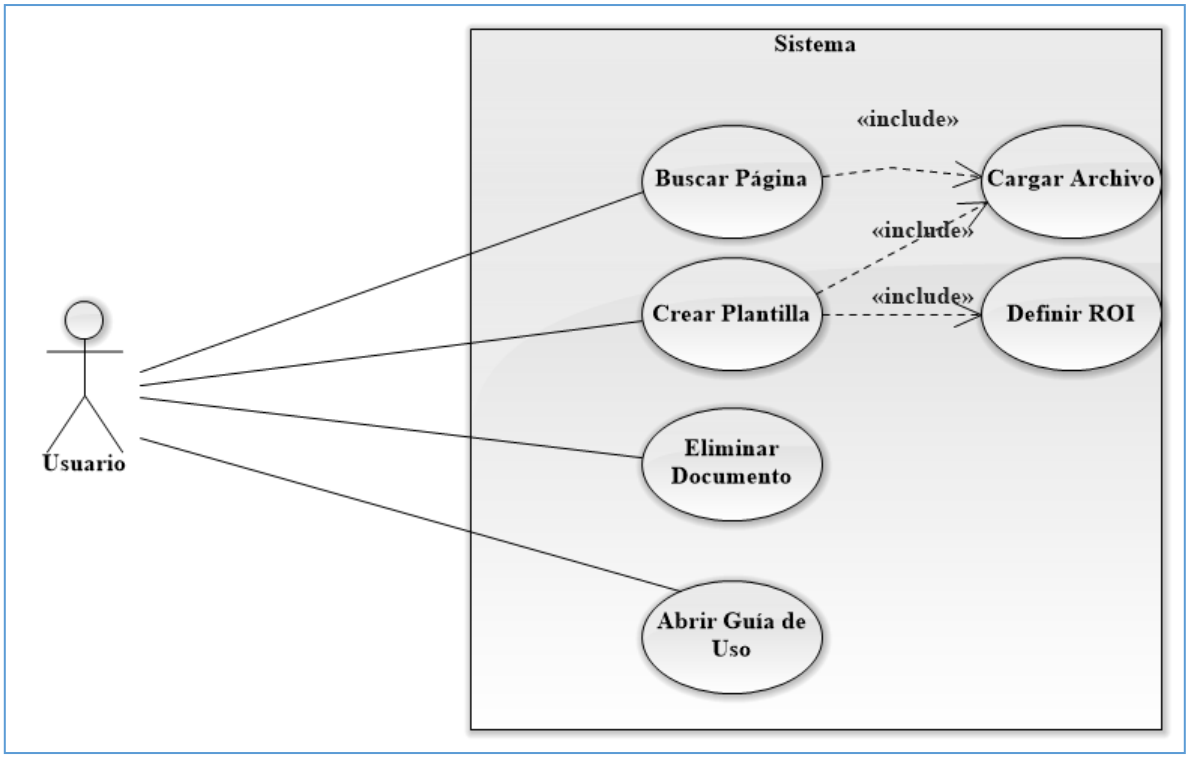


Figura 14: Diagrama de casos de Uso

Definición del Caso de Uso	
ID Caso de Uso	CU01
Nombre Caso Uso	Crear Plantilla
Actor Principal	Usuario
Descripción	Este caso de uso tiene como objetivo permitir que el usuario sea capaz de crear una plantilla
Pre-Condiciones	El documento con el cual confeccionar la plantilla debe estar cargado en el sistema y visible en el espacio de trabajo
Flujo de Eventos	1. El caso de uso comienza cuando el usuario muestra interés en crear una plantilla. 2. El usuario se dirige a la pestaña de “crear plantilla” 3. El sistema carga el espacio de trabajo donde se muestra un documento de ejemplo por defecto. 4. Incluye caso de uso Cargar Archivo. 5. Incluye caso de uso Definir ROI. 6. El usuario guarda el ROI presionando el botón “guardar”. 7. El sistema inicia la descarga de la plantilla al equipo. 8. Una vez terminado paso 7 el caso de uso llega a su fin.
Post-Condiciones	Se genera una nueva plantilla
CU Asociados	Cargar Archivo. Definir ROI.
Excepciones	1. La excepción ocurre si es que se tiene algún problema de conexión que evite el finalizar el guardado de la plantilla. 2. Se emitirá un mensaje de error y el caso de uso termina.

Figura 15: Caso de Uso 1

Definición del Caso de Uso	
ID Caso de Uso	CU02
Nombre Caso Uso	Abrir Guía de Uso
Actor Principal	Usuario
Descripción	El caso de uso tiene como objetivo abrir una guía de uso rápido.
Pre-Condiciones	
Flujo de Eventos	1. El caso de uso comienza cuando el usuario desea abrir la guía de uso. 2. El usuario selecciona la opción de “guía rápida”. 3. El sistema mostrará varias capturas de pantalla indicando paso a paso el funcionamiento del programa de una manera simple para un usuario poco experimentado. 4. Una vez que el usuario haya terminado la lectura del manual y cierra la ventana, el caso de uso llega a su fin.
Post-Condiciones	
CU Asociados	
Excepciones	• La excepción ocurre cuando no se logre cargar las imágenes de la guía de uso. • Se emitirá un mensaje de error y el caso de uso termina.

Figura 16: Caso de Uso 2

Definición del Caso de Uso	
ID Caso de Uso	CU03
Nombre Caso Uso	Eliminar Documento
Actor Principal	Usuario
Descripción	Este caso de uso tiene como objetivo eliminar un documento.
Pre-Condiciones	Se debe haber un documento cargado anteriormente en el sistema.
Flujo de Eventos	1. El caso de uso comienza cuando el usuario desea eliminar un documento. 2. El usuario selecciona la opción de “Eliminar archivo”. 3. El sistema mostrará una lista con los documentos que se encuentren cargados. 4. El usuario escoge cuál de los archivos desea eliminar y procede a presionar el botón para confirmar. 5. El sistema elimina el documento seleccionado. 6. Habiendo terminado el paso 5, el caso de uso llega a su fin.
Post-Condiciones	El documento seleccionado es eliminado del sistema
CU Asociados	
Excepciones	1. La excepción ocurre si es que se tiene algún problema de conexión que evite el finalizar la eliminación del documento. 2. Se emitirá un mensaje de error y el caso de uso termina.

Figura 17: Caso de Uso 3

Definición del Caso de Uso	
ID Caso de Uso	CU04
Nombre Caso Uso	Cargar Archivo
Actor Principal	Usuario
Descripción	Este caso de uso tiene como objetivo cargar un archivo de formato pdf al programa.
Pre-Condiciones	
Flujo de Eventos	1. Este caso de uso comienza cuando el usuario desea cargar un archivo. 2. El usuario selecciona la opción de “cargar archivo” 3. El sistema despliega una ventana con una lista de archivos que pueden ser cargados. 4. El usuario selecciona el archivo deseado de extensión .pdf y entonces escoge “subir archivo”. 5. El sistema carga el archivo seleccionado en el sistema. 6. Una vez finalizado el paso 5, el caso de uso llega a su fin.
Post-Condiciones	El archivo queda cargado en el sistema
CU Asociados	
Excepciones	• La excepción ocurre cuando no se encuentran pdfs en la carpeta seleccionada. • Se emitirá un mensaje de error y el caso de uso termina.

Figura 18: Caso de Uso 4

Definición del Caso de Uso	
ID Caso de Uso	CU05
Nombre Caso Uso	Buscar Página
Actor Principal	Usuario
Descripción	El caso de uso tiene como objetivo buscar una página específica
Pre-Condiciones	
Flujo de Eventos	1. El caso de uso comienza cuando el usuario desea buscar una página en un archivo. 2. Incluye caso de uso Cargar Archivo. 3. Una vez cargado el archivo, el usuario ingresa el número de página al cual se desea ir. 4. El sistema muestra la página deseada. 5. Terminado el paso 4, el caso de uso llega a su fin
Post-Condiciones	
CU Asociados	Cargar Archivo
Excepciones	• La excepción ocurre cuando no se encuentra la página señalada. • Se emitirá un mensaje de error y el caso de uso termina.

Figura 19: Caso de Uso 5

Definición del Caso de Uso	
ID Caso de Uso	CU06
Nombre Caso Uso	Definir ROI
Actor Principal	Usuario
Descripción	El caso de uso tiene como objetivo demarcar las áreas de interés para el OCR.
Pre-Condiciones	
Flujo de Eventos	1. El caso de uso comienza cuando el usuario desea crear un nuevo ROI 2. El usuario hace click en un área del documento arrastrando con el mouse el área de selección hasta formar un rectángulo. Esta área contendrá el texto que será procesado. 3. El sistema demarca la zona seleccionada con una línea color verde. 4. Inmediatamente el sistema muestra una ventana en donde se pide ingresar un tipo de párrafo. 5. El usuario ingresa un tipo de área acorde a sus necesidades. 6. El sistema cambia el color del área a azul demostrando que los cambios han sido guardados. 7. Terminando el paso 6, el caso de uso llega a su fin.
Post-Condiciones	
CU Asociados	
Excepciones	• La excepción ocurre cuando no se logre guardar exitosamente. • Se emitirá un mensaje de error y el caso de uso termina.

Figura 20: Caso de Uso 6

3.4 Modelo Entidad Relación

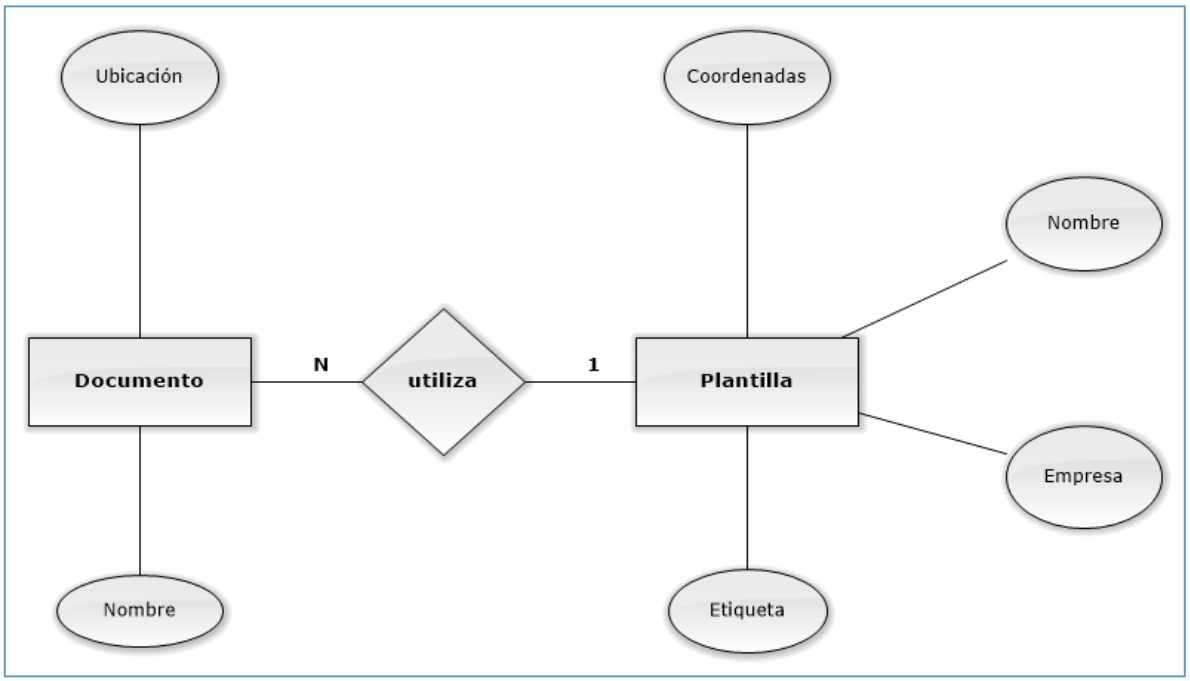


Figura 21: Modelo Entidad Relación

Especificación MER

Especificación de entidades

- Entidad Plantilla: Almacena las zonas demarcadas que serán escaneadas y posteriormente procesadas por el OCR. Tiene como atributos nombre, empresa y etiqueta.
- Entidad Documento: Representan al archivo con extensión pdf. Es el documento cargado en el programa y del que se desea extraer algún texto.

Relación entre entidades

- Plantilla-Documento: Los Documentos serán subidos al sistema y demarcados por la entidad plantilla.

3.5 Diagrama de paquetes

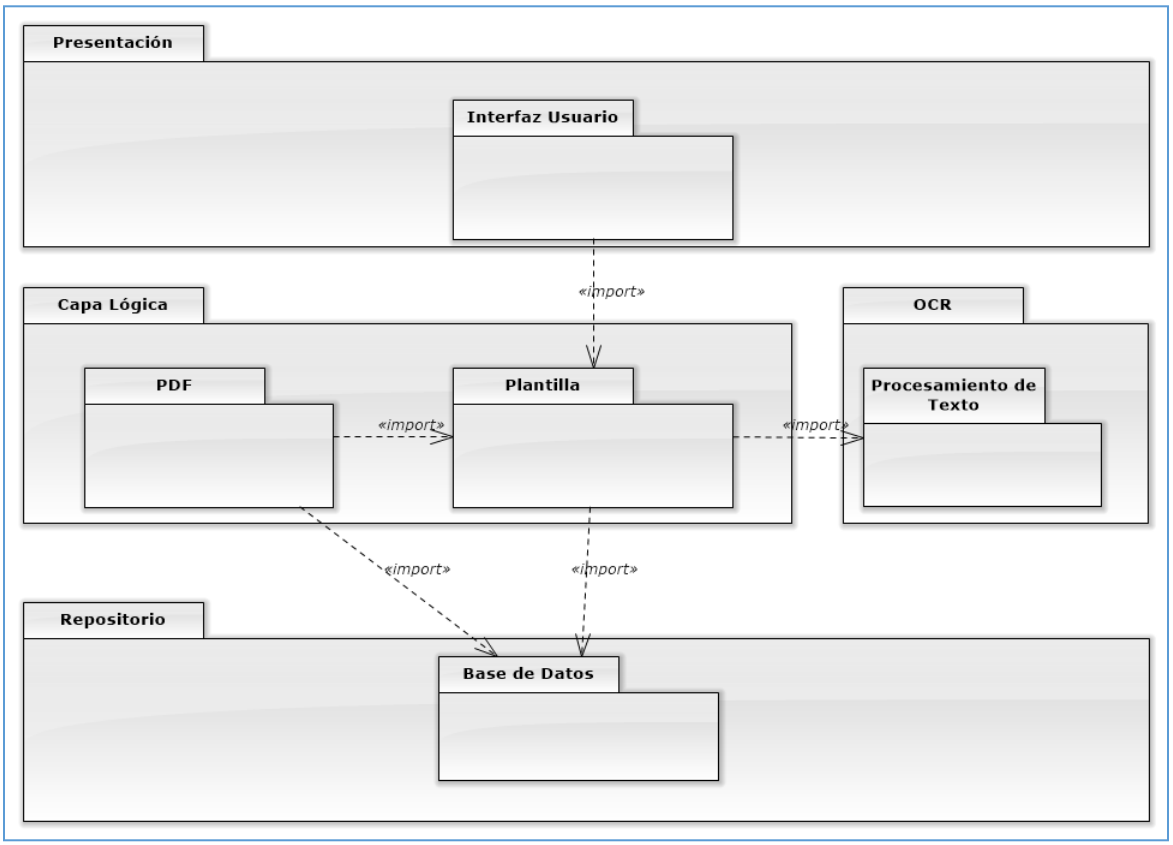


Figura 22: Diagrama de Paquetes

En este diagrama de paquetes se presentan las diferentes agrupaciones lógicas. También se puede observar las diferentes capas con las que se está trabajando.

El usuario es capaz de acceder al sistema por medio de la interfaz. De esta manera se pueden crear, editar o eliminar plantillas. Las plantillas están asociadas a un archivo PDF que se encuentra almacenado en una base de datos al igual que las plantillas creadas por el usuario. Cuando el usuario decida ejecutar la opción de reconocimiento de texto, las plantillas recurrirán al OCR para que este procese la información.

3.6 Diagrama de clases

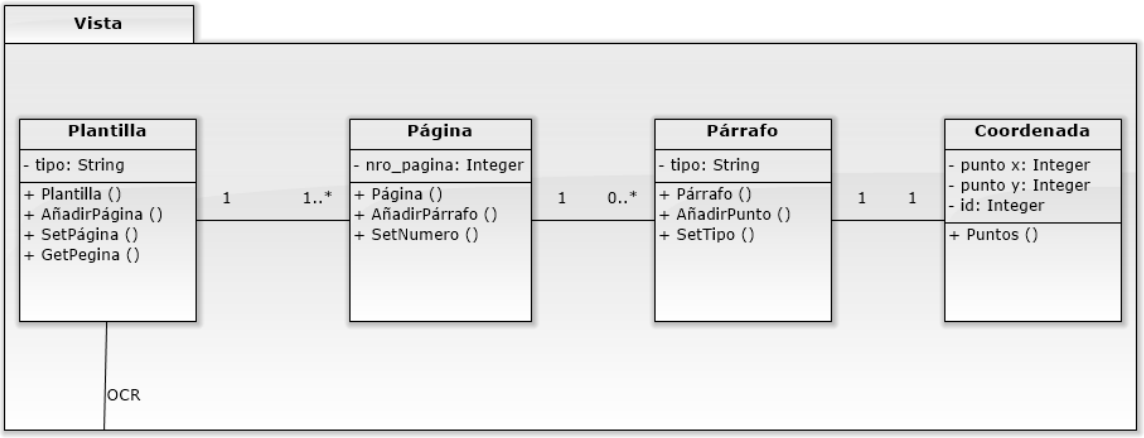


Figura 23: Diagrama de Clases

En este diagrama de clases se presentan las diferentes estructuras que componen el sistema principal. La clase Plantilla contiene un atributo tipo, esta contiene varias Páginas que a su vez poseen un número de página. Para poder determinar qué zonas deben ser escaneadas, la clase Párrafo contiene la información con la Coordenada necesaria para formar la figura que será procesada. La coordenada contiene puntos con la ubicación por pixel en donde se generará la figura del ROI.

La otra parte del sistema es lo relacionado al funcionamiento del OCR. El texto previamente delimitado del documento es mandado al OCR para que este pueda extraer todo el texto incluido en ellos. Este proceso debe ser ejecutado de manera externa al sistema principal y se debe ejecutar por medio de la consola de Windows.

3.7 Ejemplos de uso

En esta sección del informe se relatará un ejemplo de uso y las acciones a seguir para poder ejecutar el programa.

El programa comienza su ejecución en el servidor local en el cual está alojado, por medio del programa Xampp el cual configura un servidor apache para la ejecución de sistemas web. El sistema completo será implementado en un futuro como un plugin de un producto de gestión documental de la empresa Exe. Con el actual desarrollo, el programa sólo funciona de manera estable utilizando el navegador Mozilla Firefox, esto porque una de las apis utilizadas en el sistema no se encuentra optimizada para el resto de navegadores para que cargue los datos completamente.

Para hacer uso del programa se deben tener activos los servicios de Apache en Xampp, y desde la barra de dirección del navegador ingresar la URL http://localhost/ocr/new/, una vez iniciado el sistema, se presenta un saludo, luego se selecciona la pestaña "*Crear Plantilla*" para proceder como se puede apreciar en la *Figura 24 y 25*.

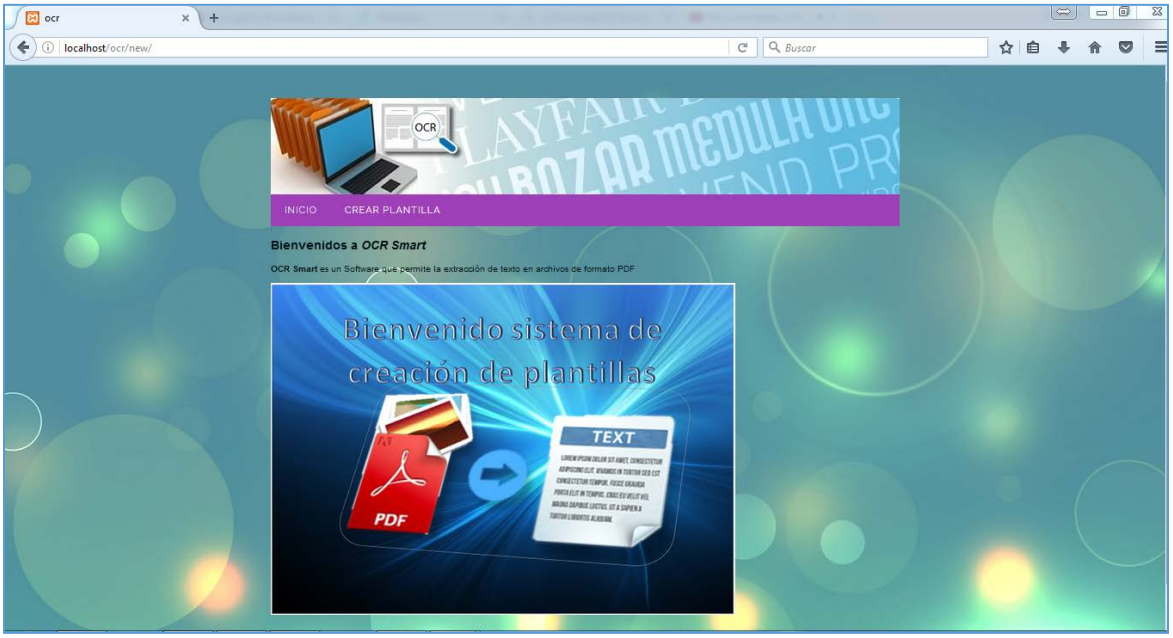


Figura 24: Inicio programa

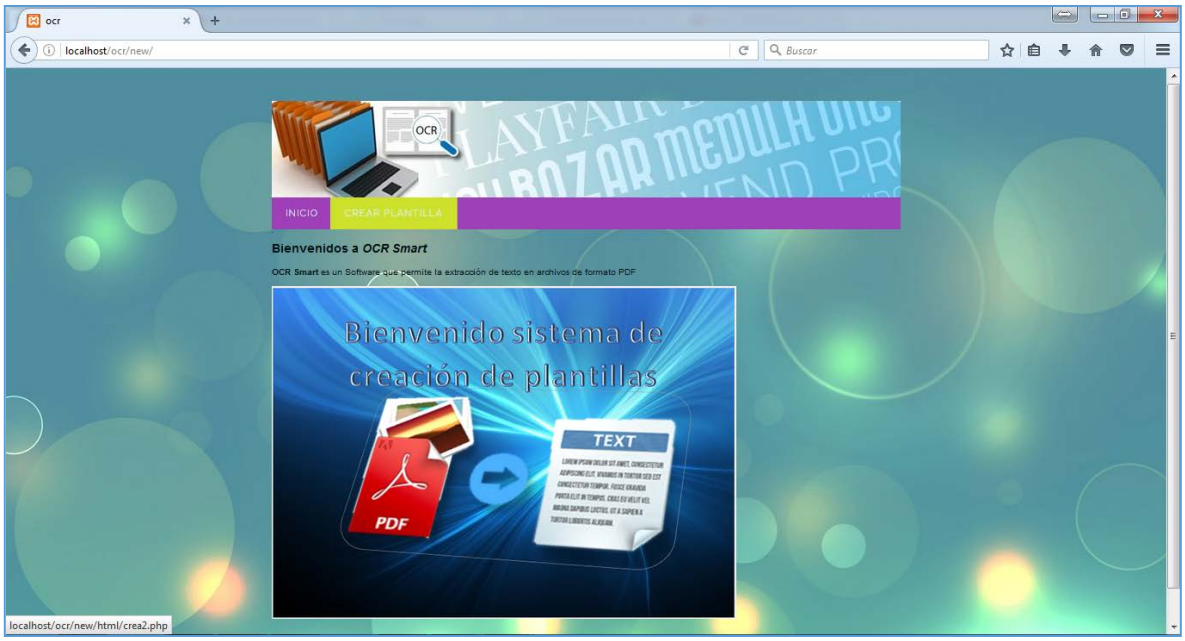


Figura 25: Selección de pestaña

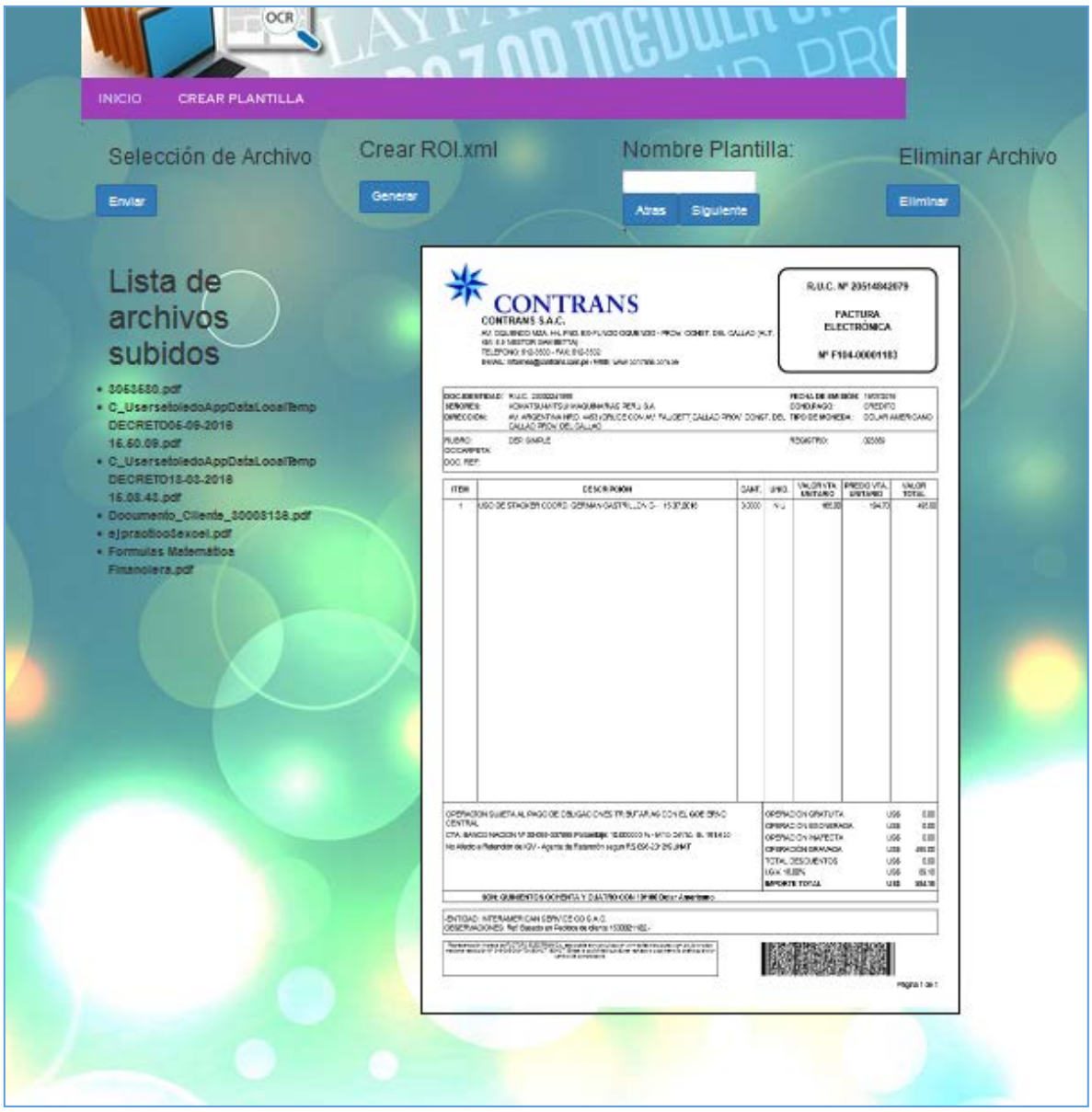


Figura 26: Creación de Plantilla

En selección de archivo se señala qué archivo se subirá al sistema, este debe ser de formato PDF, de lo contrario se emitirá un mensaje de error. Luego, utilizando el cursor se demarca con el mouse un área en el documento y haciendo click sobre el documento se arrastra hasta demarcar dentro del contorno lo que se desee analizar quedando de color verde como puede verse en la *Figura 27*. Enseguida aparecerá en la parte inferior del documento una ventana en donde se debe seleccionar un nombre de etiqueta para ese ROI, reflejado en la *Figura 28*, luego la zona demarcada cambia a color azul denotando que está guardado, lo anterior se presenta en la *Figura 29*.

Figura 27: Recuadro demarcado Factura Electrónica

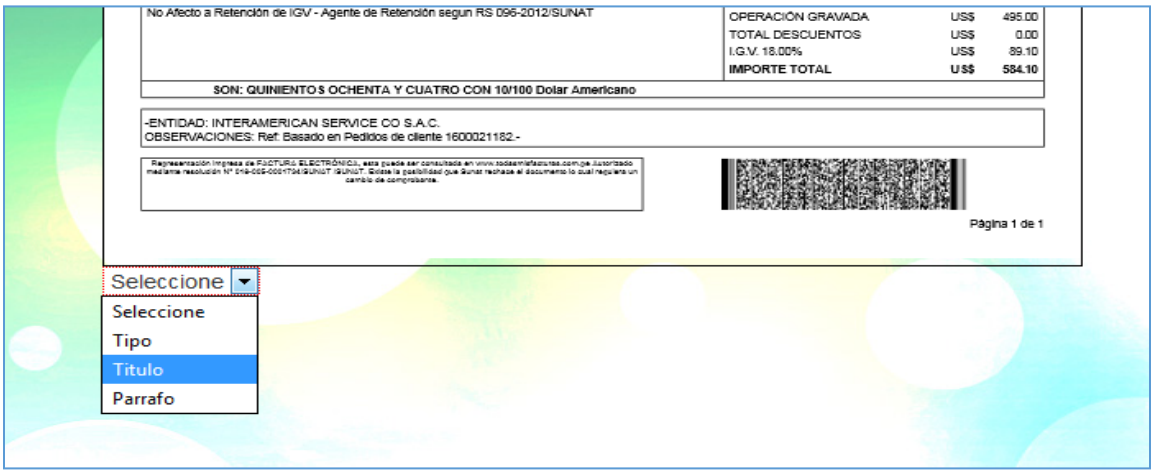


Figura 28: Ejemplo de selección 2

Figura 29: Ejemplo de selección 3

Los cambios realizados hasta ahora están almacenados de manera temporal en el sistema. Se pueden realizar múltiples áreas de selección por plantilla. Para finalizar se coloca nombre a la plantilla y se descarga un archivo con nombre ROI.xml el cual queda guardado en el disco local, como se muestra en la *Figura 30*.

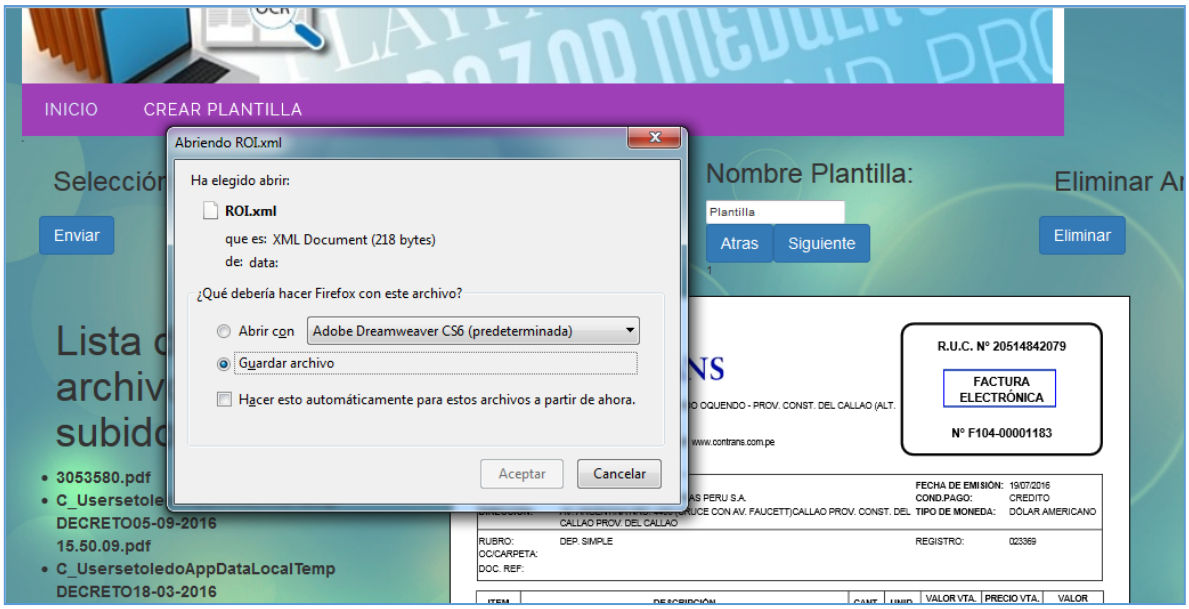


Figura 30: Guardando ROI

La ejecución del OCR se debe hacer por medio de la consola de Windows ejecutando el siguiente comando: C:\VersionOCR>java -jar ocrSmart.jar -cp ocrSmart_lib/, esto puede verse en la *Figura 31*. En este caso de ejecución, el archivo ocrSmart.jar se encuentra alojado en el disco C de nuestro sistema, en la carpeta VersionOCR, donde se hace la llamada del jar con las librerías correspondientes para su ejecución.

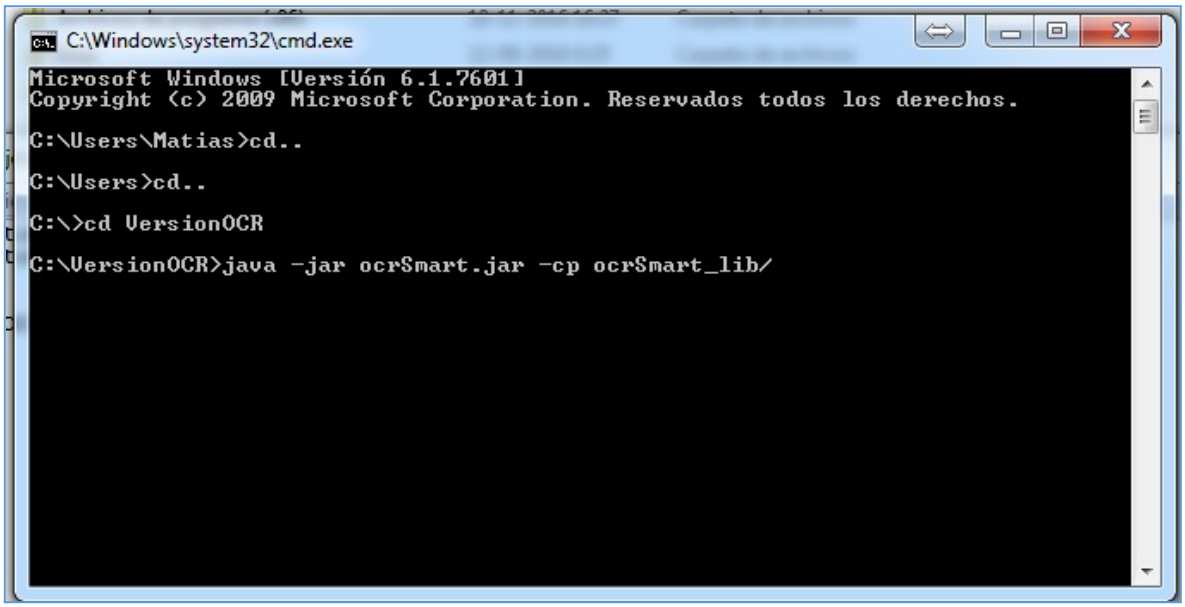


Figura 31: Inicio de Ejecución OCR

Aquí se ejecuta el programa cargando la librería asociada. La red neuronal, ha sido entrenada con anterioridad para la realización de los reconocimientos simples de los caracteres. Es importante destacar que debido al escaso grado de entrenamiento existente en la red, su eficiencia y grado de exactitud puede no ser el más apropiado. Desde la empresa Exe nos aseguraron que la optimización del OCR con un mayor número de ejercicios para su entrenamiento, será propuesto como un tema de proyecto de tesis futuro para alguien que desee abordarlo.

Para la ejecución correcta del OCR, se debe haber creado una plantilla anteriormente, esta estará almacenada en la carpeta templates del sistema que se encuentra en la ruta C:\VersionOCR\resources\templates en el computador del usuario. La biblioteca de documentos a analizar debe estar en la carpeta Archivos alojada en la ruta C:\VersionOCR\resources\Archivos.

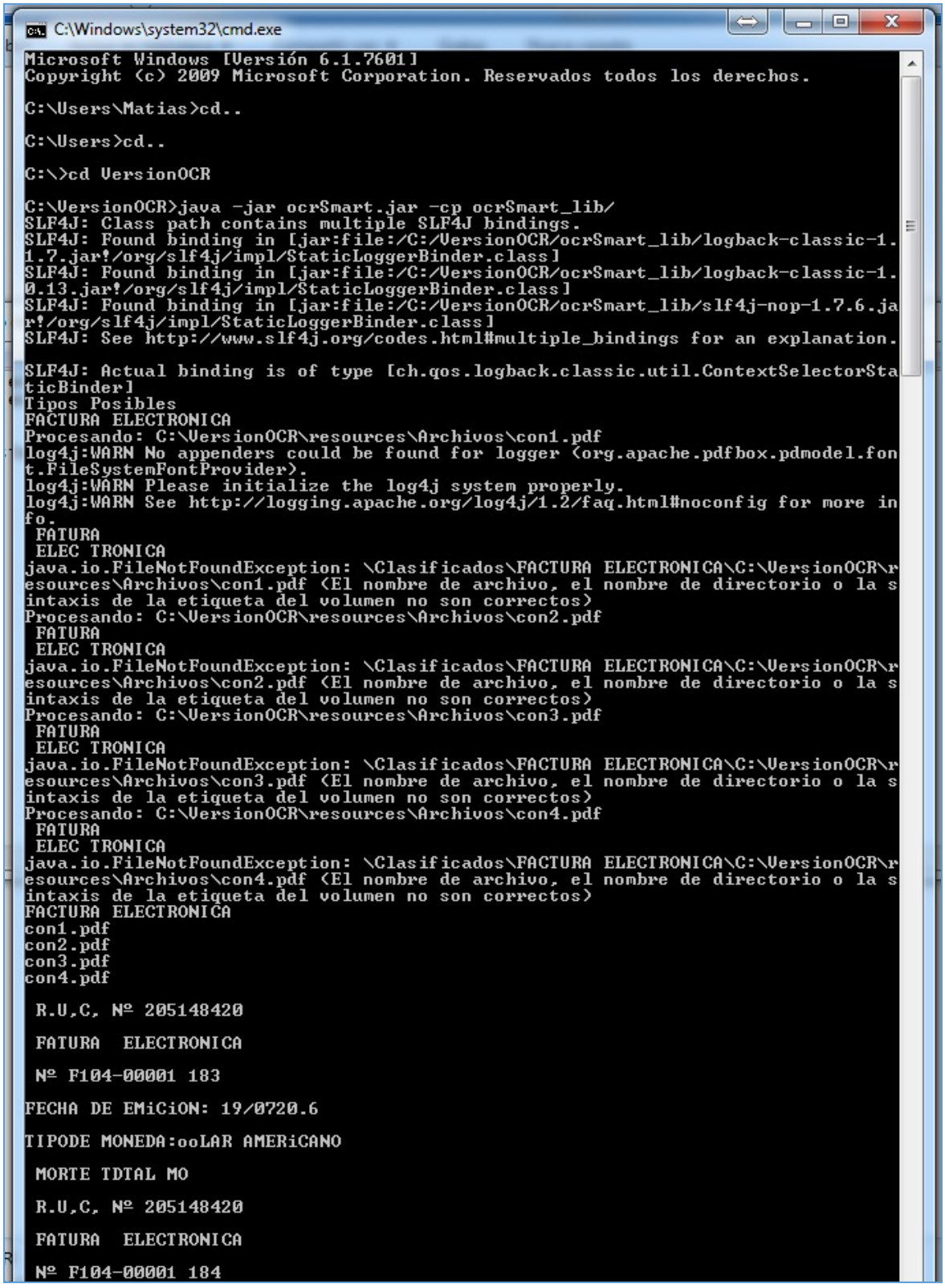


Figura 32: Resultado de Ejecución OCR 1

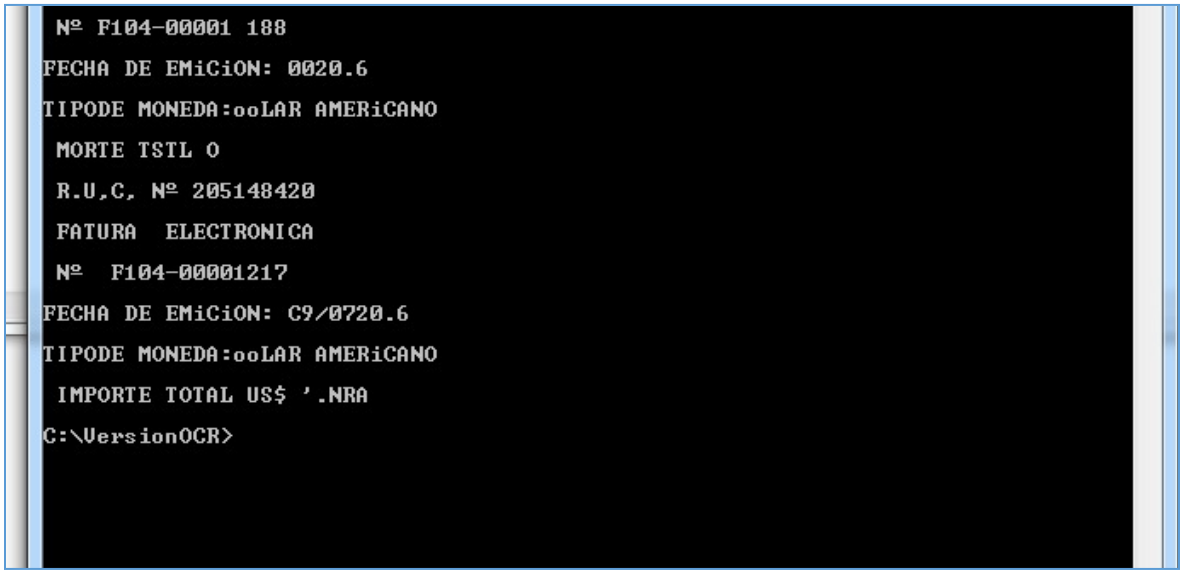
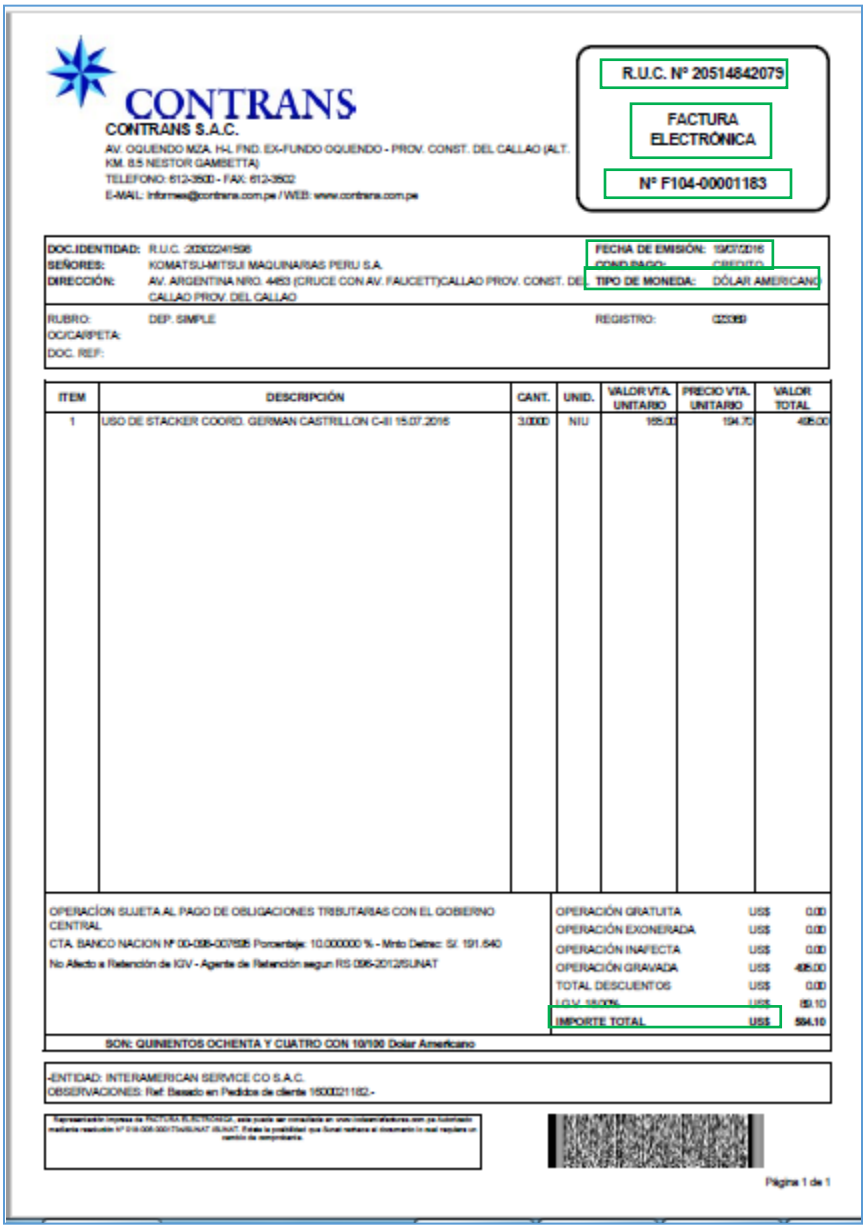


Figura 33: Resultado de Ejecución OCR 2

Como se puede apreciar en las *Figuras 32 y 33*, el resultado de la extracción del OCR es bastante acertado. A pesar de unas pequeñas fallas, se puede estimar que la exactitud actual del programa es de un 75% aprox.

El documento de control al cual se le aplicó la prueba fue el que se muestra en la *Figura 34*:



CONTRANS

CONTRANS S.A.C.
AV. OQUENDO MZA. H-L FND. EX-FUNDO OQUENDO - PROV. CONST. DEL CALLAO (ALT. KM. 8.5 NESTOR GAMBETTA)
TELEFONO: 612-3500 - FAX: 612-3502
E-MAIL: informes@contrans.com.pe / WEB: www.contrans.com.pe

R.U.C. Nº 20514842079
FACTURA ELECTRÓNICA
Nº F104-00001183

DOC.IDENTIDAD: R.U.C. 20302241598
SEÑORES: KOMATSU-MITSUI MAQUINARIAS PERU S.A.
DIRECCIÓN: AV. ARGENTINA NRO. 4483 (CRUCE CON AV. FAUCETT)CALLAO PROV. CONST. DEL CALLAO PROV. DEL CALLAO
FECHA DE EMISIÓN: 19/07/2016
CONDUCTO: CREDITO
TIPO DE MONEDA: DÓLAR AMERICANO
RUBRO: DEP. SIMPLE
REGISTRO:
OC/CARPETA:
DOC. REF:

ITEM	DESCRIPCIÓN	CANT.	UNID.	VALOR VTA. UNITARIO	PRECIO VTA. UNITARIO	VALOR TOTAL
1	USO DE STACKER COORD. GERMAN CASTRILLON C-II 15.07.2016	3.0000	NIU	165.00	194.70	495.00

OPERACIÓN SUJETA AL PAGO DE OBLIGACIONES TRIBUTARIAS CON EL GOBIERNO CENTRAL
CTA. BANCO NACION Nº 00-068-007595 Porcentaje: 10.000000 % - Mnto Detrac: S/. 191.640
No Afecto a Retención de IGV - Agente de Retención segun RS 096-2012/SUNAT

OPERACIÓN GRATUITA	US$	0.00
OPERACIÓN EXONERADA	US$	0.00
OPERACIÓN INAFECTA	US$	0.00
OPERACIÓN GRAVADA	US$	495.00
TOTAL DESCUENTOS	US$	0.00
I.G.V. 18.00%	US$	89.10
IMPORTE TOTAL	US$	584.10

SON: QUINIENTOS OCHENTA Y CUATRO CON 10/100 Dolar Americano

-ENTIDAD: INTERAMERICAN SERVICE CO S.A.C.
OBSERVACIONES: Ref: Basado en Pedidos de cliente 1500021182.-

Página 1 de 1

Figura 34: Boleta

Como se puede apreciar las zonas demarcadas corresponden al área delimitada por la plantilla que utiliza este tipo de documento.

El resultado de la ejecución nos genera un documento tipo .txt que se presenta en la *Figura 35*:

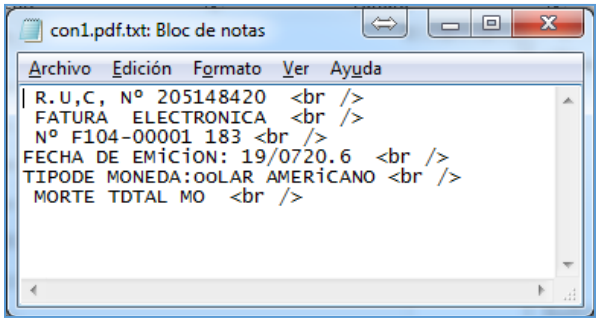


Figura 35: Resultado del OCR

Todas las líneas terminan con <*br/*>, ya que según lo señalado por Exe el OCR está pensado para ser aplicado como un web service.

Al comparar los textos resultantes nos dan los resultados que observamos en la *Figura 36 y 37.*

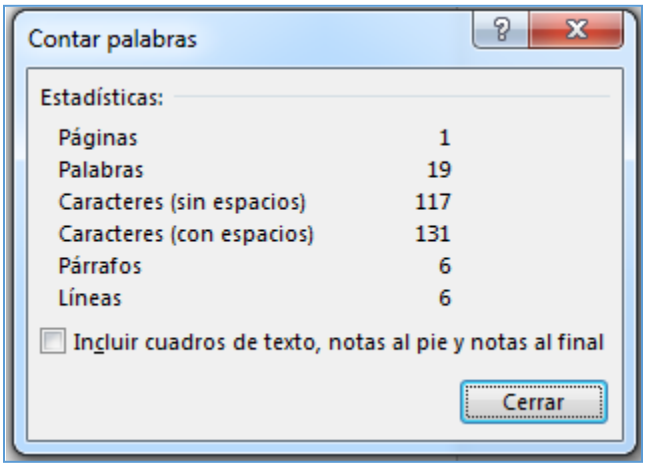


Figura 36: Contador de Palabras 1

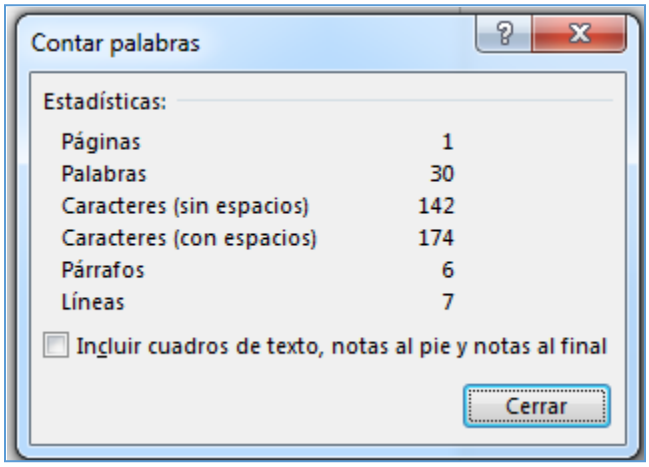


Figura 37: Contador de Palabras 2

Con estos datos podemos establecer que si bien el OCR no es perfecto, ya que aún le falta entrenamiento a las redes neuronales que se encargan del reconocimiento de patrones, tenemos la certeza de decir que:

El reconocimiento individual de las palabras aumenta un 57,89 %.

El reconocimiento de caracteres sin espaciado aumenta un 21,36%.

El reconocimiento de caracteres con espaciado aumenta un 32,82%.

Lo ideal es que estas diferencias fueran lo más cercano al 0%.

Con el fin de establecer una comparativa más completa y demostrar las bondades de las plantillas hemos utilizado Capture2text, un OCAR gratuito encontrado en línea en la página http://capture2text.sourceforge.net/, el cual presentamos en la *Figura 38*.

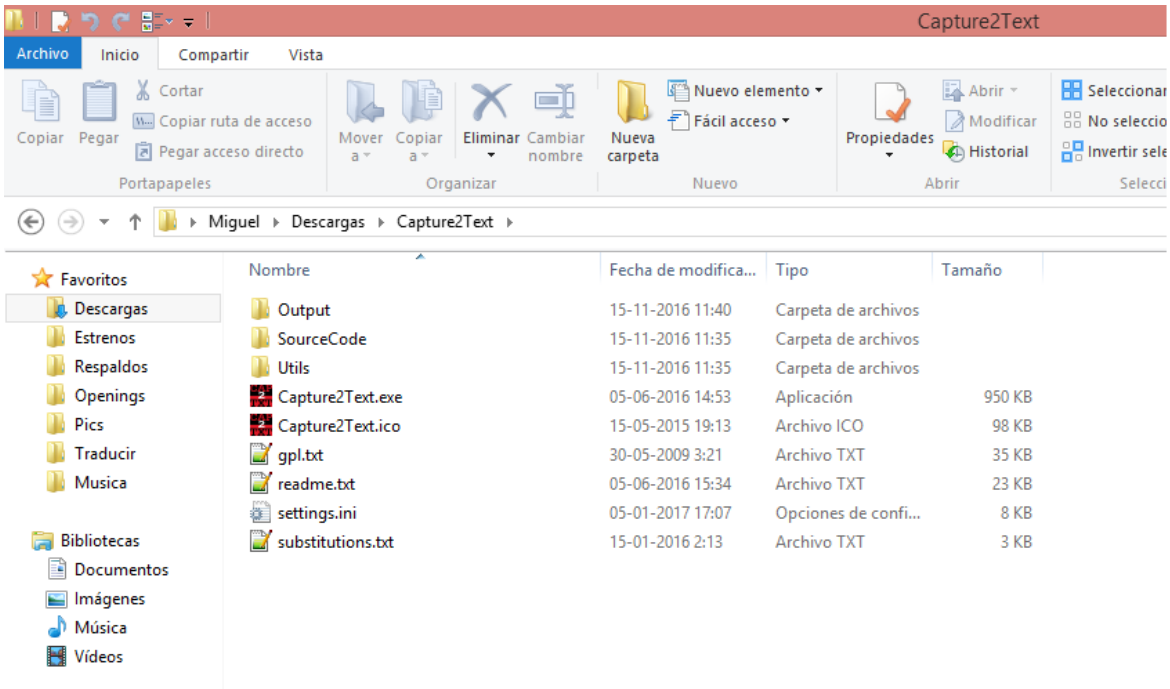


Figura 38: Capture2text

Para hacerlo funcionar basta con descargar el archivo comprimido de la página y descomprimirlo. Como se puede ver en la *Figura 38*, al descomprimir el archivo, basta con ejecutar el archivo ejecutable y presionar la tecla Windows + Q para hacer la selección de un área a analizar. Se probó como ejemplo el mismo cuadro definido en la *Figura 27* y se refleja el resultado obtenido en la *Figura 39*.

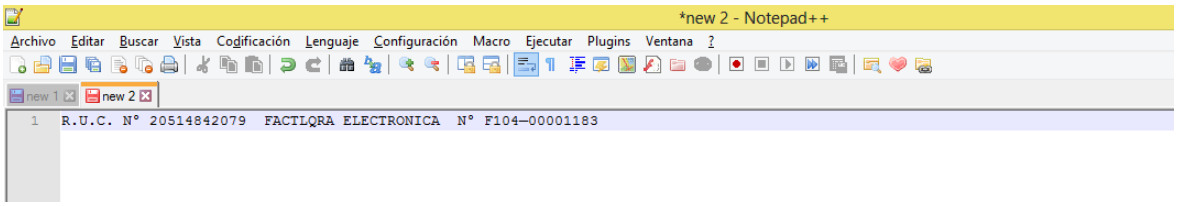


Figura 39: Resultado de Capture2text

Para probar la efectividad de las plantillas decidimos comparar escaneando el texto en el mismo documento de prueba utilizando solo el OCR para analizar la página completa, los resultados pueden apreciarse en la *Figura 40*. A diferencia de los ejemplos anteriores en donde se ve claramente el texto original, lo que se muestra ahora es bastante diferente a la imagen real, confundiendo una gran cantidad de caracteres que no corresponden y agregando un sinnúmero de errores y caracteres extraños.

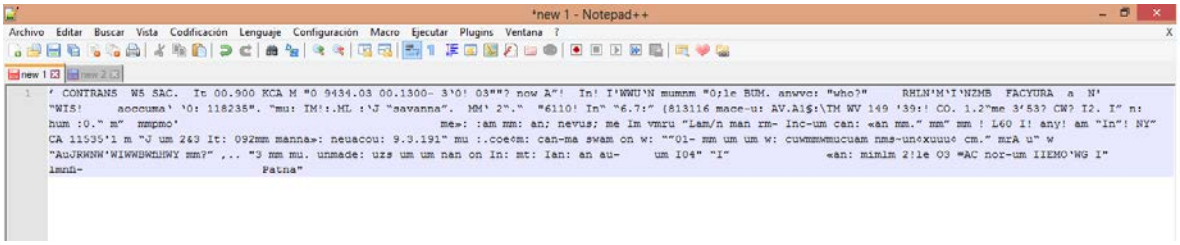


Figura 40: Resultado de lectura completa sin plantillas

Conclusiones

A lo largo del presente trabajo tuvimos la oportunidad de estudiar y comprender diversas técnicas presentes en la informática relacionadas con los OCR, su funcionamiento, implementación, y cómo estas se encuentran presentes en las tecnologías utilizadas en el día a día, la forma en que su aplicación ayuda a la realización de diversas tareas y, lo más importante, logramos proponer un método para mejorar su ejecución, como es la aplicación de plantillas.

A través del desarrollo del informe de Proyecto de Titulo, se logró observar que con el uso de plantillas que delimiten el área de proceso, la efectividad de los OCR incrementa considerablemente al eliminar procesos innecesarios y potenciales errores que estropearán las capacidades de reconocimiento del programa.

Cabe destacar que la presente tecnología era desconocida para los autores de este trabajo antes de iniciar su proceso de elaboración como Proyecto de Titulo, por lo que su abordaje en sí representó una innovación y un importante desafío que nos llenó de motivación para adquirir nuevos conocimientos en orden a lograr completar este proceso formativo a través de la investigación. Al estudiar los distintos casos en los cuales se ha aplicado esta tecnología y ver el constante avance en la digitalización de la información junto con la necesidad de documentar todos los registros posibles, se puede concluir que se garantiza la proliferación de este tipo de tecnologías.

En trabajos futuros sería interesante investigar la forma de mejorar la efectividad del OCR para mejorar su capacidad de reconocimiento, así como expandir aún más las funcionalidades para casos más específicos de uso de las plantillas para facilitar la interacción con el usuario.

Glosario

- Android: “Sistema operativo orientado a dispositivos móviles, basado en una versión modificada del núcleo Linux. Inicialmente fue desarrollado por Android Inc., una pequeña empresa, que posteriormente fue comprada por Google; en la actualidad lo desarrollan los miembros de la Open Handset Alliance”(Orozco, 2011)
- Apache: “Servidor Web más utilizado, líder con el mayor número de instalaciones a nivel mundial muy por delante de otras soluciones como el IIS (Internet Information Server) de Microsoft. Apache es un proyecto de código abierto y uso gratuito, multiplataforma, muy robusto y que destaca por su seguridad y rendimiento” (Cabello, 2012)
- Aplicación: “Programa informático que permite a un usuario utilizar una computadora con un fin específico. Las aplicaciones son parte del software de una computadora, y suelen ejecutarse sobre el sistema operativo” (Alegsa, 2010)
- Áreas de Interés (ROI): “Área o región de una imagen, delimitada por un contorno, sobre la que se evalúa un determinado parámetro.” (Barraza, 2012)
- C++: “Lenguaje imperativo orientado a objetos derivado del C. En realidad un superconjunto de C, que nació para añadirle cualidades y características de las que carecía. El resultado es que como su ancestro, sigue muy ligado al hardware subyacente, manteniendo una considerable potencia para programación a bajo nivel, pero se la han añadido elementos que le permiten también un estilo de programación con alto nivel de abstracción.” (Millán, 2013)
- Cámara de Burbujas: “Dispositivo para detectar partículas subatómicas. Fue inventada en 1952 por el físico americano Donald A. Glaser, quien en 1960, recibió por ello el Premio Nóbel de Física.” (Asimov, 2013)
- Captcha: ”Test o prueba para impedir o al menos dificultar que servicios automatizados puedan utilizar determinados recursos web y por tanto confirmar que, quien trata de hacer uso de ese servicio es un usuario humano.” (Malaga, 2011)
- Dispositivo: “Familia de teléfonos móviles que disponen de un hardware y un sistema operativo propio capaz de realizar tareas y funciones similares a las realizadas por los ordenadores fijos o portátiles, añadiéndole al teléfono

funcionalidades extras a la realización y recepción de llamadas y mensajes telefónicos.” (Ibañez, 2011)

- Distancia Euclidiana: “La fórmula euclidiana hace referencia a la distancia entre dos puntos en el plano con coordenadas (x, y) y (a, b) dada por la fórmula $\sqrt{[(x - a)^2 + (y - b)^2]}$”(Bogomolny, 2014)
- Imagen RGB: “Las imágenes RGB utilizan tres colores para reproducir en pantalla hasta 16,7 millones de colores. RGB (Siglas para Red, Green y Blue) es el modo por defecto para las imágenes de Photoshop y, por lo general, el modo en el que vienen nuestras cámaras de fotos aunque ambos perfiles pueden cambiarse.” (Nostra, 2014)
- Linux: “Se compone del núcleo importante del sistema operativo, es decir, es la parte que interactúa entre el hardware y los programas que utiliza el usuario, con las herramientas que se usan para que la interacción entre software y hardware sea la correcta.” (Segura, 2014)
- Open Source: “Movimiento que hace referencia a software distribuido que ha sido desarrollado libremente. Se enfoca en beneficios prácticos como acceso al código fuente, lo que ayuda a incrementar la calidad técnica del producto final.” (Arcos, 2014)
- Optofono: “Positivo de sonido en el cual se utiliza una célula fotoeléctrica para escanear un texto, entonces produce señales eléctricas que se convierten en sonido audible.” (Osborne, 1990)
- Transformada de Hough: “Es una técnica la cual puede ser usada para aislar funciones de una particular forma dentro de una imagen. La forma clásica es usada generalmente para la detección de curvas regulares como líneas, círculos, elipses, etc” (Fisher, Perkins, Walker, & Wolfart, 2003)
- Unix: “Sistema operativo que ofrece facilidades para la creación de programas y sistemas y el ambiente adecuado para las tareas de diseños de software.” (Benites et al., 1998)

Bibliografía

Alegsa, L. (2010). Leandro Alegsa. Retrieved from http://www.alegsa.com.ar/Dic/aplicacion.php

Arcos, E. (2014). Qué es el Open Source, explicado con Legos. Retrieved from http://hipertextual.com/archivo/imagen-del-dia/que-es-open-source-lego/

Asimov, I. (2013). Cien preguntas básicas sobre la ciencia. Retrieved from http://www.fisicanet.com.ar/monografias/monograficos5/es103_como_funciona_una_camara_de_burbujas.php

Barraza, D. (2012). Areas de Interés (ROI). Retrieved from http://www.dblaboratorios.com/index.php?option=com_content&view=article&id=1726:area-de-interes-roi&catid=46:a&Itemid=104

Benites, R., Moreyra, D., Patiño, R., Pintado, A., & Poma, A. (1998). Lynux y Unix. Retrieved from http://html.rincondelvago.com/linux-y-unix.html

Bogomolny, A. (2014). The Distance Formula from Interactive Mathematics Miscellany and Puzzles. Retrieved from http://www.cut-the-knot.org/pythagoras/DistanceFormula.shtml#

Cabello, M. (2012). ¿Qué hace un Servidor Web como Apache? Retrieved from http://www.digitallearning.es/blog/apache-servidor-web-configuracion-apache2-conf/

Campillo, I. (2010). *Sistema de Gestión Integral de Documentos de archivo para empresas de la construcción del territorio de Camagüey*. Retrieved from http://hera.ugr.es/tesisugr/19562226.pdf

Cisneros, H. (2007). Conversión de texto manuscrito a formato digital utilizando máquinas de soporte vectorial, 1–109.

Fisher, R., Perkins, S., Walker, A., & Wolfart, E. (2003). Hough Transform. Retrieved from http://homepages.inf.ed.ac.uk/rbf/HIPR2/hough.htm

Heckerman, D. (1995). *A Tutorial on Learning With Bayesian Networks*. Retrieved from https://www.microsoft.com/en-us/research/publication/a-tutorial-on-learning-with-bayesian-networks/

Ibañez, D. (2011). Smartphone - Explicación y Definición de smartphone. Retrieved from http://www.quees.info/que-es-un-smartphone.html

Malaga, D. (2011). Qué es un Captcha, como rellenarlo. Retrieved from https://basicoyfacil.wordpress.com/2011/11/17/que-es-un-captcha-como-rellenarlo/

Melorose, J., Perroy, R., & Careas, S. (2015). Reglamento general de archivos. *Statewide Agricultural Land Use Baseline 2015*, *1*. http://doi.org/10.1017/CBO9781107415324.004

Millán, A. (2013). Curso de C++. Retrieved from http://www.zator.com/Cpp/E1_2.htm

Montes, V. (2014). App móvil para reconocer texto en imágenes, 1–11.

Nielsen, M. (2016). Deep Learning.

Nostra. (2014). Modos de color: RGB. CMYK y sRGB. Retrieved from http://www.fotonostra.com/grafico/rgb.htm

Orozco, D. (2011). Definición de Android. Retrieved from http://conceptodefinicion.de/android/

Osborne. (1990). Artefactos Sonoros. Retrieved from http://www.aatespanol.cl/taa/tesauro/default.asp?a=338&Element_ID=31804

Segura, S. (2014). ¿Qué es Linux? Explicación sencilla para neófitos. Retrieved from http://noticias.mountain.es/que-es-linux-explicacion-sencilla-para-neofitos/

Valenzuela, L., & Báez, G. (2007). Reconocimiento de Patrones . Una propuesta de Corrección Automática de Test de Selección Múltiple.