UNIVERSIDAD LUTERANA SALVADOREÑA.

FACULTAD DE CIENCIAS DEL HOMBRE Y LA NATURALEZA.

LICENCIATURA EN CIENCIAS DE LA COMPUTACIÓN.

Materia: Introducción al software libre

Docente: Lic. José Luis Alvarado

Tema Desarrollado: OCR para escanear textos de imágenes.

Integrantes: Yasmin Lorena Rivas Iraeta
Emerson Elenilson Martínes Maravilla
Misael Antonio Mejia Bonilla
Karina Vanessa Molina Servellón

Fecha de Entrega: Sábado 2 de diciembre de 2017

Indice

INTRODUCCIÓN

Utilización de nuevos recursos en software libre (linux), es la nueva moda, ya que se debe a la innovación de herramientas para uso cotidiano, de fácil instalación y uso, en este caso le debemos el reconocimiento al software tesseract(en terminal) y ocrfeeder(de forma gráfica), nos permite escanear una foto o img o de solo texto, que necesitemos obtenerlo en texto, esto básicamente nos facilita el trabajo a la hora de escannear ya que no es necesario tener un scanner en casa, solo dice en la terminal el comando a utilizar y la ubicación exacta de la imagen a escanear o si desea de forma gráfica eligiendo la img y las opciones necesarias.

ANTECEDENTES

El reconocimiento óptico de caracteres (ROC), generalmente conocido como reconocimiento de caracteres y expresado con frecuencia con la sigla OCR (del inglés Optical Character Recognition), es un proceso dirigido a la digitalización de textos, los cuales identifican automáticamente a partir de una imagen símbolos o caracteres que pertenecen a un determinado alfabeto, para luego almacenarlos en forma de datos. Así podremos interactuar con estos mediante un programa de edición de texto o similar.
En los últimos años la digitalización de la información (textos, imágenes, sonido, etcétera) ha devenido un punto de interés para la sociedad. En el caso concreto de los textos, existen y se generan continuamente grandes cantidades de información escrita, tipográfica o manuscrita en todo tipo de soportes. En este contexto, poder automatizar la introducción de caracteres evitando la entrada por teclado implica un importante ahorro de recursos humanos y un aumento de la productividad, al mismo tiempo que se mantiene, o hasta se mejora, la calidad de muchos servicios.

Tesseract es un programa de computadora gratuito para reconocimiento óptico de caracteres . Originalmente fue desarrollado entre 1985 y 1995 bajo licencia de Hewlett-Packard . Después de diez años sin ningún desarrollo, Hewlett Packard y la Universidad de Nevada (Las Vegas) lo lanzaron en 2005 como fuente abierta. Tesseract ahora es desarrollado por Google y lanzado bajo la licencia Apache 2.0.

Tesseract se considera una de las máquinas de software OCR gratuitas más precisas disponibles en la actualidad.
Acerca de la máquina Tesseract OCR
Tesseract es una máquina de OCR desnuda. No tiene ningún análisis de formato de documento, formato de salida ni entorno de usuario gráfico . El único formato que puede manejar es una imagen TIFF de una sola columna de texto, desde donde produce el texto. La compresión TIFF no es compatible a menos que se instale libtiff . Puede detectar si una fuente es proporcional o no. La máquina estaba en el top 3 en 1995 en términos de precisión de caracteres. Se puede compilar y ejecutar en Linux, MS Windows y Mac OS X, sin embargo, debido a los recursos limitados, los desarrolladores solo han probado cuidadosamente MS Windows y Ubuntu Linux.
Tesseract puede procesar inglés, francés, italiano, alemán, español, portugués de Brasil y holandés, y se puede practicar para reconocer otros idiomas.
Tesseract es adecuado para usar como un programa en segundo plano, y se puede usar para realizar tareas de OCR más complicadas, incluido el análisis de diseño, en combinación con una interfaz de usuario como OCRopus . Una mayor integración con programas como OCRopus, para soportar un diseño complicado, está en el diseño.

OCRFeeder es un analizador de documentos capaz de realizar reconocimiento de caracteres ópticos del sistema es un excelente OCR para GNU/Linux.
En pasados vimos que es el OCR o reconocimiento óptico de caracteres. Además analiza

mos los usos que le podemos dar, las ventajas que nos proporciona y las limitaciones que tiene hoy en día. Sin entrar en mucho detalle también vimos que existen varios software de reconocimiento OCR libres. Para quien considere releer el post que menciono les dejo el siguiente link. En este post nos centraremos en explicar el funcionamiento de OCRfeeder que es uno de los software que en su día nombramos.

He preferido empezar por OCRfeeder ya que su funcionamiento es muy simple, requiere de pocas dependencias de instalación y además me parece una opción aceptable. Cabe decir que en GNU-Linux existen otras opciones muy aceptables e incluso me atrevería a decir de mayor calidad como por ejemplo gscan2pdf o xsane.

Objetivos

objetivo general

- ✔ El objetivo de este proyecto consiste en la implementación de un OCR, que pueda reconocer cualquier simbolo.

Objetivos especificos

- ✔ OCR debe extraer unas caracteristicas a las imagenes analizadas que permitan reconocerlas, sea cual sea el carácter que representan.
- ✔ Debe ser capaz de recordar los caracteres que ya habia analizado, para analizarlos bien en futuros reconocimientos.

Marco Teórico:

El reconocimiento óptico de caracteres (ROC), mejor conocido como reconocimiento de caracteres y expresado con frecuencia con la sigla OCR (del inglés Optical Character Recognition), es un proceso dirigido a la digitalización de textos, los cuales identifican automáticamente a partir de una imagen símbolos o caracteres que pertenecen a un determinado alfabeto, para luego almacenarlos en forma de datos. Así podremos interactuar con estos mediante un programa de edición de texto o similar.

En actualidad la demanda de digitalización de las tecnologías requieren consigo de herramientas para llevar a cabo tareas especificas en el menor tiempo posible. Ejemplo: Transcribir un libro, novelas, obras, etc.
OCR es la solución:

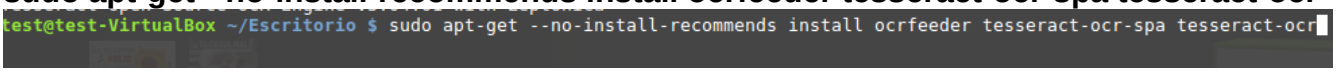
Haciendo uso de dichas tecnologías no topamos con la siguiente pregunta: ¿OCR en software libre?
La respuesta es simple. Si existe unas poderosas herramientas llamadas tesseract y ocrfeeder.

Gracias al programa teseract y al paquete que trae tesseract-ocr-spa es posible poder reconocer los caracteres españoles, y además tratar ciertas imágenes donde los niveles de color o píxeles no son los adecuados.

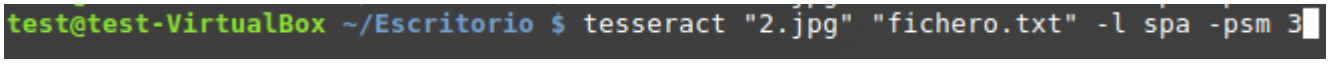
La instalación y uso de dichos paquetes es muy fácil ya que basta con ejecutar una sola linea de comando, esperar que se descargue y listo.

Instalación en Ubuntu 16

sudo apt-get --no-install-recommends install ocrfeeder tesseract-ocr-spa tesseract-ocr

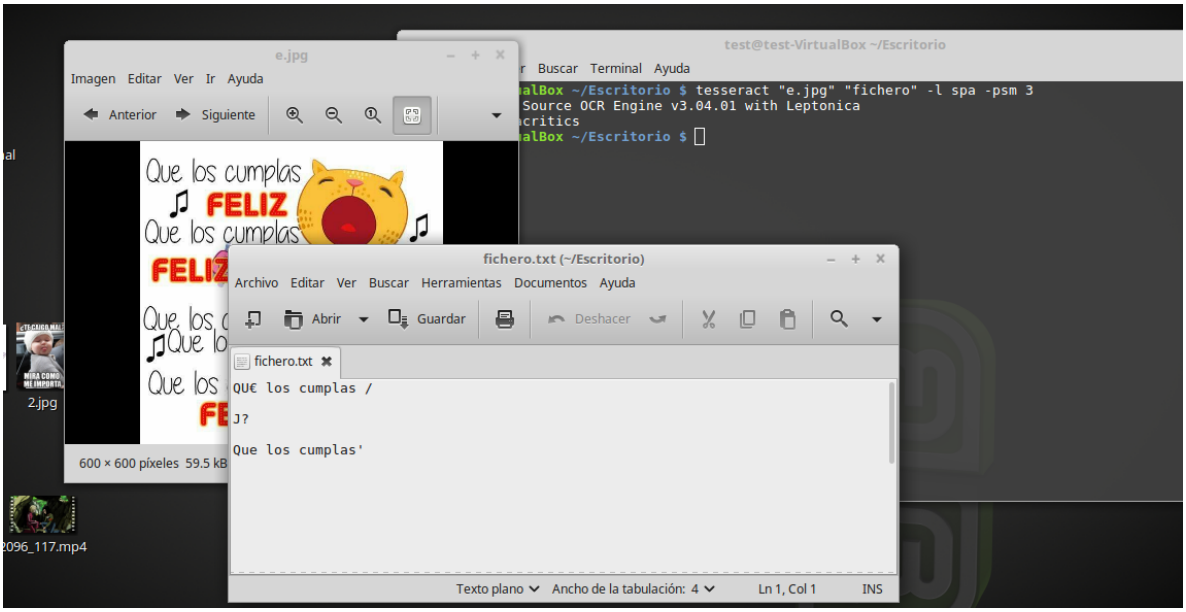


Ok. Cuando tengamos instalado el programa desde la terminal ya sea con súper usuario o usuario normal ejecutamos lo siguiente.



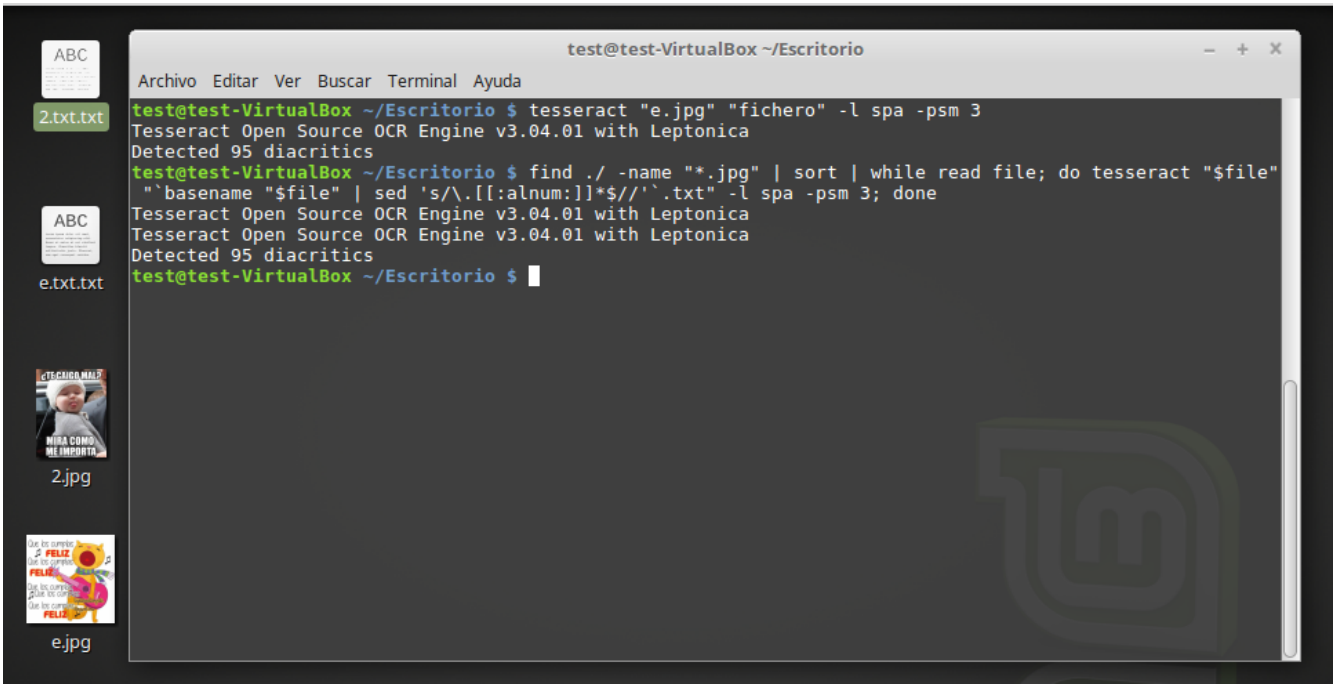
Donde “2.jpg” es la imagen que se va escanear para extraerle su texto. Y “fichero.txt” es el resultado(Aquí esta el texto de la imagen de manera texto plano).

Después de dicho proceso obtenemos los siguientes resultados.



De esa manera parte del texto de la imagen se ha convertido a texto plano en un archivo.

Múltiples imágenes por OCR



Muy importante cuando se debe de transcribir mucho y se esta buscando una solución.

Nota: El correcto funcionamiento de esta herramienta es un poco limitada, ya que mucho de este texto es de difícil lectura. Por falta de calidad o nitidez.

Conclusiones

- ✓ En conclusión, queremos decir que OCRFeeder es de vital importancia para todas las acciones que tienen que ver con el escaneo facil y rápido de imagenes.

- ✓ Concluimos que gracias a la aplicación de OCRFeeder los usuarios tienen una mayor rapidez al realizar trabajos y se ahorran mucho tiempo y recursos al convertir una imagen a texto.

Recomendaciones

- ✓ Continuar con el descubrimiento y propagación de las herramientas de trabajo innovadoras y de muy buena calidad como en este caso es OCRFeeder.

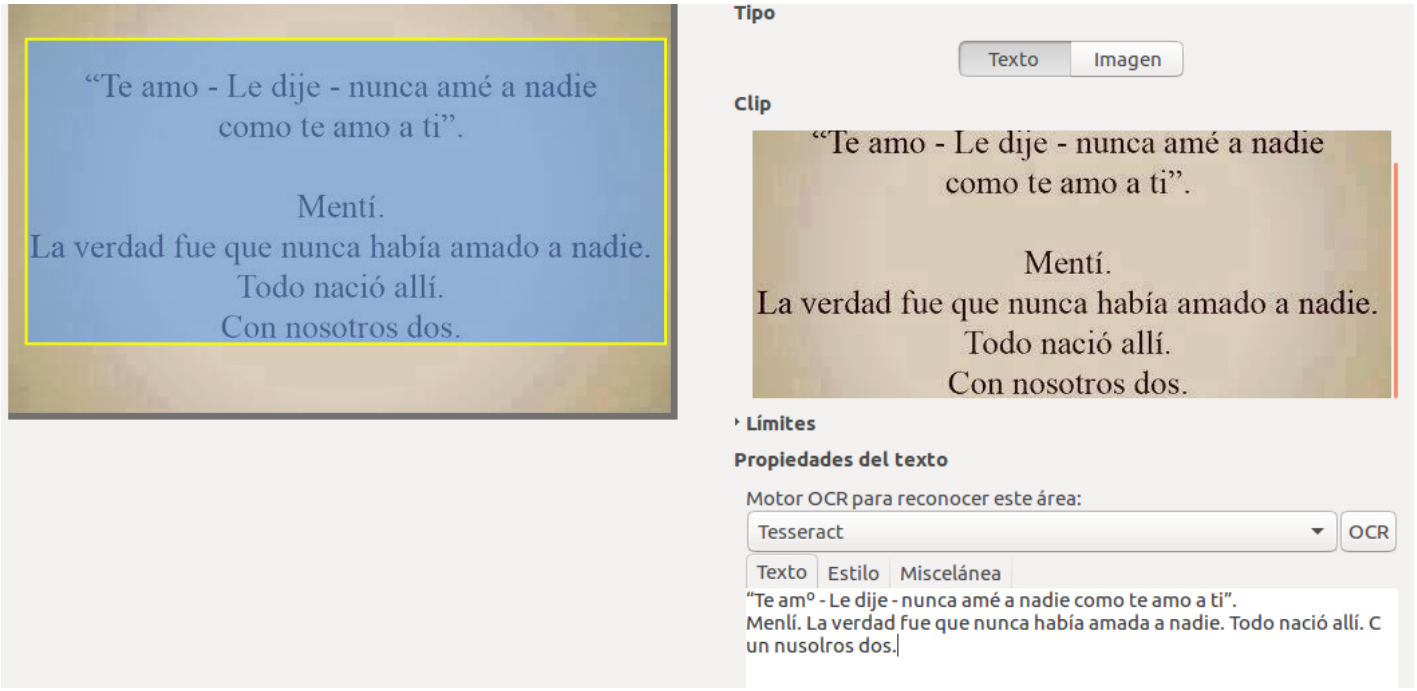
- ✓ Continuar con la misma misión de darlo todo por lograr ser independientes y tener herramientas en las cuales sabemos que es lo que estamos utilizando.

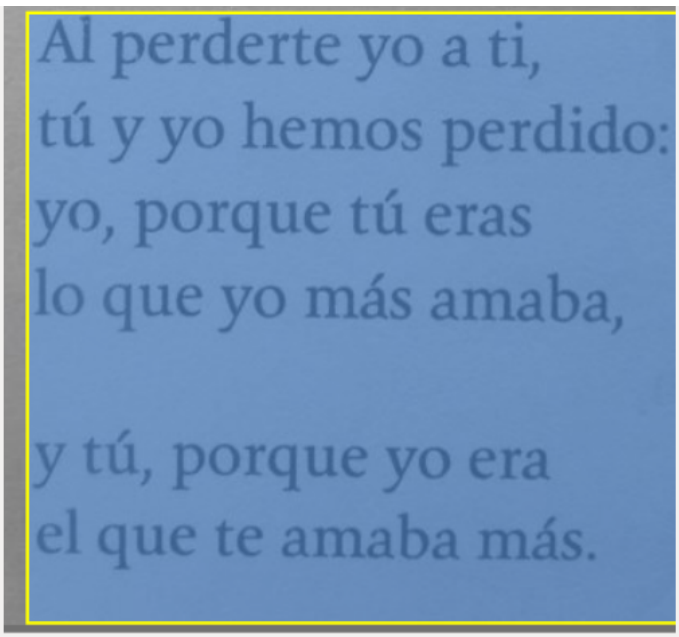
- ✓ Seguir gozando de esta herramienta de trabajo y recomendarla a personas que es una excelente aplicación y que puede serles de gran ayuda.

Blibliografia

- ✔ http://www.kacharreando.com/ubuntu/ocr-linux/
- ✔ https://www.youtube.com/watch?v=GETE8KfkbKA
- ✔ https://mislinuxapps.wordpress.com/2011/02/12/escanear-y-ocr-en-ubuntu/

Anexos

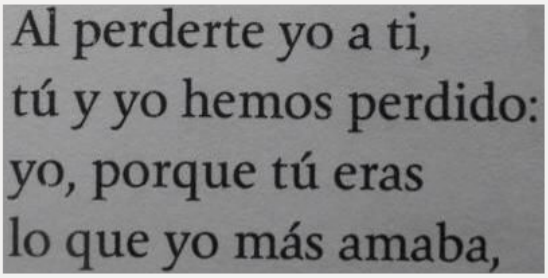

“Te amo - Le dije - nunca amé a nadie como te amo a ti”.
Mentí.
La verdad fue que nunca había amado a nadie.
Todo nació allí.
Con nosotros dos.
Tipo
Texto
Imagen
Clip
“Te amo - Le dije - nunca amé a nadie como te amo a ti”.
Mentí.
La verdad fue que nunca había amado a nadie.
Todo nació allí.
Con nosotros dos.
Límites
Propiedades del texto
Motor OCR para reconocer este área:
Tesseract
OCR
Texto
Estilo
Miscelánea
“Te amº - Le dije - nunca amé a nadie como te amo a ti”.
Menlí. La verdad fue que nunca había amada a nadie. Todo nació allí. C
un nusolros dos.



Tipo

Texto | Imagen

Clip



Límites

Propiedades del texto

Motor OCR para reconocer este área:

Tesseract | OCR

Texto | Estilo | Miscelánea

Al perderte yo a ti,
tú y yo hemos perdido: yo, porque tú eras
lo que yo más amaba,
y tú, porque yo era el que te amaba más.