

Introducción.

El presente material tiene como fin incrementar el porcentaje de eficacia al digitalizar textos mediante OCR, para ello se presentan algunas de las problemáticas más comunes y métodos sugeridos para reducirlas. Estas técnicas se pueden aplicar a una fuente de texto digitalizada ya sea mediante un Escaner o una Cámara digital, sin embargo, se detallará más ampliamente el proceso a seguir cuando la fuente es una fotografía tomada con cámara digital ya que puede presentar más problemas sino se toman los cuidados necesarios al preparar el equipo para la “captura”.

A pesar de los inconvenientes que pueden existir una cámara digital es una muy buena herramienta para digitalizar texto a demás de actualmente es un bien que muy probablemente ya este al alcance.

Este texto no pretende informar de manera exhaustiva sobre el proceso de OCR en si, sino mas bien enfocarse al aspecto practico. A continuación una introducción al OCR tomada de Wikipedia.

El Reconocimiento Óptico de Caracteres (OCR), así como el reconocimiento de texto, en general son aplicaciones dirigidas a la digitalización de textos. Identifican automáticamente símbolos o caracteres que pertenecen a un determinado alfabeto, a partir de una imagen para almacenarla en forma de datos con los que podremos interactuar mediante un programa de edición de texto o similar.

En los últimos años la digitalización de la información (textos, imágenes, sonido, etc) 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.

Todos los recursos de software utilizados son de acceso libre y funcionan en Linux, Windows o Mac por lo que están al alcance de cualquier persona y no requieren de una inversión económica.

Aplicaciones necesarias.

El proceso de digitalización del texto lo podemos dividir en captura, preparación, reconocimiento y por último edición. Es muy importante prestar atención al proceso de captura ya que repercutirá en los procesos siguientes y logrando así un gran porcentaje de eficacia o un resultado casi nulo.

GIMP

GIMP (GNU Image Manipulation Program) es un programa de edición de imágenes digitales en forma de mapa de bits, tanto dibujos como fotografías. Es un programa libre y gratuito. Forma parte del proyecto GNU y está disponible bajo la Licencia pública general de GNU.

Esta aplicación se utilizara para corregir problemas con las fotografías o imágenes del escaner.

Tesseract

Tesseract es un motor OCR libre. Fue desarrollado originalmente por Hewlett Packard como software propietario entre 1985 y 1995. Tras diez años sin ningún desarrollo, fue liberado como código abierto en el año 2005 por Hewlett Packard y la Universidad de Nevada, Las Vegas. Tesseract es desarrollado actualmente por Google y distribuido bajo la licencia Apache, versión 2.0.

Tesseract está considerado como uno de los motores OCR libres con mayor precisión disponibles actualmente.

Esta aplicación es la encargada de hacer el OCR en si, no se interactuará directamente con ella sino que utilizaremos un *front end* para acceder a sus capacidades, ya que funciona con línea de comandos.

gImageReader

gImageReader es una interfaz gráfica para Tesseract que nos permite cargar imágenes y seleccionar regiones específicas que después envía a Tesseract para su OCR.

Writer

Writer es un procesador de texto que forma parte de el paquete informático LibreOffice. Es similar a Microsoft Word y WordPerfect.

Enlaces

Gimp <http://www.gimp.org/downloads/>

Tesseract <http://code.google.com/p/tesseract-ocr/downloads/list>

gImageReader <http://sourceforge.net/projects/gimagereader/files/>

Writer <http://es.libreoffice.org/descarga/>

Recomendaciones captura del material fuente.

Como ya mencionamos hacer una buena captura del material original nos permitira llevar acabo el proceso de OCR de una manera mas eficaz, si utilizamos un escaner, existen menos variables a tomar en cuenta.

Para ello bastara con configurar la calidad del escaner al nivel mas elevado posible y presionar el material de manera que se eviten curvaturas por la perspectiva.

Ajustes para cámara digital.

Fuente de luz.

A diferencia del escaner que tiene una fuente de luz automática cuando utilizemos cámara digital es necesario buscar una fuente de luz, una forma muy sencilla de conseguirla es hacer la captura evitando espacios con luz artificial y realizarla en el exterior utilizando la luz del sol, no es necesario que esta pegue directamente puede ser en una cochera, o un espacio sin paredes.

Perspectiva.

Al tomar la fotografiá hay que cuidar que la perspectiva sea lo mas paralela y el documento este lo mas plano posible cuidando que el cuerpo del texto forme ángulos de 90 grados evitando efectos “romboides”.

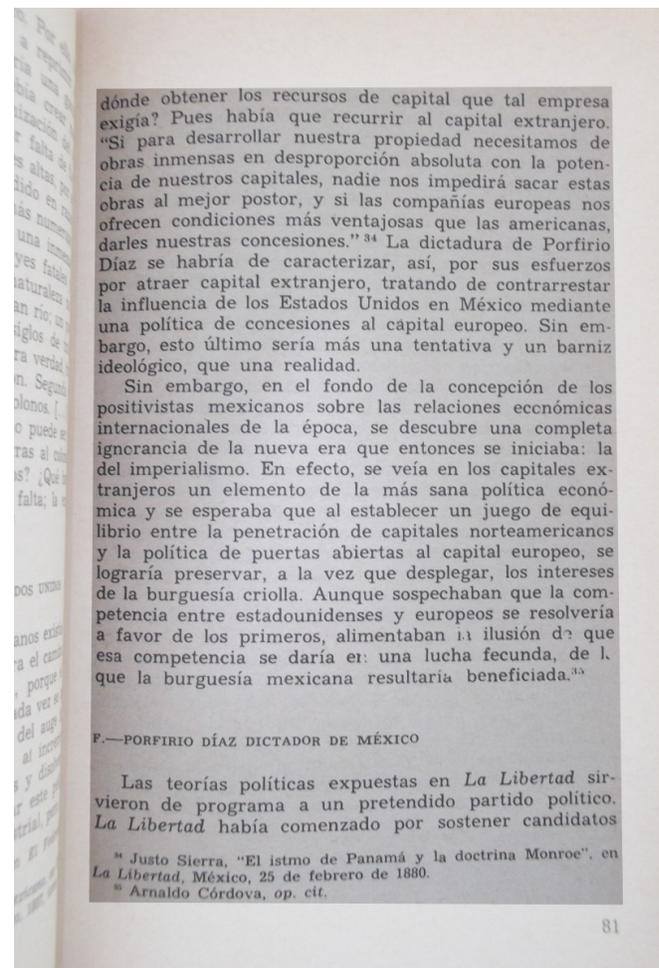
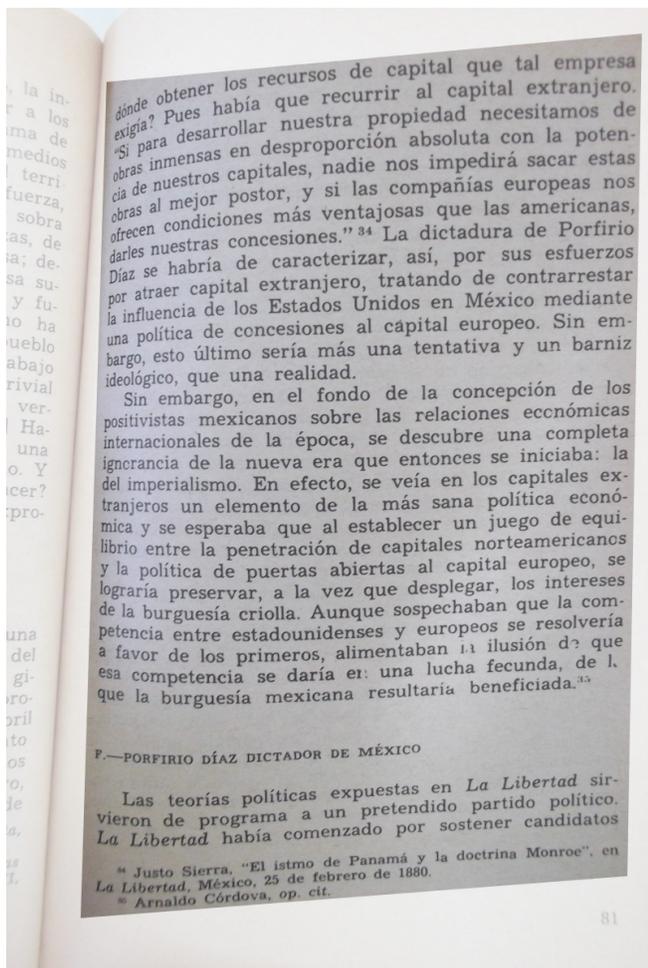
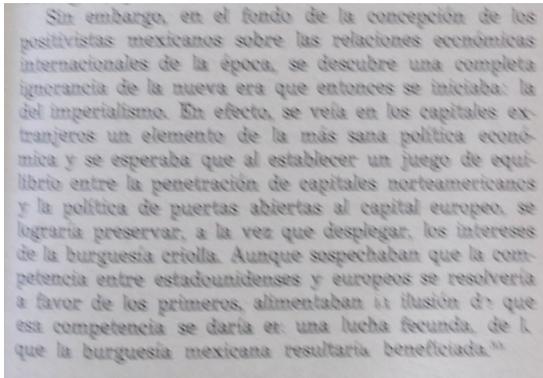
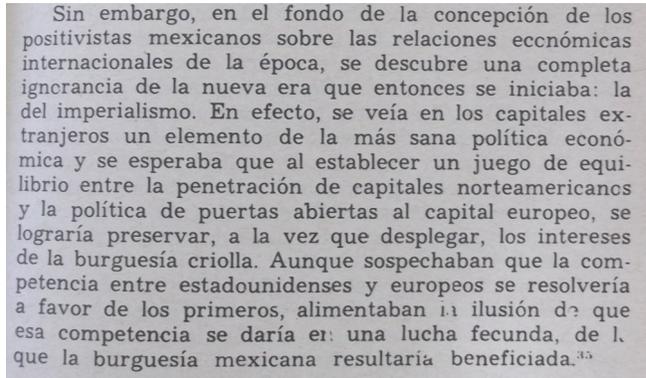


Foto en primer plano.

Debemos configurar nuestra cámara para que ajuste el foco del lente de manera que las imágenes sean nítidas evitando efectos borrosos o empañados. La configuración automática es buena y a menudo se encuentra bajo un icono de una flor.



Sin embargo, en el fondo de la concepción de los positivistas mexicanos sobre las relaciones económicas internacionales de la época, se descubre una completa ignorancia de la nueva era que entonces se iniciaba: la del imperialismo. En efecto, se veía en los capitales extranjeros un elemento de la más sana política económica y se esperaba que al establecer un juego de equilibrio entre la penetración de capitales norteamericanos y la política de puertas abiertas al capital europeo, se lograría preservar, a la vez que desplegar, los intereses de la burguesía criolla. Aunque sospechaban que la competencia entre estadounidenses y europeos se resolvería a favor de los primeros, alimentaban la ilusión de que esa competencia se daría en una lucha fecunda, de la que la burguesía mexicana resultaría beneficiada.⁴³



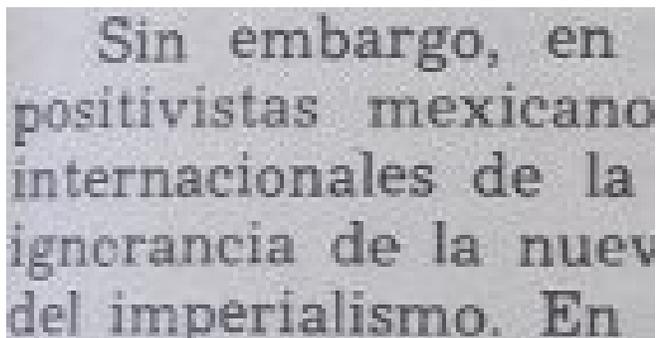
Sin embargo, en el fondo de la concepción de los positivistas mexicanos sobre las relaciones económicas internacionales de la época, se descubre una completa ignorancia de la nueva era que entonces se iniciaba: la del imperialismo. En efecto, se veía en los capitales extranjeros un elemento de la más sana política económica y se esperaba que al establecer un juego de equilibrio entre la penetración de capitales norteamericanos y la política de puertas abiertas al capital europeo, se lograría preservar, a la vez que desplegar, los intereses de la burguesía criolla. Aunque sospechaban que la competencia entre estadounidenses y europeos se resolvería a favor de los primeros, alimentaban la ilusión de que esa competencia se daría en una lucha fecunda, de la que la burguesía mexicana resultaría beneficiada.⁴³

Eliminar flash.

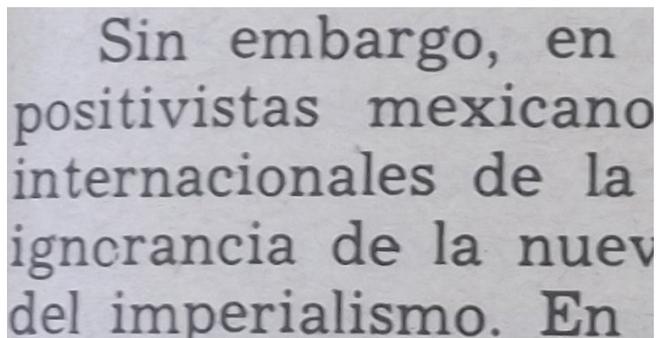
Es deseable que nuestras imágenes tengan una iluminación constante por lo que debemos eliminar el flash ya que produce efectos de reflejo y brillo que crean contrastes.

Aumentar calidad (Resolución).

Las imágenes digitales se componen de puntos llamados píxeles, para el reconocimiento OCR es necesario proporcionar la mayor cantidad posible de *píxeles que representen un carácter* por lo que aumentaremos la cantidad de píxeles o puntos por pulgada (DPI) al máximo permitido por nuestra cámara.



Sin embargo, en positivistas mexicano internacionales de la ignorancia de la nueva del imperialismo. En



Sin embargo, en positivistas mexicano internacionales de la ignorancia de la nueva del imperialismo. En

Algunas cámaras tienen un modo entrelazado el cual aumenta el número de píxeles pero de manera artificial, no debemos utilizar este modo sino la máxima calidad “natural”. Tampoco debemos utilizar zoom digital.

Preparación del material para mejorar la eficacia del software OCR.

Una vez que hemos capturado el documento en imágenes digitales será necesario retocarlas. A diferencia de los seres humanos las aplicaciones informáticas tienen gran dificultad para identificar el texto en una imagen por lo que debemos preparar las imágenes para evitar confundirlas.

Para ello debemos hacer dos cosas, una cuidar la forma de los caracteres o símbolos esto lo lograremos corrigiendo la perspectiva y rotación. y dos eliminar todo el ruido, es decir lo que no representa un carácter, como mencionamos en el apartado de resolución es necesario filtrar los *pixeles que representan a un carácter*.

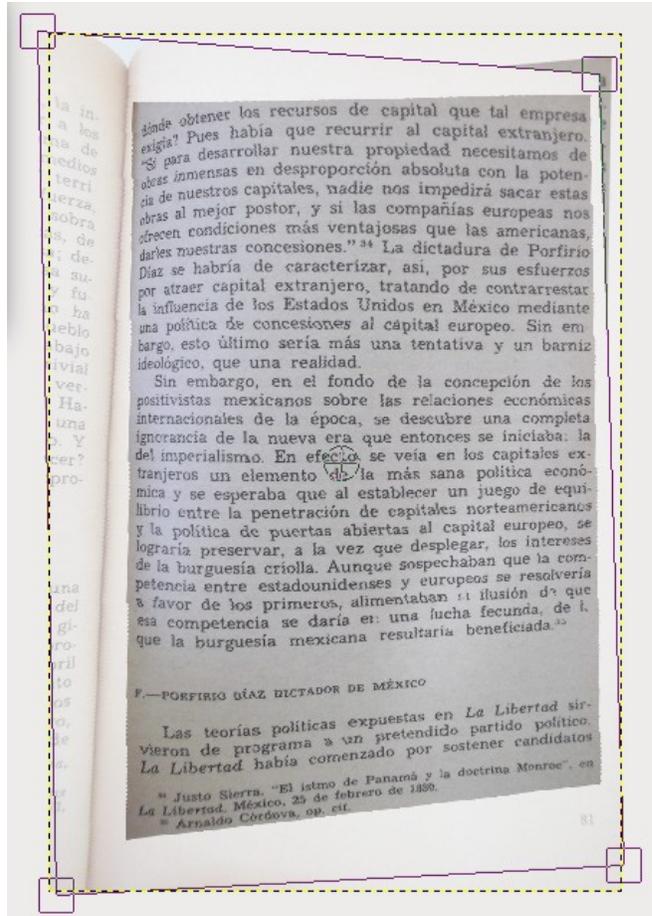
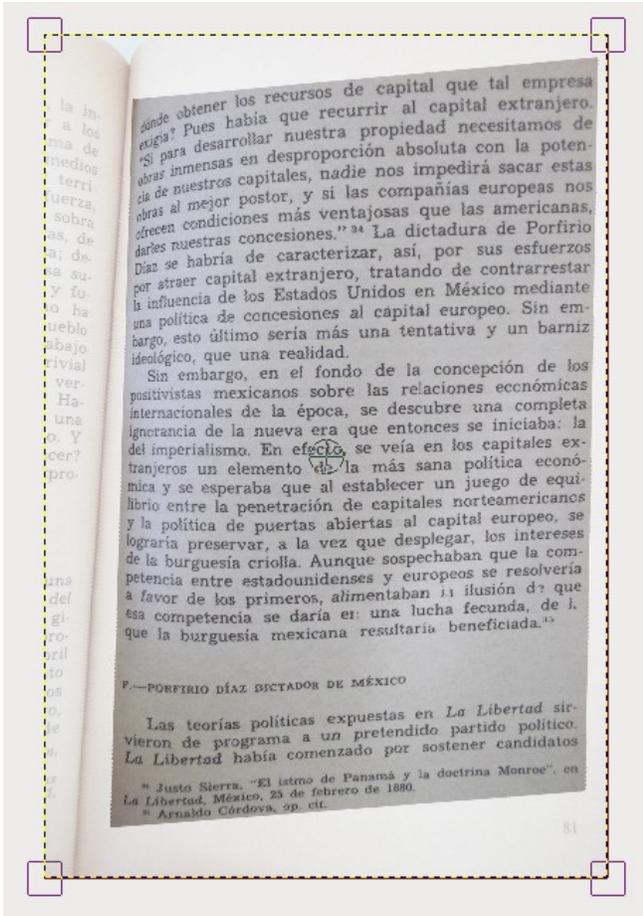
Las herramientas que se utilizaran a menudo muestran controles como deslizadores, cajas de texto o bloques que manipularemos con el ratón o ingresando valores directamente, estos ajustes los haremos utilizando el criterio propio y en relación a la calidad de la fuente.

Perspectiva.



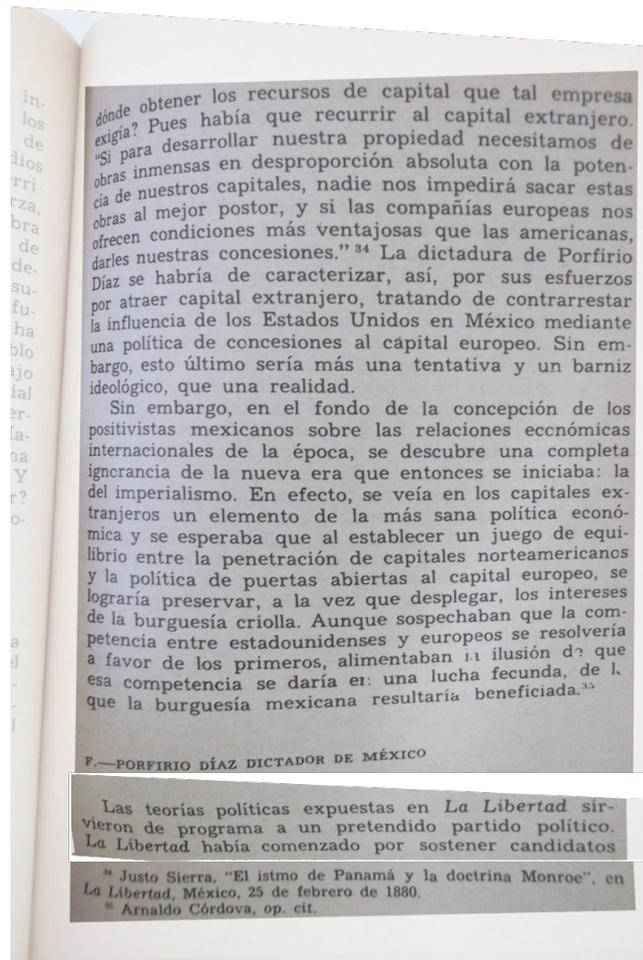
Utilizando la herramienta de perspectiva se puede corregir el efecto romboide que pudiesen tener nuestras imágenes, sin embargo si al momento de hacer la captura el material además no estaba plano aparecerá un efecto curvo en los bordes, este no es tan fácil de corregir.

Para corregir el efecto romboide seleccionamos la herramienta perspectiva, después hacemos clic en las esquinas de la imagen para hacer la corrección y sin soltar arrastramos la esquina hacia el centro o hacia afuera según la corrección que se desee dar.



Es posible que algunas áreas tengan un mayor grado de defecto que otras para ello podemos seleccionar una área específica con la herramienta de selección y aplicar el efecto solamente a esta parte. Como observamos en la imagen la parte de abajo requirió que se trataran por separado los últimos dos párrafos.

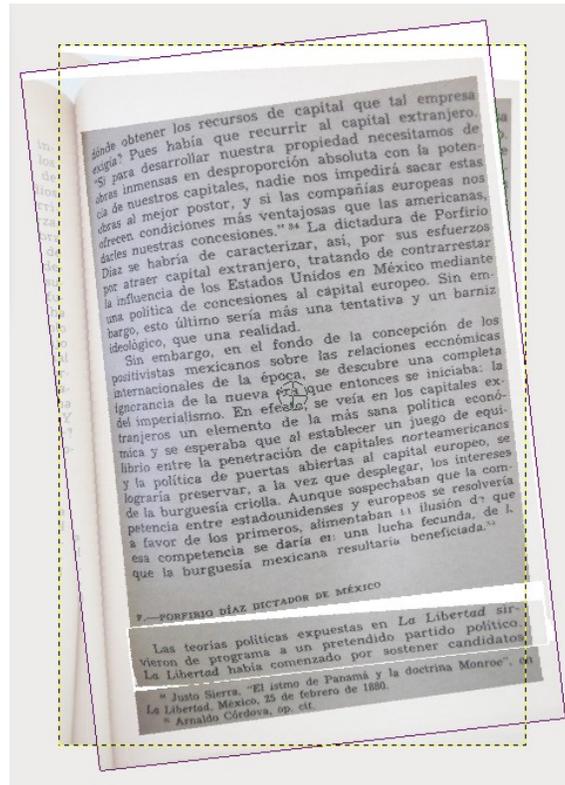
Como vemos existe cierto efecto curvo en las líneas de texto, esto puede ocasionar que se mezclen las líneas una vez que se realice el OCR pero no siempre es el caso.





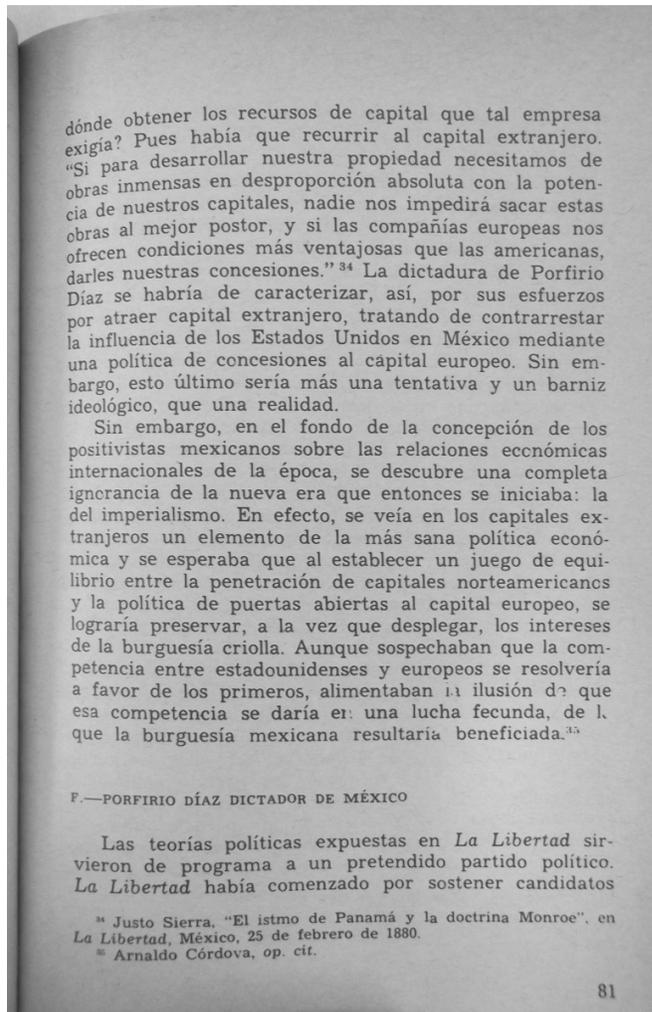
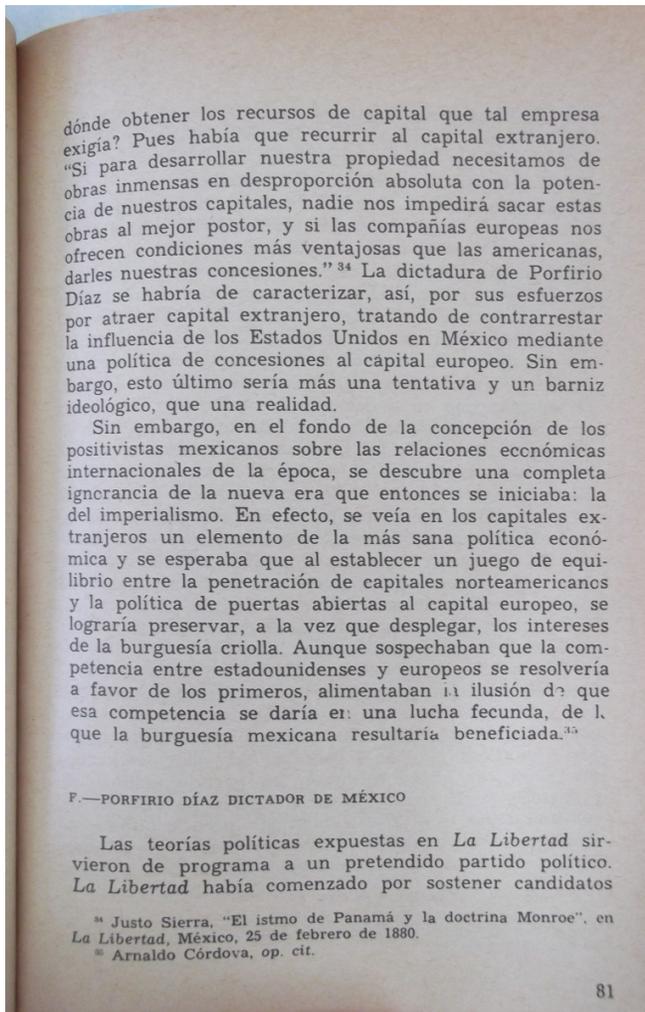
Rotación.

Al tomar las imágenes puede ser que estas estén ligeramente inclinadas, podemos corregir el ángulo, para ello podemos utilizar la herramienta de rotación y ajustarla para que el cuerpo de texto este paralelo respecto de nuestro monitor.



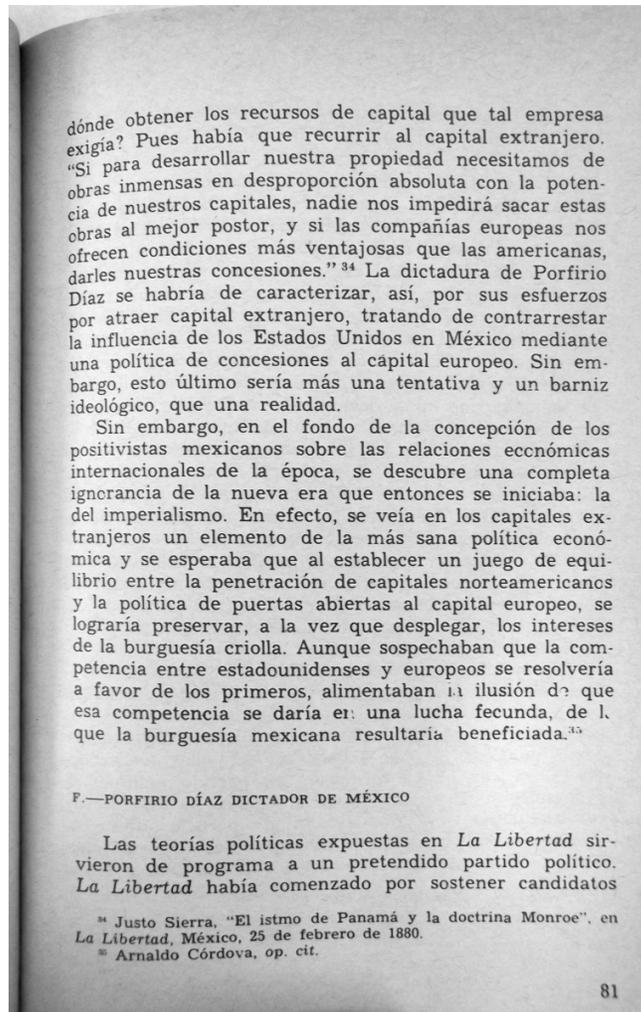
Escala de grises.

Una vez que hemos ajustado la forma y posición del cuerpo de texto, lo convertiremos en escala de grises. Para ellos nos dirigimos al menú Imagen – Modo y hacemos clic en escala de grises.



Contraste y brillo.

Ahora que tenemos una imagen en escala de grises aumentaremos el contraste para realzar la diferencia entre los caracteres y el fondo. Para ello vamos al menú Color – Brillo y contraste.



Umbral.

En ocasiones la textura del papel y la fuente de luz pueden ocasionar que los tonos de gris se mezclen con los caracteres, para ello aplicaremos un umbral al color, este tiene como objeto convertir todo lo que quede dentro del umbral en blanco y lo que quede fuera en negro.

Con este paso podemos observar como se limpian los textos que están cerca de áreas grises. Esta imagen ya podemos abrirla con gImageReader para realizar el OCR.

dónde obtener los recursos de capital que tal empresa exigía? Pues había que recurrir al capital extranjero. "Si para desarrollar nuestra propiedad necesitamos de obras inmensas en desproporción absoluta con la potencia de nuestros capitales, nadie nos impedirá sacar estas obras al mejor postor, y si las compañías europeas nos ofrecen condiciones más ventajosas que las americanas, darles nuestras concesiones."³⁴ La dictadura de Porfirio Díaz se habría de caracterizar, así, por sus esfuerzos por atraer capital extranjero, tratando de contrarrestar la influencia de los Estados Unidos en México mediante una política de concesiones al capital europeo. Sin embargo, esto último sería más una tentativa y un barniz ideológico, que una realidad.

Sin embargo, en el fondo de la concepción de los positivistas mexicanos sobre las relaciones económicas internacionales de la época, se descubre una completa ignorancia de la nueva era que entonces se iniciaba: la del imperialismo. En efecto, se veía en los capitales extranjeros un elemento de la más sana política económica y se esperaba que al establecer un juego de equilibrio entre la penetración de capitales norteamericanos y la política de puertas abiertas al capital europeo, se lograría preservar, a la vez que desplegar, los intereses de la burguesía criolla. Aunque sospechaban que la competencia entre estadounidenses y europeos se resolvería a favor de los primeros, alimentaban la ilusión de que esa competencia se daría en una lucha fecunda, de la que la burguesía mexicana resultaría beneficiada.³⁵

F.—PORFIRIO DÍAZ DICTADOR DE MÉXICO

Las teorías políticas expuestas en *La Libertad* sirvieron de programa a un pretendido partido político. *La Libertad* había comenzado por sostener candidatos

³⁴ Justo Sierra, "El istmo de Panamá y la doctrina Monroe", en *La Libertad*, México, 25 de febrero de 1880.

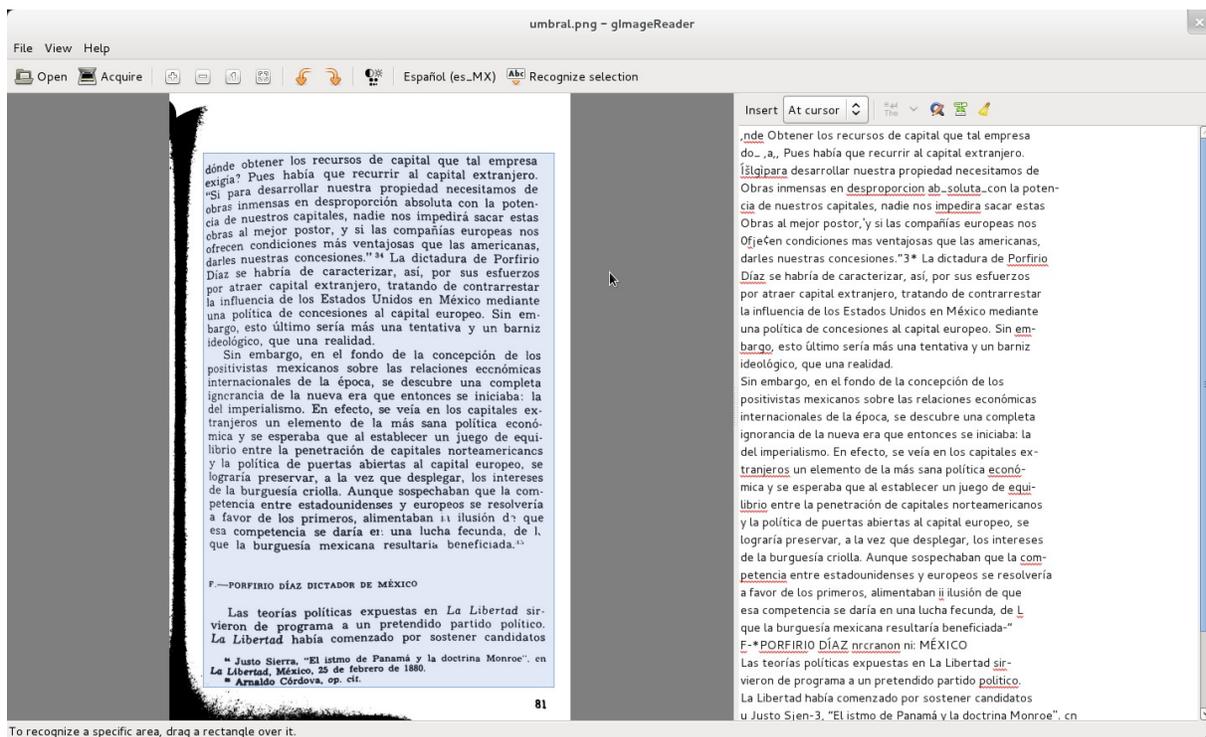
³⁵ Arnaldo Córdova, *op. cit.*

Reconocimiento del texto con OCR.

Una vez que tenemos nuestra imagen la abrimos con gImageReader, es posible hacer el OCR completo (si se eliminaron los bordes negros) o seleccionar las áreas de texto como párrafos.

Algunas veces sera mas efectivo hacer el reconocimiento completo y en otras sera necesario seleccionar por áreas, debido a diferencias en tipo y forma del carácter como por ejemplo el caso de los pies de pagina.

A continuación se presenta dos resultados de OCR uno seleccionando todo el cuerpo de texto y otro en el que se realizo por párrafos.



To recognize a specific area, drag a rectangle over it.

OCR completo	OCR por áreas.
<p>nde Obtener los recursos de capital que tal empresa do_ ,a., Pues había que recurrir al capital extranjero. Íſigipara desarrollar nuestra propiedad necesitamos de Obras inmensas en desproporcion ab_soluta_con la potencia de nuestros capitales, nadie nos impedirá sacar estas Obras al mejor postor, y si las compañías europeas nos OfreÇen condiciones mas ventajosas que las americanas, darles nuestras concesiones."3* La dictadura de Porfirio Díaz se habría de caracterizar, así, por sus esfuerzos por atraer capital extranjero, tratando de contrarrestar la influencia de los Estados Unidos en México mediante una política de concesiones al capital europeo. Sin embargo, esto último sería más una tentativa y un barniz ideológico, que una realidad.</p> <p>Sin embargo, en el fondo de la concepción de los positivistas mexicanos sobre las relaciones económicas internacionales de la época, se descubre una completa ignorancia de la nueva era que entonces se iniciaba: la del imperialismo. En efecto, se veía en los capitales extranjeros un elemento de la más sana política económica y se esperaba que al establecer un juego de equilibrio entre la penetración de capitales norteamericanos y la política de puertas abiertas al capital europeo, se lograría preservar, a la vez que desplegar, los intereses de la burguesía criolla. Aunque sospechaban que la competencia entre estadounidenses y europeos se resolvería a favor de los primeros, alimentaban ii ilusión de que esa competencia se daría en una lucha fecunda, de L que la burguesía mexicana resultaría beneficiada-" F-*PORFIRIO DÍAZ nrcranon ni: MÉXICO</p> <p>Las teorías políticas expuestas en La Libertad sirvieron de programa a un pretendido partido político. La Libertad había comenzado por sostener candidatos u Justo Sjen-3, "El istmo de Panamá y la doctrina Monroe". cn La _,ibÇfç;ijç, Méflw, 25 de febrero de 1880. , ' Arnaldo Córdova. OP- df-</p>	<p>, d Obtener los recursos de capital que tal empresa <1°?Pe., Pues había que recurrir al capital extranjero. iſigiaara desarrollar nuestra propi-edad necesitamos de 1 Emmensas en desproporci3n absoluta con la potencia- Opræe nuestros capitales, nadie nos impedirá sacar estas giras al mejor postor, y si las compañías europeas nos ofreÇen condiciones mas veitajosas .que las americanas, darles nuestras concesiones. '3* La dictadura de Porfirio Díaz se habría de caracterizar, así, por sus esfuerzos por atraer capital extranjero, tratando de contrarrestar la influencia de los Estados Unidos en México mediante una política de concesiones al capital europeo. Sin embargo, esto último sería más una tentativa y un barniz ideológico, que una realidad.</p> <p>Sin embargo, en el fondo de la concepción de los positivistas mexicanos sobre las relaciones económicas internacionales de la época, se descubre una completa ignorancia de la nueva era que entonces se iniciaba: la del imperialismo. En efecto, se veía en los capitales extranjeros un elemento de la más sana política económica y se esperaba que al establecer un juego de equilibrio entre la penetración de capitales norteamericanos y la política de puertas abiertas al capital europeo, se lograría preservar, a la vez que desplegar, los intereses de la burguesía criolla. Aunque sospechaban que la competencia entre estadounidenses y europeos se resolvería a favor de los primeros, alimentaban i.1 ilusión de que esa competencia se daría en una lucha fecunda, de L que la burguesía mexicana resultaría beneficiada-" F-*PORFIRIO DÍAZ DICTADOR mc MÉXICO</p> <p>_ Las teorías políticas expuestas en La Libertad sirvliëron de programa a un pretendido partido politiçg Lzbertad habla comenzado por sostener candidatoſ " Justo Sierra, "El istmo de Panamá y 1 (1 Ç - La Libeftad, Nléxico. 25 de febrero de 1880. a oc una Monroe ` en n ' Arnaldo Córdova, op. dt.</p>

Edición Final, Writer.

Como observamos el texto puede contener caracteres extraños, esto sucede cuando el algoritmo lo logra reconocer un carácter y trata hacer una representación de lo que capto utilizando los caracteres disponibles, utilizando un corrector ortográfico es fácil identificar cuales son las palabras problema.

Para el ultimo paso guardamos el texto digitalizado para abrirlo en Writer o lo copiamos directamente.

Una vez hecho esto podemos revisar la ortografía, hacer las correcciones necesarias y aplicar estilos. Con la digitalización completa podemos exportar el resultado en un documento PDF.