Modelo computacional basado en Inteligencia Artificial para hallar la legibilidad de un documento

Mateo Laguna G.

17 de octubre de 2018

Resumen

El presente documento presenta un bosquejo básico de las diferentes etapas que constituyen el algoritmo. Además, introduce el modelo teórico utilizado por el algoritmo para asignar el valor numérico de legibilidad a un documento. Por último, se dan instrucciones sobre cómo implementar el algoritmo computacionalmente.

1. Bosquejo del Algoritmo

Antes de iniciar el algoritmo es necesario establecer la calidad de la imagen que se va a utilizar al transformarse el archivo pdf a jpg. La medida de esta calidad lo da el parámetro dpi (*dots per inch*) que digitalmente corresponde a ppi (*pixels per inch*). Después de efectuar varias calibraciones se encontró que el dpi que mejor se ajusta para los fines prácticos del algoritmo es **200**. (Si se aumenta el dpi, aumenta el tiempo de procesamiento por imagen, pero también aumenta la calidad de los resultados)

El algoritmo se divide en tres fases:

I. Conversión

El proceso de conversión consiste en identificar la carpeta con los archivos pdf que contienen las imágenes que serán analizadas. La red neuronal que hará el procesamiento de las imágenes no recibe archivos pdf, por ende se deben convertir a un archivo de imagen. Por simplicidad se escoge jpg. En esta etapa del algoritmo entran los pdf's y salen los jpg's que serán analizados. El tiempo de ejecución de esta etapa está directamente relacionado con el parámetro de calidad dpi establecido.

II. Inteligencia Artificial

Cada una de las imágenes es introducida a la red neuronal para su respectivo análisis. La red neuronal se encarga de identificar cada una de las palabras del documento con su respectivo porcentaje de seguridad de haberla reconocido exitosamente, este porcentaje es un parámetro para cada palabra llamado *confidence*. Es decir, la confianza que tiene la red neuronal al identificar tal palabra. Se obtiene un archivo con todas las palabras identificadas, con sus respectivas medidas (altura, ancho y posición de la palabra en el documento) y confidence. El tiempo de ejecución de esta etapa depende directamente de lo que tome la red neuronal con cada imagen. (implícitamente viene dependiente de la calidad de la imagen, a mayor calidad, mayor tiempo de ejecución)

2. Modelo Teórico

III. Análisis de datos

La última de las etapas es el análisis de los datos obtenidos de la red neuronal. Un ejemplo se puede ver en la Figura (1) donde cada palabra se gráfica con su respectivo porcentaje de confianza según la red neuronal. Los valores que salen en cero es debido a que no hay elemento de texto que reconocer y por eso sale ese porcentaje (e.g. un símbolo como un escudo o bandera). El modelo que se utiliza para calcular la legibilidad del documento consiste en multiplicar cada palabra por el peso de su porcentaje, pues no aporta lo mismo a la legibilidad una palabra que la red neuronal logró identificar a un 95 % que una a un

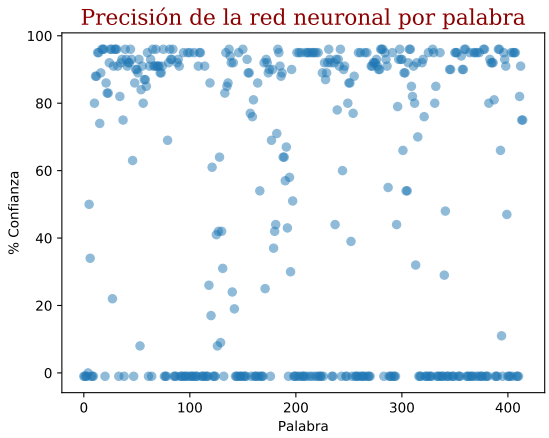

Figura 1: Precisión de la red neuronal.

20 %. Así, se pesa cada palabra por su aporte y al final se divide la suma de todos esos aportes entre la cantidad total de elementos que reconocí de la imagen (que son las palabras del documento).

Por ejemplo, suponga que tiene los siguientes datos:

Palabra	**Mejor(%)**	**Peor (%)**	**Promedio (%)**
cédula	95	10	45
nombre	95	20	55
salud	95	20	95
día	95	10	20
fecha	45	95	60

Para este caso tendríamos:

$$\sum_{\%} = 425, 155, 275 \tag{1}$$

Y al dividirlo por la cantidad de palabras que fueron identificadas (i.e. 5 en este caso), obtenemos una legibilidad de:

Mejor caso	**Peor caso**	**Caso promedio**
$\lambda = 85\,\%$	$\lambda = 31\,\%$	$\lambda = 55\,\%$

3. Instrucciones

Para implementar el algoritmo se recomienda tener los archivos a los cuales se les va a hacer el análisis de legibilidad en una carpeta. Estos archivos deben estar en formato pdf para que no tenga problemas el algoritmo en el proceso de conversión. Al final de la ejecución del programa se creará un archivo de texto por cada documento pdf analizado el cual le dará la siguiente información:

- 1: El nombre del archivo al que se le halló la legibilidad.
- 2: La legibilidad de cada una de las imágenes en porcentaje.
- 3: Los tiempos estimados en segundos de cada una de las etapas que se mencionaron previamente (i.e. conversión, inteligencia artificial y análisis).
- 4: El valor de calidad dado por parámetro para las imágenes (i.e. factor dpi).