Michael McCabe

Class: CS 521 – Summer 2

Date: 8/16

Final Project Summary: ImageOCR

The ImageOCR program module takes advantage of both the Pillow and Python-tesseract modules to enable text to be read from simple images with very good accuracy. If the image is not very simple, the ImageOCR module has a method that can modify an image to pure black and white to aid the read\_image method. The ImageOCR module also has the ability to create a unique HTML file that includes the image in question (or modified image), the output text and an optional word frequency table.

This module can be very useful in many scenarios. Let’s say you have thousands of similar images you need to extract text from - you can loop through all of the images and concatenate all of the results using the “+” magic method. Another scenario may be extracting text off of company image receipts, HR forms or other documents. Instead of manually typing what is found in an image document, one can simply loop through thousands of similarly-structured image documents instead. With that data we have leveraged, one may create a script that can parse through that data and send it into a location to be saved, such as a database, thus saving time and money for a company.