Excelsior user documentation:

Usage: $ python excelsior <input image> <output.csv>

The script is a command line tool that converts a scanned image to a csv file that may be imported by spreadsheet programs or used by a variety of applications.

The <input image> can be any scanned spreadsheet, whether the user pulls the image from a screenshot, a picture, or a scanned page. Most common image formats are supported, including: bmp, pbm, pgm, ppm, sr, ras, jpg, jpeg, tif, tiff, png, and OpenEXR formats.

The output filename should include a .csv extension, and is a simple text file with comma separated values. Only data included in cells of the scanned image are included in the output.

That is, any text outside of the spreadsheet cells will be lost, and the comma separated values will be sorted by their position in the scanned table.

The script may also be imported as a library of functions in any python program, by including excelsior.py in your build path, and adding ‘import excelsior’ to the import section. If imported as a library, the useful functions are as follows:

prepare\_image(filename)

input: filename of image to be processed

output: (bw, threshold) - a tuple of cv2 images

bw – a binary black and white version of the input image used for OCR

threshold – an inverse binary image used for finding cells

find\_cells(image)

input: binary cv2 image (bw)

output: cells. A python list of dimensions of all cells in image

The dimensions are a tuple, (x, y, x\_length, y\_length)

start\_tess\_api()

no input

output: returns an initiated tesseract api for reading text

ocr\_text(api, image)

input: api - an initiated tesseract api

image - binary image (bw) with text to be scanned

output: python string of the text contained in the image

end\_tess\_api(api)

input: initiated tesseract api

no output, function simply resolves the initiated api.