SW Test course (67778) - 2021

Exercise 3 – Test Automation

General goal: Build a simple test automation framework to test the Tesseract OCR.

Specific goal for this exercise: Find what is the smallest font size that Tesseract OCR can convert to text without errors.

Input: A CSV file that contains the following columns:

Path	**Filename**	**Text**	**Config1**	**… More config options, up to 5 more columns**

Example:

Path	**Filename**	**Text**	**Config1**	**…**	**Config6**
C:\tessTest\images	Arial_14pt.jpg	Hello!			
C:\tessTest\images	Arial_12pt.jpg	Hello!			
C:\tessTest\images	Arial_10pt.jpg	Hello!			
C:\tessTest\images	Arial_8pt.jpg	Hello!	tessedit_char_blacklist=A		
C:\tessTest\images	Arial_6pt.jpg	Hello!	tessedit_char_blacklist=A		
C:\tessTest\images	Arial_4pt.jpg	Hello World!	tessedit_char_blacklist=B	…	textord_pitch_range=3

The test automation code shall:

- Read the .csv file
- Run Tesseract on the specified file in each row, with the specified configuration parameters (if specified)
- Compare the results to the “Text” value
- Decide Pass/FAIL
- After completing the run of all the rows in the csv file, the test code shall print out what is the smallest font size that is 100% recognizable by Tesseract, for each font type.
- The printout format must be **exactly** like this (a row for each font type you had in the images):

  ```
  Arial,12
  Calibri,8
  ```

  The value next to the font type is the smallest font that is still converted 100% correctly.

  If all images for a given font failed, print only the font name and a comma. Example for Arial:

```
Arial,
```

How your submission will be evaluated

We created a folder with many image files with different fonts and different font sizes. This folder is not shared with you.

We will run your code on this folder and expect to get a correct result (correctly find the smallest font for each font type).

Supporting stuff

a) We will be testing using Tesseract v5.0.0
b) We provide an example of 5 image files and the corresponding csv file. Make sure to update the first column to point to where you placed the images.
c) Python has a built-in module to read csv files. It has functions to read a csv into a dictionary.
d) Don't assume you are guaranteed to have an image that will fail to give 100% recognition. There may be cases where ALL files are 100% converted correctly by Tesseract. In this case, just print the value of the smallest font you encountered.
e) The files we will test you on all have the text in a single row. We also will not have any Config values in the input csv file.

Submission

Submit the answers as a python file: **Ex3_<ID1>_<ID2>.py** . Submission is in pairs, so ID1 and ID2 are your two ID numbers.

Deadline:

Submit before June 21st, at 23:55

Late penalty: 10% for June 22, 20% for June 23. We will not accept submissions after June 23.