Rule14 Interview Sample Problem

Build a general parser to extract text from a simple image

Input: 5 test images of the same table and their corresponding OCR outputs

Task: Review the 5 test images in the Images folder and their corresponding OCR outputs in the OCR folder. Use key information from the images and OCR output (for example, bounding boxes, text or confidences) to build a general parser that will work properly on additional data (images of the same table with their OCR outputs).

What to submit:

- Scripts used to complete the task
  - ■ **IMPORTANT**: Use the following filename convention for your script submissions: “<last_name>_<first_name>_script”.<file_extension>
- Testing outputs: a json file of all extracted text contents from the images
  - ■ **IMPORTANT:** Use the following filename convention for your results set: “<last_name>_<first_name>_results”.<file_extension>
  - ■ The standard format of the json file should be a dictionary. Each key of the dictionary is an image name, and its value is a list of lists (containing the text entries, organized by line). Below we illustrate the expected output for the first image:

```
{‘testimage1’:[ [‘111020’,‘111020’,‘21’,‘’,‘95940’,‘59’,‘’,‘’,‘’,‘A’,‘1354’,‘00’,‘2’],
                [‘111020’,‘111020’,‘21’,‘’,‘95861’,‘TC’,‘’,‘’,‘’,‘A’,‘3460’,‘00’,‘1’],…
                ],
‘testimage2’:…}
```

(Note that the first two entries in each row are date strings, so you may need to group text in neighboring boxes.)

Due to constraints on image and OCR quality the accuracy of the parser may not be perfect, so just try your best.

How to submit:
Please upload your completed results and scripts to the following folder:
https://hgmcloud.egnyte.com/ul/L4faChip4s

Due Date:
Please submit your results no later than Sunday, March 26th at 11:59pm