Ruoying Hao

02/25/2017

This week’s progress report contains mixed findings.

I have researched the way to automate the process of converting excel sheets to images. By using a python package called “pyautogui”, I could write a script that control the keyboard and mouse. The conversion process works as follows:

1. Open an excel file
2. Using hotkeys to navigate to the target sheet
3. Copy the sheet
4. Open an empty image file with mspaint
5. Using hotkeys to copy the sheet from the clipboard to mspaint
6. Save the mspaint file as a new image file with a specified name
7. Close the image file
8. Since we are back to the excel file, navigate to the next target sheet
9. …

I’ve developed script that can successful convert ONE excel sheet to image. If adjust the script accordingly, we could massively convert excel sheets to images. The reason I have not done so is because I have several questions:

1. What would be the format of the image file? I believe it depends on which format is best suitable for TensorFlow. But we also don’t want the file that is too big. The smallest is .bmp, when comes .jpg, .bmp is the biggest.
2. How many excel sheets would be converted in a 10-Q? From a quantitative analysis perspective, the extensive sheets before “basis of presentation”, which includes balance sheet, income statement, statement of shareholder equity, etc., are the most important. The sheets include and after “basis of presentation” are mostly qualitative information.
3. How should we name each sheet? We could name them in numbers but it would be more informative if we can classify them into balance sheet, income statement, etc. And here comes the hard part: for different companies, they have different orders in ordering the sheets, and they name the same kinds of sheets differently. Even within the same company, the same kind of sheets could be names differently over time. This problem gave me inspiration for another machine learning topic: classifying financial statement based on content.