Ruoying Hao

March 14, 2017

I am pleased to announce that I have successfully developed the Python script to auto-convert the first 6 excel sheets in a 10-Q to .png images and saved them in a folder specified by dates, organized by stock tickers, with minimum human supervision. So far I am on the process of converting all the .xlsx files. For .xls files, since every time when I open them I was prompt to click a conversation box for compatibility, I would modify the current script to convert the .xls files. The Algorithms works as follows:

1. Find all the .xlsx files and store the file names in a list
2. For each item in the list, convert the 2nd to 7th excel sheets into .png images, rename the images as 1.png, 2.png, etc.
3. After extracting the images, create ticker folder (if not exist), and date folder (if not exist)
4. Move the images into the folder
5. Delete the images outside of the folder

It is worth noting that at the middle of the conversion, I decided to modify the script so that it auto-fit the cells to prevent data loss. I did it only because I discovered that among some sheets that had been converted, data loss had occurred.

Conversion of all .xlsx files for all S&P 100 tickers took about 1 day, the program occasionally terminated because of errors. I would put a sample of images folders on the lab machine.

Mid-semester review: I think the project is in a good shape and we start feeding the images to TensorFlow. I expect to finish reconstructing the data before the semester ends.