Coding test for RP position

Instructions

Please complete the two coding tasks below. The deliverable for each task should consist of (1) data files (usually CSV or Excel), (2) the code you used to generate the data file. The code should include annotations to help the reader understand what is being done. Both tasks must be completed using Python.

These tasks should be relatively straightforward; please take no longer than 12 hours to complete. Once completed, please send the result to Michelle Skinner (Michelle.Skinner@chicagobooth.edu).

TASK 1

Using the Python Requests package make a call to the S&P 100 Wikipedia page. Using Python Beautiful-Soup, extract the name, symbol, and link of the firms composing the index.

Then, for each company, find the best match in the file “names gvkeys.CSV" using the Python package “fuzzy-wuzzy".

The final dataset must have five columns: the company name, the symbol, the link to the company Wikipedia page, the matched company name and the score assigned to the match.

Deliverables:

1. Export the result as either an Excel or CSV format

TASK 2

Your task is to analyze the 10-K forms for two companies, “Apple Inc." and “Intel”, for the fiscal years 2014 through 2018.

- These forms are available on the SEC EDGAR website.
- Please note that fiscal years are not necessarily the same as calendar years.

Retrieve these documents and complete the following two sections.

<u>First part</u>

Please determine and tabulate the following for each fiscal year:

1. the company name
2. the number of words in the filing's “Risk Factors" section
3. the number of times the word “competition" is mentioned in this section

Deliverables:

1. Export the result as either an Excel or CSV format
2. Export the text of the “Risk Factors" sections as separate txt files for each company-year

<u>Second part</u>

Using the nltk Python package, remove all the stop words from the “Risk factors" section and, using regular expressions, identify all the occurrences of the word “patent" or its variations (“patents”, “patenting”, etc.). Please also retrieve the word preceding and the word following the occurrence found.

In other words, you will have to construct a dataset with the “word preceding", the “occurrence found", and the “word following".

Deliverables:

1. Export the result as either an Excel or CSV format.