**Scraping + Filtering + Fuzzymatching about CIKs and Company Names**

Editor : Yichen Zhang 7/6/2017

Script folder: "P:\PCAOB Staff\Interns\zhangy1\NAF\script\python part\_ scraping and matching "

Test folder:” P:\PCAOB Staff\Interns\zhangy1\NAF\test”

Input: Excel file with unique IssuerName from NAF

1. Read the excel, get Issuer Names, Then scrape possible CIK COMPANY STATE from SIC website.
2. Use “**01\_firmname\_cik\_scrapedata.py**” :
   1. Read data, do not change original variable, create new one to do cleaning, like remove ” INC” “ CORP”… remove punctuation and remove parentheses’ content.
   2. Scrape CIK COMPANY STATE from SIC and filter out unmatched issuernames, there are three scenarios:
      1. One issuer has many ciks, so we find all of them and save them into Sheet='Possible Companies'
      2. One issuer has too many ciks, which is too time-consuming, so we filter out them and save into Sheet=‘Wrong IssuerName’
      3. One issuer doesn’t have unmatched cik, so we filter out them and save into Sheet=‘Wrong IssuerName’
   3. Output: “**output1.xlsx**”
3. Use “**01\_wrong\_issuername\_update\_scrape.py**”, this one is to secondly match problematic issuernames:
   1. Same with 1), just add more rules to clean data, like remove “A ”,”THE ”,’AND’,’LP’,and add space before “COM ” “NET ”.
   2. Output: “**output1\_1.xlsx**”(this time problematic data need to be manually checked)
4. Read Excel “output1.xlsx” and “output1\_1”, merge them together, filter CIK by checking if it is valid and has filings which we want. And scape each valid CIKs’ formerly Names if available. Use “**02\_cik\_filter.py**”:
5. Read both “output1.xlsx” and “output1\_1.xlsx” to append them together
6. Filter out invalid ciks:
   1. Criteria 1: the number of items is greater than 10
   2. Criteria 2 :if CIK contains our aim filings=[ '10-K', '20-F', '10KSB', 'S-1','40-F', 'S-4', '1-F']
7. Scrape each valid CIKs’ formerly names if available
8. Output: **“output2.xlsx”**
9. Read Excel “output2.xlsx”, using fuzzy matching to compare Issuer Names we have with COMPANY and Formerly Names which scraped from SIC website, then choose the well matched ones and mark best matched ones. Use **“03\_fuzzymatching.py”:**
10. read the “output2.xlsx” and clean the dataset: remove punctuation, capitalize names, clean CIKs, group dataset by issuer(this step is important)
11. Conduct fuzzy matching with four scores: ratio, partial\_ratio token\_set\_ratio and token\_sort\_ratio(but token\_sort\_ratio is not important):
    1. First use “issuer” compare with one CIK’s COMPANY and Formerly Names, choose the best name by sum of four scores and get its scores
    2. Second, Compare sum of scores of best names of CIKs, mark the best matched one and keep the ones which satisfies Score1+Score2+Score3 > 260 (because I mentioned Score4 is not important)
12. Output:”**output3.xlsx”**