Scraper for SEC Form 13F-HR

Overview

Securities and Exchange Commission (SEC) describes form 13F-HR as “Quarterly report filed by institutional managers, Holdings”. SEC requires professional investment managers (who manage over 100MM USD) to disclose their holdings every quarter by filing form 13F-HR.

Scraping form 13F-HR allows us to find out the holdings of an investment firm, and track the trading pattern if we compare the holdings to previous quarters. Form 13F-HR does not track derivative exposures though. There are a few portals that aggregate 13F-HR data, but SEC website is fairly easy to scrape. SEC publishers daily reports as well as quarterly aggregates.

The scrapers in this folder scrape using quarterly index. It is trivial to adapt the scrapers to scrape using daily index files. Scraping is a two-step process: First, index file is scraped to build a list of links to follow. Index file contains a list of filings. Each entry contains the type of form filed, CIK (central index key) number of the filing firm, name of the firm, and a relative URL to the filed report. URLs lead to text files containing 13F-HR forms (with embedded XML content). Each holding entry has name of stock, number of stocks held, their USD value, and few other details. Index file for quarterly aggregates can be big (over 50MB). This file is downloaded into a local copy to reduce burden on SEC servers and to make incremental scraping faster.

Usage

Soup The scraper found in *soup* directory uses BeautifulSoup to parse tags. Specify year, quarter, and (optionally) number of links (CIKs) to follow or a list of CIK numbers to include.

```
python soup/sec13F.py -h

#> usage: sec13F.py [-h] [--year YEAR] [--quarter {1,2,3,4}] [--count COUNT]
#>                  [CIK [CIK ...]]
#> 
#> Scrape Form 13F-HR from SEC website and report current holdings of investment
#> firms
#> 
#> positional arguments:
#>   CIK                   central index key(s) (CIK) to filter, if specified
#>                         (default: None)
#> 
#> optional arguments:
#>   -h, --help            show this help message and exit
#>   --year YEAR, -y YEAR  year when report was filed (default: 2021)
#>   --quarter {1,2,3,4}, -q {1,2,3,4}
#>                         quarter (1-4) of year when report was filed (default:
#>                         1)
#>   --count COUNT, -c COUNT
#>                         maximum number of reports to parse (default: 2)
```

Scrapers in *scrapy* and *selenium* directories produce the same output as the one based on *BeautifulSoup*.

Scrapy

```
#> Usage: scrapy crawl sec13Fspider -a year=<year> -a quarter=<1-4> \
#>          -a n=<integer>  -o file.csv -t csv
#>        (place the spider file, sec13F_spider.py, inside a scrapy project)
```

Selenium Place the geckodriver executable where $PATH can find it.

```
python selenium/sec13F.py -h

#> usage: sec13F.py [-h] [--year YEAR] [--quarter {1,2,3,4}] [--count COUNT]
#>                  [CIK [CIK ...]]
#> 
#> Scrape Form 13F-HR from SEC website and report current holdings of investment
#> firms
#> 
#> positional arguments:
#>   CIK                   central index key(s) (CIK) to filter, if specified
#>                         (default: None)
#> 
#> optional arguments:
#>   -h, --help            show this help message and exit
#>   --year YEAR, -y YEAR  year when report was filed (default: 2021)
#>   --quarter {1,2,3,4}, -q {1,2,3,4}
#>                         quarter (1-4) of year when report was filed (default:
#>                         1)
#>   --count COUNT, -c COUNT
#>                         maximum number of reports to parse (default: 2)
```

Example

Berkshire Hathaway's CIK is 1067983. We can download their holdings as of 1st quarter 2021 into a local file.

```
if [ ! -f "brk_2021_1.csv" ]; then
  python soup/sec13F.py --year 2021 --quarter 1 1067983 > brk_2021_1.csv
fi
```

We can find out the top holdings by value (happens to be Apple).

```
import pandas as pd
brk2021 = pd.read_csv('brk_2021_1.csv')
gr2021 = brk2021.groupby(['issuer', 'cusip']).agg({'value':'sum', 'quantity':'sum'})
print(gr2021.sort_values('value', ascending=False).head())
#>                                               value    quantity
#> issuer                     cusip                               
#> APPLE INC                  037833100  117714016   887135554
#> BANK AMER CORP             060505104   30616150  1010100606
#> COCA COLA CO               191216100   21935999   400000000
#> AMERICAN EXPRESS CO        025816109   18331249   151610700
#> VERIZON COMMUNICATIONS INC 92343V104   12090703   205064263
```

Similarly, top holdings as of the 4th quarter of 2020 are as follows.

```
if [ ! -f "brk_2020_4.csv" ]; then
  python soup/sec13F.py --year 2020 --quarter 4 1067983 > brk_2020_4.csv
fi
```

```
import pandas as pd
brk2020 = pd.read_csv('brk_2020_4.csv')
gr2020 = brk2020.groupby(['issuer', 'cusip']).agg({'value':'sum', 'quantity':'sum'})
print(gr2020.sort_values('value', ascending=False).head())
#>                                        value    quantity
#> issuer              cusip                               
#> APPLE INC           037833100  109358868   944295554
#> BANK AMER CORP      060505104   24333323  1010100606
#> COCA COLA CO        191216100   19748000   400000000
#> AMERICAN EXPRESS CO 025816109   15198971   151610700
#> KRAFT HEINZ CO      500754106    9752763   325634818
```

It is apparent that they have sold 57160000 shares (6%) of Apple (AAPL) during the 1st quarter of 2021, when the market was hitting all-time highs!