Tutorial 2. Extracting Textual Data from 10-K

This tutorial will guide you through the process of running a set of four Python scripts to extract textual data -- the Item 1 section -- from Edgar's 10-K files.

NOTE: Before you start, you should make sure that Python 2.7 is already installed in your computer (For installation instructions, visit here: http://community.mis.temple.edu/zuyinzheng/pythonworkshop/)

We will work with four Python scripts. The purpose of the scripts is to use re (regular expression) and bs4 (Beautiful Soup) packages to extract Item 1 from firms' 10-K reports. Here is a sample of 10-K report:
http://www.sec.gov/Archives/edgar/data/1288776/000165204416000012/goog10-k2015.htm

1 Install the Beautiful Soup package

You need to have the Python package, bs4 (for Beautiful Soup), installed in your computer before executing the scripts.

To do so, typing the following command in your command line interface (On Windows it is called "Command Prompt", and on Mac it is called "Terminal"):

```
pip install beautifulsoup4
```

2 Download the Python Scripts and CompanyList.csv file

Download four Python scripts CompanyList.csv file from the following link. Make sure you download all files in the same folder.
http://community.mis.temple.edu/zuyinzheng/pythonworkshop/
There are four scripts:

- The first script 1GetIndexLinks.py extracts the URLs from each firms' search results returned by Edgar.
- The second script 2Get10kLinks.py extracts the URLs for each firm's 10-K reports.
- The third script 3DownLoadHTML.py downloads the 10-K reports as HTML files.
- The fourth script 4ReadHTML.py extracts the Item 1 section of the 10-K reports and put it into a text file.
- (Don't worry about the fifth script, 5tfidf.py at this point.)

And a csv file:

- The CompanyList.csv file contains the ticker symbols and names of three firms.

3 Change Working Directory

Find the folder where you have saved the python script in your computer.

For each of the four scripts, change the working directory to where you put the company list (CompanyList.csv).

To do so, for each of the four Python scripts:

i) Open the Python script with IDLE.

ii) Find the `os.chdir()` function. The `os.chdir()` function should be in Line 4 of all the four scripts.

iii) Change the parameter in `os.chdir()` function.

For example, I have the CompanyList.csv in the folder:
`/Users/alvinzuyinzheng/Dropbox/PythonWorkshop/scripts/`

Therefore, my `os.chdir()` function looks like this:
`os.chdir('/Users/alvinzuyinzheng/Dropbox/PythonWorkshop/scripts/')`

If you have a different folder name, make changes accordingly.

(In Windows, the folder names probably look like this:
`C:\username\Dropbox\python\workshop\Scripts`
If you are not sure how to find the folder path, check the instructions here: Copy File Folder Path in Mac OS X)

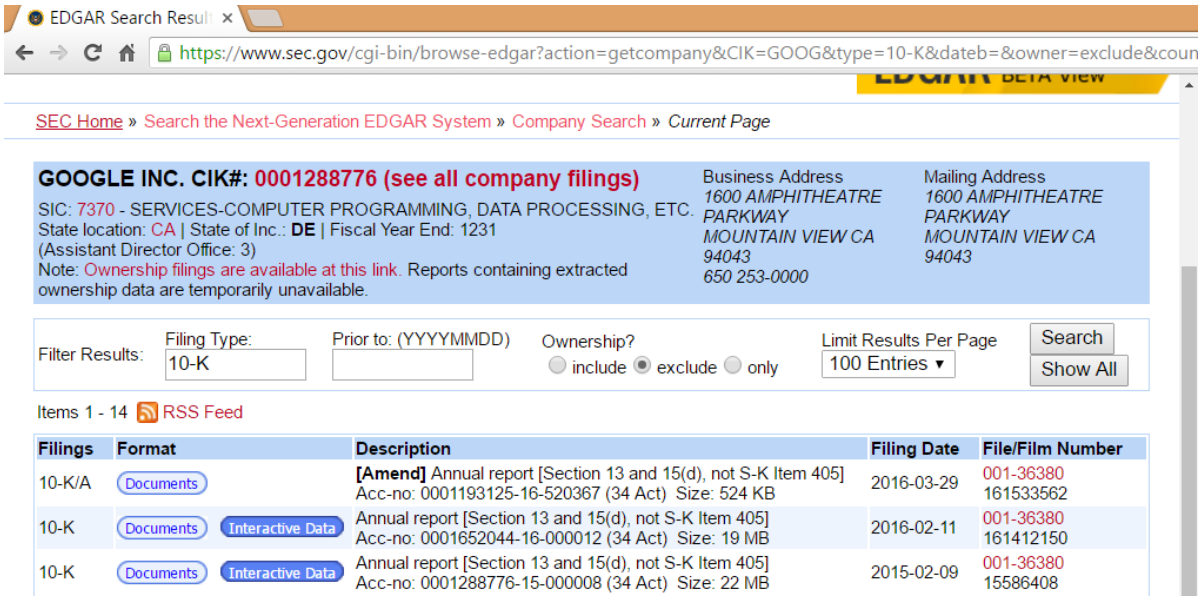
4 Run the 1GetIndexLinks.py script

The 1GetIndexLinks.py script extracts the URLs from each firms' search results return by Edgar.

When using Edgar, we often use the ticker symbol of a firm to search for the firm's 10-K reports.

Below is a sample URL for Google. Note that we in the URL we restrict to "CIK=GOOG" and "type=10-K".
https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&**CIK=GOOG**&**type=10-K**&dateb=&owner=exclude&count=100

A sample result page looks like this:



Steps to run the 1GetIndexLinks.py script:

i) Double check if you've changed the working directory in the previous step.
i) Open the python script with IDLE.
ii) Click the Run menu and choose “Run Module”.

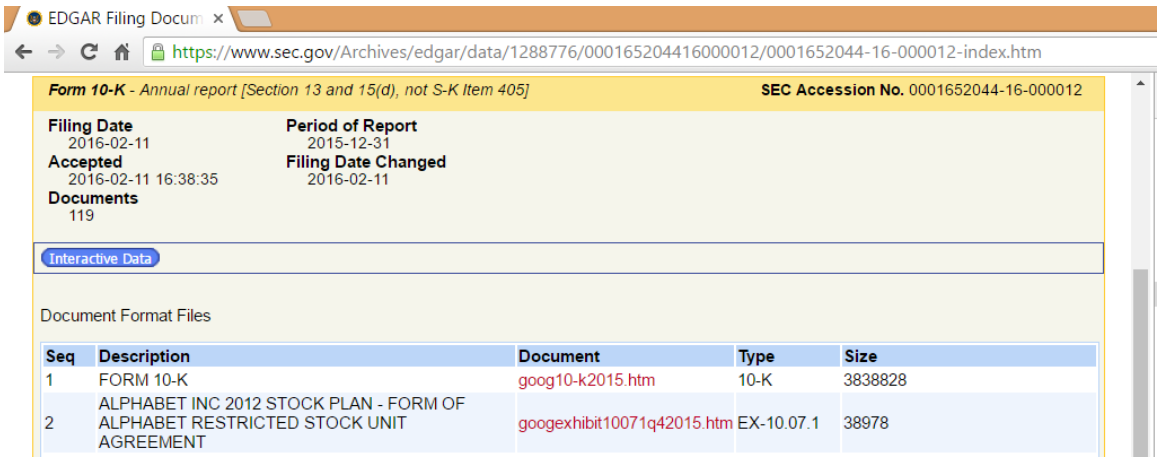
Once finished, a csv file “IndexLinks.csv” will be created in your working directory. It contains the list of index links extracted from the search result pages.

5 Run the 2Get10kLinks.pyscript

The 2Get10kLinks.py script extracts the URLs of the 10-K pages from each firms' index page.

Below is the URL for Google ‘s index page.
http://www.sec.gov/Archives/edgar/data/1288776/000165204416000012/0001652044-16-000012-index.htm

A sample index page looks like this. Essentially we want to extract the first link (goog10-k2015.htm) using Python.



Steps to run the 2Get10kLinks.py script:

i) Double check if you've changed the working directory in the script.
ii) Open the python script with IDLE.
iii) Click the Run menu and choose “Run Module”.

Once finished, a csv file “10kList.csv” will be created in your working directory. It contains the list of 10-k file links extracted from the index pages.

6 Run the 3DownloadHTML.pyscript

The 3DownHTML.py script downloads the 10-K reports as HTML files and store them in a subfolder “./HTML/”.

Steps to run the 3DownloadHTML.py script:

i) Double check if you've changed the working directory in the script.
ii) Open the python script with IDLE.
iii) Click the Run menu and choose “Run Module”.

Once finished, a subfolder “./HTML/” will be created in your working directory. It contains the 10-k files downloaded from Edgar.

Note: in line 34 of the script, I have restricted to download only 10-K files for years 2014 and 2015. For example, if you’d like to download for 2013-2015, you can modify this line to

```
FormYears = ['2013','2014','2015']
```

7 Run the 4ReadHTML.pyscript

The 4ReadHTML.py script extracts the Item 1 section of each 10-K HTML file and put it into a text file. All the text files are stored in a subfolder "./txt/".

Steps to run the 4ReadHTML.py script:

i) Double check if you've changed the working directory in the script.
ii) Open the python script with IDLE.
iii) Click the Run menu and choose "Run Module".

Once finished, a subfolder "./txt/".will be created in your working directory. It contains the text files, where each text file contains the Item 1 section of a 10-K file. The files should look like this: https://www.dropbox.com/sh/4epko0rs3gp43we/AADB_Vx7vOLlX6g_G5zlc1Fna?dl=0