**10K Filings Text Analysis Python Package**

**Steps:**

1. **Install** [**http://scipy.org/**](http://scipy.org/)

**Or directly:** [**http://python-xy.github.io/**](http://python-xy.github.io/)

1. **If Openpyxl is not supported, then install it via cmd by typing:**

$ pip install openpyxl

1. **Make sure the directories is correctly identified**
2. **Run the programs as you require. There are four python codes that do the following actions:**
   1. **1\_Edgar\_Listings:** This downloads all the listings in SEC EDGAR with the ftp addresses. This file is already downloaded and saved in a stata file as Edgar\_Listings.dta . The corresponding 10K filings (with ftp addresses) from 1993 to 2016 is derived and saved in master10k\_Final.csv
   2. **2\_Download\_Edgar:** This downloads the complete text files from the SEC edgar.
   3. **3\_Clean\_Text\_Files:** This cleans the text files from HTML tags and ASCII codes. It also removes the extensions in the filings.
   4. **4\_Count\_Number\_Words:** This counts the number of occurrences of words or phrases (case-insensitive) in the text files.

**Edgar 10K Listings:**

[**https://www.sec.gov/edgar/searchedgar/companysearch.html**](https://www.sec.gov/edgar/searchedgar/companysearch.html)

**Parsing the 10Ks:**

We take the following steps to parse the 10Ks (We follow Bodnaruk, Loughran, and McDonald, 2015,JFQA and Loughran, and McDonald, 2011, JF):

We download all the 10-K filings text files (complete text file) from the SEC Edgar. We do not include the 10-K amendment. This complete text file contains all the information (the core 10K document in HTML format, exhibits, graphic files [jpg] and XBRL files).

We focus on the textual content of the core document by removing the markup HTML tags, ASCII-encoded graphics. We do not exclude tables because Item 7 and Item 8 of the document may start with tables and moreover, the influence of tables on the general text analysis is minimal.

In abstract, we take the following steps:

1. Download the complete text file for the 10K filings of each company from 1993 to 2016
2. Remove graphics (ASCII encoded graphics)
3. Remove the SEC header
4. We remove the standard first page of the filing appearing from the beginning of the text file to </SEC-HEADER> tag.
5. Re-encode characters—We replace the HTML “encoded” characters such as &NBSP (blank space) or &AMP (&) back to their original ACSII form (blank space and & form)
6. We remove exhibits (remove all the texts between <TYPE>EX tags)
7. Remove the additional HTML tags such as <a> … </a>

Maybe: Remove the comma (,) and dot (.)