**Assignment**

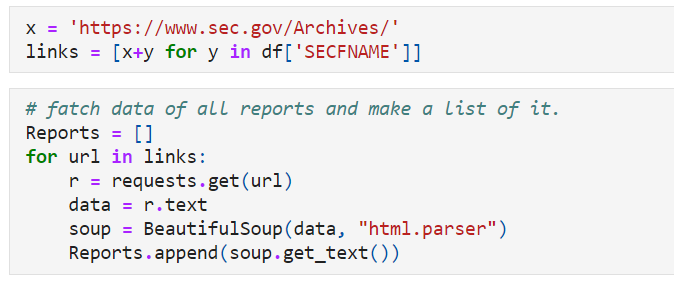
**Step 1:** Including necessary Python libraries and text,excel & csv file file using pandas .



At the end we can see that a variable named “Passive\_words”,it contain past form of all the“Irregular verb (200 values)”mentioned in table, which is there in the link given in data analysis file -->Passive word section

I have fatch that table data (Irregular verb (200 values)) in form of csv file and named file as “Passive words.csv”

**Step 2:** Data collection /fatching from given link using python’s Beautiful soup library.



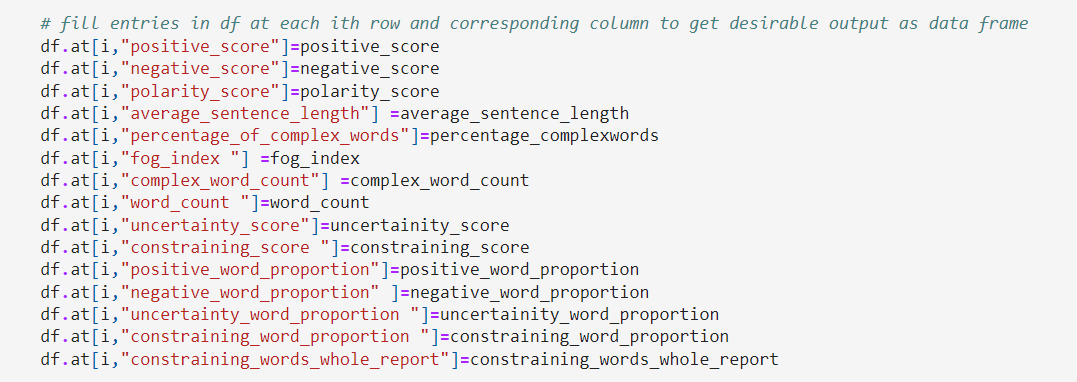
Where X = Domain name ,y = file name(present in Cik\_list file ,“SECFNAME” column)

**Step 3:** Data Preprocessing using python’s “re” and “nltk” libraries. It includes 3 basic steps

1. Removing punctuations form text data
2. Convert text data into tokens
3. Removing stopwords(Given in “StopWords\_Generic.txt”)

**Step 4** : Data Processing and filtering (For each report)

1. Positive score is the count of the words that are present in both text data and positive dictionary.where positive dictionary contain all the word of Master dictionary which has non zero positive column.
2. Negative score is the count of the words that are present in both text data and Negative dictionary.where negative dictionary contain all the word of Master dictionary which has non zero negative column.
3. Polarity score, Avg sentence length ,fog index and % complex words is calculated using given formula.
4. Complex word count is the no of words in the data that contain more than 2 syllable.
5. Uncertainity score ,constraining score and constraining words in report is calculated as the count of word that are both in text data of reports and their respective file as Uncertainity dictionary ,constraining dictionary.
6. Positive ,negative ,Uncertainity and constraining word proportion can be calculated using the given infomation and formula.
7. Fill all the derived infomation to the dataframe at ith row and respective column where i is a iterator for all the reports.



**Step 5:** Showing and storing output in excel form.

