**Data Extraction and Text Analysis**

**Objective**

Objective of this assignment is to extract textual data from SEC / EDGAR financial reports and perform text analysis to compute variables those are explained below.

**Data Source**

Link to SEC / EDGAR financial reports are given in excel spreadsheet “cik\_list.xlsx”.

Please add <https://www.sec.gov/Archives/> to every cells of column F (cik\_list.xlsx) to access link to the financial report.

Example: Row 2, column F contains edgar/data/3662/0000950170-98-000413.txt

Add <https://www.sec.gov/Archives/> to form financial report link i.e.

<https://www.sec.gov/Archives/edgar/data/3662/0000950170-98-000413.txt>

# **Variables:**

“Text Analysis.docx” you need to compute following:

Section 1.1: Positive score, negative score, polarity score

Section 2: Average Sentence Length, percentage of complex words, fog index

Section 4: Complex word count

Section 5: Word count

In addition to these eight variables, compute two more items: “uncertainty” and “constraining”. These variables are calculated similar to the ones in Section 1.1 or Section 4. Attached the lists of words that are classified as uncertain or constraining.

**For uncertainty:** “uncertainty\_dictionary.xlsx”

**For constraining:** “constraining\_dictionary.xlsx”

**4 more variables:**

# **For the variables: positive word proportion, negative word proportion, uncertainty word proportion, and constraining word proportion:**

The absolute values of “Positive/Negative Scores” are equal to the number of positive/negative words in each report of 10-Q/K; so the (Loughran-McDonald) positive/negative word proportion can be simply calculated as “Positive/Negative Scores divided by Word Count – compute these measure in addition to Polarity Score.  And, the “uncertainty score” and “constraining score” will be also just equal to the number of corresponding words and you can calculate the portion of these words as the same as above.

**1 more variables:**

**For the variable Constraining words for whole report**

Add one variable to the mix, which will be calculated for the whole report. It’s the number of “constraining” words over the whole report.

**That means you need to collect/compute 15 variables in total.**

# **Data**

For each report (financial reports, links available in excel, cik list), we would like these 15 variables calculated for the whole report.

You need to read, access and clean the financial report and KEEP TEXT ONLY from the annual report url. Clean and remove html / xml, etc. codes, syntaxes, and best of your knowledge. You can keep text (paragraphs, sections, titles, readable text, tables, etc. whichever are the part of the financial report in the given url). Remove all noise and unwanted data from the annual report urls.

Attached is the spreadsheet “cik\_list.xlsx”, which also contains the links to reports. It would be ideal if you could add 15 columns to each row, so that we would have the # rows unchanged after your data collection.

# **Output Data Structure**

**Output Variables:**

1. All input variables in “cik\_list.xlsx”
2. positive\_score
3. negative\_score
4. polarity\_score
5. average\_sentence\_length
6. percentage\_of\_complex\_words
7. fog\_index
8. complex\_word\_count
9. word\_count
10. uncertainty\_score
11. constraining\_score
12. positive\_word\_proportion
13. negative\_word\_proportion
14. uncertainty\_word\_proportion
15. constraining\_word\_proportion
16. constraining\_words\_whole\_report