**Data Extraction and Text Analysis**

**Objective**

Objective of this assignment is to extract some sections (which are mentioned below) from SEC / EDGAR financial reports and perform text analysis to compute variables those are explained below. 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”

That means you need to collect/compute 10 variables in total.

# **Sections:**

For each report (financial reports, links available in excel, cik list), we would like these 10 variables calculated for three sections. These are

“Management's Discussion and Analysis”,

“Quantitative and Qualitative Disclosures about Market Risk”, and

“Risk Factors”.

If a report does not include any of these sections, leave those fields blank.

In other words, we need 10 x 3 = 30 variables.

Attached the spreadsheet “cik\_list.xlsx”, which also contains the links to reports. It would be ideal if you could add 30 columns to each row, so that we would have the # rows unchanged after your data collection.

# **Additional Variables: positive/negative and uncertainty/constraining word proportion**

The absolute values of “Positive/Negative Scores” are equal to the number of positive/negative words in each section of 10-Q/K; so the (Loughran-McDonald) positive/negative word proportion can be simply calculated as “Positive/Negative Scores divided by Word Count (Section 5)” – 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 same as above.

# **Additional Variable: Constraining words for whole report**

Add one variable to the mix, which will be calculated only once for the whole report (i.e., not three times). It’s the number of “constraining” words over the whole report rather than in any specific section.

# **Output Data Structure**

**Notations:**

“Management's Discussion and Analysis”: MDA

“Quantitative and Qualitative Disclosures about Market Risk”: QQDMR

“Risk Factors”: RF

**Output Variables:**

1. All input variables in “cik\_list.xlsx”
2. mda\_positive\_score
3. mda\_negative\_score
4. mda\_polarity\_score
5. mda\_average\_sentence\_length
6. mda\_percentage\_of\_complex\_words
7. mda\_fog\_index
8. mda\_complex\_word\_count
9. mda\_word\_count
10. mda\_uncertainty\_score
11. mda\_constraining\_score
12. mda\_positive\_word\_proportion
13. mda\_negative\_word\_proportion
14. mda\_uncertainty\_word\_proportion
15. mda\_constraining\_word\_proportion
16. qqdmr\_positive\_score
17. qqdmr\_negative\_score
18. qqdmr\_polarity\_score
19. qqdmr\_average\_sentence\_length
20. qqdmr\_percentage\_of\_complex\_words
21. qqdmr\_fog\_index
22. qqdmr\_complex\_word\_count
23. qqdmr\_word\_count
24. qqdmr\_uncertainty\_score
25. qqdmr\_constraining\_score
26. qqdmr\_positive\_word\_proportion
27. qqdmr\_negative\_word\_proportion
28. qqdmr\_uncertainty\_word\_proportion
29. qqdmr\_constraining\_word\_proportion
30. rf\_positive\_score
31. rf\_negative\_score
32. rf\_polarity\_score
33. rf\_average\_sentence\_length
34. rf\_percentage\_of\_complex\_words
35. rf\_fog\_index
36. rf\_complex\_word\_count
37. rf\_word\_count
38. rf\_uncertainty\_score
39. rf\_constraining\_score
40. rf\_positive\_word\_proportion
41. rf\_negative\_word\_proportion
42. rf\_uncertainty\_word\_proportion
43. rf\_constraining\_word\_proportion
44. constraining\_words\_whole\_report

Checkout output data structure spreadsheet for format of your output.