Data Extraction and Text Analysis

Blackcoffer Consulting

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

1 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.

2 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.

3 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 – 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.

4 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.

5 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.

Timeline

8 days, sooner is better.

Where to submit

Submit your solutions email, including all outputs in csv / excel, source codes, documentations to run the code and other necessary details, and your updated resume.

a) Send your submission to uday@blackcoffer.com

b) Fill this google form: https://forms.gle/qj5UB8bawxAZ4mWC6

c) Inform us at uday@blackcoffer.com