**Data Extraction and NLP**

# **Objective**

The objective of this assignment is to extract textual data articles from the given URL and perform text analysis to compute variables that are explained below.

# **Data Extraction**

Input.xlsx

For each of the articles, given in the input.xlsx file, extract the article text and save the extracted article in a text file with URL\_ID as its file name.

While extracting text, program extracts only the article title and the article text. It should not extract the website header, footer, or anything other than the article text.

# **Data Analysis**

For each of the extracted texts from the article, performed textual analysis and compute variables, given in the output structure excel file. We have saved the output in the exact order as given in the output structure file, “Output Data Structure.xlsx”

# **Variables**

Definition of each of the variables given in the “Text Analysis.docx” file.

Look for these variables in the analysis document (Text Analysis.docx):

1. POSITIVE SCORE
2. NEGATIVE SCORE
3. POLARITY SCORE
4. SUBJECTIVITY SCORE
5. AVG SENTENCE LENGTH
6. PERCENTAGE OF COMPLEX WORDS
7. FOG INDEX
8. AVG NUMBER OF WORDS PER SENTENCE
9. COMPLEX WORD COUNT
10. WORD COUNT
11. SYLLABLE PER WORD
12. PERSONAL PRONOUNS
13. AVG WORD LENGTH

# **Output Data Structure**

**Output Variables:**

1. All input variables in “Input.xlsx”
2. POSITIVE SCORE
3. NEGATIVE SCORE
4. POLARITY SCORE
5. SUBJECTIVITY SCORE
6. AVG SENTENCE LENGTH
7. PERCENTAGE OF COMPLEX WORDS
8. FOG INDEX
9. AVG NUMBER OF WORDS PER SENTENCE
10. COMPLEX WORD COUNT
11. WORD COUNT
12. SYLLABLE PER WORD
13. PERSONAL PRONOUNS
14. AVG WORD LENGTH

Output data structure spreadsheet for the format of our output, i.e. “Output Data Structure.xlsx”.

a) .py file

b) output in csv or excel file as given in the output structure