**Data Extraction and NLP**

# **Objective**

The objective of this project 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.

When extracting text, I must ensure to extract only the article title and the article text. It should not include the website header, footer, or any other content unrelated to the article text..

**NOTE: USE PYTHON PROGRAMMING TO EXTRACT DATA FROM THE URLs. USE BEATIFULSOUP FOR DATA CRAWLING.**

# **Data Analysis**

For each of the extracted texts from the article, perform textual analysis and compute variables, given in the output structure excel file(text\_analysis\_results.csv).

**NOTE: USE PYTHON PROGRAMMING FOR THE DATA ANALYSIS**

# **Variables**

The definition of each of the variables given in the “Text Analysis(1).docx” file.

Look for these variables in the analysis document (Text Analysis(1).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:**

URL\_ID

URL

POSITIVE SCORE

NEGATIVE SCORE

POLARITY SCORE

SUBJECTIVITY SCORE

AVG SENTENCE LENGTH

PERCENTAGE OF COMPLEX WORDS

FOG INDEX

AVG NUMBER OF WORDS PER SENTENCE

COMPLEX WORD COUNT

WORD COUNT

SYLLABLE PER WORD

PERSONAL PRONOUNS

AVG WORD LENGTH

# **6. Data visualization**