Data Extraction and NLP

Project

1 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.

2 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, please make sure your 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.

NOTE: YOU MUST USE PYTHON PROGRAMMING TO EXTRACT DATA FROM THE URLs. YOU CAN USE BEATIFULSOUP, SELENIUM OR SCRAPY, OR ANY OTHER PYTHON LIBRARIES THAT YOU PREFER FOR DATA CRAWLING.

3 Data Analysis

For each of the extracted texts from the article, perform textual analysis and compute variables, given in the output structure excel file. You need to save the output in the exact order as given in the output structure file, “Output Data Structure.xlsx”

NOTE: YOU MUST USE PYTHON PROGRAMMING FOR THE DATA ANALYSIS

4 Variables

The 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

5 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

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