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.

Installation

List all the libraries and dependencies needed to run the project. Include installation commands if applicable.

Python
pandas
requests
BeautifulSoup
nltk
textblob
Installation command for required Python libraries:

Usage

- Upload 'Input.xlsx' containing the list of URLs to extract text from.
- Run the Python data analytics p1.ipynb
- The script will extract article text from the provided URLs and perform text analysis.
- Output will be saved in the file 'Output.xlsx'.

Project Structure

```
Submission of task1/
|
|── Extracted Files
        |-- All Extracted Files
├── MasterDictionary
        ├── Positive-words
        ├── Negative-words
├── StopWords
        ├── StopWords_Auditor
        ├── StopWords_Currencies
        ├── StopWords_DatesandNumbers
        ├── StopWords_Generic
        ├── StopWords_GenericLong
        ├── StopWords_Geographic
        ├── StopWords_Names
├── data analytics p1.ipynb
├── Input.xlsx
└── Output.xlsx
```

Code Description

task_1py: This is the main Python script that performs the following:

- Extracts article text from provided URLs.
- Analyzes the extracted text for sentiment, complexity, and other metrics.
- Generates an output file ('Output.xlsx') containing analyzed data.
- Output Description
- Describe the output file ('Output.xlsx') and its structure:

'Output.xlsx' contains the following columns:

URL_ID
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