Data Extraction and NLP

Test Assignment

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.

MAIN STEPS TO FOLLOW TO EXTRACT THE TEXT:

- For extraction of data and crawling we need to import the necessary libraries like pandas,textstat,Beautifulsoup etc for text manpulation.
- Textstat library :Textstat is a Python package designed to compute various text statistics. It can be used for a variety of text analysis tasks, including:
- **Basic statistics ,Readability scores, Lexicon analysis ,Text complexity ,Language detection.**
- Important library like pandas,Beautifulsoup etc for data extraction.

2 Data Extraction

Input.xlsx

Get the dataset from the resource for extract the data with help of pandas excel extraction to manipulate the data.

3 Function

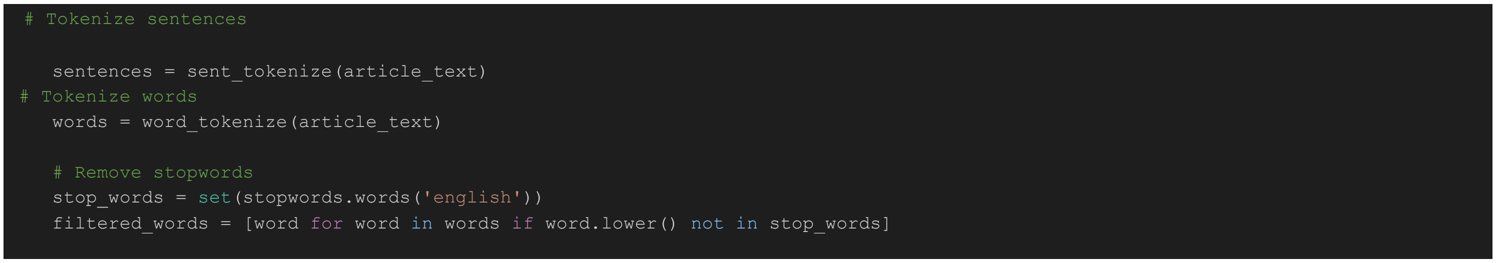
To Create a function to extract the text from the given input given by you.

After extraction text from url make it to enable the different parameter as we required.

4 To create the text function analysis

To generate a code with help of python programming and nltk.To optimize the text to get the different parameter result.With the help of tokenizer to separate a each word and generate the required parameter result.

First sentence tokenizer:

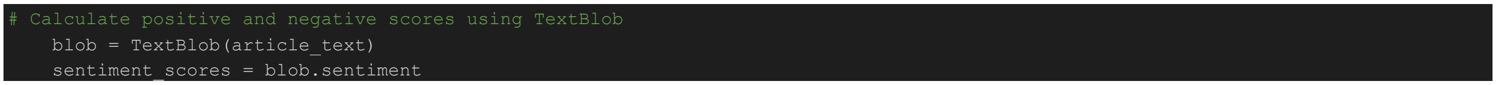


```
 # Tokenize sentences

   sentences = sent_tokenize(article_text)
# Tokenize words
   words = word_tokenize(article_text)

   # Remove stopwords
   stop_words = set(stopwords.words('english'))
   filtered_words = [word for word in words if word.lower() not in stop_words]
```

Second step: To calculate the positive and negative words



```
# Calculate positive and negative scores using TextBlob
    blob = TextBlob(article_text)
    sentiment_scores = blob.sentiment
```

5 Variable :

With the help of tokenizer to token the each word and examine into textblob.After completion of the process textstat generate the result

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

6 Output :

After getting result from the extraction of the data save the file into output file

- First create a output dataframe as our requirement.
- Insert the result data into the output file.
- Save file .py python file.