Blackcoffer

Consulting Website: https://blackcoffer.com | https://lsalead.com/
Web App Products: https://netclan.com/ | https://insights.blackcoffer.com/ | https://hirekingdom.com/ | https://workcroft.com/
Mobile App Products: Netclan | Bwstory

Data Extraction and NLP

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

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 Timeline

6 days, sooner is better.

6 Where to submit

To submit your solution, please fill this google sheet and upload your article to google drive, and share the drive url in the google sheet

Make sure your submission contains:

a) .py file

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

c) instructions

Text Analysis

Objective of this document is to explain methodology adopted to perform text analysis to drive sentimental opinion, sentiment scores, readability, passive words, personal pronouns and etc.

Table of Contents

7 Sentimental Analysis

Sentimental analysis is the process of determining whether a piece of writing is positive, negative, or neutral. The below Algorithm is designed for use in Financial Texts. It consists of steps:

7.1 Cleaning using Stop Words Lists

The Stop Words Lists (found in the folder StopWords) are used to clean the text so that Sentiment Analysis can be performed by excluding the words found in Stop Words List.

7.2 Creating a dictionary of Positive and Negative words

The Master Dictionary (found in the folder MasterDictionary) is used for creating a dictionary of Positive and Negative words. We add only those words in the dictionary if they are not found in the Stop Words Lists.

7.3 Extracting Derived variables

We convert the text into a list of tokens using the nltk tokenize module and use these tokens to calculate the 4 variables described below:

Positive Score: This score is calculated by assigning the value of +1 for each word if found in the Positive Dictionary and then adding up all the values.

Negative Score: This score is calculated by assigning the value of -1 for each word if found in the Negative Dictionary and then adding up all the values. We multiply the score with -1 so that the score is a positive number.

Polarity Score: This is the score that determines if a given text is positive or negative in nature. It is calculated by using the formula:

Polarity Score = (Positive Score – Negative Score)/ ((Positive Score + Negative Score) + 0.000001)

Range is from -1 to +1

Subjectivity Score: This is the score that determines if a given text is objective or subjective. It is calculated by using the formula:

Subjectivity Score = (Positive Score + Negative Score)/ ((Total Words after cleaning) + 0.000001)

Range is from 0 to +1

8 Analysis of Readability

Analysis of Readability is calculated using the Gunning Fox index formula described below.

Average Sentence Length = the number of words / the number of sentences

Percentage of Complex words = the number of complex words / the number of words

Fog Index = 0.4 * (Average Sentence Length + Percentage of Complex words)

9 Average Number of Words Per Sentence

The formula for calculating is:

Average Number of Words Per Sentence = the total number of words / the total number of sentences

10 Complex Word Count

Complex words are words in the text that contain more than two syllables.

11 Word Count

We count the total **cleaned** words present in the text by

1. removing the stop words (using stopwords class of nltk package).
2. removing any punctuations like ? ! , . from the word before counting.

12 Syllable Count Per Word

We count the number of Syllables in each word of the text by counting the vowels present in each word. We also handle some exceptions like words ending with "es","ed" by not counting them as a syllable.

13 Personal Pronouns

To calculate Personal Pronouns mentioned in the text, we use regex to find the counts of the words - “I,” “we,” “my,” “ours,” and “us”. Special care is taken so that the country name US is not included in the list.

14 Average Word Length

Average Word Length is calculated by the formula:

Sum of the total number of characters in each word/Total number of words