Text Insight: Extracting Knowledge from Textual Data

1 Problem Statement:

The goal of this project is to analyze articles from given website links and extract useful information from them. This involves figuring out whether the articles are positive, negative, or neutral in sentiment, understanding how easy or difficult they are to read, and identifying key linguistic features.

Data Extraction :

We had a input.xlsx file that contains URLS of the articles. For each of the articles, given extract the article text and save the extracted article in a text file with URL_ID as its file name.

Python programming, along with libraries like BeautifulSoup are used for data extraction. Each URL from the Input.xlsx file was processed to extract article titles and text, which were then saved into separate text files.

Sentiment Analysis :

Introduction:

Sentiment analysis is a technique used to determine the emotional tone of a piece of text, whether it's positive, negative, or neutral. We'll discuss the process of sentiment analysis applied specifically to financial texts. The analysis involves several steps aimed at cleaning the text, creating dictionaries of positive and negative words, and deriving variables to quantify sentiment.

Steps in Sentiment Analysis:

1. Cleaning using Stop Words Lists:

Stop words are common words like "the," "is," and "and" that are often filtered out because they don't carry significant meaning. In our analysis, we cleaned the text by removing these stop words using predefined lists.

2. Creating a Dictionary of Positive and Negative Words:

We utilized a master dictionary containing positive and negative words to create separate dictionaries for positive and negative sentiments. Only words not found in the stop words lists were added to these dictionaries.

3. Extracting Derived Variables:

After cleaning the text and creating dictionaries, we converted the text into tokens using the NLTK tokenize module. These tokens were then used to calculate four key variables:

Positive Score: This score is determined by assigning a value of +1 for each positive word found in the text and summing up all these values.

Negative Score: Similarly, this score is calculated by assigning a value of 1 for each negative word and summing up all these values. The final score is made positive by multiplying it with 1.

Polarity Score: This score indicates whether the text is overall positive or negative. It is calculated using a formula that considers both positive and negative scores, with a range from 1 to +1.

Subjectivity Score: This score indicates the subjectivity of the text, whether it's more objective or subjective. It is calculated by considering the total count of positive and negative words relative to the total number of words after cleaning, with a range from 0 to +1.

Sentiment analysis provides valuable insights into the emotional tone of financial texts, helping to understand whether they convey positive or negative sentiments. By following the outlined steps, we were able to systematically analyze text data and derive meaningful sentiment scores. This information can be used for various applications such as financial market analysis, customer feedback analysis, and sentimentdriven decisionmaking processes.

Variables that need to be calculate for each article:

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

Average Sentence Length:

Formula: Average Sentence Length = Total number of words / Total number of sentences

Explanation: It calculates how long, on average, each sentence is by dividing the total number of words by the total number of sentences in the text.

Percentage of Complex Words:

Formula: Percentage of Complex Words = (Number of complex words / Total number of words) * 100

Explanation: It measures the proportion of words in the text that are considered complex, typically those with more than two syllables, by dividing the number of complex words by the total number of words and multiplying by 100 to get a percentage.

Fog Index:

Formula: Fog Index = 0.4 * (Average Sentence Length + Percentage of Complex Words)

Explanation: The Fog Index is a readability metric that estimates the years of formal education needed to understand a piece of text. It is calculated by adding the average sentence length and the percentage of complex words, then multiplying by 0.4.

Average Number of Words Per Sentence:

Formula: Average Number of Words Per Sentence = Total number of words / Total number of sentences

Explanation: Similar to Average Sentence Length, it calculates the average number of words in each sentence by dividing the total number of words by the total number of sentences.

Complex Word Count:

Explanation: Complex words are words in the text that contain more than two syllables. The count simply involves identifying and counting these words.

Word Count:

Explanation: It counts the total number of words in the text after removing stop words and punctuation marks.

Syllable Count Per Word:

Explanation: It calculates the number of syllables in each word of the text by counting the vowels present in each word. Some exceptions, like words ending with "es" or "ed," are handled by not counting them as syllables.

Personal Pronouns:

Explanation: Personal pronouns such as "I," "we," "my," "ours," and "us" are counted to gauge the level of personalization in the text. Care is taken to exclude instances like the country name "US" from the count.

Average Word Length:

Formula: Average Word Length = Total number of characters in each word / Total number of words

Explanation: It calculates the average length of words in the text by dividing the total number of characters in all words by the total number of words.