Analysis of Word Frequency in Text Data

Sobhan Behuria

February 14, 2024

1 Objective

The objective of this analysis is to determine the frequency of each word in a given text dataset. The text data has been extracted from a file and preprocessed to extract individual words. The frequency of each word will provide insights into the usage pattern and importance of words in the text corpus.

2 Methodology

2.1 Data Collection

- The text data is loaded from a file stored in Google Drive and Databricks File System (DBFS).
- The data is then converted into an RDD (Resilient Distributed Dataset) in PySpark for further processing.

2.2 Data Preprocessing

- Regular expressions are used to split the text into individual words based on specific patterns, such as spaces, punctuation marks, and special characters.
- All words are converted to lowercase to ensure case-insensitive counting.

2.3 Word Counting

- Each word is mapped to a tuple containing the word itself and an initial count of 1.
- The RDD transformation operations are performed to count the occurrences of each word.

2.4 Result Analysis

- The total number of words in the dataset is calculated.
- The list of words and their counts is sorted in descending order based on their frequency.
- The top words with the highest frequencies are identified and presented.

3 Results

3.1 Total Number of Words

328,091

3.2 Top 10 Words by Frequency

1. 'sed': 7,575
2. 'in': 6,438
3. 'amet': 6,174
4. 'sit': 6,103
5. 'ut': 5,200
6. 'id': 5,198
7. 'eget': 5,024
8. 'et': 4,667
9. 'nunc': 4,613
10. 'vitae': 4,528

4 Conclusion

The analysis provides valuable insights into the distribution of words within the text corpus. The findings can be utilized for various purposes such as text summarization, keyword extraction, and sentiment analysis. Further analysis or visualization techniques can be applied to gain deeper insights into the text data. Additionally, the process can be scaled for larger datasets using distributed computing frameworks like Apache Spark for efficient processing.