| **INDEX**   | **S.No** | **Title** | **Page No.** | | --- | --- | --- | | 1 | Dataset Description | 2 | | 2 | Execution Process of MapReduce tasks | 3 | | 3 | Word Count using MapReduce | 4 | | 4 | Word Mean using MapReduce | 5 | | 5 | Word Median using MapReduce | 6 | | 6 | Word Standard Deviation using MapReduce | 7 |   **DATASET DESCRIPTION:**  *sentimentdataset.csv*    This dataset is a collection of social media posts, which can be leveraged for sentiment analysis and other types of social media analytics. It includes various fields that capture the content and metadata of each post. Below is a detailed description of each column in the dataset:   1. **Unnamed: 0:** This column appears to be an index column, likely automatically generated when the dataset was saved to a CSV file. 2. **Text:** The content of the social media post. 3. **Sentiment:** The sentiment label assigned to the post, indicating whether the sentiment is Positive, Negative, or Neutral. 4. **Timestamp:** The date and time when the post was created. 5. **User:** The username of the individual who posted the content. 6. **Platform:** The social media platform where the post was made (e.g., Twitter, Facebook, Instagram). 7. **Hashtags:** A list of hashtags included in the post. 8. **Retweets:** The number of times the post was retweeted (for platforms where this is applicable). 9. **Likes:** The number of likes the post received.   **EXECUTION PROCESS OF MAPREDUCE IN HADOOP USING VIRTUAL ENVIRONMENT:** *Transferring Dataset from Windows to Linux*  1. Open FileZilla and connect to your Linux environment using IP address, username, password and port of Linux environment.   The IP address of the Linux environment can be obtained using the *$ifconfig* command.   1. Navigate to the directory on your Windows machine where the dataset is located. 2. Transfer the dataset to a directory on your Linux Virtual environment using drag and drop.  *Transferring Dataset in Linux*  1. Create a local directory on Linux to store the dataset.   A directory in Linux is created using the *mkdir <directory\_name*> command.   1. Move the dataset (sentimentdataset.csv) to the local directory.  *Transferring the Dataset from Linux Local environment to HDFS*  1. Create a directory in HDFS to store the dataset.   A directory in Hadoop is created using the *hadoop fs -mkdir <hadoop\_dir>* command.   1. Move the dataset from the local directory on Linux to the HDFS directory.   Open the directory in Linux where the file is located and with the help of  *hadoop fs -copyFromLocal <localsrc> <hadoop\_dst>* command. *Executing MapReduce* Run the MapReduce tasks using the JAR files, specifying the input dataset in HDFS and the desired output directories in HDFS. Use the following command to execute MapReduce tasks in HDFS,  *hadoop jar <jar-file> <main-class> <hadoop\_dst/> <mapreduc\_output>*  The main classes of jar file used for MapReduce tasks are,   * wordcount * wordmean * wordmedian * wordstandarddeviation  *Viewing Outputs of Analysis* Check the output directories in HDFS to review the results of the MapReduce tasks.  **WORD COUNT USING MAPREDUCE:**  **Command:**  hadoop jar hadoop-mapreduce-examples-2.2.0.jar wordcount /SA\_data/sentimentdataset.csv  /SA\_Analysis/wordcount  **In terminal:**    **Output File:**    **Analysis:**  The output from the MapReduce word count process shows a list of words and their respective counts from the *sentimentdataset.csv* file. Some of the conclusions from the word count analysis are,  1. Frequency of words:  The analysis lists various words and their frequency of occurrence in the dataset.  2. Diversity of words:  The dataset contains a diverse range of words, indicating a variety of topics or sentiments covered  .  3. Sentiment Indicators:  Words like "confidence," "compassion," and "community" indicate positive sentiments, while words like "confusion" and "conflict" indicate negative sentiments.  **WORD MEAN USING MAPREDUCE:**  **Command:**  hadoop jar hadoop-mapreduce-examples-2.2.0.jar wordmean /SA\_data/sentimentdataset.csv  /SA\_Analysis/wordmean  **In terminal:**    **Output File:**    **Analysis:**  Based on the output from the MapReduce word mean process, the conclusions drawn are,  1. Sentiment Distribution:  The analysis processed a large volume of data, with a total count of **30,293,679** entries and an aggregate length of **207,919,202** characters.  2. Mean direction:  The calculated mean sentiment value is approximately **6.86**, indicating that the average sentiment expressed in the dataset is towards the positive end.  **WORD MEDIAN USING MAPREDUCE:**  **Command:**  hadoop jar hadoop-mapreduce-examples-2.2.0.jar wordmedian /SA\_data/sentimentdataset.csv  /SA\_Analysis/wordmedian  **In terminal:**    **Output File:**    **Analysis:**  Based on the output from the MapReduce word median process, the conclusions drawn are,  1. Sentiment Distribution:  The data exhibits a balanced range of sentiment values, with the median sentiment being **4**. This indicates that the sentiments in the dataset are evenly distributed, slightly towards neutrality.  2. Median variability:  The sentiment values range widely, but the majority of data points cluster around the median value  of 4.  **WORD STANDARD DEVIATION USING MAPREDUCE:**  **Command:**  hadoop jar hadoop-mapreduce-examples-2.2.0.jar wordstandarddeviation /SA\_data/sentimentdataset.csv /SA\_Analysis/wordstandarddeviation  **In terminal:**    **Output File:**    **Analysis:**  Based on the output from the MapReduce word standard deviation process, the conclusions drawn are,  1. NaN Standard Deviation:  The final values include a word count of 30,293,679, a cumulative length of words amounting to 207,919,202, and a summed square of word lengths at -482,469,106. The negative summed square value raises concerns about possible errors in the computation process. |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |