IT 609 - Big Data Processing Lab 4

Teaching Assistant - Saran Pandian P

March 9, 2023

1 Aim

- To configure PySpark in Windows
- learn about PySpark RDD and DataFrame API and how to use them to manipulate the data.

2 Problem description

Spark provides us with the functionality of processing large amount of data by using distributed computing. It abstracts the distributed data by providing API such as RDD and DataFrame. In this assignment you will work on creating and manipulating the RDDs and DataFrames, which is a prerequisite for all the machine learning and data science problems.

3 Implementation

3.1 Dataset

- For RDD task 2 you will be using the following text data. dataset here
- For the DataFrame task you will use following file. dataset here

3.2 Exercise

- RDD Tasks

  – TASK-1: Generate 100 random numbers in range 0 to 10 using numpy randint function with the seed set to 10. Create a RDD using the parallelize function using data generated in previous step. Calculate the frequency of each number (0 - 10) using appropriate function of RDD.

  – TASK-2: In this task you will calculate the frequency of each word in text8 dataset mentioned above. Create a RDD using the text8 dataset. Use appropriate functions of the RDD to get the word frequencies. Filter the RDD using appropriate function to get the frequencies of words containing the letter 'a'.

- DataFrame Task: Create a Spark dataframe using the iris json data mentioned above. Calculate Pearson Correlation between the columns petalLength and petalWidth using the appropriate dataframe API. Show the columns sepalLength, sepalWidth and species for the rows of data that has petalLength greater than or equal to 1.4 using the appropriate dataframe API.

4 Submission guidelines

You have to submit two files

- You have to submit the pdf file containing instructions for configuring PySpark in Windows.
- You have to submit your assignment in python notebook with proper comments and explanation. (Try implementing in colab notebook first then try to configure pyspark in local system)
- print the dataframes as output especially the tweets as well the corresponding sentiments.

5 References

- This is for configuring pyspark in windows
- PySpark RDD
- PySpark DataFrame
- Numpy Seed
- Numpy Random Integer Generation