DSCI 5350 Spark-Activity2 Satyasriharsha Buddha

1)Create Scala notebook in Wordcount folder

A screenshot of a computer

Description automatically generated

2) Upload Wordcount\_Activity2 text file and copy the file path

A screenshot of a computer

Description automatically generated with medium confidence

Filepath:

A screenshot of a computer

Description automatically generated with medium confidence

3)Display 600 bytes of the file

A screenshot of a email

Description automatically generated with low confidence

4)Read the file into spark context as an RDD of strings

A screenshot of a computer

Description automatically generated with low confidence

5) Perform some operations in RDD

Count-counts the number of elements in the RDD i.e., number of lines in the text file

take or collect- Displays the contents of RDD

A screenshot of a computer

Description automatically generated with medium confidence

A close-up of a message

Description automatically generated with low confidence

6) Using .cache() in RDD, so that it will exist in memory after first action

A screenshot of a computer

Description automatically generated with low confidence

7)Transforming lines to words

A picture containing text, screenshot, algebra, receipt

Description automatically generated

8) Counting the words

flatMap Break each line by the white space character “ “ and find the words

Map word=>(word,1) to initialize each word with integer count 1 and transforming each word into

(Key, value) pair

reduceByKey Count all values with same key.

Collect to display the result.

A screenshot of a computer code

Description automatically generated with low confidence

Some words have punctuation marks in the end which means same words are counted as different words. We can use regular expression to remove those marks

A screenshot of a computer

Description automatically generated with medium confidence

9) Sorting in descending order

