Batyr Charyyev

CS691 - Data Intensive Computing

HW2

Questions.

1. Launch the Spark shell. (2pt)

Answer:

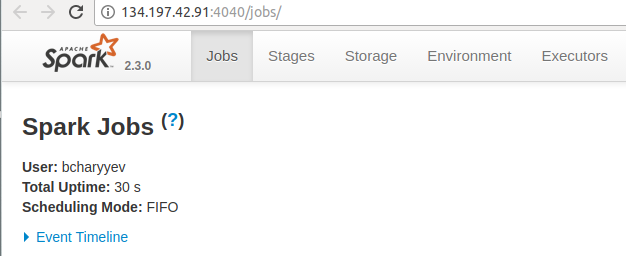
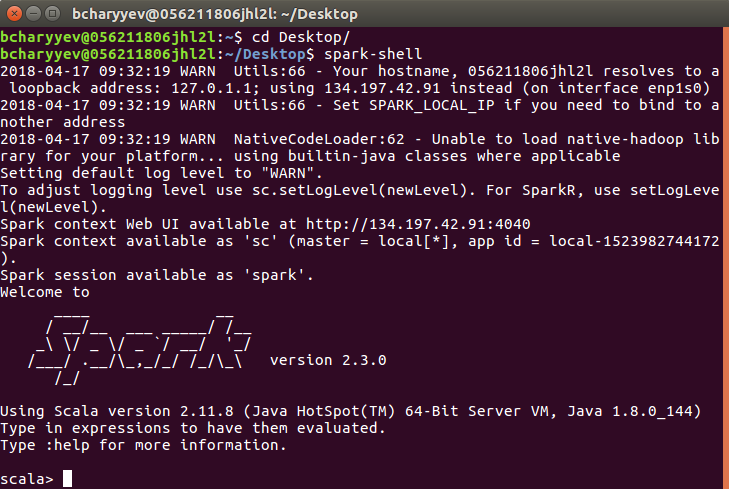


Figure 2. Web-UI

Figure 1.

2. Make a parallel collection of Array(1, 2, 3, 4, 5)  and sum up all its elements. (2pt)

Answer:

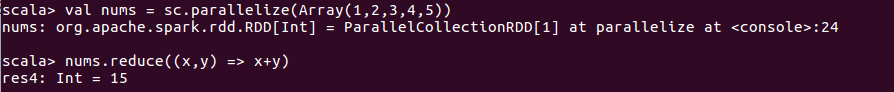


Figure 3. Implementation with Reduce

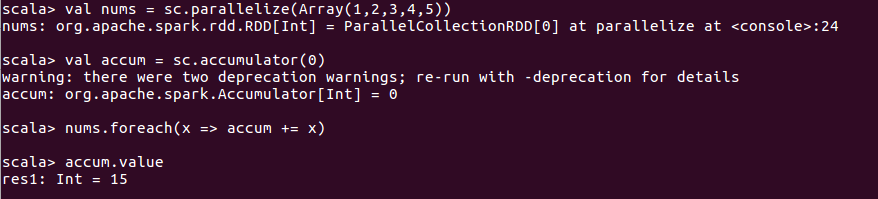


Figure 4. Implementation with accumulator

3. Create an RDD named pagecounts from the input file hamlet (3pt)

Answer:

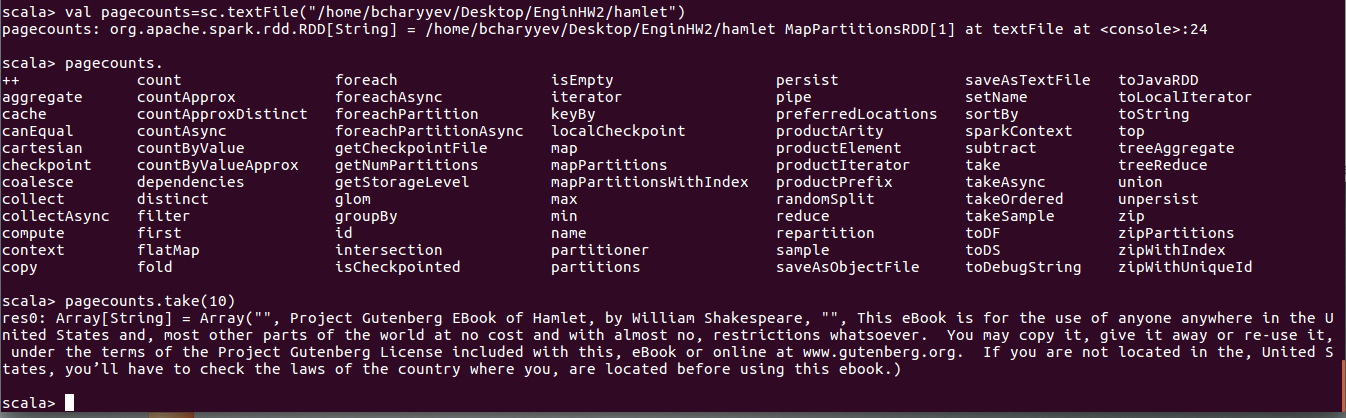


Figure 5. Pagecount and first 10-line.

4. Get the first 10 lines of hamlet (i.e., first 10 records of pagecounts). (3pt)

Answer: Provided on Figure 5.

5. Make a more readable print of the step 4. (3pt)

Answer:

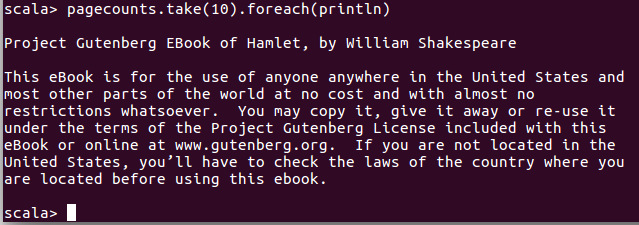


Figure 6. More readable format

6. Count the total records in the data set pagecounts, and confirm its correctness by comparing the result with the Bash wc command: wc -l hamlet . (3pt)

Answer:

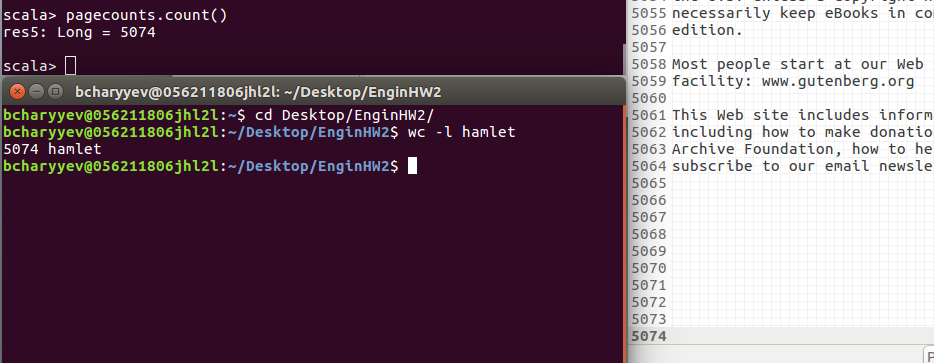


Figure 7. Count

7. Filter the data set pagecounts and return the items that have the word “this”. (5pt)

Answer: Screenshot of only few lines provided

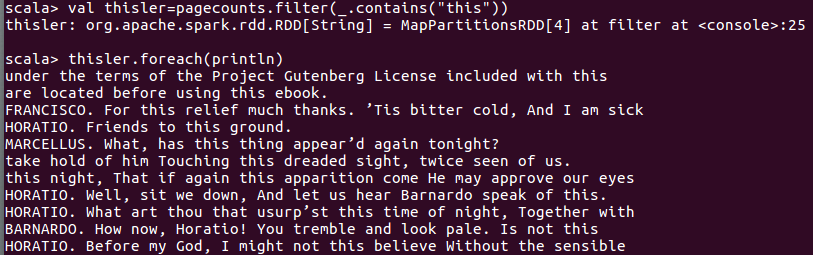
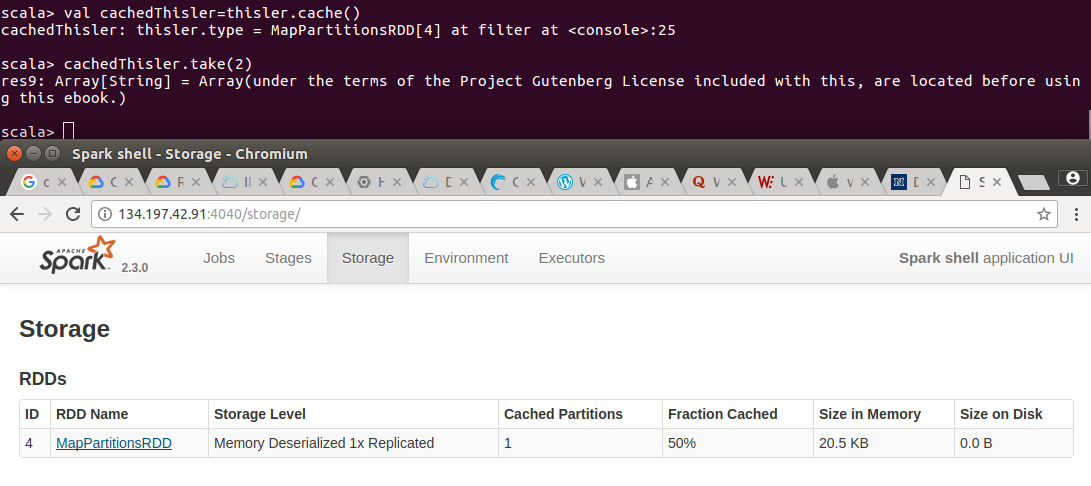


Figure 8. Filter

8. Cache the new data set in memory, to avoid reading from disks. Show cached RDD in web interface (5pt)



Answer:

Figure 9. Cache

9. Find 5 lines with the most number of words. Print lines and the number of words(6pt)

Answer:

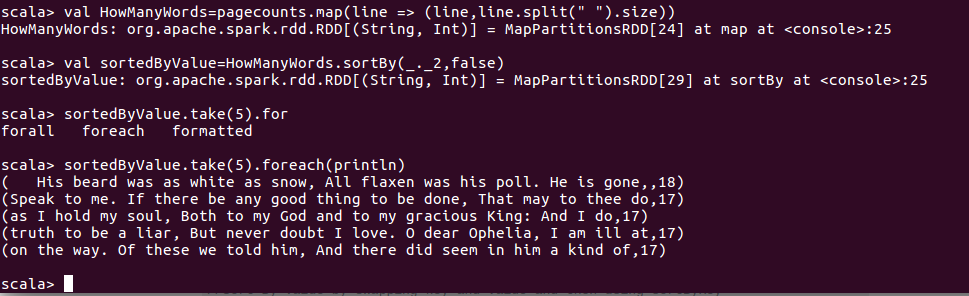
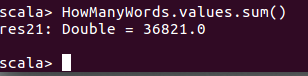


Figure 10. Top 5 lines with most word

10. Count the total number words. (3pt)



Answer: I used HowManyWords RDD from question 10. In each entry values are number of words in that line so I just add up all values.

Figure 11. Total number of words

11. Count the number of unique words. (5pt)

Answer: Here I provided both split by space and split by words. When we split by words we get more accurate results.

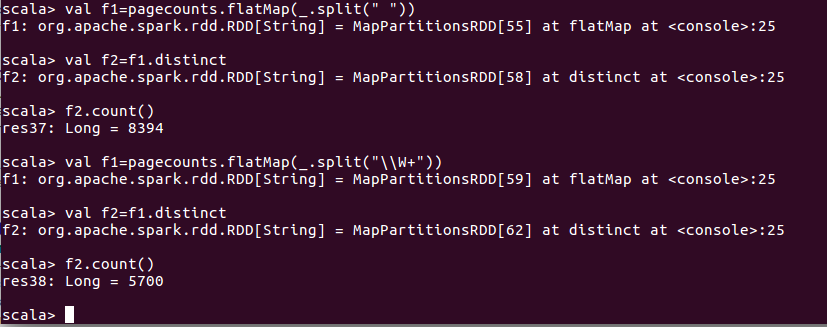


Figure 12. Unique words

12. Count the number of each word. (10pt)

Answer: I provided only 10 lines of output.

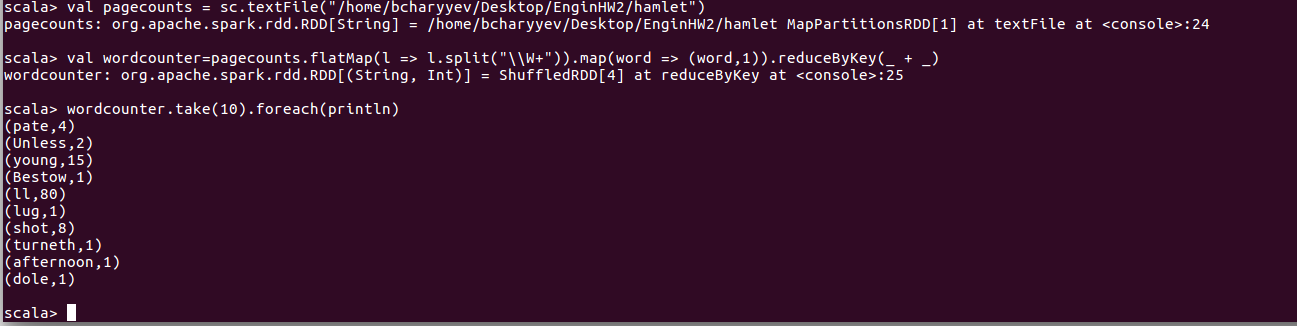


Figure 13. Number of each word

13. Show the jobs for Q12 in web interface (3pt)

Answer: Figure-14 shows job created, Figure-15 shows detailed information of that job. As you can see it provides transformations and actions performed to obtain result.

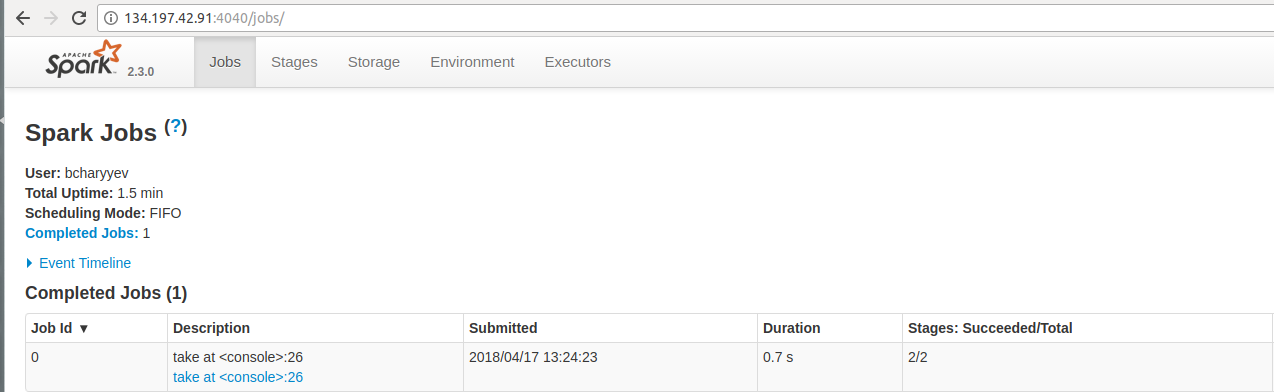


Figure 14. Job created on Web-UI

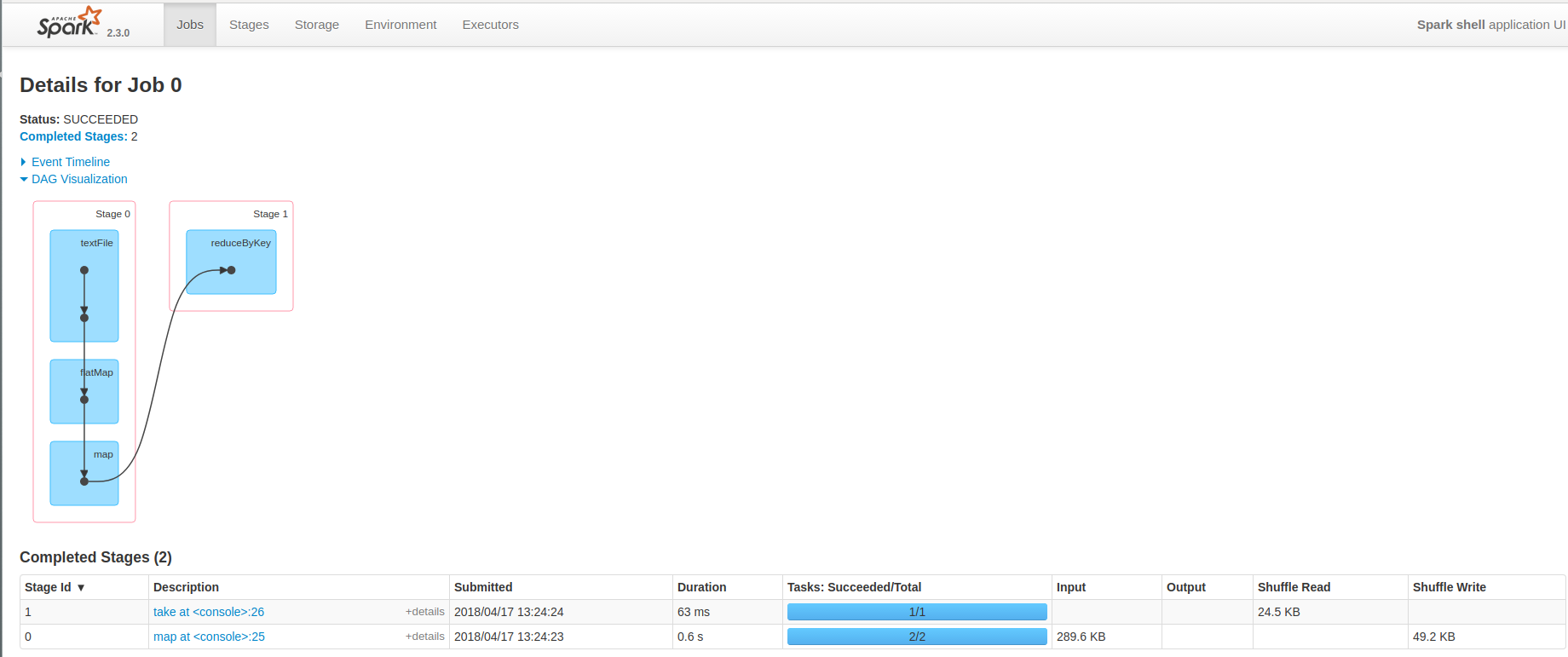


Figure 15. Job created on Web-UI detailed

14. Save the data set in a text file. (3pt)

Answer: We used command on Figure-16 to save data set in text file. You can find saved data set in “wordcountersaved” folder.



Figure 16. saveAsTextFile

15. Collect the word counts in the shell. (4pt)

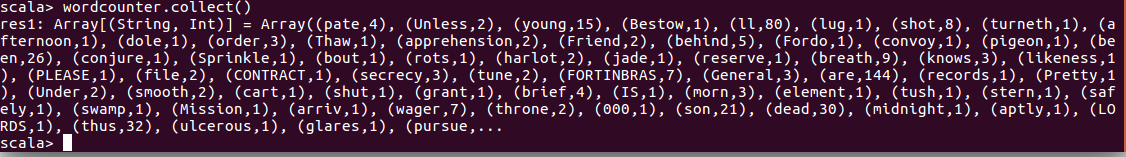
Answer:

Figure 17. Collect