PARALLEL AND DISTRIBUTED COMPUTING

ASSIGNMENT 3

SUBMITTED BY:

SANA ALI KHAN

18I-0439

CS-D

SUBMITTED TO:

SIR EHTESHAM ZAHOOR

INTRODUCTION

The assignment purpose was to implement two problems using Hadoop, and taking advantage of it and its MapReduce to execute the program rapidly, even with a large dataset. I have used pseudo-distributed Hadoop and hdfs to run my code.

All the requirements have been implemented as specified.

QUESTION 2

This question required us to find the average length of all the comments in the input files. This was done in the following steps:

- The mapper class' map method is called on each line of the file. Every line is parsed to extract the rowId and the comment, and key-value pair of (rowId, comment length) is passed to the reducer.
- Reducer class' reduce method – for every key-value pair – gets the length of every comment and adds it to the total sum
- Reducer class' cleanup method is called when the reducer has finished its work – this method calculates the average length (in characters!) and writes it to the output file.

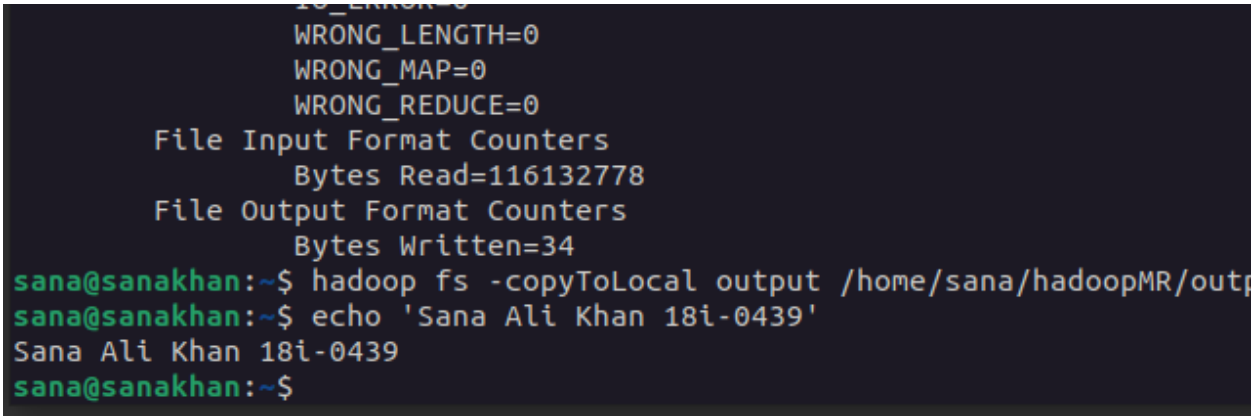

Figure 1: Executing and copying output of avgLength.jar

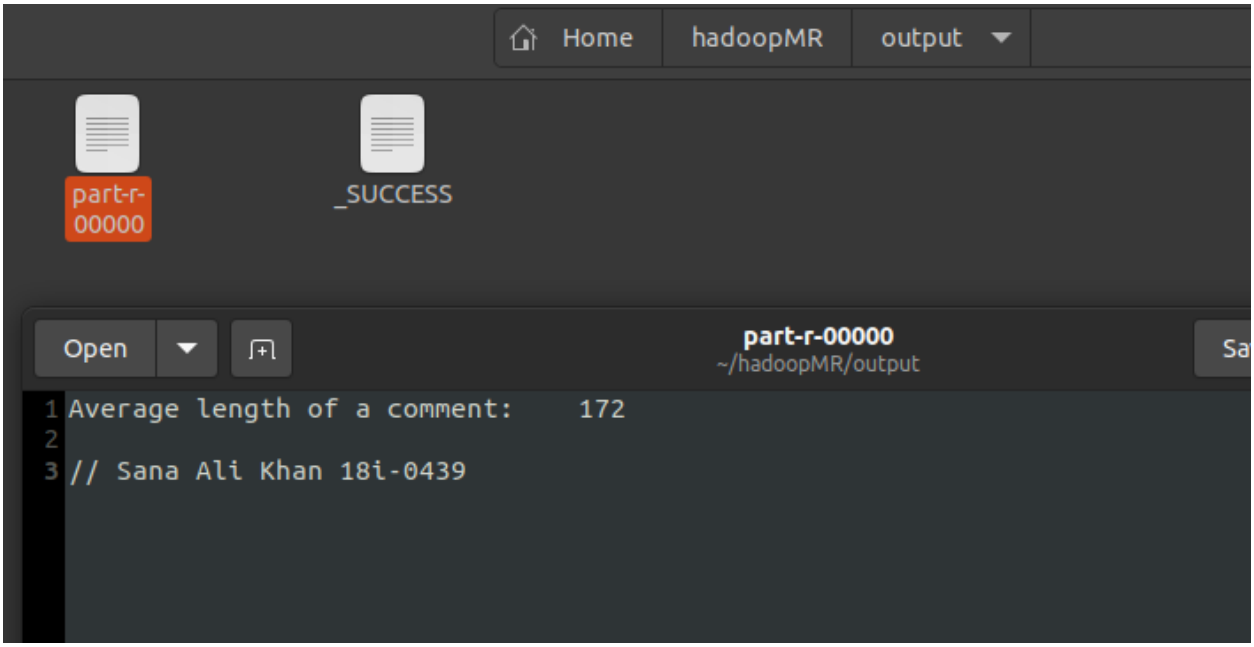

Figure 2: Output of running avgLength.jar

QUESTION 3

This question required us to identify the overall sentiment of the comments including a particular keyword. This was done as follows:

- Keyword is passed as the third argument to the program (first and second are input/output paths)

- The driver class (with the main method) is responsible for storing variables such as the chosen keyword, a list of positive words and a list of negative words. These variables are static so they can be accessed from outside this class without requiring an instantiation of it.
- Mapper class' map method receives every line in the input files, and checks if they contain a comment and checks if this comment contains the specified keyword.
- If it does contain it, then it is split into words. If a word is found in the positive-words list, then key-value pair of comment-score is passed to reducer.
- A score of '1' indicates a positive word and a score of '-1' indicates a negative word.
- Reducer combines the keys, so for every comment a corresponding array(iterators) is generated, with each value in the array signifying the score of one word in the comment.
- For every comment, the scores are summed up and an overall comment score is calculated.
- This comment score (positive/negative/neutral) is used to indicate the sentiment of the comment, and the comment and its sentiment are written to the context (aka output file)
- For every positive or negative comment, the overall score is incremented or decremented respectively.

- Reducer class' cleanup method is called when the reducer has finished its work – this method uses the overall score to determine the overall usage of the keyword, and this is written to output as well.

```
sana@sanakhan:~$ echo 'Sana Ali Khan 18i-0439'
Sana Ali Khan 18i-0439
sana@sanakhan:~$ hadoop fs -copyFromLocal /home/sana/hadoopMR/input2 input2
sana@sanakhan:~$ hadoop jar ~/hadoopMR/keyword.jar classifyKeyword input2 output2 iBooks
```

Figure 3: Executing keyword.jar

```
Comment 146: iBooks has a dark mode, but you can't read PDFs directly in iBooks on Mac. It just opens them in Preview.
Comment Sentiment: Negative as most words are found in the bag of negative words.

Comment 147: it might be worth asking on the new site http://ebooks.stackexchange.com as there will be more people knowing ePub and hopefully iBooks
Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 148: note you need to ping me.. I don't need to ping you on your post. :) Also &amp;quot;*can you also see pdfs that are just exports from Safari page prints to iBooks?*&amp;quot; I don't think I've used this feature. I download them on iOS safari, and open them in iBooks. Then its synced across devices and boom it's on Mac and I can extract it ;) . https://support.apple.com/guide/books/import-books-or-pdfs-ibkseed72068/1.16/mac/10.14.6
Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 149: we were an outsourcer, the client had a mdm installed AFTER we did the initial load. the client should have installed ibooks through the mdm, we told him that but he insisted in having iBooks right away
Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 150: with the newest version of iBooks, everything is now syncing over the air - i just open any ePub (with iBook as default application) and it's imported to _~/Library/Containers/com.apple.BKAgentService/Data/-Documents/iBooks/Books_. It then just syncs... My problem (and the reason it didn't work for me initially) was that iTunes had changed it's syncing preferences from all to only ticked items, with older, non-iBookStore ePubs unticked...
Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 151: yea, i know that we can add pdf files to iBooks, but it wud look like a pdf and letters would be too small to read :)
Comment Sentiment: Positive as most words are found in the bag of positive words.

Keyword: iBooks overall usage is positive   (Sana Ali Khan 18i-0439)
```

Figure 4: Output of running keyword.jar with keyword 'iBooks'

```
Comment 1041: unfortunately I was having difficulty finding a free to use website blocker but yesterday I found ColdTurkey which seems to do most of what I required.  The way in which the mac relies on safari means that it is not easy and is dangerous to remove safari so I agree with you it is not viable but it is certainly worth a question.  In fact I did this in Windows and removed Internet Explorer and only had Firefox and Chrome installed, this worked well.
Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 1042: well, thank again, I was aware of the osxwinebuilder but that is a quite difficult process for me to perform. I am a beginner when it comes to compiling and all this unix related stuff...I would appreciate if you could provide me with a more valuable answer.
Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 1043: with the well design UNIX as a foundation I would suppose such control is possible... it shouldn't be something that difficult to implement... the JetDrive is quite nice. It is even better than the low profile USB drive, because something is still sticking out for the low profile USB drive.  The JetDrive can be totally inserted.  Just that with my USB drive, the Macbook Air seems to run out of power quite a bit sooner.
Comment Sentiment: Positive as most words are found in the bag of positive words.

Comment 1044: yes, I know stackoverflow rules but in this link, they showed via images and it was difficult to write everything in words. So, I shared link only.
Comment Sentiment: Negative as most words are found in the bag of negative words.

Comment 1045: "There's no general mechanism to 'side-load" apps into iOS" I don't know whether you count XCode and the developer tools? I don't know how difficult it is to do so  without having the source code however.
Comment Sentiment: Negative as most words are found in the bag of negative words.


Keyword: difficult overall usage is negative (Sana Ali Khan 18i-0439)
```

Figure 5: Output of keyword.jar with keyword 'difficult'