CPSC 4330 Big Data Analytics

In-class Exercise 9 – Use RDDs to transform a Dataset

In this exercise, you will practice using RDDs in the Spark Shell. Specially, you will use Spark to explore the Loudacre (a mobile phone service provider) web server logs.

Copy your data to HDFS

1. In a docker terminal window, enter

```
# hadoop fs -mkdir  /loudacre
```

This command is to create a directory "loudacre" on HDFS.

2. Download the file "weblogs.zip" from Canvas and unzip it. The unzipped folder "weblogs" contains many log files.

Copy the folder "weblogs" to Docker (e.g. docker cp /Users/linli/Documents/data/weblogs bda:/hadoop-data/)

3. On Docker, enter into the directory where the folder "weblogs" locates (e.g. hadoop-data), then type

```
# hadoop fs -put weblogs  /loudacre/
```

to put the log files on HDFS.

4. Use the HDFS command line to review one of the log files in the HDFS /loudacre/weblogs directory, e.g. 2014-03-15.log

```
# hadoop  fs  -cat  /loudacre/weblogs/2014-03-15.log
```

Note the format of the lines, e.g.

Explore the Loudacre web log files

1. In a terminal window, start the pyspark shell:

```
# pyspark
```

2. Set a variable for the data file so you do not have to retype it each time.

```
>>>logfile = "/loudacre/weblogs/201*.log"
```

201*.log means all the files which name start with "201" and end with ".log".

3. Write code to create an RDD from the data file.

4. Write code to create an RDD containing only those lines that are requests for JPG files. (Hints: You can use filter)

5. Write code to view the first 5 lines of the data using take.

6. To print out the results of step 5 nicely, you can use a loop. Below is an example, where "jpglogs" is a RDD.

```
>>> for  line  in jpglogs.take(5):
        print(line)
```

7. Sometimes you do not need to store intermediate objects in a variable, in which case you can combine the steps into a single line of code. Combine the code of step 3 and 4 into a single line of code to count the number of JPG requests.

8. Write code to create an RDD containing the length of each line in the log file. (Hints: You can use map). To check the result, you can write code to view the result (e.g. first 5 lines of the result) – your code will produce an array of five integers corresponding to the first five lines in the file.

9. Write code to create an RDD of arrays of words on each line. That is to say, for each line, map to an array of words in that line. Again, to check the result, you can write code to view the result (e.g. the first line) – your code will produce an array containing the words in the first line.

10. Now that you know how map works, write code to create a new RDD containing just the IP addresses from each line in the log file. (The IP address is the first "word" in each line).

11. Write a loop (similar to step 6) to print out the first 5 IP addresses from the result of step 10.

12. Finally, save the list of IP addresses as a text file to HDFS, e.g. /loudacre/iplist. (Hints: You can use the function saveAsTextFile() and pass the directory as the argument.) You then should see multiple output files under the directory /loudacre/iplist.

13. You can exit the shell by typing exit().