Sami Emre Erdogan
2019280513

HW10 Spark Streaming

To run this assignment, I have followed the instructions and just modified the "SimpleApp.py" there were a lot of references online to look from and this assignment wasn't that hard since the skeleton code was already given. We just had to add a couple of lines of code in order to make it work.

The code I had was simply splitting the lines and adding value of 1, It uses "reduceByKey" to combine the values with same word and at the end, it updated the word count looking to our previous values which is in our case was our history. Of course, it was supposed to look for the top 100 but that would take a long time for the Linux server to compute.

I had to "generate.sh" to generate the words. After that, once I have modified the code and used ./submit.sh to get the results. The terminal was accurately calculating. Here are some screenshots:

```
Time: 2020-05-29 10:11:00
-------------------------------------------
('i24', 8)
('l26', 7)
('k20', 6)
('o38', 6)
('k15', 6)
('o17', 5)
```

```
-------------------------------------------
Time: 2020-05-29 10:12:00
-------------------------------------------
('m12', 12)
('i24', 11)
('m19', 11)
('m34', 10)
('l26', 10)
('l16', 10)
```

```
-------------------------------------------
Time: 2020-05-29 10:13:00
-------------------------------------------
('i24', 15)
('m1', 14)
('l24', 13)
('m12', 13)
('m34', 13)
('l13', 13)
```

```
-------------------------------------------
Time: 2020-05-29 10:14:00
-------------------------------------------
('l29', 19)
('m10', 18)
('m12', 17)
('m1', 17)
('l13', 17)
('o2', 17)
('o0', 17)
```

```
Time: 2020-05-29 10:15:00
-------------------------------------------
('l29', 22)
('o27', 21)
('m30', 21)
('m34', 21)
('l13', 21)
('o2', 21)
```

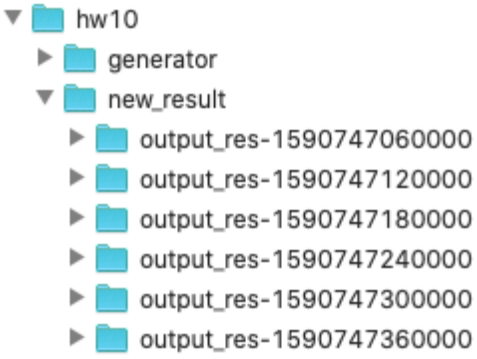
```
-------------------------------------------
Time: 2020-05-29 10:16:00
-------------------------------------------
('l13', 27)
('l29', 24)
('m30', 23)
('m34', 23)
('k34', 23)
('l24', 22)
```

After confirming my results were stored in the hdfs directory by checking with:

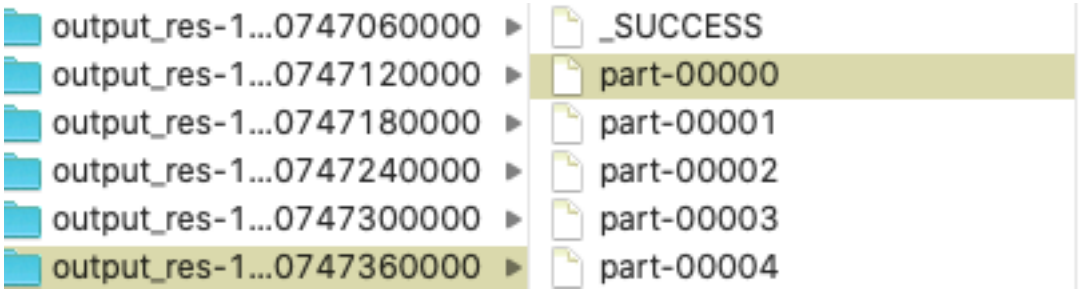
```
hdfs dfs -ls:
drwxr-xr-x   - 2019280513 2019280513          0 2020-05-29 10:11 output_res-1590747060000
drwxr-xr-x   - 2019280513 2019280513          0 2020-05-29 10:12 output_res-1590747120000
drwxr-xr-x   - 2019280513 2019280513          0 2020-05-29 10:13 output_res-1590747180000
drwxr-xr-x   - 2019280513 2019280513          0 2020-05-29 10:14 output_res-1590747240000
drwxr-xr-x   - 2019280513 2019280513          0 2020-05-29 10:15 output_res-1590747300000
drwxr-xr-x   - 2019280513 2019280513          0 2020-05-29 10:16 output_res-1590747360000
```

Sami Emre Erdogan
2019280513

I had to get the results from my hdfs directory to my local directory. I was struggling on how to do it but the TA helped me with it. I created a folder to store my results. I have used the following command to retrieve it:
hdfs dfs -get output_res-* /home/2019280513/hw10/new_result



The results can be carefully observed in my assignment folder. Example screenshot:



Result:

```
('l13', 27)
('l29', 24)
('m30', 23)
('m34', 23)
('k34', 23)
('l24', 22)
('m12', 22)
('o34', 22)
('o2', 22)
('m19', 22)
('o0', 22)
('o27', 21)
('m10', 21)
('m13', 21)
('p2', 21)
('n7', 21)
('l7', 20)
('l34', 20)
('l26', 20)
('o6', 20)
('l23', 20)
('k15', 20)
('o22', 20)
```