Big Data Analytics Techniques and Applications

Homework III

0316313 張逸群

Q1. Implement a program to calculate the average occurrences of each word in a sentence in the attached article (Youvegottofindwhatyoulove.txt).

A. The top 30 most frequent occurring words and their average occurrences in a sentence.

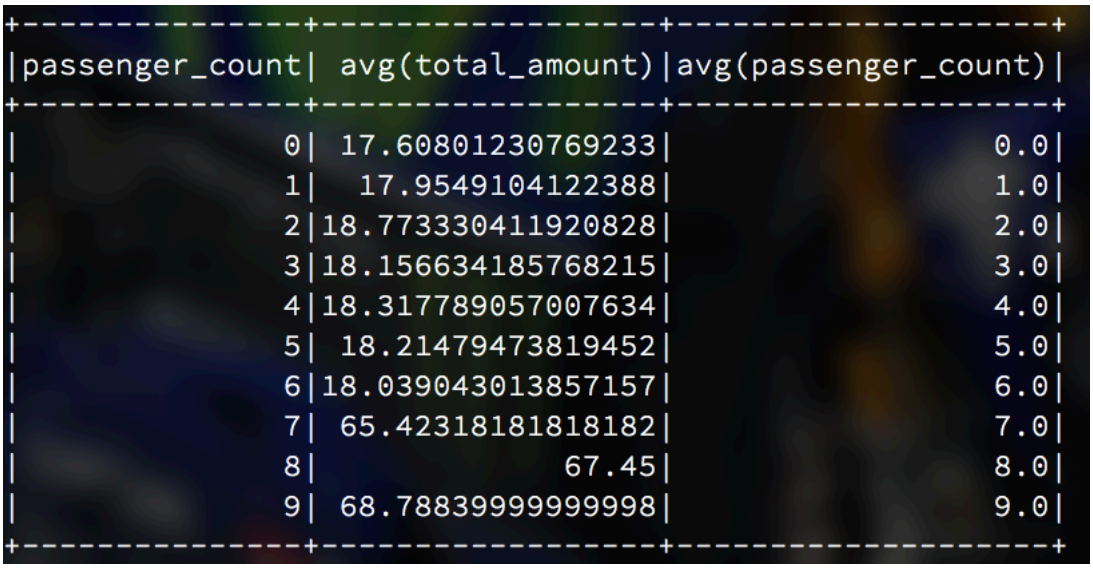
word	counts	average_occurrences
the	98	0.671233
i	93	0.636986
to	71	0.486301
and	67	0.458904
it	55	0.376712
was	48	0.328767
a	46	0.315068
of	42	0.287671
that	39	0.267123
you	35	0.239726
in	34	0.232877
my	30	0.205479
is	29	0.19863
had	22	0.150685
out	20	0.136986
with	19	0.130137
t	19	0.130137
me	18	0.123288
for	17	0.116438
life	17	0.116438
have	17	0.116438
so	17	0.116438
all	16	0.109589
your	16	0.109589
as	15	0.10274
what	15	0.10274
on	15	0.10274
college	14	0.0958904
but	14	0.0958904
be	14	0.0958904

B. According to the result, what are the characteristics of these words?

大部分皆為連接詞、介系詞以及代名詞，大多沒有太多的意義，主要是協助組成一個句子，表達句子中的相互關係。

Q2. Implement a program to calculate the average amounts (payment) in credit card trip for different number of passengers which are from one to four passengers in 2017.09 NYC Yellow Taxi trip data. In NYC Taxi data, the "Passenger_count" is a driver-entered value. Explain also how you deal with the data loss issue.

Ans.



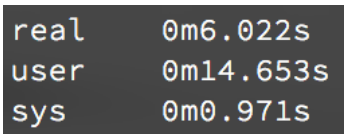
passenger_count	avg(total_amount)	avg(passenger_count)
0	17.60801230769233	0.0
1	17.9549104122388	1.0
2	18.773330411920828	2.0
3	18.156634185768215	3.0
4	18.317789057007634	4.0
5	18.21479473819452	5.0
6	18.039043013857157	6.0
7	65.42318181818182	7.0
8	67.45	8.0
9	68.78839999999998	9.0

資料中可能會有 NA 亦或是資料缺失的狀況，在讀取的時候會使用：DROPMALFORMED 關鍵字，以丟去不正常的資料。

Q3. For each of the above task 1 and 2, compare the execution time on local and yarn client. Also, give some discussions on your observation.

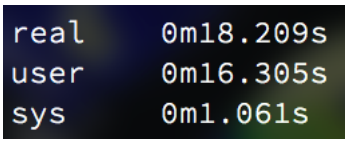
Task 1.

Local.



```
real    0m6.022s
user    0m14.653s
sys     0m0.971s
```

Yarn-Client.



```
real    0m18.209s
user    0m16.305s
sys     0m1.061s
```

Discussion.

可以發現在 Local Mode 執行時間是小於 Yarn-Client Mode 的，考量到這個 Task 檔案較小，且所需的處理較少，以 Yarn-Client Mode 執行的話會因為分送檔案，以及回傳結果的 overhead 過大，而導致執行效率低於直接在 Local Mode 下執行。

Task 2.

Local.

```
real    2m48.369s
user    4m10.895s
sys     0m8.198s
```

Yarn-Client.

```
real    0m46.727s
user    0m37.684s
sys     0m2.216s
```

Discussion.

在 Task 2 下，可以發現以 Yarn-Client Mode 執行速度是快上不少的，應是因為檔案較 Task 1 大，且所需的處理較多，可以克服 Overhead 的問題，使得 Yarn-Client Mode 發揮真正實力，得到較快的處理速度。

Program Workflow & Execution commands.

Q1.

1. 先使用 filter 過濾非可印的字元
2. 計算句子數量
   1) 使用 flatMap 將每行依照空格拆開
   2) 將包含句點、驚嘆號或問號的詞，Map 為 1，其餘為 0
   3) 使用 reduce 將值全部相加即可得到句子數量
3. 計算每個詞出現的數量
   1) 使用 flatMap 將每行依照標點符號及空格拆開
   2) 過濾空白的字元
   3) Map 為 word, (1, 1/句子數量)
   4) 使用 reduceByKey 將結果相加即可以得到詞出現的數量，以及平均出現數量
   5) 使用 sortBy 即可得到排序後的結果
4. 使用$SPARK_HOME/bin/spark-submit --master local hw3.py 以 Local Mode 執行，使用 $SPARK_HOME/bin/spark-submit --master yarn --deploy-mode client hw3.py 以 Yarn-Client Mode 執行。

Q2.

1. 使用 com.databricks.spark.csv package 讀取 csv 檔案得到 df，並使用 DROPMALFORMED option 將不正常的資料丟去
2. 使用 df[df.payment_type == 1] 得到使用信用卡付款的車次
3. 以 df[['total_amount', 'passenger_count']] 取出車資以及乘客數量
4. 以 df.groupBy('passenger_count').avg() 計算各乘客數量下的車資平均值
5. 以 df.show() 印出結果
6. 以 $SPARK_HOME/bin/spark-submit --packages com.databricks:spark-csv_2.11:1.5.0 --master yarn --deploy-mode client hw3.py 以 Yarn-Client Mode 執行，以 $SPARK_HOME/bin/spark-submit --packages com.databricks:spark-csv_2.11:1.5.0 --master local hw3.py 以 Local Mode 執行。