PRACTICAL-7

AIM: Perform the following apache spark program in DATABRICKS.

1. Find the average number of friends by age. (avgfriends.csv)

id	name	age	friends
0	Will	33	385
1	Jean-Luc	26	2
2	Hugh	55	221
3	Deanna	40	465
4	Quark	68	21
5	Weyoun	59	318

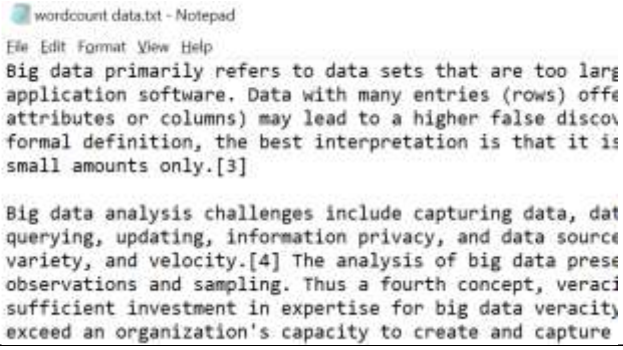
2. Use the dataset given and write the code to find the minimum temperature by the location (each whether station) and understand it and modify it to find maximum temperature by the location. (temp.csv)

Weather stationID	Date	Temp Type	Temp Value
ITE00100554	18000101	TMAX	-75
ITE00100554	18000101	TMIN	-148
GM000010962	18000101	PRCP	0
EZE00100082	18000101	TMAX	-86
EZE00100082	18000101	TMIN	-135

3. Use a given dataset of customers and their spending; find how much amount is spent by the individual customer in total, creating proper RDD in the databricks python notebook and sort out result based on the total spent. (customerorders.csv)

CustID	TrID	Amount
44	8602	37.19
35	5368	65.89
2	3391	40.64
47	6694	14.98
29	680	13.08

4. Use a text-file given as dataset and count the number of words occur in it. Also, use regular expressions to clean and count the number of words and sort out your output. (wordcount data.txt)



```
Big data primarily refers to data sets that are too larg
application software. Data with many entries (rows) offe
attributes or columns) may lead to a higher false discov
formal definition, the best interpretation is that it is
small amounts only.[3]

Big data analysis challenges include capturing data, dat
querying, updating, information privacy, and data source
variety, and velocity.[4] The analysis of big data prese
observations and sampling. Thus a fourth concept, veraci
sufficient investment in expertise for big data veracity
exceed an organization's capacity to create and capture
```