PySpark Assignment:

Problem 1-

1. Created RDD with input file and mapped it to unique buckets with keys as specified by case number.
2. Used A-priori algorithm to get frequent item-sets in a partition with a threshold of support by number of partitions to get if the itemset is frequent in at least one partition.
3. Then they are passed for checking count and then filter out by threshold equal to support.

Run command :
“bin\spark-submit Snehal_Shirgure_SON.py <case> <inputfile> <support>”

Spark version used : 2.2.1

Problem 2 –

CASE1		CASE2	
SUPPORT THRESHOLD	EXECUTION TIME	SUPPORT THRESHOLD	EXECUTION TIME
120	23.14 SECS	180	231.5 SECS
150	24.92 SECS	200	366.4

Problem 3-
Runtime exceeding