Assignment 2:

Execution Steps:

1. Copy the file to be tested and the python script in the same folder.
2. Open the command prompt in the same path of the folder.
3. To run the program on any file , type the following in your command line.
4. *spark-submit Anmol_Chawla_SON.py <Case Number> <Name_of_file.csv file> <Support Threshold>*

- First parameter is 1 or 2 for the Case Number.
- Second parameter is the .csv Input File.
- Third parameter is the Support Threshold.

Approach:

Apriori algorithm was implemented. Map phase1 returned frequent itemset to the reduce function from all the partitions, the phase 1 reduce function gathered a collection of candidates that would later be tested.

Phase 2 map, took all the candidates and gave their frequency in all the partitions , the reduce function then compared the frequency of each candidate with the given support to yield truly frequent itemsets.

Execution Times:

Problem 2: Movie-Lens Small Dataset

CASE 1		CASE 2	
Support Threshold	**Execution Time**	**Support Threshold**	**Execution Time**
120	18.31	180	196.46
150	12.026	200	126.11

Problem 3: Movie-Lens Big Dataset

CASE 1		CASE 2	
Support Threshold	**Execution Time**	**Support Threshold**	**Execution Time**
30000	394.42	2800	567.17
35000	154.23	3000	345.53