Two tasks are handled in both Python and Spark.

Version:

Spark- 2.2.1

Scala- 2.11

Python- 2.7

NOTE: Path directory containing ratings.csv and tags.csv need to be passed as program arguments

Folder Structure:

```
- ml-latest-small
   - ml-latest-small
     -ratings.csv
     -tags.csv
```

Note:

In snippet: Replace "/" with "\" depending on the system environment

1. Two Python files one for each task are Prashanth_Manja_task1.py and Prashanth_Manja_task2.py

2**. Guidlines to run the code: (Python_code)**

a. Task1: (Prashanth_Manja_task1.py)

```
bin\spark-submit Prashanth_Manja_task1.py <Inputpath> <Outputpath>
```

<Inputpath> is path containing 'ml-20m' containing big datasets or 'ml-latest-small' containing small dataset. (filename- 'ratings.csv' not to be included)

<Outputpath> path to save the output file (filename not to be included)

Example: Inputpath -> G:\DM_hw1\ml-latest-small\ml-latest-small

Outputpath -> G:\DM_hw1\result_files

Example:

```
bin\spark-submit Prashanth_Manja_task1.py G:\DM_hw1\ml-latest-small\ml-latest-small G:\DM_hw1\result_files
```

Performing the above steps would output 'Prashanth_Manja_result_task1_small.txt' or 'Prashanth_Manja_result_task1_big.txt' for small and large datasets respectively.

b. Task2:(Prashanth_Manja_task2.py)

```
bin\spark-submit Prashanth_Manja_task2.py <Inputpath_for_ratings.csv> <Inputpath_for_tags.csv> <Outputpath>
```

<Inputpath> is path containing 'ml-20m' containing big datasets or 'ml-latest-small' containing small dataset.(filename- 'ratings.csv' not to be included)

<Outputpath> path to save the output file (filename not to be included i.e 'Prashanth_Manja_result_task1_result_small)

Example:

```
bin\spark-submit Prashanth_Manja_task1.py G:\DM_hw1\ml-latest-small\ml-latest-small G:\DM_hw1\ml-latest-small\ml-latest-small G:\DM_hw1\result_files
```

Performing the above steps would output 'Prashanth_Manja_result_task1_small.txt' or 'Prashanth_Manja_result_task1_big.txt' for small and large datasets respectively.