Homework 4

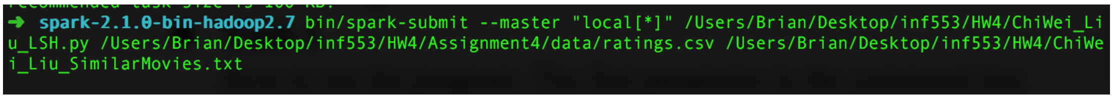
The version of Spark is version 2.1.0

The version of Python is version 2.7.10

- ## Task 1

  How to run the program: The first parameter in the command line should be the file path of the source code (ChiWei_Liu_LSH.py), and the second parameter should be the file path of the *ratings.csv*. The third parameter should be the output path and file name (ChiWei_Liu_SimiliarMovies.txt).

  For example:



```
recommended task size is 100 KB.
➜ spark-2.1.0-bin-hadoop2.7 bin/spark-submit --master "local[*]" /Users/Brian/Desktop/inf553/HW4/ChiWei_Li
u_LSH.py /Users/Brian/Desktop/inf553/HW4/Assignment4/data/ratings.csv /Users/Brian/Desktop/inf553/HW4/ChiWe
i_Liu_SimilarMovies.txt
```

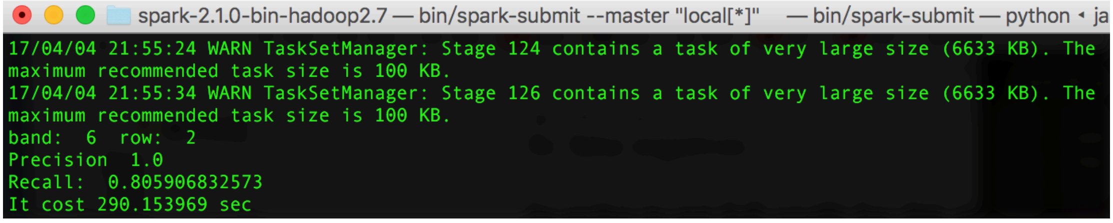
- ## Task 2

  I hashed totally 12 times in the program. I used 6 bands and 2 rows to find the similar items.

  * Precision = tp / (tp + fp) = 1.0

  * Recall = tp / (tp + fn) = 0.81

  Screenshot of the result:



spark-2.1.0-bin-hadoop2.7 — bin/spark-submit --master "local[*]" — bin/spark-submit — python ◂ ja

```
17/04/04 21:55:24 WARN TaskSetManager: Stage 124 contains a task of very large size (6633 KB). The
maximum recommended task size is 100 KB.
17/04/04 21:55:34 WARN TaskSetManager: Stage 126 contains a task of very large size (6633 KB). The
maximum recommended task size is 100 KB.
band:  6  row:  2
Precision  1.0
Recall:  0.805906832573
It cost 290.153969 sec
```

- ## Task 3

  r = 2, b = 6

s	$1-(1-s^r)^b$
0.2	0.217
0.3	0.432
0.4	0.648
0.5	0.822
0.6	0.931
0.7	0.982
0.8	0.9978

- **Task 4**

  When deciding the threshold for finding similar pair. We can observe from the testing data to get the result if we have larger r, we will get larger threshold. If we have larger b, we will get lower threshold.

  * Larger r => larger threshold

  * Larger b => lower threshold