Project_1

October 9, 2019

```
In [1]: import numpy as np
        import pandas as pd
        import matplotlib.pyplot as plt
        import scipy as sp
```

```
In [2]: ratings = pd.read_csv('./ml-latest/ratings.csv', header=0)
```

```
In [3]: ratings.head()
```

Out[3]:

	userId	movieId	rating	timestamp
0	1	307	3.5	1256677221
1	1	481	3.5	1256677456
2	1	1091	1.5	1256677471
3	1	1257	4.5	1256677460
4	1	1449	4.5	1256677264

0.0.1 Select 8000 users with most ratings

```
In [35]: NUM_TOP_USERS = 8000
```

```
In [20]: selected_users = ratings.groupby(['userId']).size().reset_index(name='num_rated')\
             .sort_values(by='num_rated', ascending=False)[0:NUM_TOP_USERS]
```

```
In [21]: selected_users.head()
```

Out[21]:

	userId	num_rated
123099	123100	23715
117489	117490	9279
134595	134596	8381
212342	212343	7884
242682	242683	7515

```
In [22]: selected_ratings = ratings.merge(selected_users, left_on='userId', right_on='userId',s
         selected_ratings = selected_ratings[['userId', 'movieId', 'rating', 'timestamp']]
```

```
In [28]: selected_ratings.shape
```

Out[28]: (8144389, 4)

0.0.2 Items Analysis

```
In [32]: unique_movies = ratings.movieId.unique()
         unique_movies

Out[32]: array([   307,    481,   1091, ..., 117857, 133409, 142855])

In [33]: num_unique_movies = len(unique_movies)

In [36]: matrix_completion_rate = selected_ratings.shape[0] / (NUM_TOP_USERS * num_unique_movie
         matrix_completion_rate

Out[36]: 0.01889158501735048
```