Project of Recommender System

Describe and clean the dataset

This dataset is about the reviews of movies. It contains 1 million ratings from 6000 users on 4000 movies and was Released 2/2003. It can be download from https://grouplens.org/datasets/movielens/1m/ (https://grouplens.org/datasets/movielens/1m/)

First we clean the dataset.

In [1]:
```
import gzip
from collections import defaultdict
import scipy
import scipy.optimize
import numpy
import random

path='ratings.dat'
f = open(path,"rt",encoding="utf8")

dataset=[]
header = ['uid','mid','star']
for line in f:
    fields = line.strip().split('::')
    d = dict(zip(header,fields))
    d['uid'] = int(d['uid'])
    d['mid'] = int(d['mid'])
    d['star'] = int(d['star'])
    dataset.append(d)
```

In [2]:
```
dataset[:10]
```

Out[2]:
```
[{'uid': 1, 'mid': 1193, 'star': 5},
 {'uid': 1, 'mid': 661, 'star': 3},
 {'uid': 1, 'mid': 914, 'star': 3},
 {'uid': 1, 'mid': 3408, 'star': 4},
 {'uid': 1, 'mid': 2355, 'star': 5},
 {'uid': 1, 'mid': 1197, 'star': 3},
 {'uid': 1, 'mid': 1287, 'star': 5},
 {'uid': 1, 'mid': 2804, 'star': 5},
 {'uid': 1, 'mid': 594, 'star': 4},
 {'uid': 1, 'mid': 919, 'star': 4}]
```

Firstly use Jaccard function to find similarities