Project 3: **Movie Recommendations**
Comp 475 – Web Science
100 points

The goal of this project is to use the basic recommendation principles we have learned for user-collected data. You will modify the code given to you which performs movie recommendations from the MovieLense data sets.

The MovieLense data sets were collected by the GroupLens Research Project at the University of Minnesota during the seven-month period from September 19th, 1997 through April 22nd, 1998. It is available for download from http://www.grouplens.org/node/73

There are three files which we will use:

1. **u.data** – 100,000 ratings by 943 users on 1,682 movies. Each user has rated at least 20 movies. Users and items are numbered consecutively from 1. The data is randomly ordered. This is a tab separated list of user id | item id | rating | timestamp. The time stamps are unix seconds since 1/1/1970 UTC.

Example:

```
196     242     3       881250949
186     302     3       891717742
22      377     1       878887116
244     51      2       880606923
166     346     1       886397596
298     474     4       884182806
115     265     2       881171488
```

2. **u.item** - Information about the 1,682 movies. This is a tab separated list of

movie id | movie title | release date | video release date | IMDb URL | unknown | Action | Adventure | Animation |Children's | Comedy | Crime | Documentary | Drama | Fantasy | Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi | Thriller | War | Western |

The last 19 fields are the genres, a 1 indicates the movie is of that genre, a 0 indicates it is not; movies can be in several genres at once. The movie ids are the ones used in the u.data data set.

Example:

```
161|Top Gun (1986)|01-Jan-1986||http://us.imdb.com/M/title-
exact?Top%20Gun%20(1986)|0|1|0|0|0|0|0|0|0|0|0|0|0|1|0|0|0|0
162|On Golden Pond (1981)|01-Jan-1981||http://us.imdb.com/M/title-
exact?On%20Golden%20Pond%20(1981)|0|0|0|0|0|0|0|0|1|0|0|0|0|0|0|0|0|0|0
163|Return of the Pink Panther, The (1974)|01-Jan-1974||http://us.imdb.com/M/title-
exact?Return%20of%20the%20Pink%20Panther,%20The%20(1974)|0|0|0|0|0|1|0|0|0|0|0|0|0|0|
0|0|0|0|0
```

3. u.user - Demographic information about the users. This is a tab separated list of

user id | age | gender | occupation | zip code

The user ids are the ones used in the u.data data set.

Example:

```
1|24|M|technician|85711
2|53|F|other|94043
3|23|M|writer|32067
4|24|M|technician|43537
5|33|F|other|15213
```

The code for reading from the u.data and u.item files and creating recommendations is described in the book *Programming Collective Intelligence* and is on cs1. You are to modify recommendations.py to answer the following questions. Each question your program answers correctly will award you **10 points**. You must have the question answered completely correct; partial credit will only be awarded if your answer is *very* close to the correct one. You do not have to answer all the questions. For example, you could answer 3, 5, 7-10 which would award you 60%. In all the questions that involve comparing movie ratings, use the Pearson correlation coefficient to compute similarity.

1. What 5 movies have the *highest* average ratings? Show the movies and their ratings sorted by their average ratings.
2. What 5 movies received the *most* ratings? Show the movies and the number of ratings sorted by number of ratings.
3. What 5 movies were rated the highest on average by *women*? Show the movies and their ratings sorted by ratings.
4. What 5 movies were rated the highest on average by *men*? Show the movies and their ratings sorted by ratings.
5. What movie received ratings *most* like *Top Gun*? Which movie received ratings that were *least* like *Top Gun* (negative correlation)?
6. Which 5 raters rated the most films? Show the raters' IDs and the number of films each rated.
7. Which 5 raters most agreed with each other? Show the raters' IDs and Pearson's *r*, sorted by *r*.
8. Which 5 raters most disagreed with each other (negative correlation)? Show the raters' IDs and Pearson's *r*, sorted by *r*.
9. What movie was rated highest on average by men over 40? By men under 40?
10. What movie was rated highest on average by women over 40? By women under 40?

Your output should clearly indicate the answers from the question you answered. Use the formatting of the example output shown below. Note that the answers shown below are not necessarily correct, they are just shown to indicate the desired format for your answers.

```
Question 1:
   1. Top Gun (1986) - 4.9483
   2. Star Wars (1977) - 4.9232
   etc…

Question 2:
   1. Die Hard (1988) - 223
   2. Return of the Jedi (1983) - 209
   etc…

Question 9:
   Men over 40: Godfather, The (1972) - 4.9321
   Men under 40: Die Hard (1988) – 4.9114
```

Submit to Easel your altered recommendation.py script before it is due. I will run your script on the ml-100k data set from MovieLense and compare your solution to mine.

McChallenge: You may receive an additional 1% added to your final grade by creating a 2D image that shows a subset of the movies that are placed on the image using the multidimensional scaling technique discussed in class. The clusers.py program is currently producing an image based on blog data. You will alter this program to produce an image based on the movie data set. You may want to keep the number of movies below 50 as the algorithm used in the program tends to have difficulty with larger values. You can use any similarity metric you'd like based on any features of the movie data set you would like. For example, you could use average men and average women ratings to determine the distance between each movie. Email me your image and code if you complete the extra credit.