CS46N Project #1: Movie-Rating Prediction
Due Sunday October 25 before midnight, no lates permitted

The goal of the project is to use past movie ratings to predict how users will rate movies they haven't watched yet. This type of prediction algorithm forms the underpinning of recommendation engines, such as the one used by Netflix. We're giving you a large set of ratings from real movie-rating data, but holding back 50 ratings for you to predict. (In machine-learning parlance, the data we provide is "labeled training data"; you will use it to come up with rating predictions for 50 "unlabeled" data items.)

The student or team who comes up with the most accurate predictions wins a … well… we're not sure yet but we'll think of something! You're welcome to use any techniques and any tools to come up with your 50 rating predictions. (There's a bit more discussion on this below.) In addition to the 50 predictions you will submit when the project is due, we will have an in-class session on October 27 where students will use their methods to make 10 more predictions on the spot, giving us an opportunity to compare results again, and discuss the variety of approaches.

The Data

On the class website under the **Assignments and Projects** tab you will find these instructions, along with the following five files available for download in tsv format (tab-separated values).

- **users.tsv** -- Information about 943 movie watchers. Each line has three fields: *userID, age, gender*
- **movies.tsv** -- Information about 1,682 movies. Each line has seven fields: *movieID, name, genre1, genre2, genre3, genre4*. If a movie has fewer than four genres the extra fields are blank.
- **ratings.tsv** -- 17,556 movie ratings. Each line has three fields: *userID, movieID, rating*. The userID and movieID correspond to those in the **user.tsv** and **movie.tsv** files, respectively. Ratings are integers in the range 1 to 5 (from worst to best).
- **allData.tsv** -- For those who prefer having everything in one place, this file contains the combined information from the previous three files. Each line has 10 fields: *userID, age, gender, movieID, name, genre1, genre2, genre3, genre4, rating.*
- **predict.tsv** -- Ratings for you to predict. Each line has two fields: *userID, movieID*. There are no ratings for these pairs in **ratings.tsv**.

This data is real: it's a subset of the movie ratings data from MovieLens, collected by GroupLens Research, which we anonymized for this assignment.

Some Details

- The project can be done individually or in a team of two.
- You can use any method you like to make your predictions. At one end of the spectrum you could implement sophisticated modeling of users and movies based on their "features," and/or apply machine-learning techniques like the ones we discussed in class. At the other end, you could base your predictions on simple statistics or even just eyeballing/exploring the data. Most students will probably do something in between.
- You are also free to use any tools that you like; we'll do our best to help you find and use tools suited to your approach.
- If you've tried a few different methods and are torn about which one might be best, don't despair: we're allowing (but not expecting) each student or team to submit up to three solutions.
- We're not worried about the honor code -- feel free to discuss your ideas with other students, the course staff, or anyone else. Of course in the end you need to come up with the 50 predicted ratings. We are asking you to document your approach, and you'll need reproduce your efforts to come up with 10 more rating predictions during the in-class session.

What to Turn In

Before midnight on Sunday October 25, each student or team should send an email with the following files attached to Akash at akashds@stanford.edu.

- **lastname(s)_readme.txt** -- A description of your approach. Your description should be detailed enough for us to understand your methodology, but don't go overboard -- in most cases 1-2 short paragraphs should be enough.
- **lastname(s)_predict.tsv** -- A copy of **predict.tsv** except now with three fields: *userID, movieID, rating.* Values in the rating field should be integers in the range 1 to 5.
- Optionally if you want to submit up to two additional solutions, please name the files **lastname(s)_predict2.tsv** and **lastname(s)_predict3.tsv.**
- If your approach included coding, or you generated intermediate data files on the way to your final solution, submit one or more additional files containing the code and/or data. Please include a list of the additional files and what they contain in your **lastname(s)_readme.txt** file.