**Project Proposal: Recommender System for Movies**

***The Problem:***

For my capstone project, I’d like to develop a robust recommender system for movie goers, one that uses their existing ratings to develop recommendations for movies that they would enjoy regardless of the general population’s aggregate view. Whether this comes as a personal bias towards a specific genre or appreciation of the entertainment value of “terrible” movies, I hope to be able to use this recommender system to predict what rating a user would give to a movie.

***The Client:***

While a recommender system for movies is not a novel concept (imdb, Netflix, etc.), to me it is a verifiable proof of concept that can be tailored to a large variety of other products and services.

Speaking very generally, the client would be any provider of a product or service (**content**) for which large ratings data can be accumulated. Although the providers are the client for this project, this project’s purpose is to help users of that provider decide on content by forming a recommendation based on their preferences. The users see the direct benefit in the sense that they are empowered to decide or do something based on the analysis. However, the client (providers) can reap tangential benefits by promoting their content to interested users. Content providers may also be able to directly benefit from this project by using clustering to try and predict the popularity of new content (products or services).

More specifically, a recommender system can be used in all kinds of entertainment products from movies to books to games and in services, such as selecting travel destinations.

***The Data Source:***

For this project, I will use data from movielens.org, aggregated by GroupLens Research (<http://grouplens.org/datasets/movielens/>). They provide data sets of movie ratings ranging from 100k ratings to the full data set of over 22 million ratings. The smaller dataset or a subset of the large dataset should prove valuable during development of the recommender system in terms of speed and ease of use. The full data set will be used for the final project submission. The datasets are both readily available from the site mentioned above.

***The Plan:***

To my knowledge, developing a recommender system will require a few steps:

1. Remove or otherwise handle movies with few ratings. A movie with less than a certain threshold number of ratings will likely fail to develop appropriate features.
2. Restructuring the existing data set from a list of user ID, movie ID, and ratings to a matrix containing the ratings from a user for a specific movie.
3. Apply mean normalization to all of the ratings by movie.
4. Applying a collaborative filtering algorithm with gradient descent should be able to resolve features for both the movies and the users.
5. Use features to form recommendations by either predicting a user’s rating or determining similarity between movies.

With this dataset, I also see the possibility of a few stretch goals:

1. To inform users about their genre preferences. This could be used to help people transition between different forms of media (i.e. from movies to TV, books, games) in the absence of an all-encompassing dataset.
2. Clustering of user population to try and predict viewership. Applied to user features, we can determine which raters are similar in much the same way we find similar movies.

***The Deliverables:***

* Code for recommender system
  + Processing input into a workable matrix
  + Collaborative filtering with gradient descent
  + Predict ratings for a given user
  + Determine ‘distance’ between movies
* Stretch goal 1
  + Assess genre classifications
  + Combine with predicted ratings
* Stretch goal 2
  + Clustering algorithm (k-means)
* Accompanying documentation for all work
* Paper or presentation