

A Recommender System

Weighted Alternating Least Squares

Sandra Giorgetta | Data Science Intensive | 5/2/2017

Capstone Project Summary/Milestone Report

# Introduction - Implementing a Recommender System

### 

There’s no doubt that we live in a data driven world and to be competitive in it, a business must have digital access to its consumers via computer and hand-held devices. Through the collection of information regarding the consumer’s behavior, businesses can now provide a more personalized experience when targeting its customers. It can utilize this information to measure performance, enhance marketing efforts and drive business decisions. One of the methodologies used for targeting specific users with relevant material such as a recommended purchase, subscription suggestion, or online advertisement is the recommender system. With the implementation of a recommender system, a company can provide a personalized online experience that has the ability to drive consumer engagement, increase sales, and improve business outcomes.

With an average of 2,577 films produced per year around the world, it is no surprise that a movie recommendation system would be a popular idea. For companies like Netflix, a movie rental service, a movie recommendation engine is a must. With the average person watching 20 to 30 movies a year, there has been the advent of a plethora of movie recommendation services such as RottenTomatoes, Taste Kid, Jinni, Criticker, Flixter, and Movielens, to name a few. Taste Kid not only provides a movie recommendation but goes further by adding recommendations for books, bands, music, and other items related to the movie. The Movielens organization has generously provided access to datasets for educational purposes to encourage the development of better recommendation engines. The recommender system presented and studied in this report utilizes one of the datasets donated by Movielens.

# The Data Set

The stable Movielens dataset to be studied in this report can be found at the website: <http://grouplens.org/datasets/movielens/>. This dataset consists of movie ratings collected from the website <http://movielens.com> between the dates of January 9, 1995 and October 16, 2016. The data includes ratings by 671 users and information of 9123 movies, which have at least one rating or tag. Timestamps of the rating and tag are provided as well as link information for each movie for two movie websites, [www.imdb.com](http://www.imdb.com) and [www.themoviedb.org](http://www.themoviedb.org). Specifically, the dataset studied is named “ml-latest-small”, generated on October 17, 2016. This is a very sparse data set due to the sheer number of movies in comparison to the number of movies rated by a user.

The dataset is composed of four files in csv format: movies.csv, ratings.csv, tags.csv, and links.csv. The movies file includes a movieId, title, and genre field for 9125 movies. The ratings file includes userId, movieId, rating, and timestamp. The tags file includes userId, movieId, tag, and timestamp. And finally, the links file includes movieId, and identifiers for links to the movies as listed on the websites www.imdb.com and www.themoviedb.org.

# Data Clean Up and Organization

## Clean up

Due to the fact that some of the data is hand entered and some of the data is scraped from the web, there are bound to be some not so ideal entries. For this dataset, several irregularities were discovered. There were duplicate entries with the same movie title with different movie Id’s. There were four movies for which no release date was included in the title. And there were two movies that are in fact “made for TV” movies and not what would be considered your traditional box office movie. The two website url identifier columns from the links file were checked for null data. The column with data from the www.imdb.com site was complete. However, there were several null entries in the column for the www.moviedb.org identifiers.

Although going back into the original csv file and making manual edits is laborious, it is felt that in a real world scenario, it would be advisable to have a “cleaned up” version of the file ready for use at a later time. If at some time in the future, the potential client had the desire to revisit the dataset, this would be most efficient. For these reasons, when possible changes were made to the original file. However, for some cases such as changing null values to zeroes, the changes were made in the pandas dataframe itself.

## Data Organization

The data from the four csv files is merged into several different pandas dataframes depending on the requirements for the use of the data.

The first dataframe includes data from the ratings and movies files and is used to create a matrix, denoted as Q, where the columns are the movie Id’s, the rows are the user Id’s and the contents are the ratings. This matrix is the basis for the recommendation algorithm.

A second dataframe is the movies file merged with the links file. It contains the movie Id’s, the movie titles, the movie genres and the identifier information for creating the url link to information on the movie for the two different movie websites. The movieId in this dataframe is set as the index and this dataframe is used as a lookup table for retrieving the movie title and url that is the output of the recommender system.

The third dataframe is the ratings and movies files merged onto the tags file. It contains the userId, movieId, rating, tag, and timestamp of the tag. Some users provided multiple tags (in some cases, as many as 10) for the same movie. For this reason, the ratings and movies files needed to be merged onto the tags file as opposed to the other way around to preserve all of the tags. The use of the tag as a way of “inferring” user interest was considered as the act of creating a tag could indicate positive interest in the movie. A scan of the tags indicates that it would not be trivial to glean any information from the tags since they are user produced and somewhat random in how each user uses them. With some tags such as “dark”, “gloomy”, “violent”, it is clear that tagging the movie does not necessarily imply a positive interest in the movie. The information in the tags file is useful in a user-based or item-based collaborative filtering system. At this time, since the implementation of choice is a model based approach, the tags are not being used as part of the recommendation system.

# The Algorithm – Model Based Recommender System

Since there is not a unique way of implementing a recommender system, in what follows we briefly discuss the main possible approaches.

The two approaches for designing a recommender system are content-based filtering and collaborative filtering. The content-based approach requires specific knowledge of the features of an item and uses that information to recommend items with similar features. Collaborative filtering does not require prior knowledge of the features of an item. It builds a model based on users’ past preferences and uses that model to predict future preferences. There are two types of collaborative filtering algorithms: memory-based and model-based. Memory-based algorithms find a group of users that have similar preferences and use those preferences to make a prediction to a given user. A similarity measurement is used to find a correlation between users. The model-based algorithm uses a subset of the data as “training data” to learn a predictive model. There are advantages and disadvantages to the various approaches and algorithms. Some advantages of model-based collaborative filtering are:

1. No prior knowledge of the item or user is required
2. It handlse the problem of sparsity (portion of matrix that is empty) better
3. Scalability. The model typically is smaller than the entire dataset so can be used even for very large datasets.
4. Are typically faster than memory-based
5. Can handle the problem of over-fitting

There are many model-based collaborative filtering algorithms. Several research articles indicate that the algorithm with the highest accuracy is matrix factorization. However, it has also been shown that the “best” algorithm for a given data set may depend on the nature of the data set itself. The sparsity of the MovieLens Data set lends it well to the use of the Alternating Least Squares Algorithm.

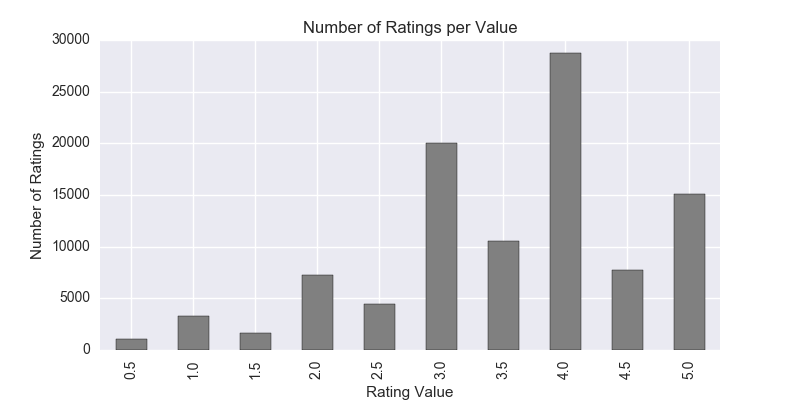
# Initial Findings

The data exploration phase of the project involved finding the answers to questions that would be helpful and relevant in making decisions regarding the design of the recommender system and in understanding trends in user engagement that would be helpful in driving business decisions for the potential client.

## Distributions of Ratings

After the recommender engine is implemented and the matrix of values for predicted ratings is defined, a decision needs to be made as to what values should be considered high enough for making purposeful recommendations. For this reason, it is helpful in investigating the distribution of ratings per value, the distribution of ratings per user, and the distribution of ratings per movie.

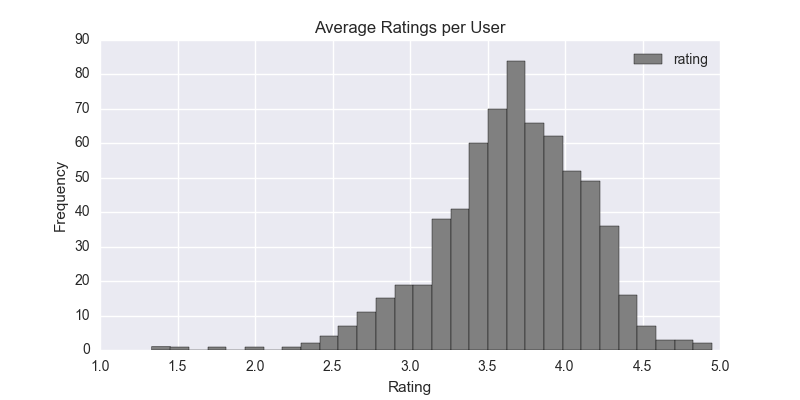
The rating values range from 0.5 to 5.0 in 0.5 increments. Figure 1 shows the distribution:



*Figure 1 Distribution of Ratings per Value*

The most frequent rating given was 4.0, followed by 3.0. The least most frequent rating was 0.5. More frequently than not, a higher rating is given than a lower rating. Is this because movie goers are not very discriminate in the way they rate movies? Or is it possibly the fact that a user is more likely to rate a movie they enjoy versus a movie they don’t enjoy?

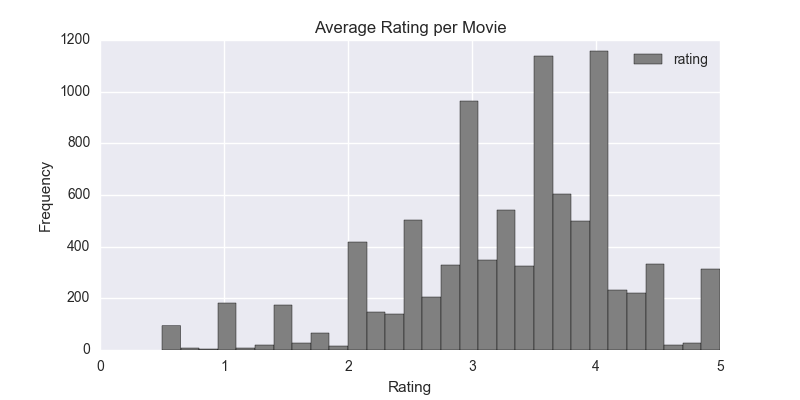
Knowing that there are more “high” ratings assigned than not, it may be helpful in knowing what the average rating is per user. To analyze the mean rating per user, the distribution was plotted and is shown in Figure 2:



*Figure 2 Average Rating per User*

As can be seen in the plot, the most frequent mean rating by a user was between 3.5 and 4.0.

Another piece of information that may be helpful is to look at the distribution of the average rating per movie. This distribution was plotted and can be seen in Figure 3. The distribution is not normal and is skewed to the left. More mean movie ratings tended to be higher than 2.5 than lower than 2.5.



*Figure 3 Distribution of Average Rating per Movie*

After the distribution of movie ratings has been analyzed, the “cut-off” point for deciding whether a predicted rating shall be considered positive or not can be made. One could choose an arbitrary value such as 2.5 to use as the “cut-off” in determining whether the predicted rating should be considered for a positive recommendation or not (only predicted ratings higher than the cut off will be output). Knowing that the model will be utilizing the data in the rating matrix and knowing that the ratings tend to be on the higher side, we can see that choosing a rating of 2.5 may not produce the most useful results. For this reason, we will use the mean as the cut-off point. Any movie with a predicted rating for the user that is greater than the mean rating will be considered for output.

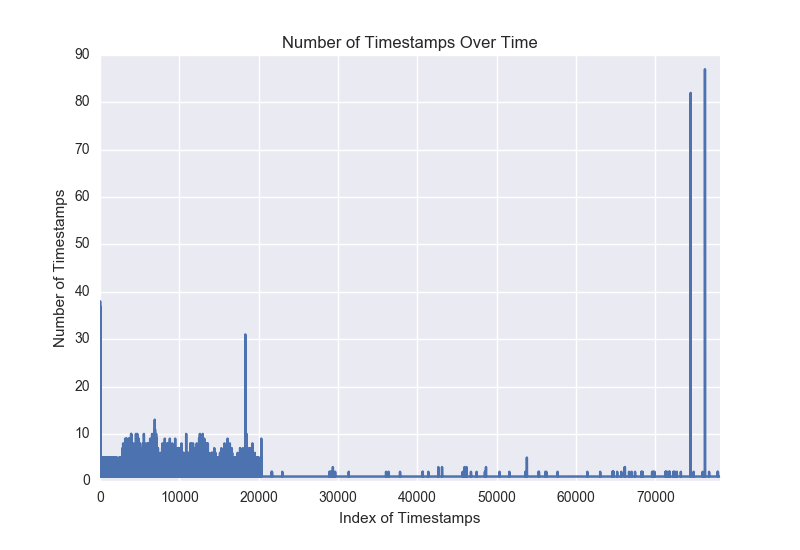
Knowing the sample size or the number of ratings a movie received is useful. If the movie has received at least 30 ratings, then the mean rating of the movie is likely to be the true mean of the population. The analysis showed that the most ratings received by a movie was 341 and the mean number of ratings was 11. For a larger data set where the mean number of ratings is larger, say above 30, then we could utilize statistical analsyis to compute probabilities of a movie receiving a given rating. These probabilities can be used to initialize the Y vector for a more hybrid approach.

## User Engagement

User engagement is a topic of concern to any business. The first item considered concerning user engagement is to know how many movies on average a user rated. We found that the largest number of movies rated by a user was 2391. The smallest number of movies rated by a user was 20 and the average number of movies rated by a user was 149. If the minimum number of movies rated were 30, then having a sample size of 30 we would use statistical analysis to determine the probability of a user assigning a given rating value. This information could be useful in creating a hybrid approach where the X vector is initialized with those probabilities as opposed to random assignments.

The next item of concern is the duration of engagement. The timestamp entries for the ratings were analyzed. The timestamp is the number of seconds since midnight Coordinated Universal Time (UTC) of January 1, 1970. The number of timestamps entered over time was plotted and is shown in Figure 4. The x-axis is the index of timestamps and the y-axis is the number of timestamps per timestamp value. There were a total of 78,141 timestamps over the approximately 21 years of data collection. As evidenced in the plot, the user engagement dropped off considerably around the timestamp corresponding with 20,000 which corresponds to the date of May 1, 2003. It was also observed that near the end of the data collection period, there was a large surge of engagement.

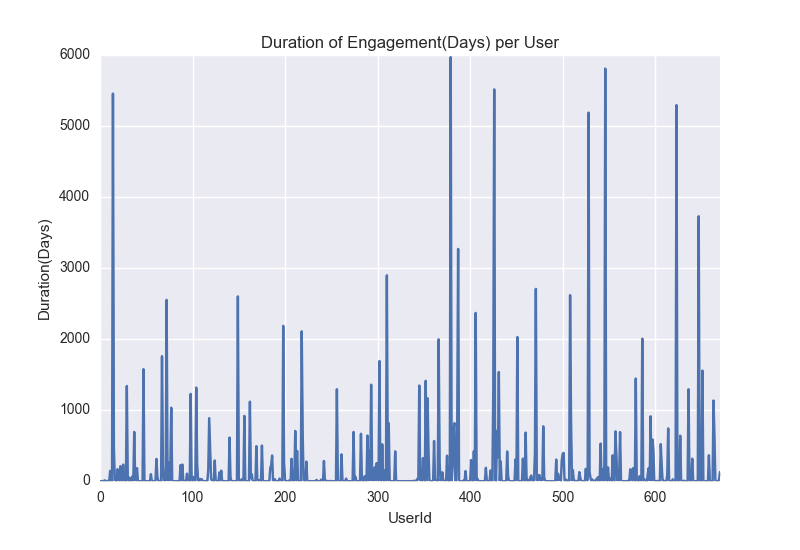
Another way of analyzing user engagement is to look at the duration of engagement by user. The minimum tiemstamp was subtracted from the maximum to determine the time of engagaement in days per user. The data was plotted and can be seen in Figure 5. UserId 35 had the least duration of engagement which was 29 seconds while userId 380 had the longest duration of engagement which was 16.36 years. The average duration of engagement was 201 days which is approximately 7 months. The average length of



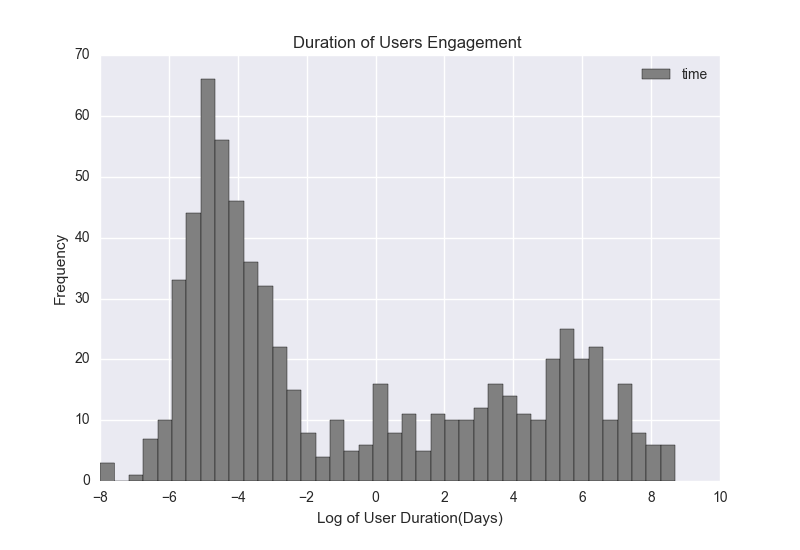
*Figure 4 User Engagement over Time*

engagement is most likely related to the fact that the movielens project is conducted in an educational setting where students are required to use the system during an academic calendar year.

Because there are some users whose duration was less than a minute and other users whose duration was over a twenty year period, it is useful to plot the log of the distribution to be able to view more detail. As a reminder, when looking at the plot, log0 corresponds to 1, log1 corresponds to 10, log2 corresponds to 100, and so on. The distribution of the log of the user duration in days was plotted and can be seen in Figure 6. There is a large proportion of users whose duration of engagement was less than 1 day (values <= 0).

The number of users that dropped out before the end of one year was 588. The number of users that stayed for at least one year and up to two years was 36, at least two years and up to three was 10, at least three years and up to four was 13, at least four years and up to five years was 5, and those staying over five years was 19. 

*Figure 5 Duration of User Engagement by UserId*



*Figure 6 Log of User Duration in Days*

## TAGS FILE ANALYSIS

The tags file was analyzed and considered as a vehicle for inferring user interest. There are 1296 tags, 582 of which are unique. There are 689 movies that have at least one tag provided by 61 users over a 10 year period beginning in January of 2006 and ending at the end of the collection process which was October 2016. Of the movies that were tagged, 266 were not rated. The maximum number of tags a movie received was 25 and the mean was 1.88. The maximum number of tags submitted by a user was 401 and the mean was 21.25. The tags information may be useful in a user-based collaborative filtering system. With the mean number of tags per movie less than two, using tags for an item-based collaborative filtering is not advised.

# Experiment

In addition to implementing the algorithm, five-fold cross validation is used for tuning the regularization parameter λ and the K value that determines the sizes of the X and Y factor vectors. The root mean squared error is calculated after each fold and then the parameters are re-adjusted. After the final test set run, the X and Y vectors will be used to create the Q\_hat matrix with the predicted ratings. The program looks for the predicted ratings that are above the average rating and have release dates within the last five years. If no predicted rating is found, the year is reduced by one until a rating is found that is above average. For each user, three items are printed; the title of the user’s top rated film, the title of the user’s top recommended film regardless of release date, and the title of the user’s top recently released recommended film. For the recommended films, in addition to the title, a link to the www.imdb.com website is provided.

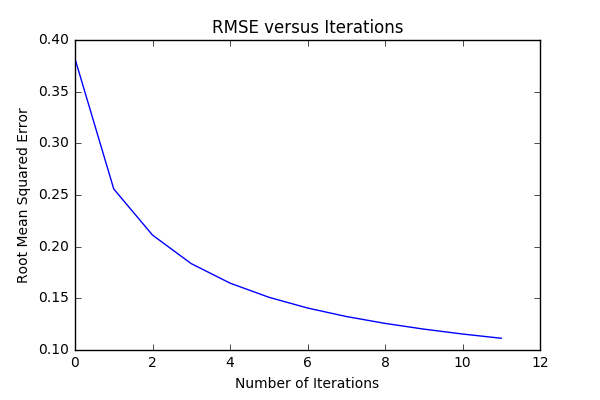
The code representing the first four folds of the five-fold cross validation will be timed. At a later date, the code will be run using Amazon Web Services and the results will be compared.

# Results

The initial values for the lambda and K values were chosen. Lambdas = [0.01, 0.1, 1.0]. K\_values = [10, 13, 16]. The middle value of K was chosen to be 1/10 the size of the training sets which contain 134 samples. The number of iterations was chosen as 12 to allow the program to complete overnight. The five-fold cross validation was executed. The time of execution of the first four folds was 23 hours, 38 minutes, and 12 seconds. After the final test run, the best parameters were identified as K=20 and lambda=0.011875. The RMSE results were shown to be better with larger K values and the smaller lambda values. The lowest validation run RMSE was 0.1194 while the lowest RMSE for the test run was 0.1113. Here are the RMSE outputs of the four validation test runs:

{0: [0.44187657971522604,  
 0.2743703871025651,  
 0.22630041651000407,  
 0.19741593471695518,  
 0.17779109784405184,  
 0.16332371808955432,  
 0.15203698208763555,  
 0.1429313514696443,  
 0.13545171949689128,  
 0.12921119681672213,  
 0.12392114379401133,  
 0.11936594799632551],  
 1: [0.38502403049444456,  
 0.27187147692347347,  
 0.2307206318250529,  
 0.20593540960749782,  
 0.18898367405654404,  
 0.17608496350457745,  
 0.16563081110357722,  
 0.15707613496766368,  
 0.15007527715956268,  
 0.14429838555204857,  
 0.13947136667188711,  
 0.13536666278445095],  
 2: [0.39612831487288885,  
 0.26988116933741907,  
 0.22636149092996705,  
 0.200593765086749,  
 0.18268755859124952,  
 0.16921210122592328,  
 0.1585861466867265,  
 0.14996638786564398,  
 0.14285651323898294,  
 0.13692715088649576,  
 0.1319312344086906,  
 0.1276766036042474],  
 3: [0.40069839670278123,  
 0.2749496024563838,  
 0.22831521616876996,  
 0.2005680108579301,  
 0.1820474720360778,  
 0.16865174829720736,  
 0.15834075955709703,  
 0.15007260440655523,  
 0.14329028444796918,  
 0.13758324026445332,  
 0.1326905461743255,  
 0.12841804657596217]}

RMSE was plotted versus the number of iterations for the final test run. Refer to Figure 7.



*Figure 7 Final Test Run RMSE vs Number of Iterations*

The recommendations for each user were printed to a file. Here’s an example of the output:

userId: 18.0:  
 Top Rated Movie was: Twelve Monkeys (a.k.a. 12 Monkeys) (1995).  
 Top Recommended Movie is: Clockwork Orange, A (1971).  
 Link:<http://www.imdb.com/title/tt66921/>  
 Top Recommended Recent Release Movie is: Grand Budapest Hotel, The (2014).  
 Link:<http://www.imdb.com/title/tt2278388/>  
  
userId: 19.0:  
 Top Rated Movie was: Raising Arizona (1987).  
 Top Recommended Movie is: Lethal Weapon (1987).  
 Link:<http://www.imdb.com/title/tt93409/>  
 Top Recommended Recent Release Movie is: Django Unchained (2012).  
 Link:<http://www.imdb.com/title/tt1853728/>  
  
userId: 21.0:  
 Top Rated Movie was: Man Who Would Be King, The (1975).  
 Top Recommended Movie is: Willy Wonka & the Chocolate Factory (1971).  
 Link:<http://www.imdb.com/title/tt67992/>  
 Top Recommended Recent Release Movie is: Her (2013).  
 Link:<http://www.imdb.com/title/tt1798709/>  
  
userId: 38.0:  
 Top Rated Movie was: Fireworks Wednesday (Chaharshanbe-soori) (2006).  
 Top Recommended Movie is: Gattaca (1997).  
 Link:<http://www.imdb.com/title/tt119177/>  
 Top Recommended Recent Release Movie is: Avengers, The (2012).  
 Link:<http://www.imdb.com/title/tt848228/>

# Further Study

Items to consider before final project report is submitted:

Evaluate the algorithm used for tuning the parameters.

Only tune the parameter lambda. Choose one value for K. For larger values of K, the results are better, but the trade-off is computation time since the value of K determines the sizes of the X and Y factor vectors. For larger datasets, it will be necessary to use small values of K.

Take a closer look at the code for implementing the five-fold cross validation for improvements in speed.

Consider using sentiment classification on the tags of the unrated movies and then use that information for initializing the Y vector for those movies with an “initial” rating rather than a random rating.

Utilize the Amazon AWS service to run the program on a faster more powerful machine. Could increase the number of iterations to 15.

Try loading in the larger movielens dataset using the AWS service and compare the results.

Final Report will include an Appendix with more detailed information on the algorithm and the five-fold cross validation process used for parameter tuning..