IST718 Big Data Analytics

Martin Alonso, LaRue Brown, Rashad Davis

What problem are you attempting to solve?

We want to build a movie recommendation system capable of suggesting new movies to watch based on a user's previous viewing experience, the movies the user rates highest, and the tags used for each movie. Our objective is to build a better recommendation system that takes more input aside from what the viewer's watched.

We also want to understand how movie tags evolve over time, both in usage and the number of movies that are tagged. This would help us understand how genres evolve over time and how it affects movie taste.

What have you observed in the data so far?

So far, we have found two interesting trends in the data set:

1. As time has passed, the average movie rating has declined, going from 3.4 in 1893 to 2.9 in 2015. However, this trend is mired with spikes, with years with a lot of highly rated movies followed by years with poorly rated movies.
2. Another interesting pattern we observed in the data is that reviews are mostly skewed positive, with the average review around 3.53 (median comes in at 3.5). However, it is also interesting that poorly-rated movies have low rating counts, while highly-rated ones have more counts. This makes sense as bad movies will be watched less than acceptable or good movies.

What modeling techniques are you using?

We will start by using clustering methods, grouping the data points in order to observe how movies are bunched together. Outside of K-means, we will also explore other potential clustering techniques such as mean-shift and DBSCAN. For these methods, we will build clusters using a combination of tags and genres.

For the second part of our project, regarding tag prediction for movies, we will try to create a prediction system using a mixture of Support Vector Machines and Neural Networks.

What work do you still have to do?

We've already completed most of the data analysis phase and are moving into the visualization and modeling part of the project. Since we have a clear idea of what we want to analyze and predict, choosing the models was easy.

Over the coming weeks, we'll be fine-tuning the rest of the analysis, building more graphs, and working on the models and their parameter tuning.