IST718 Big Data Analytics

Martin Alonso, LaRue Brown, Rashad Davis, Hana Kim

For our project, we want to build a movie recommendation system capable of suggesting new movies to watch based on a user’s previous viewing experience and movie tastes. Creating this system that not only considers movie genre but is also capable of taking into account user tags will be useful since it will help us understand movie tastes over time.

Additionally, we want to analyze how tags change over time, as movies of same genres may have different tags applied to them at different points of time. We feel that this input is important since movies within the same genre may evolve, showing similarities across several genres that may expand on viewer’s movie experience.

The data for this project were acquired from the GroupLens website (<https://grouplens.org/datasets/movielens/20m/>) and consists of 20 million movie ratings and 465,000 tags applied to over 27,000 movies. The tags were assigned by 138,000 users over a span of 21 years, with each user reviewing at least 20 movies to provide consistency.

The data has been spread out over six comma-separated value files that include data on movie genome scores, genome tags, tags, ratings, links, and movies. The data sets can be merged on movie id, user id, and tag id, making any data merges easy to handle.

The following table has number of rows, columns, and number of variables for each table:

|  |  |  |
| --- | --- | --- |
| Data set | Rows | Columns |
| genome-scores | 11,709,768 | 3 |
| genome-tags | 1,154 | 2 |
| tags | 465,564 | 4 |
| ratings | 20,000,263 | 4 |
| links | 27,278 | 3 |
| movies | 27,278 | 3 |

The tentative assignments for this project will have LaRue Brown and Rashad Davis cleaning the data. LaRue will also explore the data, noting any interesting relationships; Hana Kim will provide data and trend analysis. Rashad, along with Martin Alonso, will build the model frameworks and evaluate the results.