Analyzing User Preferences Through Genome Tags

Alek Petuskey, Ben Huckell, Jenny Yang, Mayank Pandey

Highlights

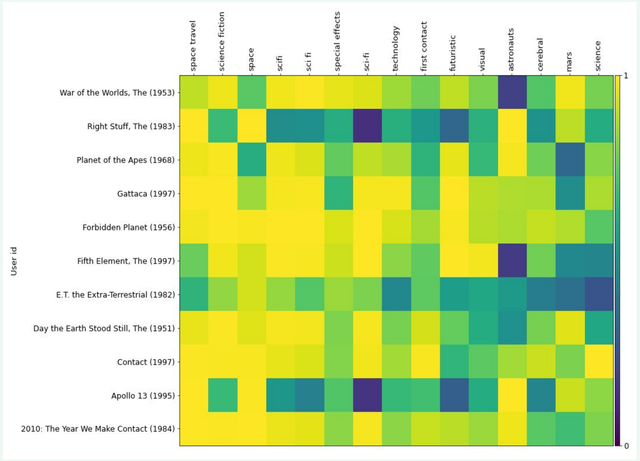
- Accurately classify and cluster users and movies into meaningful groups and show statistical significance of these clusters.
- Extrapolate user tag descriptions beyond movies by comparing comparing their similarity to products with a given set of tags.

Background

In this project, we explored the question of whether or not movies and users could be clustered into meaningful groups based on the ratings given by viewers. We also explored the extent to which these groups evolved over time.

One technique used to model the spectrum of users and movies came from the genome tag relevances provided. Before normalization, the relevance scores directly provide an associated high-dimensional point indexed by genome tags for movies. A point associated to users was also generated from a weighted average of relevance scores over the movies the user rated, weighing highly rated movies more highly. After normalization, the space of user points and movie points become comparable. Furthermore, clustering on these points provides a way to compare users and movies, with closer points indicating users/movies that are similar. Below and to the right are some visualizations of some of what follows, as well as a table validating the superiority of recommendations based on this clustering over a baseline random model.

Heatmap for a subset of movies and genome tags, representing the points on which the movies were clustered



Performance Scores				
# of clusters	Model Cluster		Random Cluster	
	Mean Relevance	Mean Weighted Std	Mean Relevance	Mean Weighted Std
50	0.661	1.215	0.399	1.309
100	0.734	1.117	0.405	1.348
150	0.713	1.074	0.410	1.385
200	0.729	1.003	0.419	1.439
250	0.742	0.973	0.427	1.479
300	0.744	0.956	0.430	1.500
350	0.752	0.935	0.441	1.577
400	0.751	0.886	0.45	1.630

An example of a movies most highly rated by users in a particular cluster with 7 users:

User 1: Strawberry and Chocolate; Love and a .45; Cocoon: The Return, Peter's Friends

User 2: Rumble in the Bronx; Lassie; Seventh Seal, The; MURDER and murder; Black Mask

User 3: Rumble in the Bronx; Tarantula; Romeo and Juliet; Clockwise; Minority Report

User 4: Batman; Ninotchka; Die Hard; Trees Lounge; Shadow Conspiracy

User 5: Shanghai Triad; Across the Sea of Time; Immortal Beloved; Ladybird Ladybird; Naked

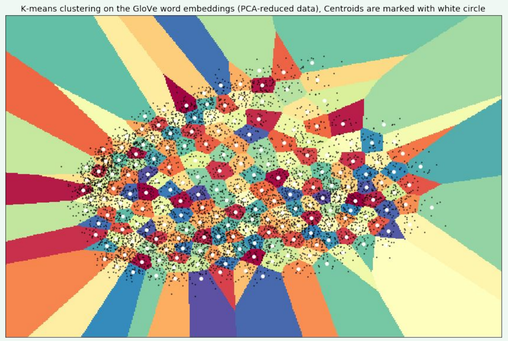
User 6: Pocahontas; Before and After; Ladybird Ladybird; Priest; Loaded

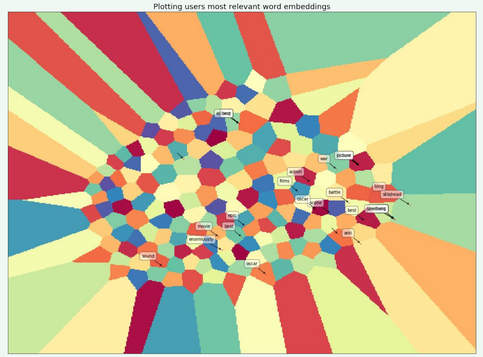
User 7: Man with Two Brains, The; Indochine; Lassie; Scout, The; Beauty and the Beast

Analyzing Beyond Movies

In order to extrapolate genome tags beyond word associations within the given data set we used a pre trained GloveWord Embedding Model. GloVe is an unsupervised learning algorithm for obtaining vector representations for words. In example below we used K-means to cluster the model on PCA reduced data.

A specific user's tags shown in the clusters





We classified users into clusters by only considering their the most relevant tags from their favorite movies. By doing this we gained a description of the users potential intrest. We can compare a users tags to other products to determine the potential that a user will like a certain product with a certain set of tags.