Machine Learning & Data Mining
Miniproject 3

Caltech CS/CNS/EE 155
Released March 3^{rd}, 2017

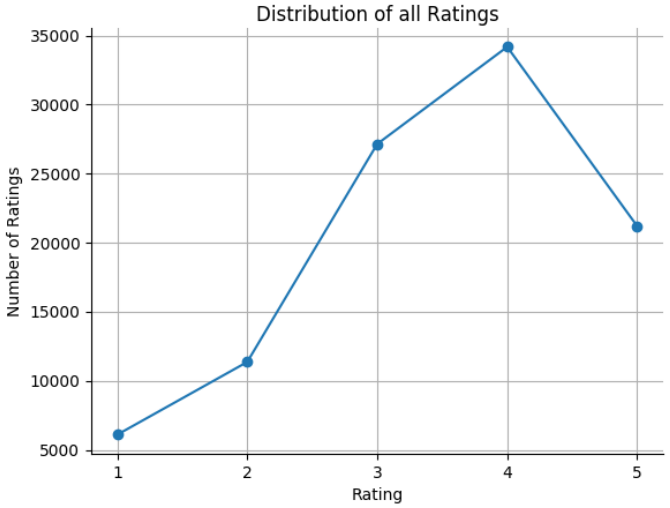
---

1 Introduction

- **Group members:** Enrico Borba, Claire Goeckner-Wald
- **Team name:** Papa Mart's Mini Gary - The End
- **Division of labour:** Enrico Borba: Programming, ideas, report visualization. Claire Goeckner-Wald: Programming, ideas, report assembly.

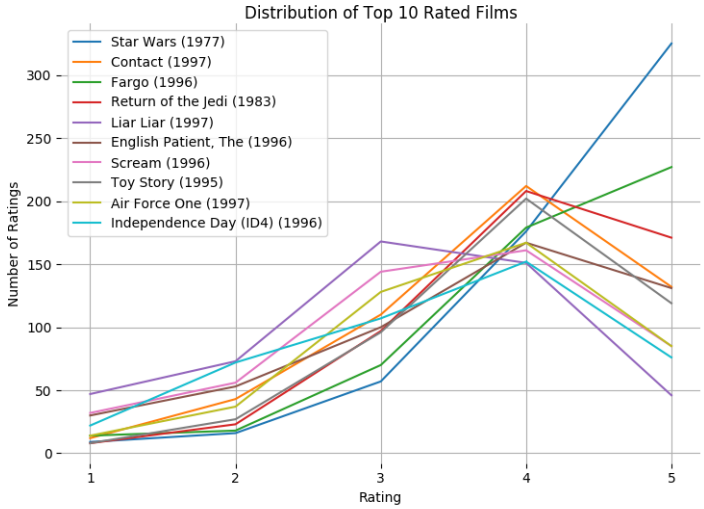
2 Basic Visualizations

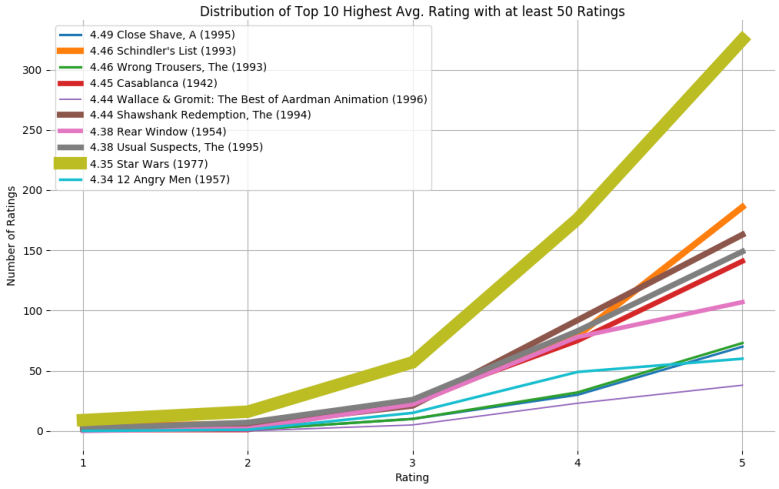
- **Choice of visualization method:** We used a line graph to show visualize the dataset. We didn't use histograms, because the amount of data shown on the chart would have overwhelmed the reader using this method. For visualizing our three genres, we used the size of the dot to indicate the number of genres the movie is in (larger dot implies more genre crossover); the color of the dot indicated the standard deviation of the reviews (yellow implies higher standard deviation, blue implies lower standard deviation.)
- **Observations:** We observed that 3 was not the average rating, perhaps unsurprisingly. While one might expect a uniform, or normal distribution, center around 3 stars, 4 stars was in fact the most common rating given. On second thought, this is perhaps more expected, because people will tend to watch movies they enjoy - so that would skew the distribution towards higher ratings.



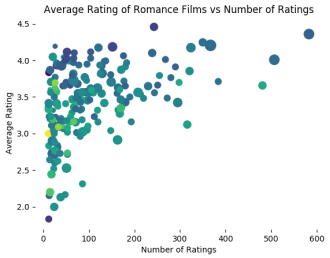
- **Results:** The results approximately matched what we expected.
- **Best ratings:** With the exception of Star Wars, none of the top 10 movies with the highest number of ratings overlapped with the top 10 highest average-rated movies. This is perhaps unexpected. To visualize the highest average-rating movies, we decided to use the size of the line to indicate the

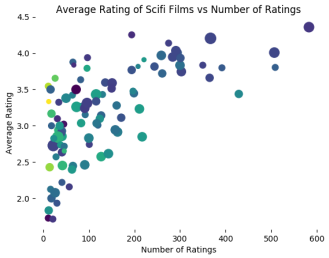
number of ratings for the film. This visualization method really shows how much Star Wars stands out. In order to acquire this list, we did not include films with fewer than 50 ratings, which removed films receiving one/two five star rating (and thus having a very high average of five stars.)

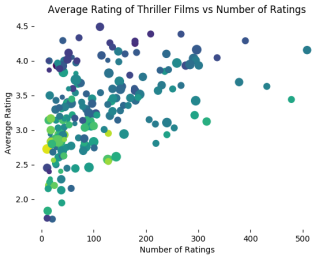




- **Three genres:** We note that the SciFi genre has no movies rated over 4 stars with less than about 200 ratings. We postulate that this is because SciFi's are more difficult to produce, and thus the better scifi films (those with funding) would correlate with more/better advertising (and thus more ratings). As a reminder, we used the size of the dot to indicate the number of genres the movie is in (larger dot implies more genre crossover); the color of the dot indicated the standard deviation of the reviews (yellow implies higher standard deviation, blue implies lower standard deviation.)





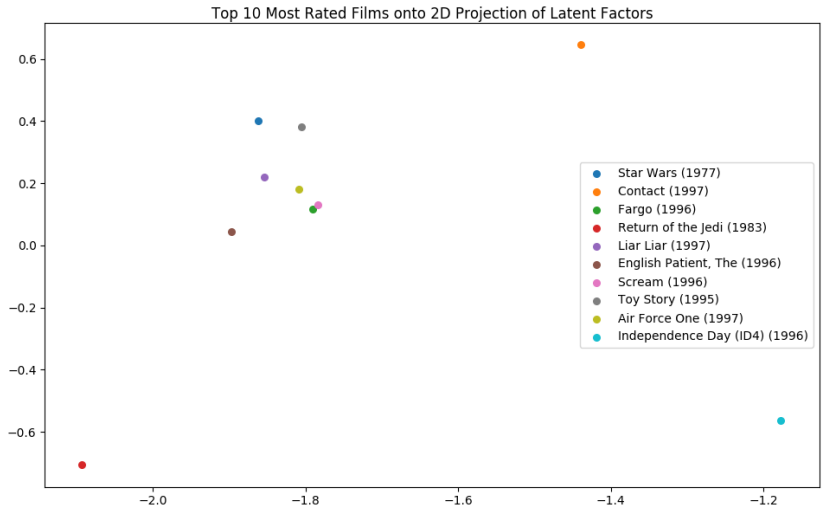


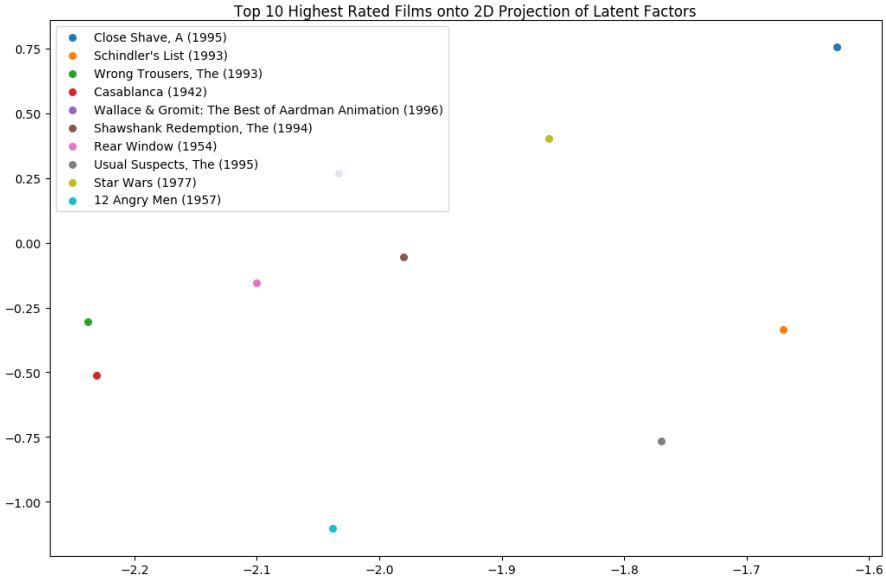
3 Matrix Factorization Algorithm

- **Adjustment of parameters:** We used an η of 0.01, and regularization constant of $\lambda = 10^{-3}$.
- **Justification of parameters and stopping criteria:** We attempted a grid search with $\eta = 0.1, 0.01, 0.001$, and $\lambda = 10^{-2}, 10^{-3}, 10^{-4}$. And found that the lowest loss was encountered at $\eta = 0.01$ and $\lambda = 10^{-3}$. We chose to center the grid search at $\eta = 0.01$ and $\lambda = 10^{-3}$ because those were the parameters we used for $K = 20$ on assignment 6.
- **Significant modifications:** No significant modifications were made. We used the homework solutions as our code to obtain U and V.

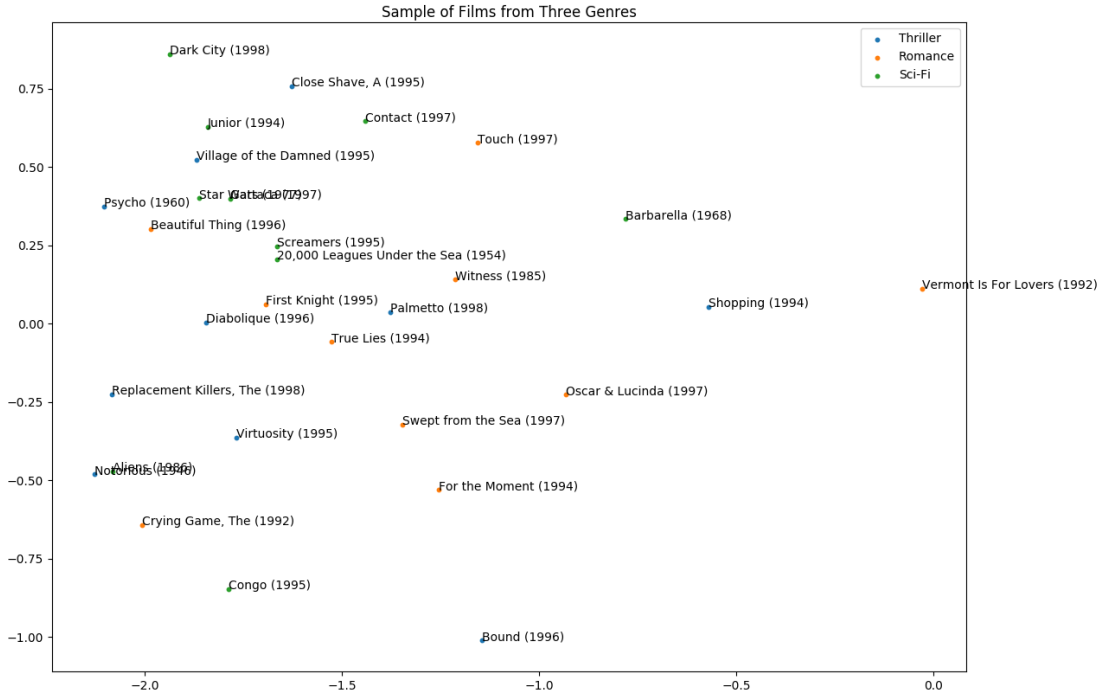
4 Marix Factorization Visualization

- **Observations:** We observed that the matrix factorization method revealed unexpected relationships between some movies.
- **Best/most ratings:** In the top 10 most rated films, there is a marked clustering of 8 of the 10 movies - *Return of the Jedi* and *Contact* are outliers. The distance of *Return of the Jedi* may not be unexpected, due to the remarkable popularity and high ratings that the Star Wars franchise has enjoyed, in general. *Contact* is a 1997 science fiction drama film based on the writings of Carl Sagan. In retrospect, it might be better to plot the two following graphs on the same chart, to tell us more about their respective positions. Either that, or fix the axis.





- **Three genres:** We sampled films from Romance, SciFi, and Thriller. There appears to be some mild to moderate clustering of points, but the films are overall hard to distinguish from location alone, by human eye.



- **Expected, and unexpected:** We expected more demonstrable clustering of points in the three genres. We also noticed that *Gattaca* and *Star Wars* were very, very close together, as well as a handful of other movie pairings. We guess that this makes a statement on the similarities of the movies. However, this one makes more sense than some other film pairings. Some pairs are:
  - *Gattaca* and *Star Wars* (A 1997 science fiction biopunk and a 1977 space opera epic)
  - *Screamers* and *20,000 Leagues Under the Sea* (A 1995 dystopian science fiction and a 1954 Technicolor science fiction)
  - *Aliens* and *Notorious* (A 1986 science fiction action horror and a 200biographical drama about Notorious B.I.G.)

  Weirdly, all of the above are SciFi, with the exception of Notorious, which was a Thriller.

5 Conclusions

- **Summary:** Our main observations were that 'top-rated' and 'best-rated' movies tend to be two very different things. Moreover, the matrix factorization revealed unusual relationships between certain movies.

- **Did it help?:** The visualizations helped us better understand the MovieLens dataset.