Anime Recommendation System

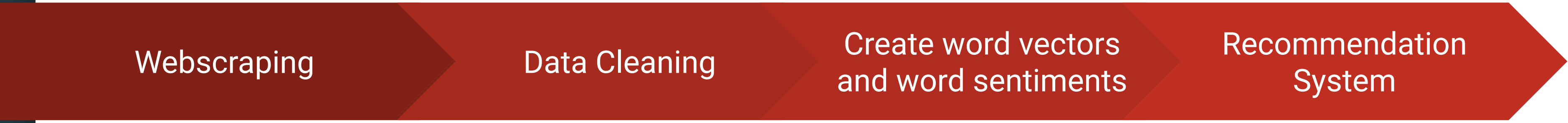
By: Antonio Hila
&
Zawar Ahmed

Why Anime?

Task: Create an Anime Recommendation system based on user ratings and user review text data.

Purpose: Generate recommendations for people who like anime and want to find something to watch.

Process



Webscraping

Data Cleaning

Create word vectors and word sentiments

Recommendation System

Gather reviews and ratings for over 1000 anime from myanimelist.net.

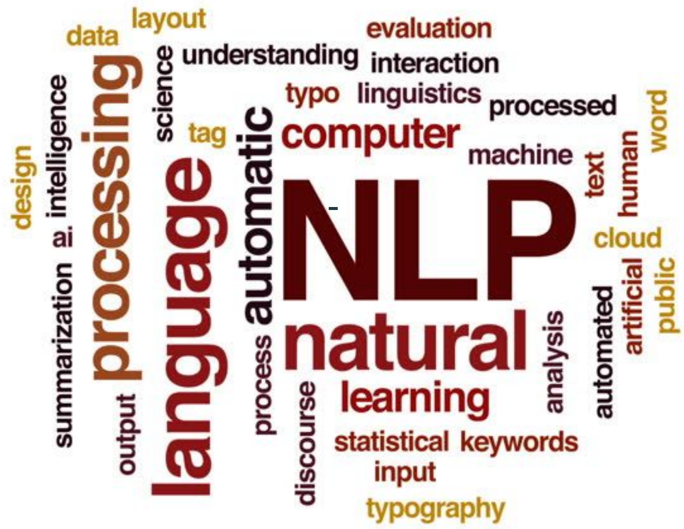
Clean the NLP data using regex and tokenizer

Use Spacy and TextBlob to convert the user reviews into a single vector score or sentiment score

Use all the different rating scores to make different models and gridsearch to tune the models.

Data Processing

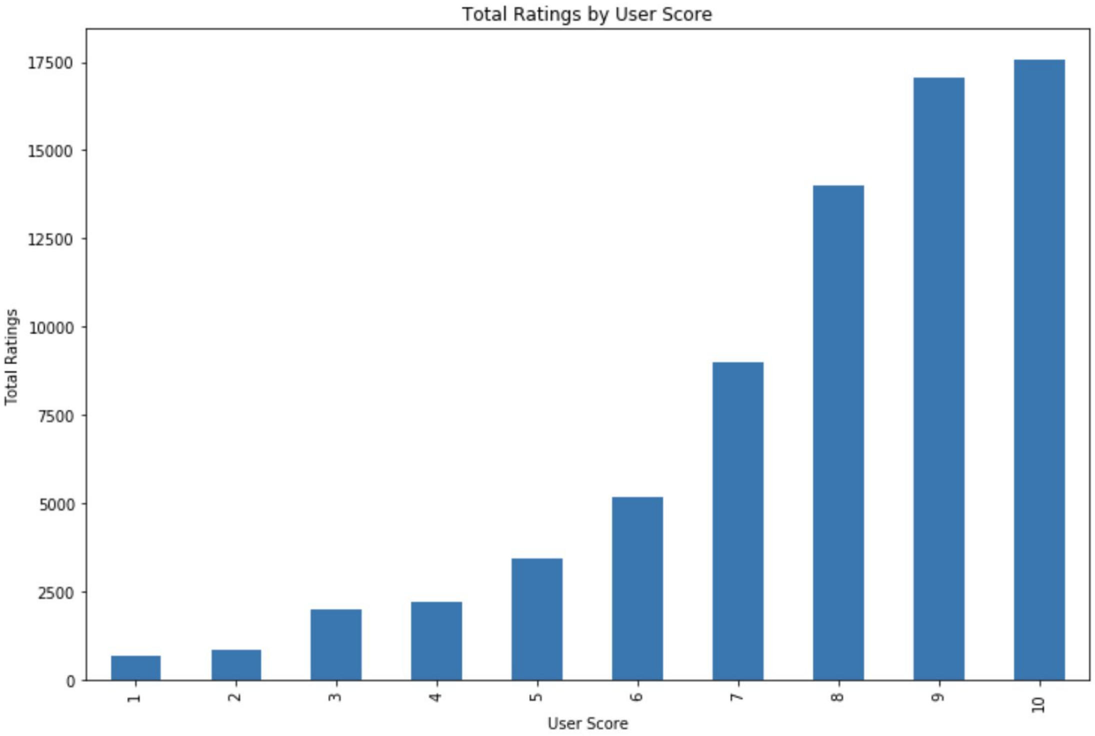
- Made separate Dataframes for ratings and reviews
- Cleaned up reviews, taking out punctuation, white space and numbers.
- Final DataFrame included ratings along with sentiment and SpaCy scores



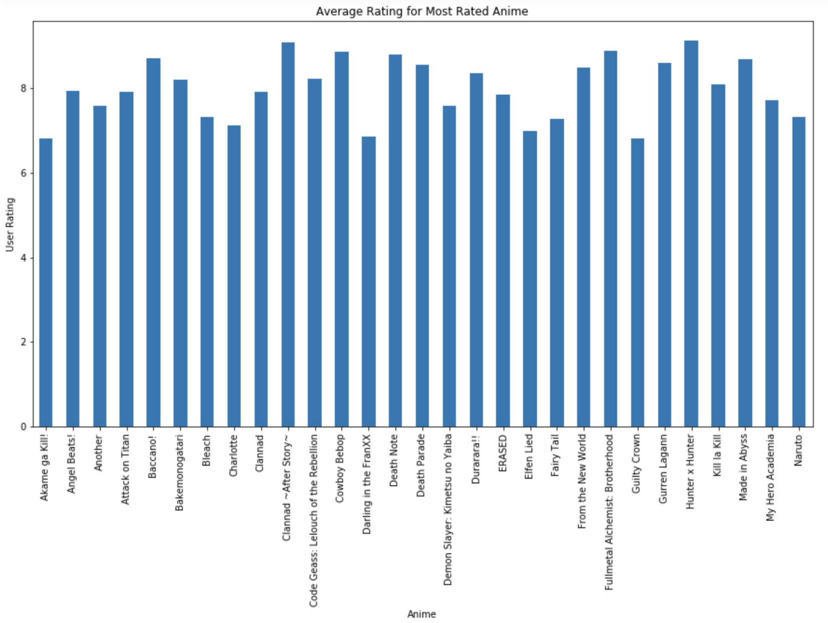
EDA

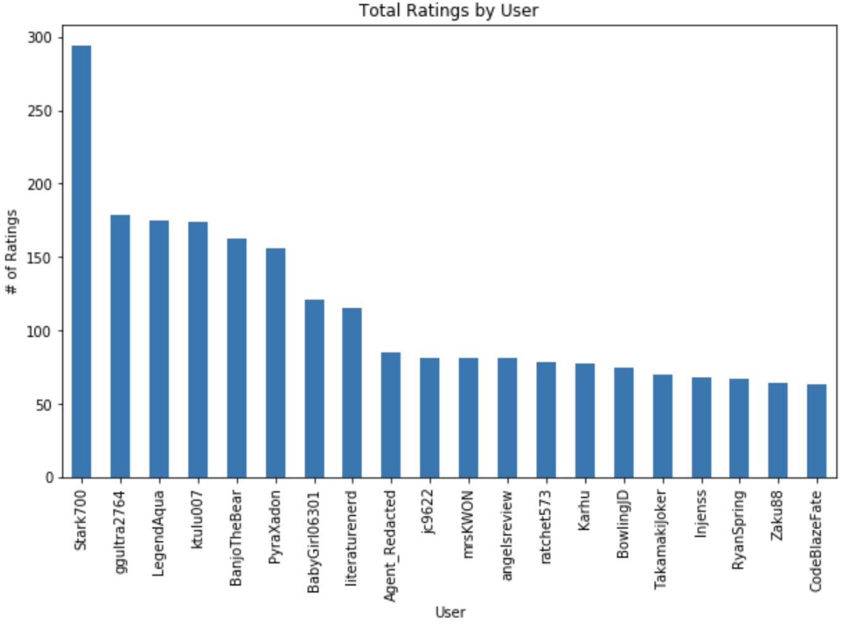
Most anime rated above a 7

Ratings go from 1 - 10



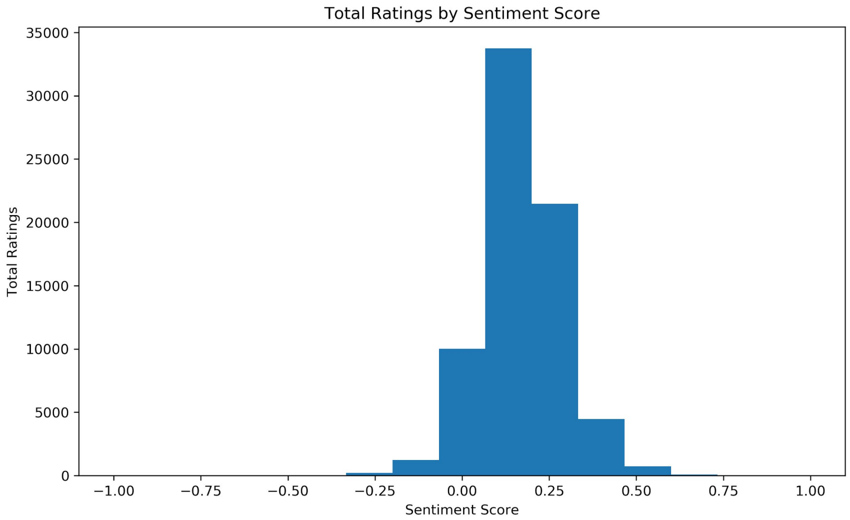
EDA

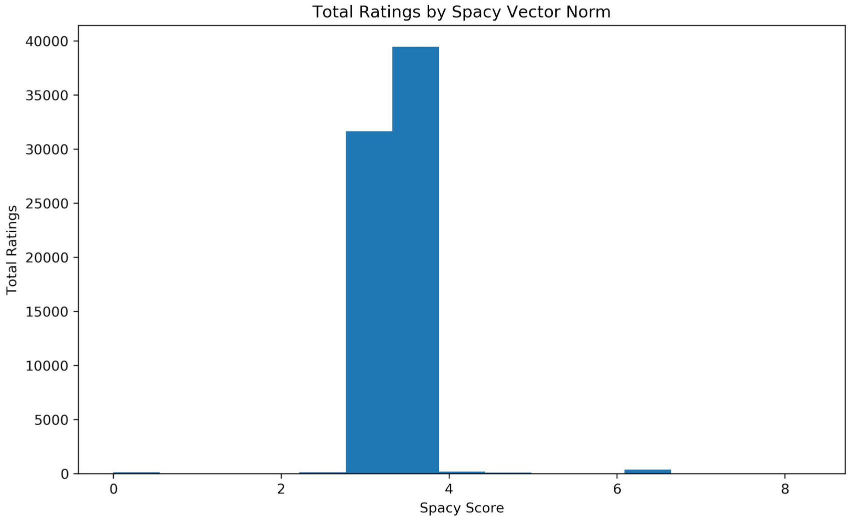




- Most of the anime that is rated often has above a 7 score.
- The amount of shows rated by a single user was between 1 and 285

NLP EDA





- Even though scores range from -1 to 1 for the sentiment and 0 - 8.5 for the spacy vector score the bulk of the scores are in a much tighter range

Baseline, KNN & SVD Models for Ratings

RMSE: 1.833

Baseline only Model

-Takes into account user and item bias
-Has regularization factors to limit

- SGD
- Reg=0.02
- N_epochs = 25

RMSE: 1.895

KNN-Baseline Model

- Item-Item collaborative filtering system
- Also uses baseline biases

- Baseline:
  - SGD
  - reg=0.02
- Similarity:
  - Pearson baseline
  - Shrinkage=50
  - Min_support=5
- k=15

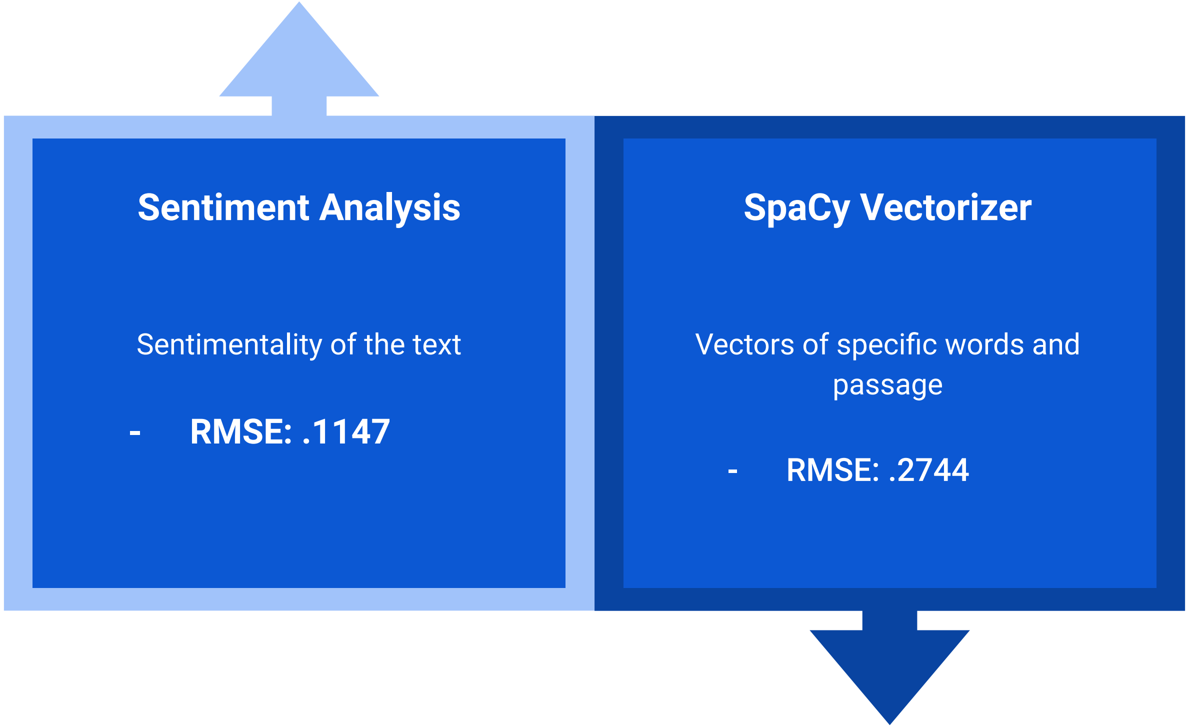
RMSE: 1.8326

SVD Model

- Collaborative filter system
- Technique used to reduce dimensionality
- Also takes Bias into account

- N_factors = 200
- N_epochs = 15
- Lr_all = 0.05
- Reg_all = .1

Sentiment Analysis vs SpaCy vectorizer



Sentiment Analysis

Sentimentality of the text

- **RMSE: .1147**

SpaCy Vectorizer

Vectors of specific words and passage

- **RMSE: .2744**

Best Models

01

SVD model with User Ratings

- **RMSE: 1.8326**

- Scores range between 0 and 10
- More diverse in recommendations

02

Baseline Model for User Ratings

- RMSE: 1.833

- Scores range between 1 and 10

03

SVD with Sentiment Data

- **RMSE: .1147**

- Scores range from -1 to 1
- Most scores clustered between -.25 to .5
- Lean towards the highest rated anime recommendations

04

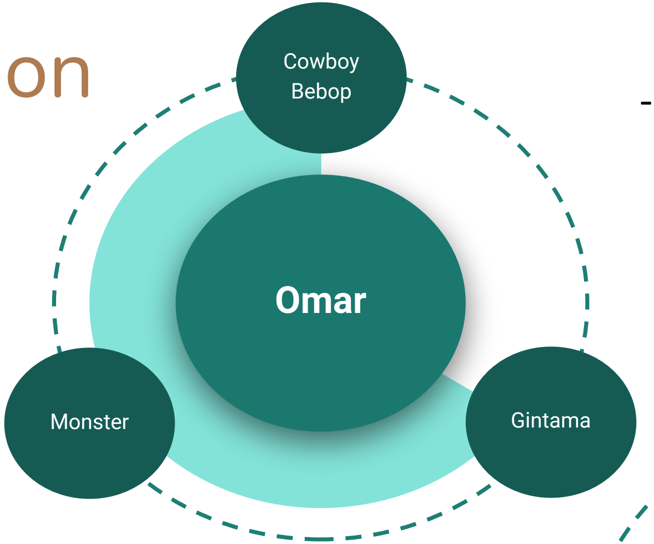
SVD with SpaCy vector

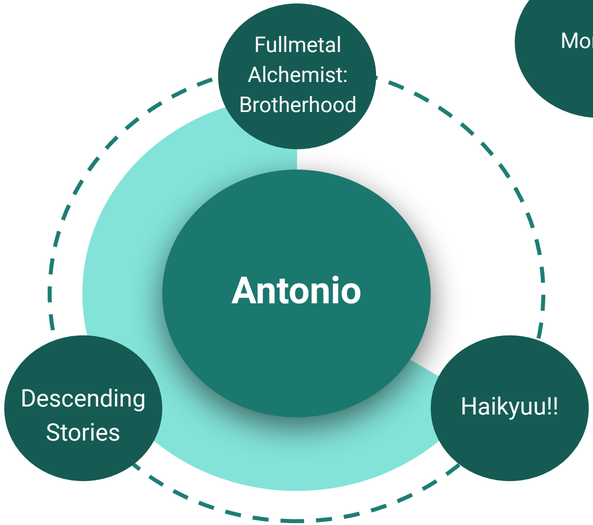
- RMSE: .2744

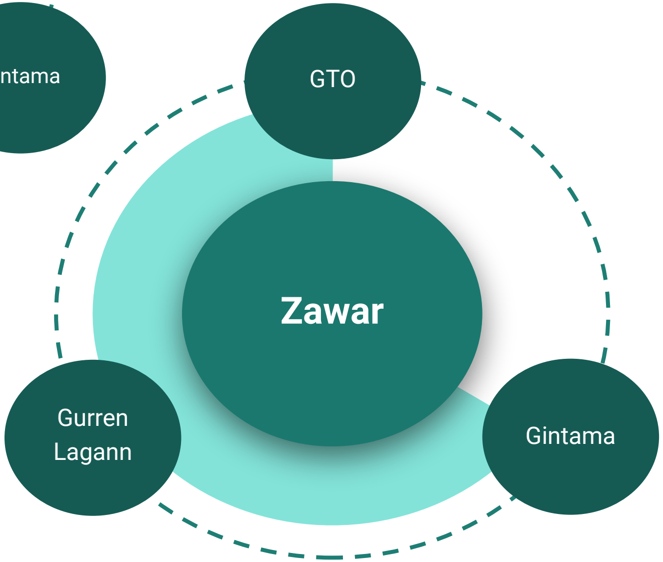
- Scores range from 0 to 10
- Most scores clustered between 2.5 and 4

Recommendation systems Predictions

- SVD Model Predictions using only ratings for the anime







Conclusion

- SVD using sentiments had the best overall RMSE
- Sentiment scores require user reviews which is not good for new users to get recommendations quickly, harder to test
- SVD using user ratings was most practical and performed best among rating only models
- Easier and faster to rate a few anime you have watched and get recommendations
- Did well in our tests

Appendix

- Things to improve upon
  - Adding together all the seasons of a show
  - Finding different libraries that can use the NLP data in conjunction with the ratings

- https://github.com/Ahila700/Anime_recommendation_system
- https://docs.google.com/presentation/d/1q0oUCRAW5AnT161LhBWXp6nwBUl1_xfbZ8tHjAsNlSQ/edit?usp=sharing