Midterm Report: Movie Recommendation

by Grant Doan, Prayash Joshi, Reagan Orth, Ved Patel

Problem definition / motivation

Motivation:

- Create a recommendation system that accommodates for the entirety of a user's interest
- Create a model that focuses on the similarities between different users

Problem:

- Recommendations are often centered around one movie
- In other cases, recommendations are centered around search history, not preference
- Popular movies are far more highly considered
    - Sunset Blvd. and The Godfather are often recommended together
- Users are never asked for preferences
- Recommendations cannot reliably improve

Existing solutions and limitations

Cold Start:

- Initially, websites like IMDB have no prior data on a user
- This makes recommendations difficult

Data Sparsity:

- Difficult to find users that have rated the same movies

Scalability:

- With the large amount of data being used on popular websites, inaccurate results are inevitable

Filter Bubble:

- Users are often diverged from having different perspectives

Proposed approach

Web scrape two datasets from IMDb:

- Dataset one: Reviews of top movies
- Dataset two: Title, Description, Genre, Directors, and Actors of top 1000 movies

NLP preprocessing:

- Cleaning up punctuation and creating a sparse matrix on the reviews
- DF/TDF to remove insignificant words
    - Give special attention to names mentioned in reviews
- Pretrained BERT for Aspect Based Sentiment Analysis
- TF/IDF for finding specific film elements that are preferred by users

Graph Neural Network:

- Use a Graph Neural Network to learn high-dimensional mappings for movies

Proposed approach continued...

Application of our GNN:

- Large, undirected, unweighted graph
- Movies are nodes
- Edges represent similarities
- Similar movies will be close together

Use user ratings and/or reviews to check user preferences:

- Check movies most similar to movies the user already likes
- Check descriptive words in any user reviews provided
- Combine BERT pretrained-model and TF/IDF to find what user’s find interesting in movies
- Combine movie preferences

Expected impact

- Find movies more suited to individual taste
- Reduce decision time
- Save money
- Increase scope of search
- Take forgotten movies into account
- Leverage the opinions of similar people
- Find joint preferences for group viewings
- Discover new movies

Data and Features

Two datasets:
- Scraped movie information
- Scraped user reviews

Graph Creation
- Scraped movie information dataset
  - Year, actors, director, genres
  - Modified TF/IDF from plot synopsis

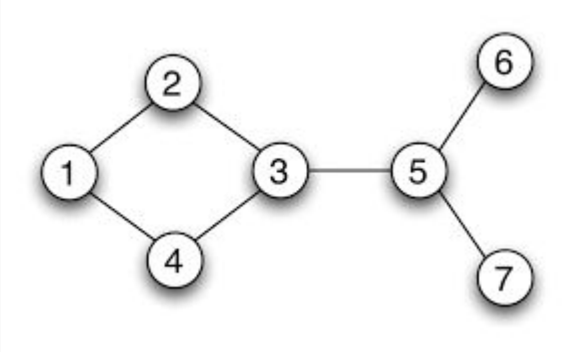
User Recommendation
- User reviews dataset
  - Modified TF/IDF from reviews
  - User ratings
  - Other movies each user has seen

title	Action	Adventure	Crime	Drama	John Huston	Martin Scorsese	Al Pacino	Bette Davis	Brad Pitt
Life Is Beautiful	False	False	False	True	False	False	False	False	False
It's a Wonderful Life	False	False	False	True	False	False	False	False	False
Seven Samurai	True	False	False	True	False	False	False	False	False
Harakiri	True	False	False	True	False	False	False	False	False
Parasite	False	False	False	True	False	False	False	False	False
The Departed	False	False	True	True	False	True	False	False	False
Whiplash	False	False	False	True	False	False	False	False	False
Gladiator	True	True	False	True	False	False	False	False	False
Back to the Future	False	True	False	False	False	False	False	False	False
The Prestige	False	False	False	True	False	False	False	False	False
Alien	False	False	False	False	False	False	False	False	False
Léon: The Professional	True	False	True	True	False	False	False	False	False
The Lion King	False	True	False	True	False	False	False	False	False

	the	movie	satire	cat	superhero	imperceptible
The Menu	3252.0	640.0	68.0	1.0	NaN	NaN
Antman	3465.0	774.0	NaN	NaN	20.0	NaN
Puss In Boots	3157.0	872.0	NaN	54.0	NaN	1.0
Totals	9874.0	2286.0	68.0	55.0	20.0	1.0

	I	Great	Disney	Horror	Fiennes	Marvel	Dreamworks
The Menu	571.0	2.0	3.0	2.0	76.0	1.0	NaN
Antman	648.0	4.0	36.0	NaN	NaN	144.0	NaN
Puss In Boots	642.0	3.0	29.0	NaN	NaN	NaN	49.0
Totals	1861.0	9.0	68.0	2.0	76.0	145.0	49.0

Experimental Methodology and Evaluation



Data preprocessing
- Modified TF/IDF preprocessing

Sentiment analysis
- Considering using external library with set words
- Leaning towards using RNN

Graph creation
- Using PyG
- Undirected graph in high-dimensional space modeling similarities
- GCN for edge prediction

User mapping
- Considering mapping users directly onto the graph
- Considering using the graph with an external prediction algorithm

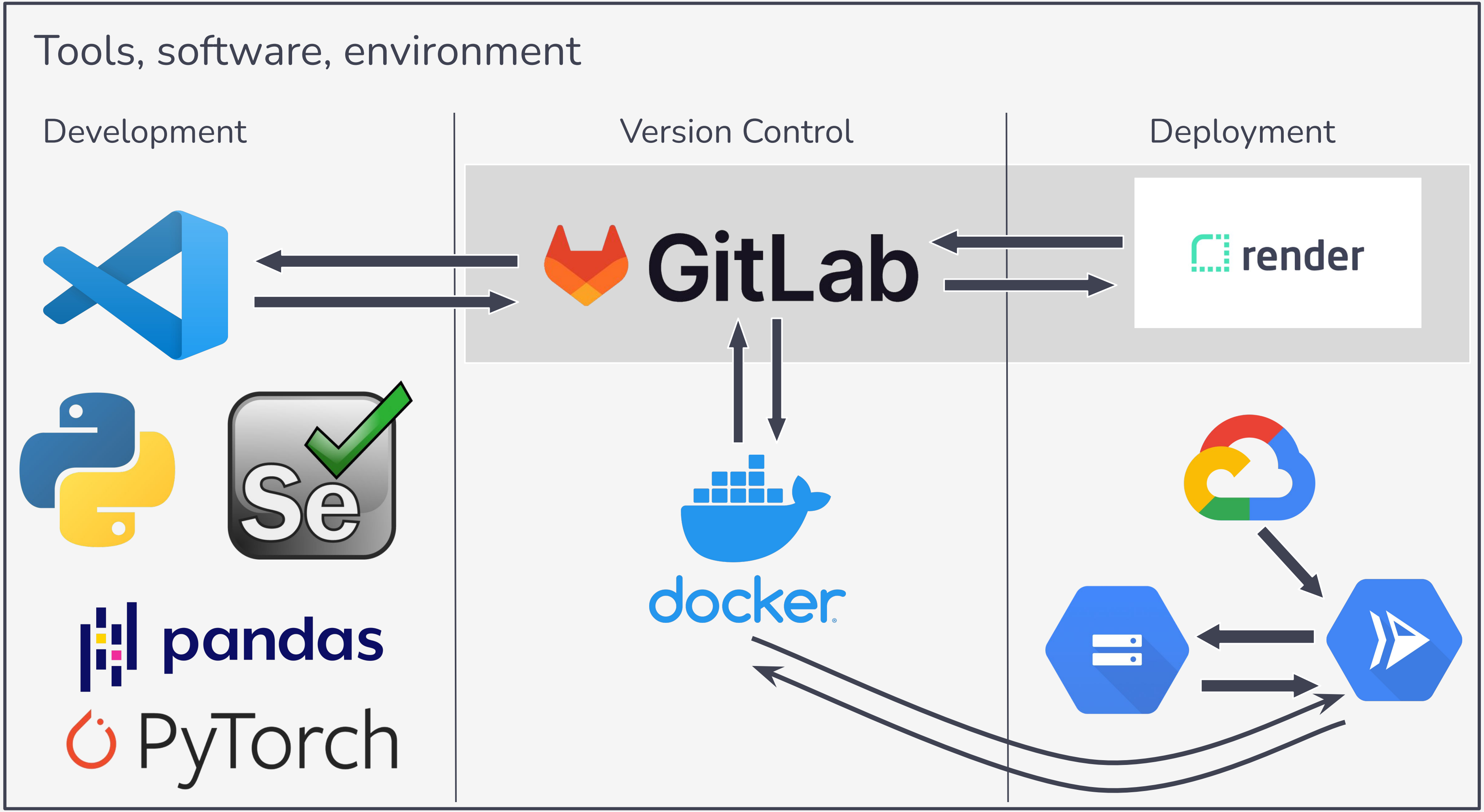
Potential risks and mitigation strategies

- R-rated movies can currently be recommended to children
    - Can take rating into account

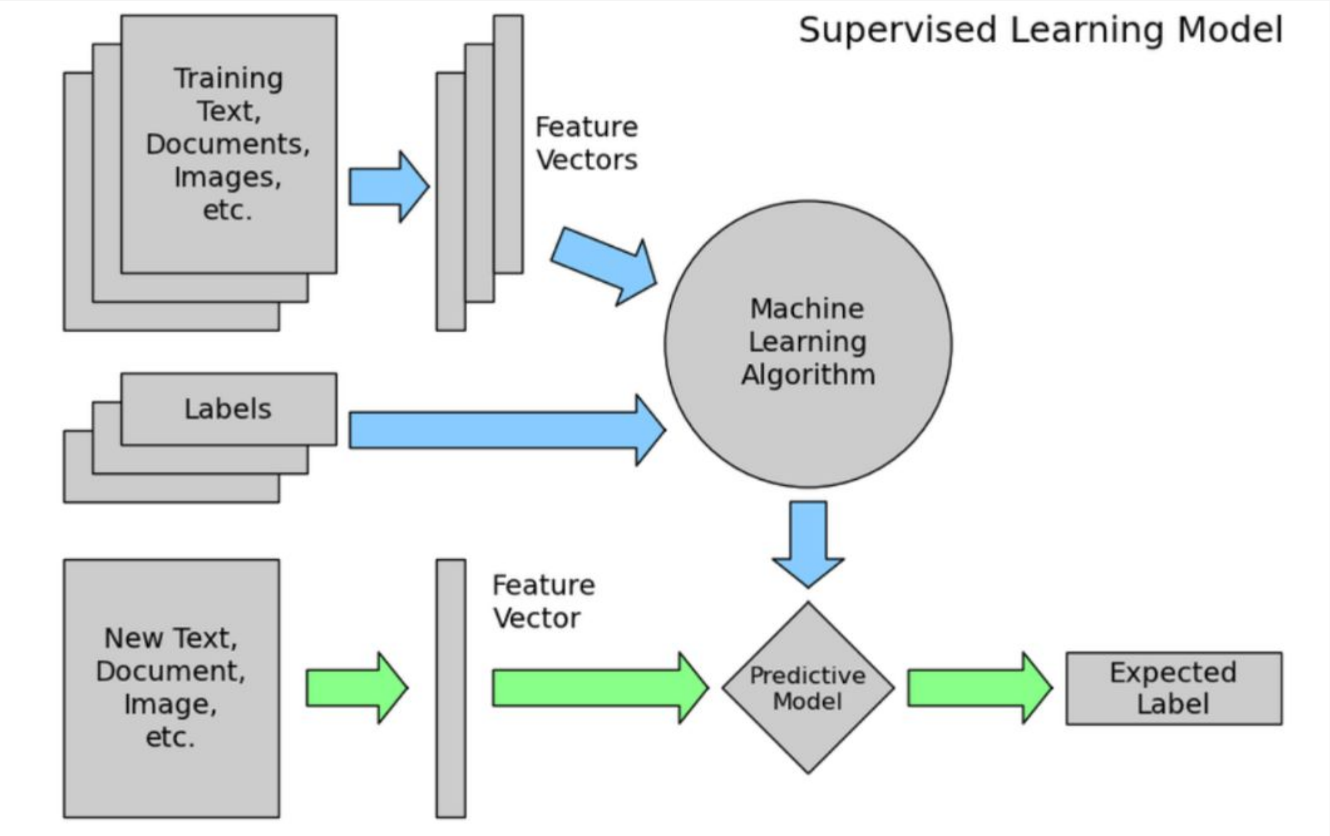
- Movies might not be available for streaming or may cost money
    - Could web scrape justwatch or IMDb to check

- Some movies share a name
    - Use unique ID within the program
    - Add year for users

- No guarantee of accuracy of model


Tools, software, environment
Development
Version Control
Deployment
GitLab
render
pandas
PyTorch
docker

Pipeline diagram



Progress To Date

Classification

- ✅ Literature Review
- ✅ Preliminary Website
- ✅ Feature Selection
- ✅ Data Cleaning
- ✅ TF/IDF preprocessing
- ✅ Initial Sentiment Analysis
- ✅ Neural Net Sentiment Analysis
- ✅ Preliminary WebApp
- ✅ Preliminary Database
- ☐ Precision/accuracy optimization

Recommendation

- ✅ Naive Recommender System
- ✅ Prepare data for GNN
- ✅ Create initial GNN
- ✅ Map user preferences
- ☐ Improve GNN
- ☐ Evaluate recommendations
- ☐ Storing data in GCP buckets
- ☐ GCP Cloud Run of Container
- ☐ Integration with App

Project Timeline

03/17

Begining of webscrapping

03/24

Webscrapping completed

03/31

- webscraped data was cleaned
- NLP preprocessing
- put data in format for GCN
- prelimanry webapp created

04/02

created larger dataset of movie description (1000 movies)

04/7

- Improve GNN
- Evaluate Recommendations

04/14

Mean user prediction rating within three points of real ratings

04/21:

Mean user prediction rating within two points of real ratings

Tasks breakdown and team members contributions

Prayash: Static Website, Firebase Backend Database, Dynamic WebApp

Reagan: Movie dataset preprocessing, Graph Neural Network

Grant: Sentiment analysis, scraping user reviews

Ved: Scraping movie dataset, Website graphics

References

https://www.researchgate.net/figure/An-undirected-graph-with-7-nodes-and-7-edges_fig3_265428782