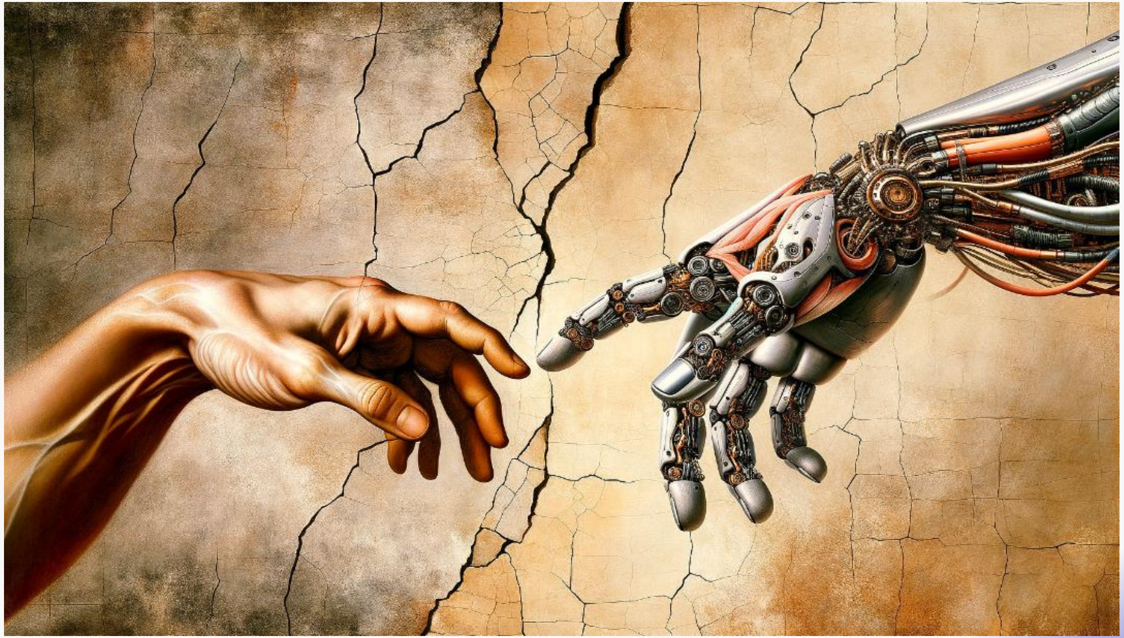
ML based Movie Recommendation System

Capstone Project: Group 4

Tell me Oh AI Lord! Which movie to watch next?



Team Members

TANUMAY DATTA	Content based filtering using TF-IDF, model validation and documentation
ANURAG SHARMA	KNN and Softmax based Collaborative filtering, Github repo maintenance, CI/CD pipeline, Docker
MAHALAKSHMI	SVD based Collaborative filtering, model validation and documentation
SUDHANSHU PATHRABE	Content based filtering using NLP
SHREYA DAS	TMDB 5000 data pre-processing
SATENDAR KUMAR TIWARI	

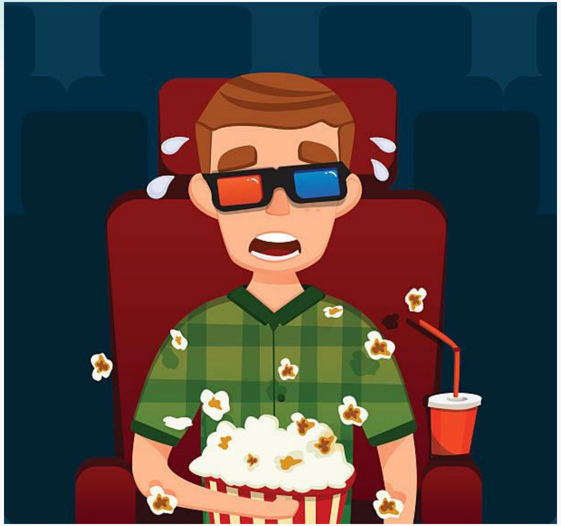
Introduction

- Algorithm suggests a user movies based on their viewing history, suggestions, and ratings
- Widely used in different streaming platforms
- Variants of the same can be used for book and music recommendations as well

Dataset used for the project

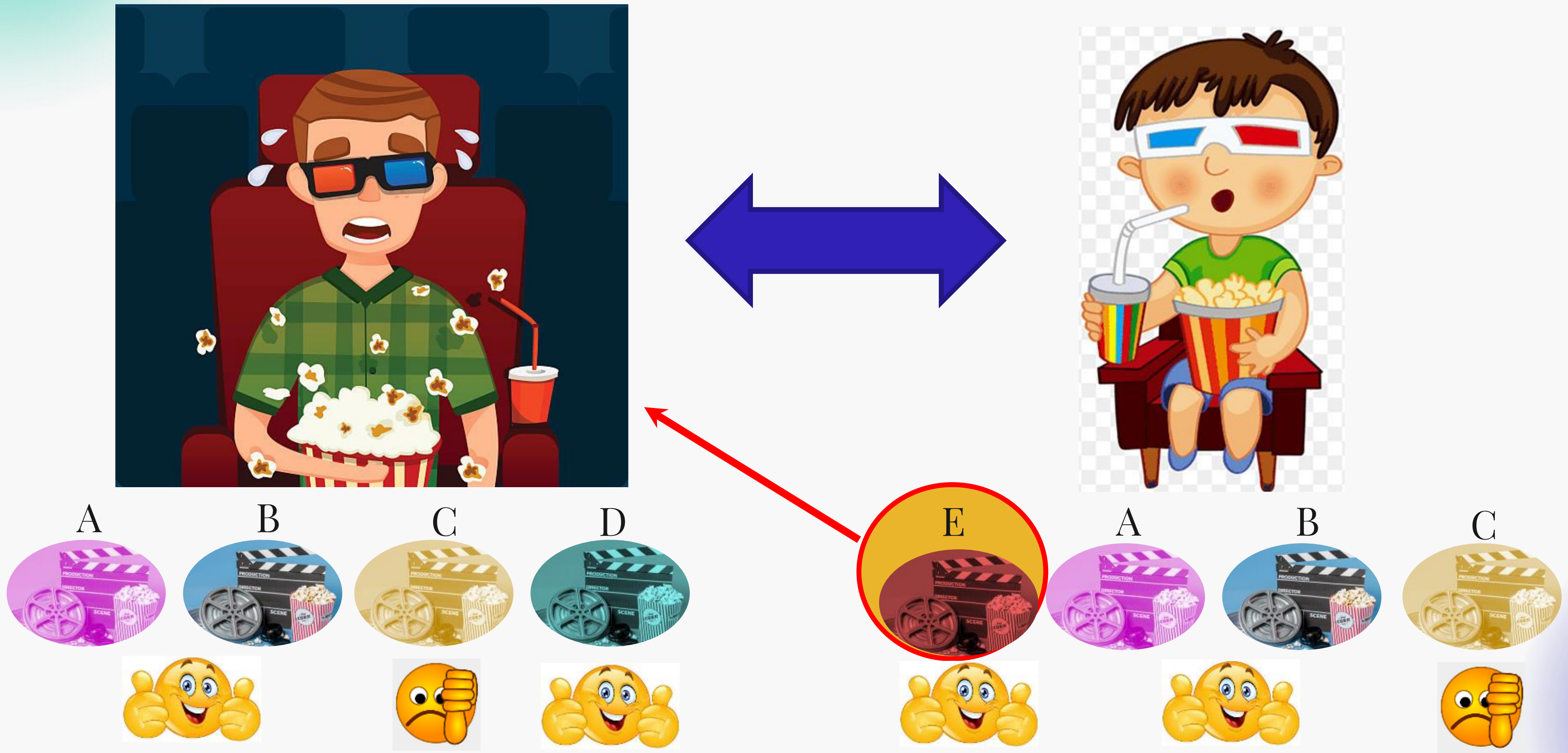
- KMDB 5000 database
  - 4804 movies with 23 input features
- Movielens 100K database
  - 100,000 ratings from 1000 users on 1700 movies

Content based Filtering



"I see dead people..."

Collaborative Filtering



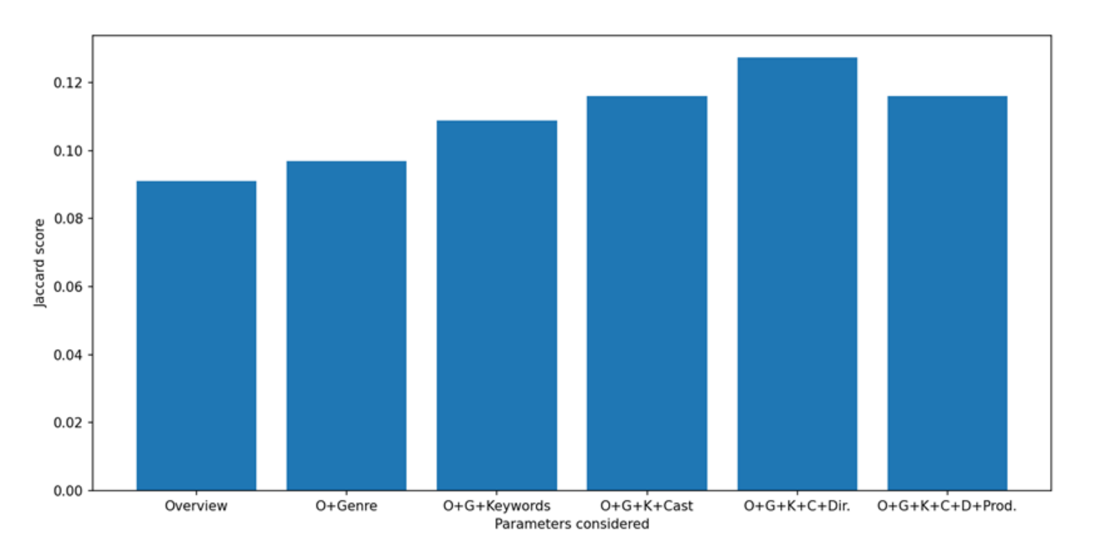
ML Methodologies used

Content Based Filtering

- TF-IDF based vectorization of movie characteristics and Cosine similarity to compute similarity between movies
- TF-IDF and Count Vectorization is compared. TF-IDF is shown to perform better.
- Output is validated using Jaccard Similarity index with Tastedive.com output

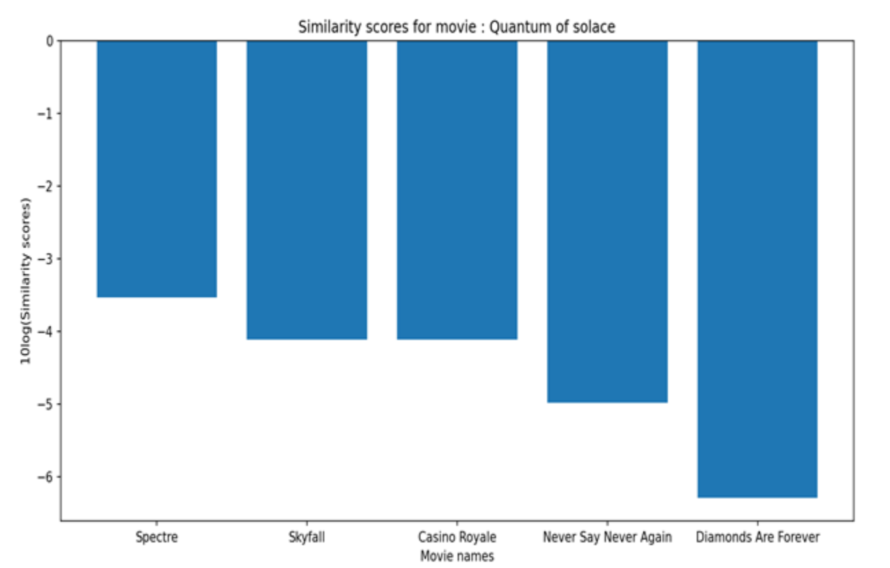
Model Validation with Content-based Filtering

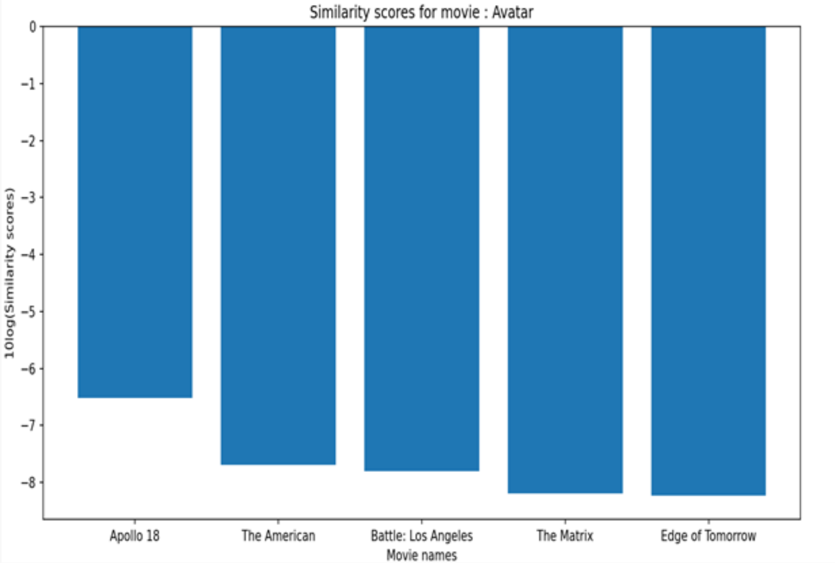
- Model output is compared with output from Tastedive.com, which is a standard website for movie recommendations
- Accuracy of our recommended set is computed by comparing with tastedive output using Jaccard Similarity.
- For example, a average Jaccard index is computed for a set of 10 movies, and compared for different feature combinations used in the model.
- From this result, we select a combination of overview, genres, keywords, cast and director features for the model training.



Results with Cosine-Similarity

- TF-IDF is used to vectorise overview, keywords, cast, director, genre





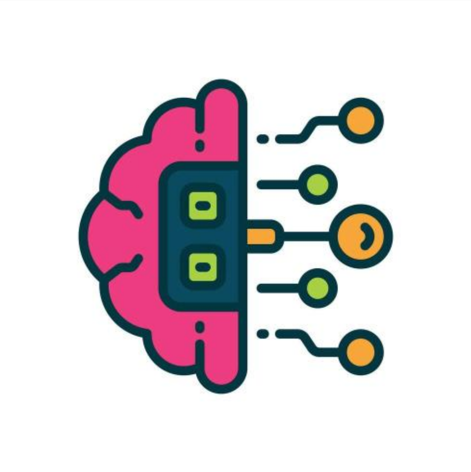
- Algorithm seems to work well with movies that are in a series like Harry Potter, James Bond etc.
- For Stand-alone movies, the recommendation is not that useful.

ML Methodologies used

Collaborative Filtering

SVD Approach :

- SVD is applied to the normalized utility matrix, breaking it down into three matrices (U, Σ, V^T), which reduces dimensionality and reveals hidden patterns such as user preferences and movie characteristics.
- SVD uses user ratings to uncover latent factors, enabling the system to identify similarities between movies based on their ratings and recommend the most related ones.



SVD Metrics

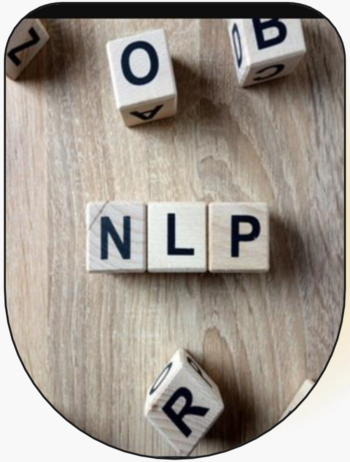
5-fold cross-validation with a 25% train-test split

Metric	SVD (Mean ± Std)
RMSE (test set)	0.9349 ± 0.0038
MAE (test set)	0.7366 ± 0.0028
Fit time	1.52 ± 0.16
Test time	0.18 ± 0.08

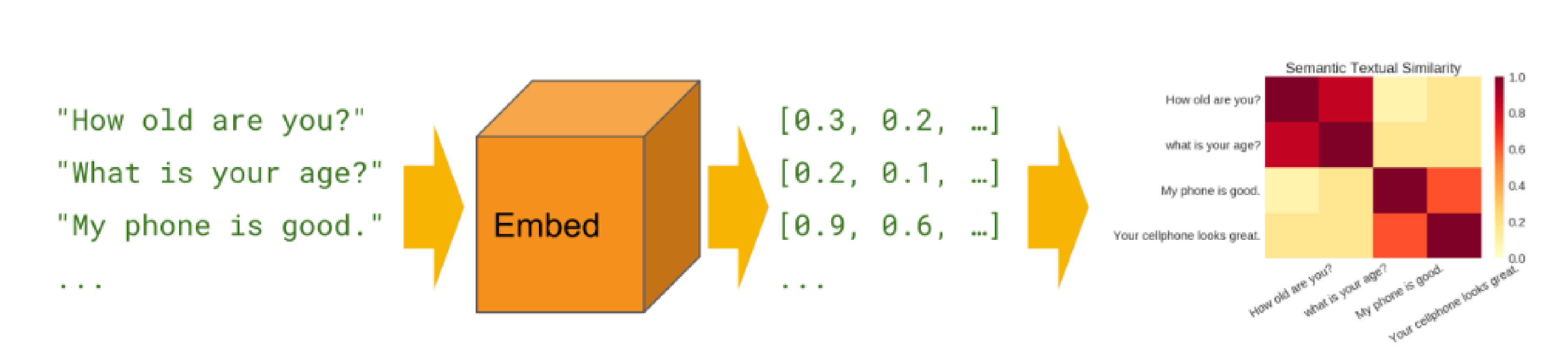
ML Methodologies used

Content Based Filtering using NLP

- Used TMDB 5000 Dataset
- Encodes text into 512-dimensional vectors for tasks like text classification, semantic similarity, and clustering.
- Uses a Deep Averaging Network (DAN) encoder to model the meaning of word sequences rather than individual words.
- Used PCA for data- visualisation.
- KNN algorithm is used to find most similar movies.



Metrics used NLP



Results with NLP

Model	SST 1k	SST 2k	SST 4k	SST 8k	SST 16k	SST 32k	SST 67.3k
Sentence & Word Embedding Transfer Learning							
USE_D+DNN (w2v w.e.)	78.65	78.68	79.07	81.69	81.14	81.47	82.14
USE_D+CNN (w2v w.e.)	77.79	79.19	79.75	82.32	82.70	83.56	85.29
USE_T+DNN (w2v w.e.)	85.24	84.75	85.05	86.48	86.44	86.38	86.62
USE_T+CNN (w2v w.e.)	84.44	84.16	84.77	85.70	85.22	86.38	86.69
Sentence Embedding Transfer Learning							
USE_D	77.47	76.38	77.39	79.02	78.38	77.79	77.62
USE_T	84.85	84.25	85.18	85.63	85.83	85.59	85.38
USE_D+DNN (lrn w.e.)	75.90	78.68	79.01	82.31	82.31	82.14	83.41
USE_D+CNN (lrn w.e.)	77.28	77.74	79.84	81.83	82.64	84.24	85.27
USE_T+DNN (lrn w.e.)	84.51	84.87	84.55	85.96	85.62	85.86	86.24
USE_T+CNN (lrn w.e.)	82.66	83.73	84.23	85.74	86.06	86.97	87.21
Word Embedding Transfer Learning							
DNN (w2v w.e.)	66.34	69.67	73.03	77.42	78.29	79.81	80.24
CNN (w2v w.e.)	68.10	71.80	74.91	78.86	80.83	81.98	83.74
Baselines with No Transfer Learning							
DNN (lrn w.e.)	66.87	71.23	73.70	77.85	78.07	80.15	81.52
CNN (lrn w.e.)	67.98	71.81	74.90	79.14	81.04	82.72	84.90

ML Methodologies used

Collaborative Filtering & Deep Learning using Softmax

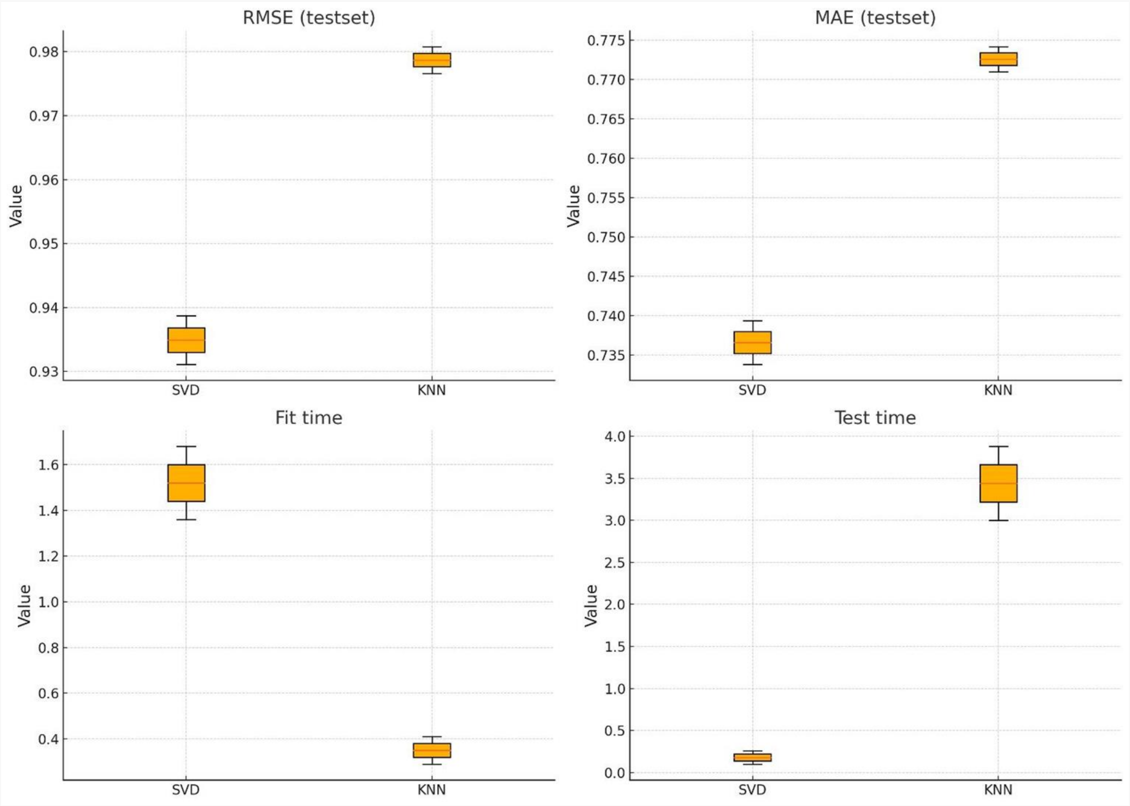
- KNN identifies similar users based on rating patterns using cosine similarity
- Deep learning with softmax adds sophisticated pattern recognition through:
    - User and movie embeddings that capture latent features
    - Neural networks that learn complex relationships
    - Softmax activation for probabilistic recommendations
- RMSE & MAE used as validations metrics.

Using Softmax

Layer (type)	Output Shape	Param #	Connected to
user_input (InputLayer)	(None, 1)	0	-
input_layer_1 (InputLayer)	(None, 1)	0	-
embedding_2 (Embedding)	(None, 1, 150)	141,450	user_input[0][0]
embedding_3 (Embedding)	(None, 1, 150)	249,600	input_layer_1[0]…
reshape_2 (Reshape)	(None, 150)	0	embedding_2[0][0]
reshape_3 (Reshape)	(None, 150)	0	embedding_3[0][0]
concatenate_1 (Concatenate)	(None, 300)	0	reshape_2[0][0], reshape_3[0][0]
dropout_3 (Dropout)	(None, 300)	0	concatenate_1[0]…
dense_3 (Dense)	(None, 32)	9,632	dropout_3[0][0]
activation_3 (Activation)	(None, 32)	0	dense_3[0][0]
dropout_4 (Dropout)	(None, 32)	0	activation_3[0][…
dense_4 (Dense)	(None, 16)	528	dropout_4[0][0]
activation_4 (Activation)	(None, 16)	0	dense_4[0][0]
dropout_5 (Dropout)	(None, 16)	0	activation_4[0][…
dense_5 (Dense)	(None, 9)	153	dropout_5[0][0]
activation_5 (Activation)	(None, 9)	0	dense_5[0][0]

Results

Comparison between KNN and SVD



Conclusion

- Both Content based filtering and Collaborative filtering methods are studied.
- Various parallel approaches are taken for each of the techniques.
- TF-IDF and word embedding techniques are used content based filtering.
- SVD, KNN and NN with Softmax approaches are adapted for Collaborative filtering.
- Simplified CI/CD pipeline has also been implemented.

Contribution

	TANUMAY (TF-IDF Content based)	ANURAG (KNN & Softmax)	MAHALAKSHMI (SVD Collaborative based)	SUDHANSHU (NLP Content based)	SHREYA	SATENDER
Data cleanup and acquisition	100	100	100	100		
ML Model Selection/ Training	100	100	100	100		
Hyper parameter tuning	100	100	100	100		
Metrics	100	100	100	100		
Presentation	40	20	20	20		
Documentation	10	70	10	10		

Thank You

And enjoy your next movie!