Joke Recommendation System

Using Jester Dataset

Team 21 Members:

Animesh Pareek *(2021131)*
Sameer Gupta *(2021093)*
Abhijay Tiwari *(2021439)*
Gunjapalli Sravani Reddy *(MT22098)*

Problem Statement

Building a joke recommendation system using classical machine learning techniques.

About the Dataset:

- Values from (-10.00 to +10.00) of 100 jokes from 73,421 users, and a total of 4.1 million ratings.
- A file containing all 100 jokes.

Goal:

*Given a file with 20 jokes, predict rating for each joke.
Predict ratings for sparse users, and recommend 20 jokes.

Exploratory Data Analysis

"Exploratory Data Analysis (EDA) is an initial phase in data analysis, involving the use of statistical methods and visualizations to understand patterns, relationships, and characteristics within a dataset."

Jokes Dataframe:

- Words and Frequencies
- Articles and Determinants

Ratings Dataframe:

- Description before and after normalization
- Histograms
- Box Plots
- Violin Plot

Methodology

"Methodology is the structured approach used to gather, interpret, predict, and draw conclusions from data or information."

Preprocessing:

joke_id	joke	Processed_joke	cluster	sorted_topics	main_topic
1	A man visits the doctor. The doctor says '	man visit doctor doctor say news havecancer alzhe	[(0, 0.002279222), (1, 0.002276016'	[(4, 0.97790796), (6, 0.0033486	4

user_id	number_of_jokes_rate	joke_1	joke_2	joke_3	joke_4	joke_5	joke_6	joke_7	joke_8	joke_9	joke_10	joke_11	joke_12	joke_13	joke_14	joke_15	joke_16	joke_17
1	74	-7.82	8.79	-9.66	-8.16	-7.52	-8.5	-9.85	4.17	-8.98	-4.76	-8.5	-6.75	-7.18	8.45	-7.18	-7.52	-7.4

Model Training:

Models: Rating Prediction, User-User Based, Content Based

Results

Input:	A text file with 20 jokes.
Operations:	For each joke, it is preprocessed, and using a model, a rating is calculated for it used on training.
Output:	Ratings for each joke,

Comparisons

Existing Works:

Surgoku Abbi163 Junolee JoeDockrill

How ours is different:

Our work, includes only classical machine learning methods, while having several detailed results.

Conclusions

- Some of our key learning from the project was on the prepossessing techniques used in text and numerical based data.

- We learnt how to lemmentize texts, tokenize them, POS tag them, and figure out primary words from a line of texts.

- We learnt about collaborative filtering techniques and what types of recommendation systems and data for them are present, and how the methodology for each differs.

Contributions

Contribution of work from each member, over the project.

Animesh:

Recommendation system
Latex Report
EDA

Sameer:

Recommendation system
Latex
EDA
Presentation
Organisation

Abhijay:

Phase 1:
Pre-existing works
Latex

Sravani:

Phase 1:
Pre-existing works
Latex