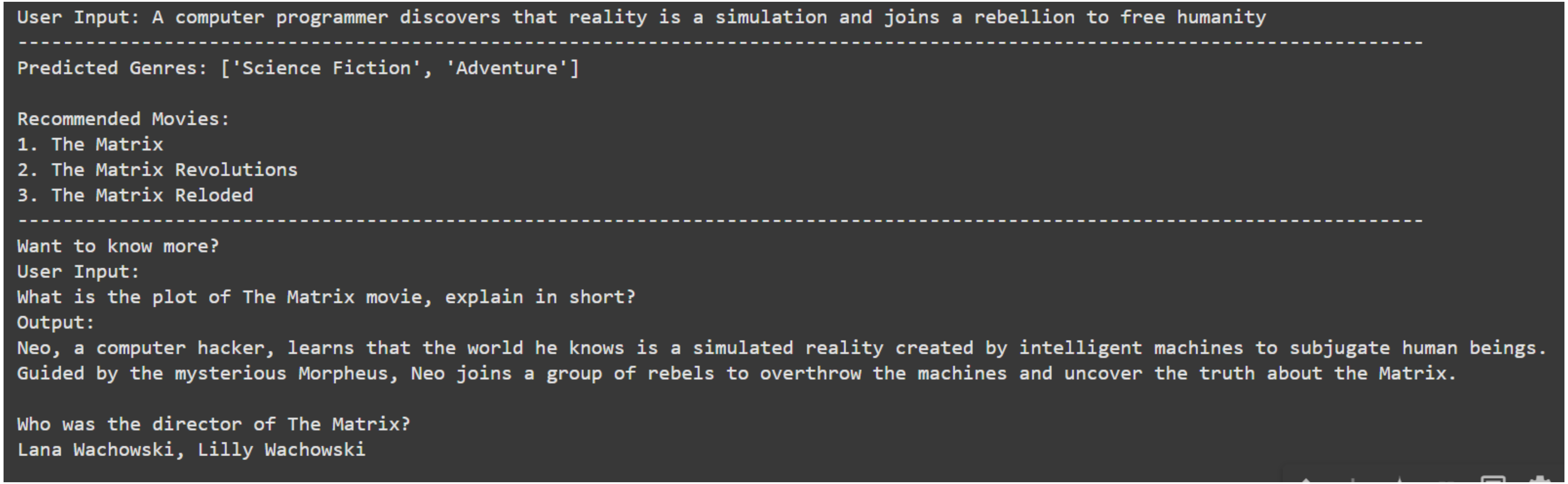
Movie Recommendation System Using Transformers

- Team 19:
- Arvind Chary Padala (ap2522)
- Ram Sampreeth Budireddy (rb1424)
- Tanishq Sharma (ts1266)

Problem Statement and Data Set

- **Objective:** Build a system that uses plot summaries and metadata to recommend movies based on genre and narrative similarity.
- **Data Set Overview:** CMU Movie Summary Corpus
  - 42,306 movies with plot summaries.
  - Metadata: genres, release date, runtime, language.

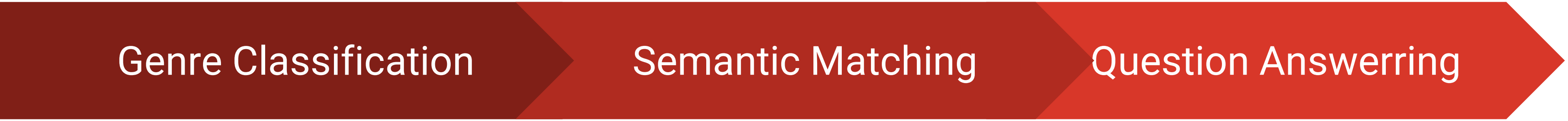


```
User Input: A computer programmer discovers that reality is a simulation and joins a rebellion to free humanity
---------------------------------------------------------------------------------------------------------------
Predicted Genres: ['Science Fiction', 'Adventure']

Recommended Movies:
1. The Matrix
2. The Matrix Revolutions
3. The Matrix Reloaded
---------------------------------------------------------------------------------------------------------------
Want to know more?
User Input:
What is the plot of The Matrix movie, explain in short?
Output:
Neo, a computer hacker, learns that the world he knows is a simulated reality created by intelligent machines to subjugate human beings.
Guided by the mysterious Morpheus, Neo joins a group of rebels to overthrow the machines and uncover the truth about the Matrix.

Who was the director of The Matrix?
Lana Wachowski, Lilly Wachowski
```

Model Pipeline



Genre Classification

Task: Classify the user's natural language query into a movie genre.
Approach: Fine-tune a pre-trained DistillBERT model for multi-class classification using labeled data mapping movie plots to genres.
Baseline Model: Logistic Regression and Naïve bayes with TF-IDF features.

Semantic Matching

Task: Perform semantic search using the user query and the predicted genre to recommend the top 2-3 movies.
Approach: Use DistillBERT embeddings of the query concatenated with the predicted genre to search plot summaries.
Baseline Model: Cosine similarity with TF-IDF vectors and simple bag-of-words representations.

Question Answerring

Task: Answer user questions about the recommended movies.
Approach: Fine-tune DistilBERT for question answering using a subset of the movie metadata as the context.

Transformer Models and Optimizations

Model Used: **DistilBERT,** a lightweight transformer model

Genre Classification: Fine-tuned DistilBERT predicts genres from plot summaries.

Key Challenges:

- **Class Imbalance:** Genres like "Cult" and "Musical" were underrepresented.
- **Text Truncation:** Plots longer than the tokenizer’s max length were truncated
- **High Computation Costs:** Fine-tuning - significant time and resources.

Hyper Parameters:

Learning rate: 5e-5, Batch Size: 32, Epochs: 3

Metric	Baseline Model (naïve bayes)	Transformer Model (Distill BERT)
Accuracy	0.21	0.61
Precision (weighted)	0.14	0.67
Recall (weighted)	0.22	0.61
F1-Score (weighted)	0.15	0.62

Transformer Models and Optimizations

- **Optimizations:**
  - **Learning Rate Scheduler:** Linear decay with warm-up steps to stabilize training
  - **Early Stopping:** Prevented overfitting by monitoring validation loss
  - **Weight Balance Implementation:** ensuring fairer contribution from underrepresented classes.

Movie Recommendations based on Semantic Matching

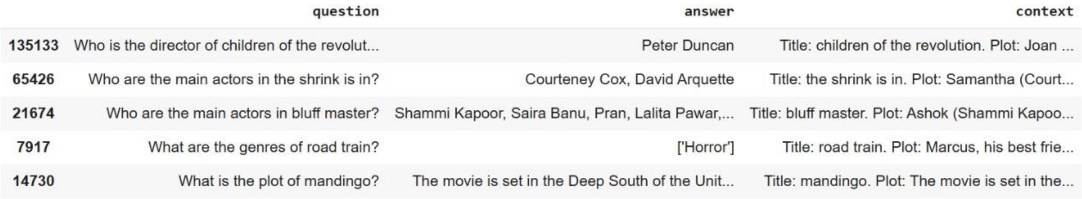
- **Objective**: Recommend movies based on user queries by matching them with movie plot summaries.

- **Model**:
  - **DistilBERT Embeddings**: Convert user queries and movie plots into dense vector embeddings.
  - **Query Embeddings**: Concatenate the predicted genre with the user query to generate embeddings.
  - **Cosine Similarity**: Compare the embeddings of user queries and movie plot summaries to find the closest matches.

- **Baseline**: **TF-IDF Vectorization**: A traditional method where queries and plot summaries are represented as sparse vectors, and cosine similarity is used for matching.

- **Evaluation**:Top-3 movie recommendations are retrieved, with accuracy compared between the **DistilBERT** and **TF-IDF** approaches.

Question Answering (Contextual Answer Extraction)



	question	answer	context
135133	Who is the director of children of the revolut...	Peter Duncan	Title: children of the revolution. Plot: Joan ...
65426	Who are the main actors in the shrink is in?	Courteney Cox, David Arquette	Title: the shrink is in. Plot: Samantha (Court...
21674	Who are the main actors in bluff master?	Shammi Kapoor, Saira Banu, Pran, Lalita Pawar,...	Title: bluff master. Plot: Ashok (Shammi Kapoo...
7917	What are the genres of road train?	['Horror']	Title: road train. Plot: Marcus, his best frie...
14730	What is the plot of mandingo?	The movie is set in the Deep South of the Unit...	Title: mandingo. Plot: The movie is set in the...

Created a synthetic **QA** dataset using the movie plot and other movie metadata(actors, directors, plot summary etc.)

Concatenated all movie metadata and plot summary to create a context.

Fine tuned **DistilBERTForQuestionAnswering**(a variant of DistilBERT which is finetuned on fine-tuned on SQuAD (Stanford Question Answering Dataset))

Evaluation Metrics:	
Exact Match	0.7457
F1 Score	0.7992
Precision	0.7457
Recall	0.8160