**TEAM MEMBERS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S. No.** | **Name** | **Branch** | **Sem.** | **Contact** | **Email** |
| 1. | Kushagar Mahajan | CSE | 3rd | 7889308271 | kushagar.33-cse-17@mietjammu.in |
| 2. | Mohit Sharma | CSE | 3rd | 9149777180 | mohit. 112-cse-17@mietjammu.in |
| 3. | Danish Soma | IT | 3rd | 9596492447 | danish.36-it-2017@mietjammu.in |
| 4. | Deepanshu Chajgotra | CSE | 3rd | 7006435558 | deepanshu.29-cse-17@mietjammu.in |

**PROBLEM STATEMENT**

**Movie dataset analysis**

The challenge is aimed at making use of machine learning and artificial intelligence in interpreting Movie dataset. The dataset made available to participants is on the Scripts of the movies, Trailers of the movies, Wikipedia data about the movies and Images in the movies.

In this project, we aim to impart the ability to get rid of biases in a machine or an AI system. Specifically, we will aim to go beyond information retrieval to do reasoning over the multimodal dataset and develop algorithms to remove the bias. The dataset is available at: <https://github.com/BollywoodData/Bollywood-Data>

For ease of use we have made available pre-processed versions of these datasets. We have applied Watson NLP API and Open IE to produce more enriched text. Similarly, for previews, we have identified emotions in selected frames along with metadata for the movies. Participants are at liberty to use one or more of these datasets to interpret, predict, and draw intelligence of any sort from the dataset provided. The following section outlines few potential problems that can be taken up.

**Description**

Enable multi modal Question Answer system and help in capturing information about the dataset.

Stage 1 - Extract the data from Wikipedia-Data folder and extract plot text for each Bollywood movie. Using this data, one should be able to query the dataset and ask natural language query and the output of the query should be in natural language or an image. This image can be extracted from image Data in the corresponding folder on GitHub.

Stage 2 – Extract the data from image-data folder on GitHub as an input and the output should be text or natural language corresponding to the image. This text can be taken from Wikipedia-data containing plot of each image.

**Description:**

Convert the movie plot into entity- relationship graph where each path traversal provides a different story arc of the movie.

Stage1- Extract the data from Wikipedia data folder and extract plot text for each Bollywood movie. Using the data, one should be able to summarize the movie plot on 5 lines.

Stage2- Use this text data to construct entity-relationship graphs. Further using these entity-relationship graphs find out various arcs of the movie story.

**Description:**

Design and Develop algorithm to remove gender bias in text.

Stage 1- Extract Wikipedia plots data from Wikipedia-data folder and try to construct a different and unbiased version of a story.

Stage 2- Use attention model to pin point various parts in the story and then debias those parts. Further show these nodes in an interactive visualization.

**Description:**

Interesting visualization to explore the dataset.

Stage 1: To explore the whole dataset, we look for Google Big Query to explore the whole dataset. This also provides an interface to user to be able to navigate at the relevant parts of the dataset.

Stage 2: Google Big Query have the capability to flag the relavant parts of the dataset.

**About Dataset**

The dataset represents a large multimodal dataset derived out of multiple sources. The data consists of:

**Wikipedia Data -** Contains text from plots of all movies from 1970 – 2017. The plots are taken from Wikipedia.

**Image Data –** Posters of all movies from 1970-2017.

**Scripts Data –** PDF scripts for 13 movies. The scripts contain complete dialogues.

**Preview Data -** Previews of around 880 movies from 2010-2017.

**TOOLS & TECHNOLOGY**

IBM Cloud § IBM Watson

App development framework for desktop (e.g. Python, Java) and mobile (e.g. Android, iOS)