PROBLEM STATEMENT - 1

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, draw intelligence of any sort from the dataset provided. The following section outlines few potential problems that can be taken up.

PROBLEM DESCRIPTION

1. Probable Use case to implement- Enable multi modal Question Answer on top of this dataset.

   - User should be able to ask questions (in text format), and the output should be text and/or image.
   - User may also provide an image as an input, and the output should be the plot/points relevant to that image.

   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.

2. Probable Use case to implement- Convert the movie plot into entity-relationship graph where each path traversal provides a different story arc of the movie

   - Use this graph to summarize the movie plot on 5 lines.

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

**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 summarize the movie plot on 5 lines.

**Stage 2** – Use this text data to construct entity-relationship graphs. Further using these entity-relationship graphs find out various arcs of the movie story.

3. Probable Use case to implement- The data set has been used to show bias present in Bollywood [http://proceedings.mlr.press/v81/madaan18a/madaan18a.pdf](http://proceedings.mlr.press/v81/madaan18a/madaan18a.pdf)

   - Develop algorithms to remove/reduce such biases

**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 pinpoint various parts in the story and then debias those parts. Further show these nodes in an interactive visualization.

4. Probable Use case to implement—Develop interesting visualization to interactively explore this dataset.

**Description:**
Develop interesting visualization to explore this dataset.

**Stage 1** – To explore the whole dataset, we look for innovative ideas and applications which allow a user to explore the whole dataset. This also includes providing an interface to user to be able to navigate at relevant parts of the dataset.

**Stage 2** – The application should have the capability to flag the relevant parts of the dataset and show those in the form of an interactive viz.

**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.

The dataset is available as- [https://github.com/BollywoodData/Bollywood-Data](https://github.com/BollywoodData/Bollywood-Data). For ease of use we also provide 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 movie.

We encourage participants to propose interesting problems and novel solutions.

**EXPECTATION**

- Solution should be AI driven.
- Participants should demonstrate through system demo at least some useful application.
- Outcome should have document explaining thought process and design approach to arrive at solution.

**EVALUATION CRITERIA**

The evaluation criteria are listed on the hackathon page.

**TOOLS & TECHNOLOGY**

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

**RESOURCES & REFERENCES**

- [https://personality-insights-livedemo.mybluemix.net/](https://personality-insights-livedemo.mybluemix.net/)
- [https://tone-analyzer-demo.ng.bluemix.net/](https://tone-analyzer-demo.ng.bluemix.net/)
- [https://www.ibm.com/watson/services/discovery/](https://www.ibm.com/watson/services/discovery/)
FAQ’s

Q: What are the programming languages?
A: Python, Java

Q: What are mobile platforms allowed?
A: Android, iOS

Q: Where to get free access to IBM Cloud?
A: Sign up on - https://www.ibm.com/cloud/

Q: Is there any documentation available to use IBM Cloud?
A: Yes, Each service comes with elaborate documentation with step by step illustration to use the services available on IBM cloud, follow the VIEW DOCS, link available on each service.

Q: Is the knowledge of ML/DL is required?
A: No

Q: Is there any dataset provided?
A: Yes, the dataset is hosted on the link https://github.com/BollywoodData/Bollywood-Data

Post your technical queries on slack:
https://join.slack.com/t/ibmhackchallenge/shared_invite/enQtNDA5NDk4OTUxNjMyLWZmNGQ1Mzc5YzQzYTg1ODEwNGY3MTVlMzZhNGJiYTk4ZWI2MzQ1NjA4YzI5YTNiMTU3NjNkMDJkNDY1ZDUzYTI