CS5242 Project Proposal

Topic: Genre-detection from IMDb reviews

Description:

IMDb is an online database of information related to films, television series, and streaming content online – including plot summaries, ratings, and critical reviews. As of March 2022, the database contained 10.1 million titles. Our aim is to train a model that predicts the genre of the provided media (movie/series) based on its corresponding text reviews. Such a model can be used to categorize content based on its genre for easy information retrieval and recommendations.

Proposed Solution:

1. Scrape IMDb data that contains the movie title, movie review and its genres (truth value) to create a labelled dataset.
2. Visualize and analyse the dataset distribution to get early insights and reduce bias in the dataset (e.g., most movies have a genre as “crime”). Clean the dataset based on custom rules in order to reduce noisy predictions.
3. Perform supervised learning using MLP/CNN/RNN.
4. Compare and analyse model performances for each method.

Project Milestones:

1. Build custom web scrapper using Python libraries such as Beautiful Soup to scrape and store IMDb data.
2. Create visualizations of the dataset to validate a uniformly distributed dataset across genres.
3. Preprocess the data using Python libraries such as nltk and text-preprocessing.
4. Train a baseline model using an MLP.
5. Train a CNN model with regularization.
6. Train an RNN model with optimization.
7. Compare, analyse results and report conclusions. Create slides and record the final video.

Member Contribution Breakdown:

Task	Member 1	Member 2
Build scrapper	Niharika	Shreyas
Preprocess data	Shreyas	Sri
Data visualization	Sri	Niharika
Train MLP	Niharika	Shreyas
Train CNN	Shreyas	Sri
Train RNN	Sri	Niharika