PROJECT PROPOSAL

About Dataset:

Cornell Movie--Dialogs Corpus

Related corpus: Cornell Movie-Quotes Corpus

DESCRIPTION:

This corpus contains a large metadata-rich collection of fictional conversations extracted from raw movie scripts:

- 220,579 conversational exchanges between 10,292 pairs of movie characters

- involves 9,035 characters from 617 movies

- in total 304,713 utterances

- movie metadata included:

    - genres

    - release year

    - IMDB rating

    - number of IMDB votes

    - IMDB rating

- character metadata included:

    - gender (for 3,774 characters)

    - position on movie credits (3,321 characters)

- see the documentation for details

Data Processing and Purpose:

1. Text Analysis: Utilize natural language processing techniques to analyze the dialogues, including sentiment analysis, topic modeling, and entity recognition.
2. Network Analysis: Construct networks of characters based on their interactions to study character relationships and centrality.
3. Genre and Rating Analysis: Investigate how different genres correlate with dialogue characteristics and how IMDB ratings are influenced by dialogue quality and other factors.
4. Character Attribute Analysis: Explore how character attributes such as gender and position on movie credits influence dialogue patterns and interactions.
5. Machine Learning: Train machine learning models to predict dialogue outcomes or character attributes based on contextual information.

Overall, the project aims to leverage the rich dataset provided by the Cornell Movie-Dialogs Corpus to uncover valuable insights into movie dialogues and character interactions, contributing to the broader understanding of storytelling in the cinematic domain.