Paper data set introduction document

Abstract

The purpose of this document is to introduce the Movie data sets used in the research of this paper, including Popular Dataset, Unpopular Dataset, Test Dataset, Movie review dataset and Movie rating dataset. These data sets will be used to evaluate and optimize the performance of the movie script generation model.

1. Data set overview

1.1 Popular Dataset

- Quantity : 10

- Source : High number of review and high score movies in Movie review dataset and Movie rating dataset

- Purpose: To serve as a baseline data set for model testing to compare the quality of model-generated scripts. Verify the performance of the model when dealing with known Popular movie scripts.

1.2 Unpopular Dataset

- Quantity : 10

- Source : Low number of review and low score movies in Movie review dataset and Movie rating dataset

- Purpose : To evaluate the model's ability to produce film scripts with less data support.

Provide opportunities for model optimization and improvement.

1.3 Test Dataset

- Quantity : 5

- Source : Selected from the Popular Dataset and Unpopular Dataset

- Composition : 2 Popular movies, 3 Unpopular movies

- Objective : To verify the generalization ability of the model when dealing with different types of movie scripts. Improve the recognition and reliability of test results.

1.4 Movie review dataset and Movie rating dataset

- Quantity : 1,355,791

- Source : https://www.kaggle.com and data crawl

- Purpose : Used to determine the popularity of a movie, to assist in screening Popular and Unpopular movies.

2. Data set fields

The dataset contains the following fields:

- Film title : The official title of the film.

- Corresponding novel and script : The original novel and script of the film.

- Movie Genre : Categories of movies, such as action, comedy, etc.

- Film word count : Word count for film scripts.

- Novel word count : Word count of the original novel.

- Number of scenes : Number of scenes in the movie.

- Film Reviews : Audience and critic reviews of the film.

- Rating : The film's rating, usually on a scale of 1 to 10.

- Number of Viewers : The number of viewers of the movie.

3. Data set selection process

3.1 Data set classification

- Step 1 : Use the Movie review dataset and Movie rating dataset to classify movies.

- Step 2 : Divide movies into Popular and Unpopular categories based on their ratings and number of review.

3.2 Selecting the Test Dataset

- Step 3 : Select 2 representative movies from the Popular Dataset.

- Step 4 : Pick 3 representative movies from the Unpopular Dataset.

- Step 5 : Make sure you choose a movie that covers a wide range of movie genres and ratings.

4. Use of data set

- Test Model : Use Test Dateset to evaluate the performance of the model in different situations.

- Optimize the model : Optimize the model based on the test results, especially for Unpopular and low-scoring movie scripts.

5. Conclusion

The film data set introduced in this document will provide a solid data foundation for the thesis research. With these data sets, we can comprehensively evaluate and optimize the movie script generation model to improve its production quality across different movie genres and rating situations.

---