Extracting Genre Information from Movie Plot Summaries

Project Proposal

TECHNISCHE UNIVERSITÄT DARMSTADT

Volker Knauthe, Benedikt Lins, Stefan Thaut

Idea and Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Prediction of the Movie Genre based on its Plot Summary
- Plot Summaries provide information to catch peoples interest

  → Hence, also contains information about the Genre
- Evaluation of Plot Summary or Genre Tags

Example:

[1]

Idea and Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT

Plot Summary:

„Thomas A. Anderson is a man living two lives. By day he is an average computer programmer and by night a hacker known as Neo. Neo has always questioned his reality, but the truth is far beyond his imagination. Neo finds himself targeted by the police when he is contacted by Morpheus, a legendary computer hacker branded a ***terrorist*** *by the government. Morpheus awakens Neo to the real world, a* ***ravaged wasteland*** *where most of humanity have been captured by a* ***race of machines*** *that live off of the humans' body heat and electrochemical energy and who* ***imprison*** *their minds within an* ***artificial reality*** *known as the Matrix. As a* ***rebel against*** *the machines, Neo must return to the Matrix and* ***confront the agents****:* ***super-powerful computer programs*** *devoted to snuffing out Neo and the entire human* ***rebellion****.“*

Genre Tags:

*„**Action**“, „**Sci-Fi**“*

[1]

Data

TECHNISCHE UNIVERSITÄT DARMSTADT

- Internet Movie Database (IMDb) [2] provides such information for a variety of movies and series
  - Dataset containing about 5.5 million Titles with respective Genres
  - Plot Summaries also available for many movies but are not contained in the dataset

- Open Movie Database (OMDb) [3] provides an API to access similar Plot Summaries

  Allows us to use the IMDb Ids

- Such huge Dataset needs to be restricted to allow faster crawling in the limited time

Machine Learning Idea

TECHNISCHE UNIVERSITÄT DARMSTADT

- Can be framed as a multi-label Classification Task
  Each Movie has up to three Genres

- Two Approaches:
  - Extract Features (e.g. Sentiment, Keywords, Sentence length) from the whole Plot Summary and use them for Learning
  - Use a Sequence Model on the Summary Text

- Ertugrul et al. [4] use a LSTM Network to assign Genres to Sentences of Plot Summary
  - Restricted to 4 Genres (Thriller, Horror, Comedy and Drama)
  - Also use Plot Summaries from OMDb [3]

Evaluation

TECHNISCHE UNIVERSITÄT DARMSTADT

- Split Dataset into Train- and Test-Set
  - Use Cross-Validation on Train-Set
- Different Metrics are possible:
  - Accuracy
  - F1-Measure
  - (Area under ROC-Curve)
- Compare Results with:
  - Ertugrul et al. [4]
  - (Kaggle Challenge [5])

Micro F1	Macro F1
0.6761	0.6768

[4]

Thank you for your Attention!

TECHNISCHE UNIVERSITÄT DARMSTADT

[1] https://www.imdb.com/title/tt0133093/ [accessed on 16.01.2019]

[2] https://www.imdb.com [accessed on 16.01.2019]

[3] http://www.omdbapi.com/ [accessed on 16.01.2019]

[4] A. M. Ertugrul and P. Karagoz, "Movie Genre Classification from Plot Summaries Using Bidirectional LSTM," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, 2018, pp. 248-251.

[5] https://www.kaggle.com/c/miia4406-movie-genre-classification [accessed on 16.01.2019]