The Elephant Man in The Room: An analysis of good and bad movies

[Project Assignment B in 02806: Social Data Analysis and Visualizations] *

Søren Howe Gersager
s094557
syrelyre@gmail.com

ABSTRACT

This report and the associated website[1] is the result of the final project in DTU Course 02806: Social Data Analysis and Visualizations. The course subject was both gathering data and analyzing data using visualizations, machine learning, sentiment analysis and network theory.
This final project investigates potential differences in movies with high IMDB ratings and low IMDB ratings.

1. MOTIVATION

Our dataset is created from the top and bottom 1000 movies sorted by IMDB rating[2][3]. The motivation for the analysis stems from a personal interest in movies, and being curious on what makes a movie good or bad and perhaps eliminating prejudice especially regarding bad movies and their characteristics. An example is the comedy genre, it's our belief that people often have the notion that a lot of bad comedies and romantic comedies exist. By doing the analysis we can investigate whether that hypothesis holds true or not. It is our goal to either confirm or reject the end-users belief in what makes a movie good or bad.

2. IMPLEMENTATION

For creating the dataset used in the analysis, we used BeautifulSoup[4] to scrape the top 1000 movies and bottom 1000 movie pages of IMDB[6]. Afterwards we used the ID's of the movies extracted from IMDB to get more data about each movie from The Open Movie Database API[5].
The data exposed through the API is: title, release date, MPAA-rating, runtime, genre, director, writers, names of first-billed actors, language spoken, country of origin, short description of plot, awards received, poster image url, Metascore, IMDB rating and number of imdb votes.

3. THEORY

We have used several of the tools learned in the course for the website and report, below is a list of tools we have used.

3.1 Visualization

We have used matplotlib for visualizing differences between good and bad movies.

3.2 Machine learning

We have used machine learning with cross-validation to try to classify good movies from bad movies. We tried this with several datasets: One dataset with basic features of all 1389 movies, one with a bag of words representations of 100 movie subtitles and one with the ANEW scores of the bag of words representations. We used several models: Decision Tree, Naive Bayes and Logistic Regression. We used K-fold cross-validation with $k = 10$.

3.3 Natural Language Processing

We have used sentiment analysis by calculating the ANEW scores for each movie subtitle by using its bag of words representation. From this we tried to visualize the scores to find differences, however we did not find any significant differences between the two groups.

3.4 Python and relevant libraries

For performing the analysis we have relied on Python, its standard library and the external libraries: scikit-learn, requests, beautifulsoup.

4. DISCUSSION

5. REFERENCES

[1] `http://blank.dk` *Website showcase*
[2] `http://www.imdb.com/search/title?=&groups=top_1000&sort=user_rating,desc&view=simple` *IMDB Top 1000 movies*
[3] `http://www.imdb.com/search/title?=&groups=bottom_1000&sort=user_rating,asc&start=301&view=simple` *IMDB Bottom 1000 movies*
[4] `http://www.crummy.com/software/BeautifulSoup/` *BeautifulSoup*
[5] `http://www.omdbapi.com/` *The Open Movie Database API*
[6] `http://www.imdb.com/` *The Internet Movie Database*

*A website showcasing the project is available at `inserturl.dk`