Final Project Proposal: Movie Classification by Script

Gregory McCord
Princeton University '20
gmccord

Abstract

Every year, movies attract billions of consumers around the globe to the theater. Some years draw much larger crowds than others, but oftentimes, each year or grouping of years has its own style that separates it from the rest. Much like a genre, these eras can be used to classify movies and help identify trends in the data. While some genres are more popular than others, they all attract their own audiences. It is well known that the popularities of certain genres and movie archetypes directly relates to the occurrences of real world events that cause spikes in popularities. However, we have found no evidence that research has been done to investigate the language tendencies used in a particular genre or era of movie. Our research aims to classify movies by their dialogue into different categories dependent on their production year and their genre. We will use a bag of words model to analyze the contents of each movie's script of dialogue and from their will attempt to classify the movies into a number of categories based on their similarities and differences. This is a novel approach to classifying movies by their scripts that could unveil new uses for big data in the movie industry or indicate to movies in certain genres can tailor their language to appeal to broader or more focused groups of people. In our experiment, we will be using a dataset that Cornell researchers used to analyze whether or not movie dialogue replicated general human conversations. For the purposes of classification, we expect LDA to perform best at this task, but we also expect logistic regression (both binary and multinomial) to provide a strong baseline.

1 Motivation

Movies are one of the top forms of entertainment, and companies are always interested in different characteristics that will push their productions above the competition. A perfect example of this is the Netflix hit show *House of Cards*, where many of the major decisions such as choosing actors were made using big data [3]. We want to approach the problem in a novel way - by analyzing dialogue to see if it is strongly indicative of particular genres or eras of films. This study would indicate that big data could potentially be used to generate script bases for shows and movies.

2 Data

The data comes from a Cornell dataset used to analyze movies based on their dialogue. The dataset was obtained from the Kaggle dataset website. It contains 220,579 conversations from over 600 different movies. There are over 300,000 different sentences included from different conversations in all of these movies. The metadata includes both genres and release year (both of which are relevant to our study) but also includes IMDb rating, number of IMDb votes, and gender data on the involved characters (which present other opportunities to include other features into our data) [1].

3 Methods

We plan to use several models from scikit-learn including logistic regression, multinomial logistic regression, and clustering models (like Gaussian Mixture Model) [2]. We also plan to use a probabilistic topic model such as Latent Dirichlet Allocation, however we haven't chosen a library to use yet (MALLET seems to be the most popular but is only available in Java, so its usage depends on the availability of a Python wrapper).

4 Evaluation Metrics

Because we aim to make this a classification problem, it is fitting to use the misclassification rate to determine whether certain movies are 'older' or 'newer'.

$$MC = 1 - \frac{\sum_i C_i}{N}$$

where C_i represents whether the i^{th} movie was classified accurately or not and N is the size of the test set.

We will follow the same general principal for classifying genre except that we will now have a matrix A that represents the similarities between different genres. The indices of this matrix will be estimated using the inverse of the distance between clusters (where each cluster represents a genre). This matrix will be used to weight incorrect predictions based on how similar it was to the correct answer.

$$MC = \frac{\sum_i A_{i_j,i_k} \cdot M_i}{N}$$

where A_{i_j,i_k} represents the similarity between the predicted genre (j) of movie i and the true genre (k) of i. M_i is 1 if misclassified and 0 otherwise.

References

[1] Cristian Danescu-Niculescu-Mizil and Lillian Lee. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011*, 2011.

[2] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, and et al. Grisel O. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] Greg Petraetis. How netflix built a house of cards with big data, Jul 2017.