Classfying Movie Genres With Plot Summaries

Brian Kimmig, Jesus Zarate

1 Introduction

According to Wikipedia a "genre is the term for any category of literature or other forms of art or entertainment" [1].

Currently, being able to classify what content in language is a hot topic in data science. Something about sentiment analysis. We thought a good way to be introduced to this tool set that allow these analyses was to try and classify writing with pre-made and somewhat objective labels.

Our first thought was to try to categorized some of the topics in twitter data, but this proved to be hard and not as objective as we would have been responsible for the original categories, which are easily more subjective.

Using this movie data gives us the opportunity to classify documents with 'correct' answers, allowing us to use a good number of techniques seen in the real world but have more clear cut answers and allow us to make sure that these tools can be used in this manner.

The above may be the key idea?

2 Data

We obtained the base of our data from a dataset published on Kaggle. The dataset contained information on $\sim$ 5000 movies. We used this dataset for the movie list and the IMDB IDs. With IMDB IDs it is easy to automate gathering the synopses of each movie via GET requests to the OMDB API. The OMDB API allows you to search for movies, and gather information via the title or the IMDB ID. To ensure we get the correct information for every movie we performed GET requests querying with the IMDB ID.

The OMDB API allows a user to specify the length of the plot summary it returns, with either 'full' or 'short'. We chose to gather the 'full' synopses for every movie we queried.

From the Kaggle data we used the OMDB API to compile title, plot summary, and genres for all $\sim$ 5000 movies. The data was stored in a JSON file, with each entry (or movie) containing the fields ['title', 'plot', 'genres'].

There were considerably more genres than expected, in total there were 26. Figure 1 shows every genre and its percent occurrence. In our data, there were genres that occurred less than 1% of the time, and generally they tended to be more obscure genres. Specifically, the genre 'Game-Show' occurred 0.02% of the time. Because of the rare genres we decided to limit our genre classification labels to those that occur more often. In the end we settled on a cut of 5% (shown by the red dotted line in Figure 1). This was to ensure that we had captured the major, or most common, genres. The cut of 5% also allowed us to ensure that every movie had at least 1 label, or genre, associated with it. If the cut was higher, we found that some movies did not have a genre associated with it. We also wanted to avoid throwing away data.

The genre labels set up with a one hot encoding. This was done to allow us to easily classify them with either a multi label classifier, or by individually, by column, with other classifiers (discussed

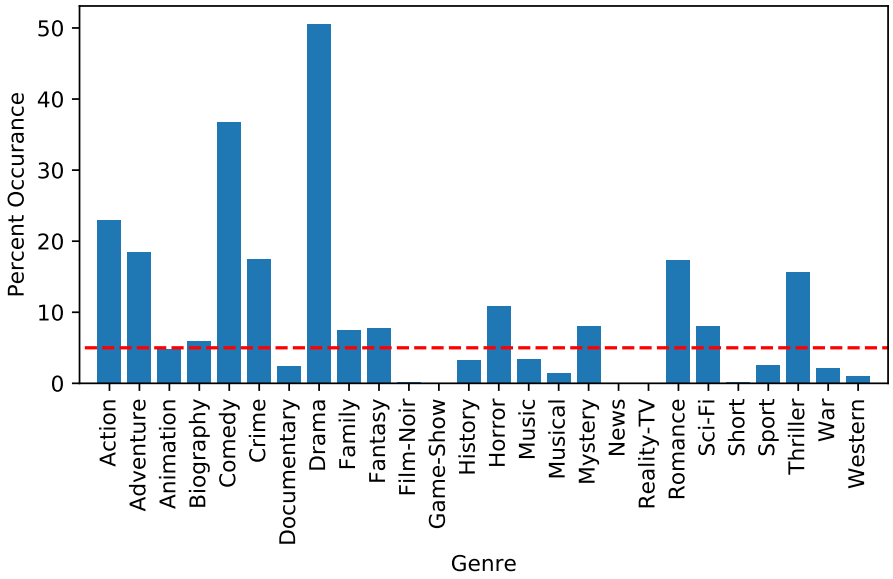

Figure 1: The percent occurrence for each genre. The red dotted line represents the cut made to get the list of genres (5%).

further in §4). Figure 2 shows the genres for every movie in the one hot encoding. A yellow line indicates the movie is associated with that genre, from this we can clearly see that a majority of the movies are associated with the 'Drama' genre.

The synopses of the movies live in long strings or documents. We have a list of these that we will processing to create the features that we will then use to build models to classify genres. Extracting features from these documents is the main part of the project and will be discussed §4.

3 Key Idea

4 Methods and Results

4.1 Term Frequency - Inverse Document Frequency (TF-IDF)

4.1.1 Linear Regression

Classify one genre at a time.

4.1.2 Logistic Regression

Classify one genre at a time.

4.1.3 Random Forrest

Classify all genres at once.

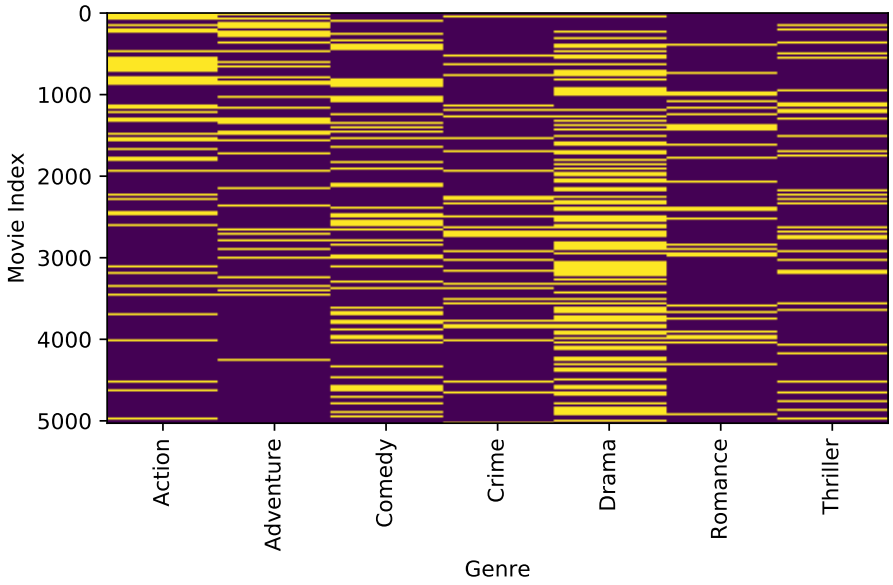

Figure 2: The one hot encoding of the genres for every movie. Each row represents a movie and a column represents a genre. Yellow indicates the movie has that genres associated with it.

4.2 Latent Dirichlet Allocation (LDA)

4.2.1 Linear Regression

4.2.2 Logistic Regression

4.2.3 Random Forrest

4.3 TF-IDF + LDA

4.3.1 Linear Regression

4.3.2 Logistic Regression

4.3.3 Random Forrest

5 Discussion

5.1 Brian's Thoughts

5.2 Chuy's Thoughts

References

[1] Wikipedia. List of genres — wikipedia, the free encyclopedia, 2017. [Avaliable at https://en.wikipedia.org/w/index.php?title=List_of_genres&oldid=774783669].