QA & Information Retrieval:

Popularity by Quora in the Movie Industry

Jinha Hwang
New York University
Jh3275@nyu.edu

Sang Jun Yum
New York University
Sjy269@nyu.edu

Abstract

In this project we attempt to analyze the popularity of movie genres by comparing language characteristics of specific genres against a database of queries extracted from Quora. After creating a query data set from Quora and a data set of movie synopses by genre with IMDB, we attempt to compare/contrast the queries against the movie synopsis data set by calculating TFIDF values and the average cosine similarity. By analyzing our results against other movie popularity statistics, we hope to find if specific characteristic language of movies used in Quora can be used to determine the popularity of genres.

Our biggest challenge was creating a substantial enough dataset to run our tests on. Creating a substantial data set proved to be significant as the results ran with a smaller database was substantially different from the result ran with a larger dataset of movies per genre.

Keywords: genre, popularity, cosine similarity, Quora, information retrieval

1. Introduction

Figuring out which movie genre is popular is no new thing. Numerous statistics on the popularity of movie genres are done on a monthly and yearly basis. However, despite the large number of analysis done on this subject, there is very little (if not any) analysis that is done with a focus on the characteristic language aspect that a movie as a medium encompasses.

It is evident that each different genre of movies has a characteristic set of languages that are used specifically or more often than that of other genres of movies. This research focuses on analyzing the popularity of movie genres based on the frequency of the language of different genres in queries extracted from Quora. It is our underlying assumption that if a certain genre of movie is talked or asked about frequently among the audience, that genre movie is safely considered to be popular. Thus, by comparing a substantial dataset of queries from Quora and comparing it against movie synopses encompassing the characteristic language of a genre, we can find which genres are more popular than others.

2. Defining ‘Popularity’

Since the concept of ‘popularity’ is a vague one, and this study attempts a different approach compared to the more conventional method of how movie popularity is analyzed, it is necessary to give a clear definition. More conventionally, the definition of ‘popularity’ when ranking movies either by genre or individually refers to how much revenue that genre or movie has generated in a certain time frame. To illustrate this point, Figure 1 shows a certain study done by Statista in which they rank movie genres according to the amount of total box revenue the genre generated from movies between 1995 and 2016. Note that Statista names this statistic as a listing of the most ‘popular’ movie genres.

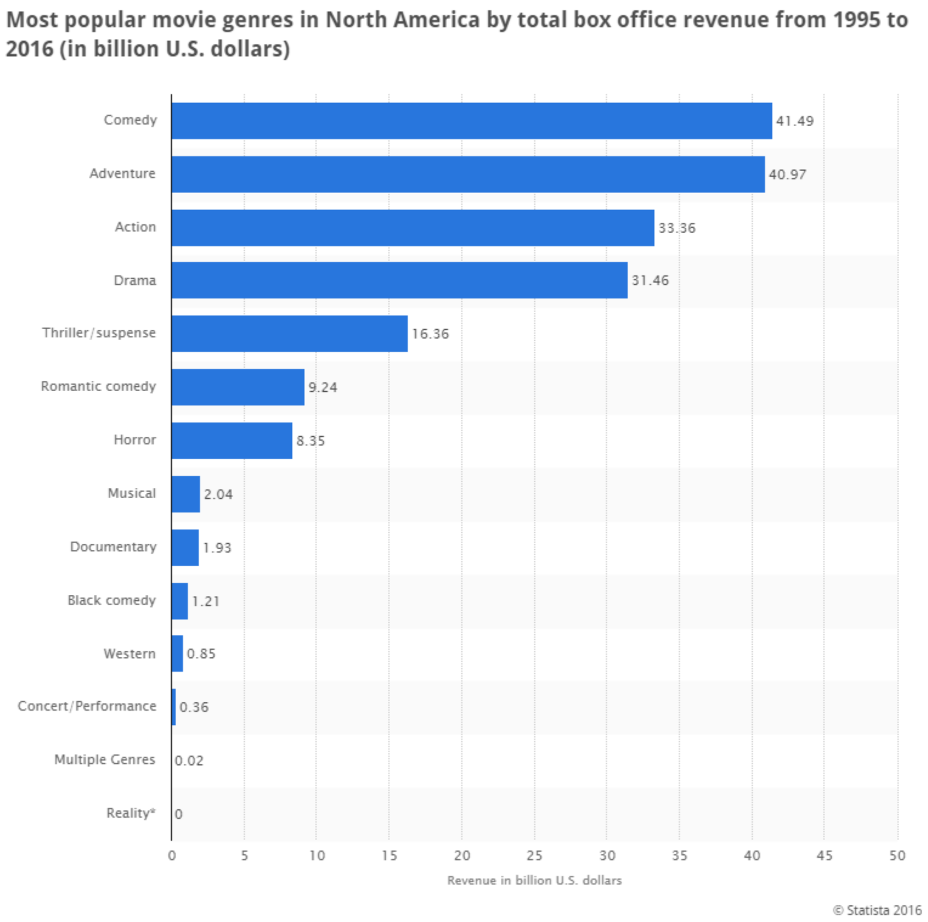

Figure 1: Popular Movie Genres by Revenue

In this study, we use a completely different definition of 'popularity'. This study works under the assumption that if a certain movie genre is talked and asked about more often than other genres, that genre is said to be 'more popular' that the genre that is discussed less frequently. Thus it is the frequency of the language of a certain genre used that determines popularity, not the amount of revenue it generates. It is essential that the term 'popularity' is understood as this definition and not the conventional definition in the scope of this study.

3. Implementation

The first thing that had to be done to measure the popularity of different movie genres was to set the mean of measurement. In real world, this measurement is the revenue. Nonetheless, as mentioned earlier, in this project, a different type of measurement is used. In this project, it is assumed that there are wordings and terms that are characteristic to some specific movie genres. For instance, the term, 'murder weapon' would be characteristic to horror or crime movies. The hypothesis is made from the expansion of this idea; if some words appear often in the web, the genre of movies of which synopsis contain the words would be popular in the real world.

In this project, two classes of data are necessary. The first class of data is the set of movie-related questions from Quora. The second class of data is three different text files, each representing different movie genres: horror, action, and fantasy. Each text file of the second class is comprised of portions of different movie synopses scraped from IMDB.

Each query (question) in the first class of data is transformed into a vector format such that each entry in the query is the TFIDF value of each word in the query. For example, consider the question, ‘Which are some of the best serial killer films in English all time?’ This question converts into a vector such as < 5, 7, 10, 34, 23, 64, 44, 25, 98, 105, 24, 100, 12 > where each entry is the TFIDF value of each word in the question.

The second class of data is also converted into numerical values. Each synopsis of each text file in the second class of data is converted into vector such that each entry is the TFIDF value of each word in the synopsis. Considering that synopses are generally much longer than each query of the first class of data, the dimension of each vector that represents a synopsis is considerably larger than the vector that represents a query. For example, the vector that represents the synopsis of the movie, “OUIJA” is represented as following: < 8,3,23,4,5, … n > where 350<= n <=500(note that only a portion of the synopsis is used instead of the whole synopsis)

Consider that the synopsis of “OUIJA”, of which vector representation is shown above contains “which”, “killer”, and “time”. Then, the vector of the synopsis is transformed such that dimension equals the dimension of the query vector, and the entry position of words above also equals the entry position of the words in the query vector. Consider the TFIDF of “which”, “killer”, and “time” in the synopsis are 9,7, and 3. Then, the vector < 8,3,23,4,5, … n > is transformed into the vector, < 9,0,0,0,0,0,0,7,0,0,0,0,3 >. Note that the entries representing non-overlapping words are all 0s. Then, the cosine similarity between the query vector, < 5, 7, 10, 34, 23, 64, 44, 25, 98, 105, 24, 100, 12 > and < 9,0,0,0,0,0,0,7,0,0,0,0,3 > is computed. This process is done for every query for every synopsis in three different genres. The following is the pseudo code representation of the algorithm.

```
FOR QUERY IN QUERIES,
  FOR SYNOPSIS IN GENRE ‘1’
    COMPUTE COSINE SIMILARITY
    PUSH THE COSINE SIMILARITY INTO LIST ‘1’

FOR QUERY IN QUERIES,
  FOR SYNOPSIS IN GENRE ‘2’
    COMPUTE COSINE SIMILARITY
    PUSH THE COSINE SIMILARITY INTO LIST ‘2’

FOR QUERY IN QUERIES,
  FOR SYNOPSIS IN GENRE ‘3’
    COMPUTE COSINE SIMILARITY
    PUSH THE COSINE SIMILARITY INTO LIST ‘3’

COMPUTE THE AVERAGE VALUES OF ENTRIES IN
LIST 1, 2, 3
```

Finally, the program would return three different average cosine similarity values; these values represent the popularity of each movie genre in Quora.

4. Results

Initially, we ran the program with 421 queries and 50 different partial synopses per genre. The resulting average cosine similarities were the following.

Between Queries and Horror synopsis =
0.38197429752

Between Queries and Action synopsis =
0.334282502826

Between Queries and Fantasy synopsis =
0.25138603121

Compared to the statistical dataset, “Most popular movie genres in North America by total box office revenue from 1995 to 2016” provided by Statista.com, our result does not reflect that popularity. The chart representation of the dataset from Statista.com is shown in Figure 1. For the second try, we increased the size of our data. We ran the program with 801 queries and 100 different partial synopses for genre. The resulting average cosine similarities were the following.

Between Queries and Horror synopsis = 0.400900964622

Between Queries and Action synopsis = 0.41057781809

Between Queries and Fantasy synopsis = 0.346830117579

Note that Action movies are now more popular than Horror movies. It seemed that with an increased amount of data, our model was proving our hypothesis.

5. Extensions

There are several challenges that should be resolved. Firstly, as mentioned earlier, the biggest challenge of this project is gathering data sets that are big enough to prove our hypothesis. We succeeded in gathering up to 1441 queries from the Quora web site. However, manually scraping from IMDB website strongly limited us from gathering the high number of the movie synopsis. If we can find a data source or data base that provides movie script/synopsis data differentiated by genres, it is possible that our program will prove our hypothesis.

Secondly, the quality of data should improve. The Quora is full of movie-related questions, however, a big portion of query data from Quora is comprised of general movie question, not a movie question that strictly concerns the plot or synopsis of movies. For example, one of the queries was, “How do the actors know where to go on casting?” It is important that we find data source that can provide more questions that are synopsis/plot related.

Thirdly, it is possible that converting the texts to numerical values can improve. In our project, we used TFIDF to score each word in the data sets, and used cosine similarity to measure the similarities between texts: specifically, between queries and synopses. An academic paper, “Efficient Features for Movie Recommendation System”, written by Suvir Bhargav suggests another mathematical equation called Hellinger Distance that can be used to measure the similarity between texts. The Hellinger Distance can be explained as following. For probability distributions P and Q such that

$P = \{p_i\}_{i \in [n]}$, $Q = \{q_i\}_{i \in [n]}$ supported on [n],
The Hellinger distance is defined as

$$h(P,Q) = \frac{1}{\sqrt{2}} \cdot ||\sqrt{P} - \sqrt{Q}||_2$$

Note that unlike cosine similarity, smaller Hellinger Distance means higher similarities between texts.

In conclusion, for the improvement of this project, it is imperative that we find data sources that will provide significantly bigger data sets with better quality. In addition, by implementing other measures of similarities, we can improve our proof and minimize the numerical and logical errors that disproved our hypothesis.

6. References

Bhargav, Suvir. "Efficient Features for Movie Recommendation Systems." (2014).

Blackstock, Alex, and Matt Spitz. "Classifying Movie Scripts by Genre with a MEMM Using NLP-Based Features." (2008).

the-numbers.com. "Most Popular Movie Genres in North America by Total Box Office Revenue from 1995 to 2016 (in Billion U.S. Dollars) ." Statista - The Statistics Portal. Statista. January 2016. Web. 18 Dec 2016.