MOVIES WE LIKE

User Content Tags Among IMDB’S Top 250 Movies

By Dalton Simancek

Individual Project Report

SI 601 Winter 2016

**MOTIVATION**

The IMDB top 250 provides a snapshot of insight into the internet’s all-time favorite movies. However, aside from the list those 250 movies share, it’s difficult to discern what else these movies have in common, particularly in terms of their content. What makes a movie an all-time favorite? Are there underlying patterns or characteristics that are more popular/favored over others? With this project, I am personally interested in what I would describe as “Narrative Analytics” or running analysis on story content for the purpose of understanding viewer/consumer trends as well as inform production of future content. Companies such as Netflix, whom use content tagging methods to understand customer trends and tastes in their develop recommendation and content management systems, would also benefit from the results of this project.

To complete this project, I will use datasets from MovieLens, which contains 582149 content tags from their users for 33869 movies - ranging from “beautiful scenery” to “smart writing” to “French.”  From there, I will examine the frequency of content tags for IMDB’s top 250 movies in order to further understand the characteristics of the movies that the internet has chosen to admire most.

**DATA SOURCES**

The data for this project came from two different sources: web pages from the internet movie database(IMDB.com) and comma separated values files from MovieLens, a movie ratings dataset created and maintained by the GroupLens research lab in the Department of Computer Science and Engineering at the University of Minnesota.

**Internet Movie Database Top 250 Movies**

<http://www.imdb.com/chart/top>

**Background:** The IMDB top 250 lists the first 250 movies in the database with the highest IMDB rating. The IMDB rating is scored on a scale of 1.0 to 10.0 and is calculated based on a weighted average that considers A) the total number of user IMDB votes, B) whether or not a particular movie has minimum number of votes (currently 20,000) and C) the mean of user ratings. Furthermore, only *movies* are eligible for the top 250 list, not television shows, documentaries, miniseries or shorts.

**Format:** Data for this project was retrieved directly from the html of IMDB’s top 250 web page. The html document contained title strings and imdb id numbers for every movie on the top 250 list. The python code for this project retrieves the html, parses it and returns a list of 250 dictionaries with a field for both the movie title and the imdbid. This data will be used to cross-reference these movies in the MovieLens data sets and retrieve the associated user content tags.

**MovieLens Latest Datasets**

<http://grouplens.org/datasets/movielens/>

**Background:** The movie recommendation service MovieLens monitors and collects tagging activity 247753 users between January 9, 1995 and January 29, 2016 for a total 586994 content tags across 34208 movies. Tags range from genre – dark comedy, drama, anime – to thematic – time-travel, dystopia, Japanese culture – to purely descriptive – Bill Murray, ridiculous, exaggerated communists.

**Format:** The large latest dataset offers four files: links.csv,  movies.csv,  ratings.csv and tags.csv. Links.csv is a spreadsheet that maintains the identifiers and ids for a particular movie. This project uses links.csv to retrieve the movie id associated with the imdbId of the same movie. Each movie Id is added to the corresponding python dictionary in the list of movies. tags.csv is a spreadsheet that has rows of content tags ascribed to a particular movie by a particular user. The python code for this project pulls all of the content tags for any tag associated with the movieId that are representative of the movies in IMDB’s top 250. Tags are retrieved and counted by frequency into a list of content tag dictionaries that record the frequency of content tags in IMDB’s top 250 movies.

**DATA MANIPULATION METHODS**

**Step 1:** HTML code is retrieved from the URL <http://www.imdb.com/chart/top> and used to write a html file that will be parsed in Step 2.

**Input:** URL

**Output:** HTML file

**Step 2:** Using BeautifulSoup, the html file in Step 1 is parsed to retrieve the HTML tags containing the title of movie and its imdbId. The movie title and imdbId are then written into a dictionary for a total list of 250 dictionaries. Additionally, The ‘tt’ in the imdbIds is removed for comparisons with Ids used in the MovieLens data.

**Input:** HTML file

**Output:** List of Movie Dictionaries

**Variables:** Movie title, imdbID

**Step 3:** Links.csv is opened and the associated MovieLens MovieID for every imdbId is retrieved and added to each movie dictionary. However, at the time of this project, MovieLens did not have a corresponding movieID for every movie on IMDB’s top 250(Deadpool, for example, was released on February 12, 2016 and, therefore, outside of MovieLens’s most recent and available dataset) To account for this, the source code creates an empty field for movieID and is therefore passed over when retrieving content tags in Step 4.

**Input:** links.csv, List of Movie Dictionaries

**Output:** Updated List of Movie Dictionaries

**New Variable:** movieID

**Step 4:** Tags.csv is opened and the associated content MovieLens content tags for every valid movieId is retrieved and placed in a new list of tag dictionaries. Each tag dictionary contains a counter that increases by one for each successive repeat of a particular tag. The source also accounts for duplicate tags with varying cases by handling all tags in lower case. The resulting list of dictionaries is sorted by the tag frequency – highest to lowest.

**Input:** tags.csv,List of Movie Dictionaries

**Output:** Dictionary of Tag Frequencies

**Variables:** content tag, frequency counter

**Step 5:** Creates a comma separated value (CSV) file out of the information in the dictionary of tag frequencies.

**Input:** Dictionary of Tag Frequencies

**Output:** CSV file

**Columns:** Content tag, Counter

**ANALYSIS AND VISUALIZATION**

DISCUSION

This project sought to combine data sets from IMDB and MovieLens to yield a contextual overview of the Internet’s favorite movies. The final output for this project’s visualization considered the top 25 most frequent content tags in IMDB top 250 movies, which are listed below:

|  |  |
| --- | --- |
| **TAG** | **COUNTER** |
| sci-fi | 2497 |
| classic | 1830 |
| twist ending | 1681 |
| atmospheric | 1632 |
| psychology | 1339 |
| action | 1294 |
| fantasy | 1137 |
| imdb top 250 | 1119 |
| surreal | 1092 |
| thought-provoking | 1078 |
| dark comedy | 982 |

|  |  |
| --- | --- |
| stylized | 965 |
| violence | 920 |
| space | 893 |
| adventure | 884 |
| based on a book | 805 |
| social commentary | 778 |
| disturbing | 749 |
| quentin tarantino | 748 |
| visually appealing | 687 |
| dystopia | 682 |
| quirky | 674 |
| drama | 653 |
| dark | 629 |

These 25 tags provide a contextual profile of movie elements included in some of the internet’s favorite movies. In terms of specific genres, IMDB audiences prefer sci-fi/fantasy as well action/adventure, classic and dramas. IMDB audiences also enjoy films that are psychology engaging or surprising: twist ending, surreal, thought-provoking, atmospheric, disturbing. IMDB users enjoy a movie that is aesthetically appealing: atmospheric, stylized, visually appealing. Several tags also suggest that IMDB users prefer some darkness/dysfunction: dark, disturbing, violence, dark comedy, dystopia, quirky. Perhaps because his movies are dark, dysfunctional, psychologically compelling and violent, Quentin Tarantino showed up as the only named director or filmmaker in the top 25 tags. Finally, IMDB audiences also appreciate movies based on a book as well as those that express some degree of social commentary. Based on the results of this project, I have distilled the top 25 tags down to 4 primary characteristics; the data suggests that all time favorite movies tend to be 1) psychologically engaging, 2) dark/dysfunctional and 3) aesthetically appealing and 4) fall into one of the falling genres: sci-fi/fantasy, action/adventure, classic or drama. Finally, the data also suggests that Quentin Tarantino is well in tune with what IMDB audiences are looking for in a movie.

From this output, the resulting word cloud visualization was produced:

****

LIMITATIONS

This final output of this project comes with some limitations. Most notably, this project’s methods do not consider the number of content tags for each movie. Lesser known or classical movies may have a lower count of content tags which would therefore bias the selected tags for the word cloud towards blockbuster movies with a greater number of tags. Future projects could account for this limitation by limiting the number of tags that each movie can contribute and making those contributions reflect a representative sample of the actual number of tags.

Another limitation is the variation in tags with virtually the same meaning. For example, MovieLens users wrote tags for Science Fiction in both short (“sci-fi”) and long (“science fiction”) forms, both of which are tags in the top 100 of the final output. Or, as another example, the frequency list includes all of the following tags: “disney”, “best disney movie”, “disney animation”, “disney studios” and ”walt disney studios”.