

Department of Media and Communication

**COM5507 Social Media Data Acquisition and Processing**

Project Report Paper

Instructor: Dr. Xinzhi ZHANG

QIU Sha  55437944

QIN Jiali  55356740

CHEN Weizhou  55500730

ZHAO Feier  55035408

**Comparing and Analyzing Low-rated movies on Douban and IMDb.**

*First, this paper mainly focuses on the full procedures for our project. That means we want to explain why we choose this topic and how we scrap and process the data from Douban and IMDb. Second, after reviewing the literature and research written by others and analyzing the results from our visible tables, we try to make some conclusions about our puzzle and figure out the main reason which caused the negative comments on domestic movies in China.*

*Keywords: Douban, IMDb, negative comments, domestic movies*

**Introduction**

*Background*

In the information age, we can easily get comments and rates of a movie and decide whether to watch it or not. There are some sayings like “the golden age of Chinese movies has passed” and “film industry in China not only falls behind American but also can’t catch up with Korean and Japanese”, similar negative comments could be easily found on the internet. After an overview of online comment, we will find that Chinese audiences and mass media prefer to hold a pessimistic attitude towards domestic movies. On the contrary, the number of domestic film production and total box office are gradually increasing, which implicates the flourishing of the entertainment industry. According to a report published in 2017, the total number of Chinese film production will exceed 930, and the number of total box office will increase to 559.11 million yuan.

Two polarized attitudes(or trends)made us really confused. We decided to find more evidence to both sides although we had no idea about who was right. So we started the project of data collection and processing two months ago, trying to give a fair and objective evaluation to domestic movies from a new perspective. After several-day discussion, we decided to collect information of all low-rated movies on the internet and find out something in common among these “bad movies”. Here are some details about what we had done during such a tough but meaningful period:

Firstly, we scraped genre pages of IMDb, a website which was launched in 1990 and has been a subsidiary of Amazon.com since 1998. Moreover, IMDb has approximately 5.3 million movie titles (including episodes) in its database as well as 83 million registered users (Data Source: Wikipedia, data updated to 2018), above two features make IMDb a good information source for data collection.

After a discussion with our instructor, the instructor reminded us to compare the difference between a Chinese movie website and IMDb in recent years, by which we could get some interesting and convincing findings from a large amount of data. So we selected Douban, the largest online community of movie fans in China, which has more than 200,000,000 monthly unique visitors (Data Source: TechWeb, data updated to 2013), as another object for our project.

*Research question*

How to define the low-rated movies? After scanning all the rates on both two websites, we defined the quantities of movie graded 4 points and below in recent 5 years as research objects (DV). From the data of these two websites, we tried to figure out the distribution of our objects. And we measured the distribution from three different dimensions: time(date), genre and country (IV). We tried to find out if the low-grade movies were intensively released in a certain year and if one certain kind of movies were likely to be graded lower than other genres. As Chinese always said that China is the cradle of garbage movies, we also wanted to verify the credibility of this kind of claim, so we compared the quantities of low-rated movies in different countries and to be specific, we would show the distribution of low-grade movie in China, e.g. director, actor, genre and etc.

1. Time Distribution

We took “time(date)” as independent variables to test if there was any relationship between the time and the quantities of low-grade movies. We had compared the quantities of the object in recent 5 years - 2013, 2014, 2015, 2016, 2017。

2. Genres Distribution

We classified all the movies we had scraped according to their genres. And we calculated the quantities under each genre to see under which genre the most low-rated movies were produced. We got different results in two different platforms so that we could move further to compare the output on Douban and IMDb, like whether different audiences had different tastes on movies, whether there were some genres that easily to be rated below 4 points on both two platforms, and so on.

3. Country Distribution

Both websites contain the movies almost from all the countries, and after glance the countries classifying, we found the highest production countries were same. So we’d like to process the data from two platforms.

4. Distribution in China

Not much difference between the effort above, we would present the distribution of China low-grade movies regarding time and genre.

**Research method**

*Douban*

**Data collection.** Firstly, we tried to use the code which was taught in our class to realize our goal. Our data range of year was 2013 to 2017 since we could not select movies by every year in Douban, the closest data source was movies of 2010s, so we decided to scrap all the target movies of 2010s, and then narrowed the range from 2013 to 2017. There were only movie titles and scores shown on the first webpage, which meant we had to use selenium to imitate the click action to click every target movie titles in order to get details like movies all. URLs of the first page, and then looped every URL to click and gain every movie detail inside. We found the URL that includes all URLs in the first page, copied the xpath and then we tried the flow, but it reminded us there were something wrong with the xpath. However, when we focused the exact xpath, it did not show all the urls as we expected. Things were so strange, but we had no idea. After googled the solution, we downloaded a Plug-in which was designed to analyze every xpath on the website. But the strange thing was that it also proved our xpath was correct. When we added it into our code, it still didn't work. After asking friends of the related field, we knew the problem was the website itself, without any solution, so we turned to our second attempt.

we found several reference codes on the website like JianShu.com and Zhihu.com. We tried to find the code aiming the same goal as ours. Then we decided to use a package called pyquery, which could analyze all the HTMLs on the website so that we could find all the target URLs. Besides, pandas as well as so were also imported. After import and definition, we began to loop the URLs on the web pages and scrape all the details we wanted. Due to pyquery was a new package for us, we cited, integrated, and amended other codes from different website, the debugging process was tough since we didn’t know the package well. So we had to consult some related students to explain some code for us. At the same time, we were prevented by the anti-scrap mechanism of Douban. So we googled again, finding that adding: headers = {'user-agent': 'Mozilla/5.0 (iPhone; CPU iPhone OS 11\_0 like Mac OS X) AppleWebKit/604.1.38 (KHTML, like Gecko) Version/11.0 Mobile/15A372 Safari/604.1'} could imitate a browser visit instead of crawler, which was extremely useful. Although all these attempts and difficulties were time consuming and energy consuming, the process meant a lot to us.

***Data processing.*** After reviewing all the raw data, we got from websites, we found there were many duplicate data and missing values. Integrating the codes from teachers’ lecture and some online notes, we dropped all the useless values.

1.    First, we cleaned the type index and the production index. Both of them contained more than one word, and we used “for” as a determine statement which helped us keep the first type and producer in every cell.

2.    Then we applied “data. isnull().sum()” to count the vacancy variables and we deleted useless rows by the step of “data.dropna(axis = 0,inplace=True)”.

3.    The “director” was a useless column for the next analysis, so we deleted it

4.    The last step was coded to process the “publish date” column, reserving the years，selecting the movies published during 2013 and 2017, and removing other rows.

After processing by python, the rows of our table declined from 1680 rows to 879 rows. The remained CSV was easier to be visualized. To be honest, it is too hard to visualize this text table by Python, so we used Dishu.com as our main visualizing tool.

*IMDb*

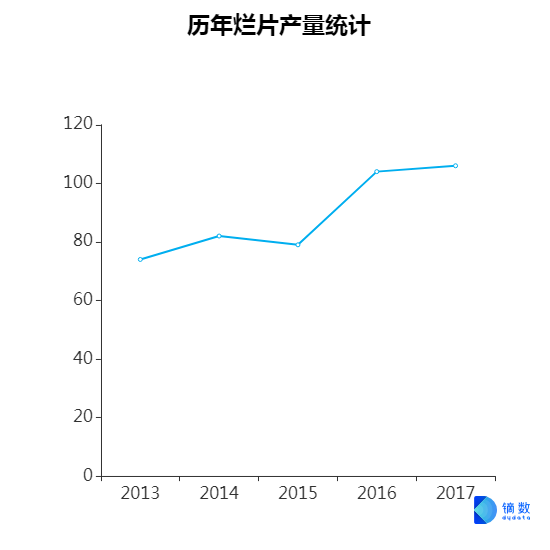
**Data collection.** We started the work of IMDb part by scraping data from specific pages of different genres, such as comedy, horror, etc. We wrote a set of code to automatically enter each page and rank all movie items according to their rates. Then we used Python to collect all movies that scored with 4 points and below. If releasing date of a low-rate movie was within the range from 2013 to 2017, relevant information of this movie (including titles, countries, and genres) would be recorded. After completing the previous step, we used Panda to frame all the results so that we could create a new file as CSV; This one was considered as the longest as well as the hardest step of IMDb part. In the process of creating the new large file, we were reminded by our instructor to divide results into some sub files, for time-saving consideration. Thanks to all efforts we made in several days, 5426 items were successfully recorded and printed in different 5 CSV sub files, while some items appearing multiple times had not been removed at that time.

**Data processing.** As one movie might fall into different genres, it was inevitable to record the same items several times. Too many duplicates would bring troubles to further processing, for this reason, we still needed to clean the raw data which we had acquired. So we merged above CSV sub files and imported Numpy to find all duplicates then remove them. Finally, 2969 items retained, the number of which was equivalent to 54.7% of the raw data. Both RE module and Counter module helped us to work on word frequency statistics. But when we tried to use Word Cloud to visualize above results, we were warned by a “module-not-found error” and it pointed that “no module named 'wordcloud.query\_integral\_image'”. Such an unexpected problem made us so confused because all of us had checked the path for several times, however, it also drove us to seek more details and solutions. Finally, the reason became known to us: There must be something wrong with the version of Python, resulting in the incompatibility with the Win 10 operating system. A problem solved, another came: All of us were not familiar with the usage of Regular Expression, so we experienced a really tough time to learn about it. We spent several days to have hundreds of attempts, but frankly, the final result failed to reach our anticipated efficiency. To save more time for follow-up work, we decided to complete the rest part of calculation and visualization by Excel.

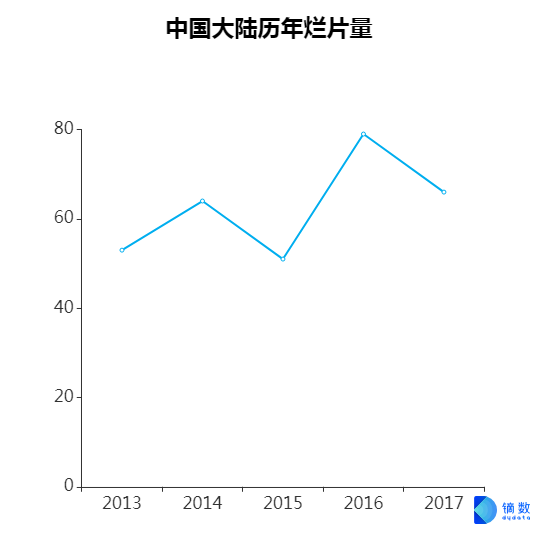
**Results**

*Douban*

Our independent variables were countries, years, and movie genres while the dependent variables were movies rates. The statistics were one-dimension, and we analyzed them in China and the whole world perspectives.



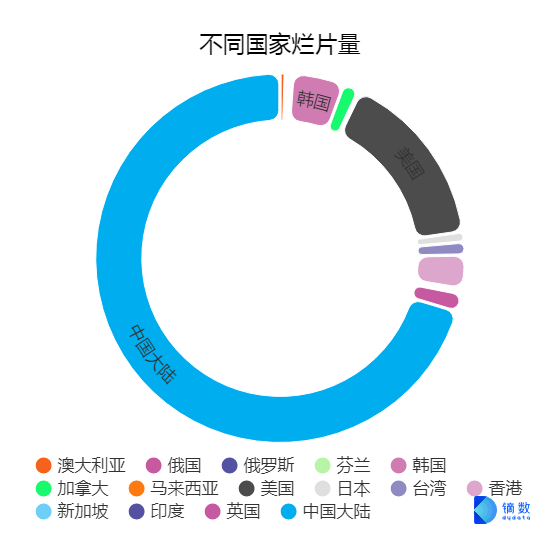
**Figure 1.** The line chart of year distribution of low-rated movies



**Figure 2.** The line chart of year distribution of low-rated movies in China

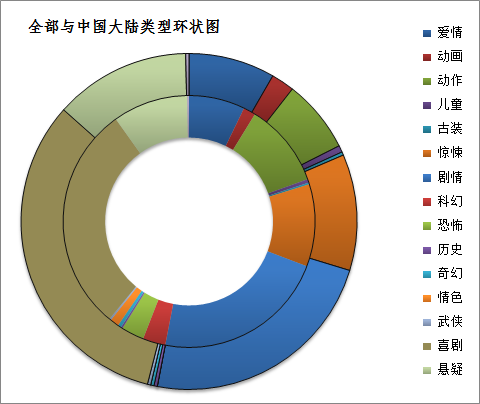
Generally speaking, the year trend of world’s bad movie output experienced a continuous increase from 2013 to 2017, first slightly and then a little bit significantly.

While year trend of China was a little bit different. The numbers of bad movies fluctuated within 5 years, and its total amount increased, reaching a peak in 2016.



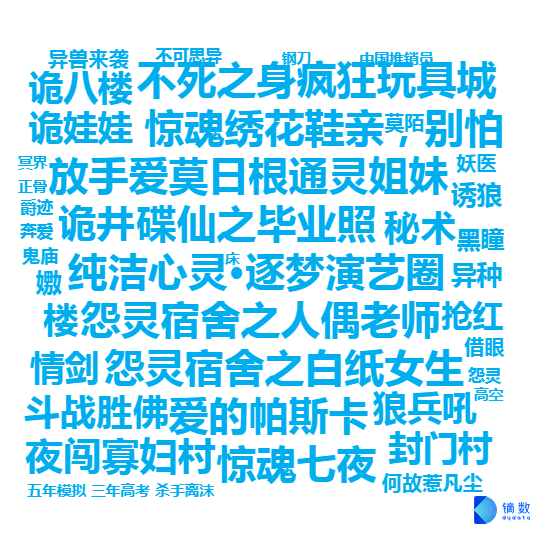
**Figure3.** The countries distribution pie chart of low-rated movies on Douban

As we could see clearly from this chart, mainland China made up the largest percentage of bad movies output from 2013 to 2017, which was approximately three forth, followed by America and Korea.



**Figure4.** The types distribution pie chart of low-rated movies on Douban.com.

In the movie genres distribution part (the outside circle), we could see comedy accounts for largest percentage of the whole world bad movie output, after which was romantic movies. It was quite surprising to find that Chinese and world’s movie genres followed the same distribution (the inside circle).

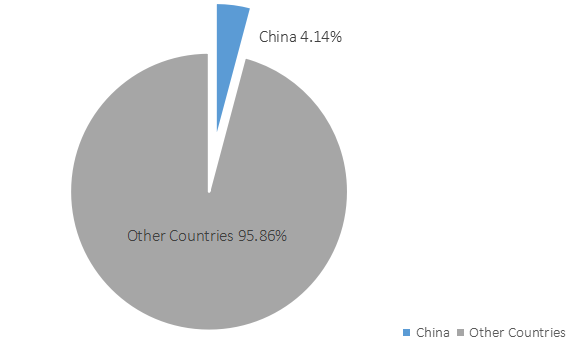


**Figure 5.** The titles distribution word-cloud of low-rated movies on Douban

In this graph, the bigger size of words, the lower rates for these movies. From this chart, we found that titles of those horror films or teen movies were always easy to get a lower grade on Douban when it compared with other genres.

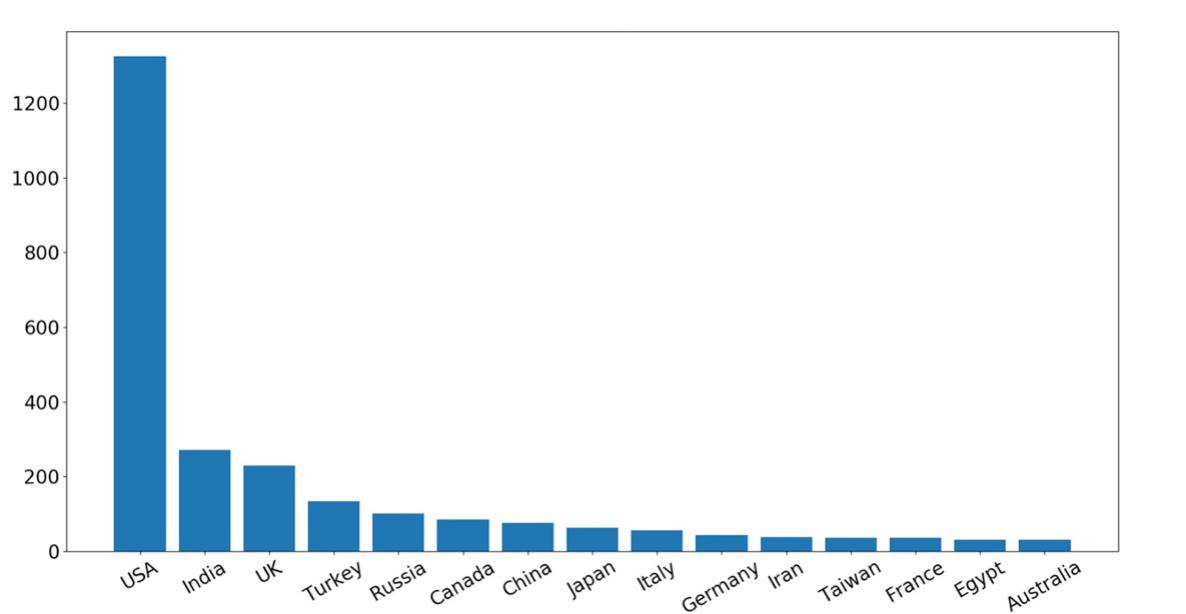
*IMDb*

It might be a common sense that most of IMDb users were those people who came from English spoken countries, which made IMDb so different with Douban, an online community for Chinese movie fans. As we had hypothesized above, movie preference of IMDb users could be found by some indicators on pages, for example, the genres of most popular movies might show some trends to follow. At the same time, to support the result we acquired from Douban, we collected a series of statistics from IMDb as a reference group. The details were as follow:

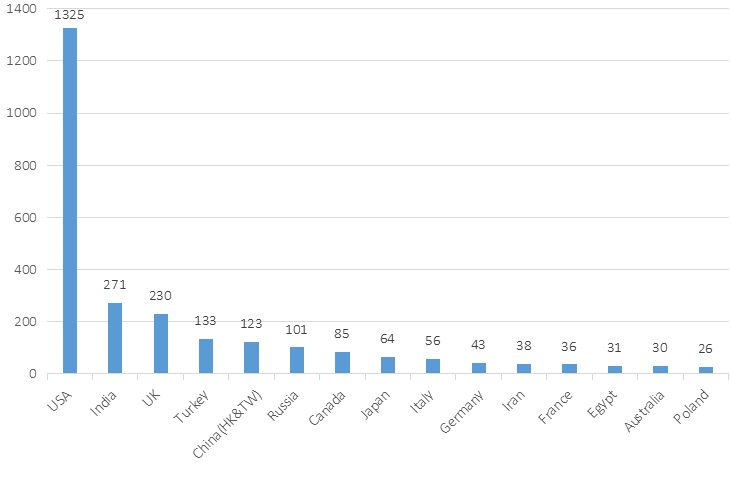


**Figure 6.** The Proportion of Chinese Low-rated (4 Points and Below) Movies, Including HK and TW. (Created by Excel)

123 Chinese bad movies (including movies of Hong Kong and Taiwan) released in recent five years had been recorded on IMDb pages, of which number made up 4.14% of the total number of world’s bad movies, by comparing with the same period. To enhance visual effect, we used a brighter color to highlight the share of Chinese bad movies, which occupied a little proportion of the pie. Could we directly say Chinese movies performed better than many countries? Probably not. Rather, this aroused us to think about whether Chinese movies weighed too little to bring attention from world’s film industry.

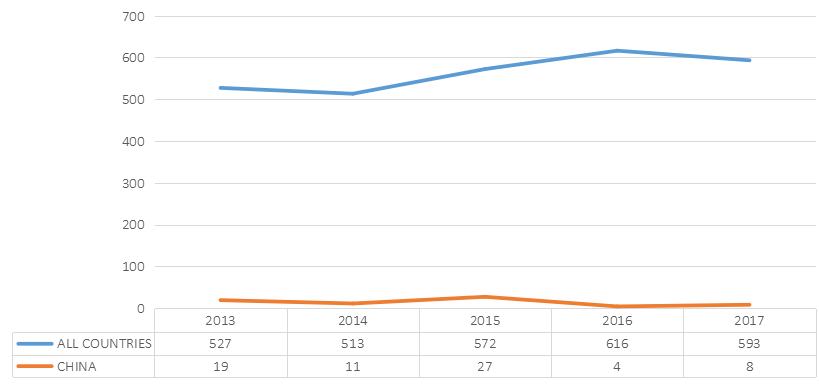


**Figure 7.** World’s Top 5 Low-rated Movies on IMDb. (Created by Python)



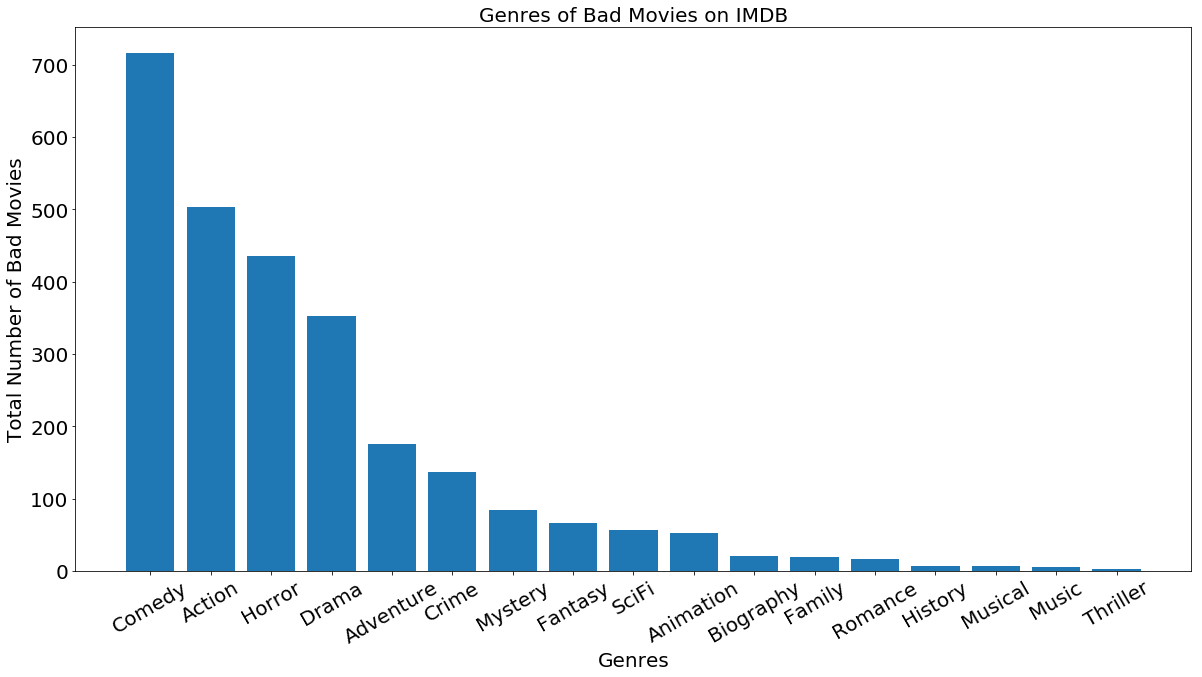
**Figure 8.** A Revision of Figure 7-1. Movies of Hong Kong and Taiwan Were Included to Chinese Movies. (Created by Excel)

The two charts showed more details than the above pie chart. As we had known already, when it came to the number of bad movies, China did not rank first on the world’s list from 2013 to 2017, even could not enter the Top 3. Which country had produced much more bad movies than other countries? The answer was evident enough: USA. According to the theatrical market statistics published by Motion Picture Association of America (MPAA), 708 movies released in 2015, some of which were made by MPAA members (8%) and most of others were produced by non-MPAA affiliated independents(MPAA, 2016). Compared to the number of American low-rated movies, probably we could say IMDb users were not satisfied with the overall performance of American movies, which might be so different with our impression.



**Figure 9.** A Comparison of Bad-movie Production between China and All Countries. (Created by Excel)

Every year bad movies came out, while some years bad movies were more than other years, although the specific reason was unknown to us. We could only speculate those yearly changes in world’s bad movies might result from the global political situation, new cultural trends or some social events. Compared to the world’s number from 2013 to 2017, the number of Chinese low-rated movies changed smoothly, mainly because of the small sample size of Chinese movies we got from IMDb.



**Figure 10.** 17 Common Genres of Bad Movies on IMDb. (Created by Python)

We could learn about a fact that Chinese audiences were more likely to show less patience to comedy than other genres, according to the findings we got from Douban. Regardless of so-called “cultural difference” between China and the West, at this time, we found a striking similarity that people came from English spoken countries also rated hundreds of comedies with low points in the past five years which made “comedy” rank first of the worst list of genres. The second and third positions belonged to “action” and “horror” with 503 and 435 movies respectively, although both of them were popular around the world. We considered that movie genres like comedy, action and horror had more audiences than others, so people would like to pay more attention to the above three options.



**Figure 11.** Word cloud of High-frequency Words in Titles (Created by Python)

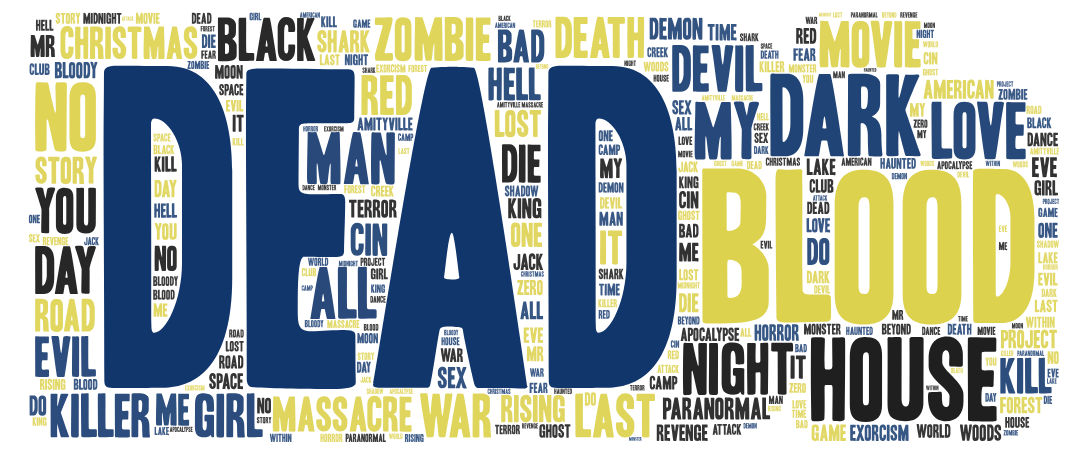
**

Figure 12.  A Revision of Figure 10-1 Only English Words Were Retained.

(Created by Word art)

People familiar with IMDb might know that the website editors did not name all movies with their official English titles. For example, “Once Upon a Time in the Northeast”, a Chinese low-rated movie, was recorded on IMDb with its unofficial name in Pinyin as “Po Ma Zhang Fei”, which was meaningless in English. So we used Python to count all words in titles and list top 100 in a TXT file, then manually selected English words.

After removing those stop words and non-English words, a meaningful graph jumped out: The word “dead” occupied a relatively center place with the largest size, which meant it appeared more frequently in titles of low-rated movies. As what we could find from the word cloud, some non-mainstream words like “massacre”, “zombie” and “killer” also commonly used by bad movies.

**Conclusion**

Back to our purpose of this topic, we think the output charts and graphs can help us to conclude the real status of Chinese movies.

**Chinese movies are not bad as we imagined before.** On one hand, we have to realize the gap when comparing with others, on the other hand, we can’t look down upon domestic movies. From the results from IMDb, we can find that China isn’t a country with the highest level of bad movies.

**Chinese movie industry should admit its shortcomings.** We can’t be cheated by the increasing number of movie production in China and can’t recognize the gap between China and the west. From the distribution of the type of bad movies in China, the number of comedies is the highest. Comparing with other genres of movies, the comedy is a kind of typical commercial movies. The investment in comedy always be low although the audiences of comedy are numerous. High income and low spend, this feature really appealing in the film market. To attract audiences and get more feedback, using those idols with high popularity is another way for filmmakers to earn money. The largest part of investment is spent on actors, which definitely have to cut the spends on screenwriters and producers. Low quality of scenarios and production are two fatal reasons for these bad movies. The technology of film production is another gap between China and the west. Many people would like to watch Hollywood Action Movies, who enjoys the sensory stimulation of superb film special effects. It seems that Chinese movie industry still has a long way to go to catch up with other developed countries.

**Discussion and limitation**

*From the perspective of technique*

To be honest, we have applied some third-party tools like Excel, to help us to process and visualize our data. We mainly want to figure out the distribution of these movies on aspects of years, countries and genres. So the line charts and pie charts are two main kinds of chats in out report. It is difficult for us to draw these charts by Python, however, we can get them easily by Excel. After trying a lot of time, we had to give up the code and choose an easier way to get our graphs.

*From the level of measurement*

Another pity of our research is that the independent variables and dimension of measurement are limited. If we can deeply research this topic in the future, we’d like to extend the range of movies to get a more obvious relationship between years and numbers of low-rated movies. Second, we also want to get more detail information about these movies, like the directors, actors, to find if there are some same parts, which caused the movie rated low.

**Reference**

Press Room. (2018). Retrieved from

https://www.imdb.com/pressroom/stats/

豆瓣宣布月覆盖用户数达2亿 同比增长一倍. (2013). Retrieved from <http://www.techweb.com.cn/internet/2013-11-13/1356287.shtml>

Booming Global Box Office Takes Center Stage at CinemaCon, Reaches Record $38.3 Billion in 2015. (2016). Retrieved from

<https://www.mpaa.org/press/booming-global-box-office-takes-center-stage-at-cinemacon-reaches-record-38-3-billion-in-2015/>