# Cinema: Analysis of Genres and Plot Texts and Their Impact on 'Box Office' Performance

*Maxim Novoseltsev*
*under the guidance*
*Prof. Dr. Martin Braschler and Dr. Mark Cieliebak*

Project Work
DAS Data Science
School of Engineering
Zurich University of Applied Sciences (ZHAW)
October 6, 2015

# Introduction

The growth of interest in the text and data analysis as well as in informational retrieval in the last years from the commercial institutions is partially based on the desire for better understanding of the customer or user preferences.

Along with the media monitoring, the sentiment analysis of the user reviews and social media is playing more and more important role.

The sentiment analysis is mainly based on the building classification models on the manually labeled "positive" and "negative" texts.

The motivation for the current work is to consider the impact of the customer perception related attributes such as movie genre and movie plot, on the "box office" performance, and inspired mainly by the sentiment analysis.

As a data source the data from International Movie Database (imdb.com) was used. The data [1] can be downloaded from ftp server [2] and is free for non-commercial use [3].

# Problem description and previous work

There is a rather lot of statistical analysis works performed on IMDb data. Some of these works are performed by academical institutes, other by enthusiasts (see the *Used Sources* section).

Big part of them focused either on the prediction the user ratings of the movies (e.g. [5], [6]) or social network analysis of the involved staff (actors, directors etc.), e.g [9], [8].

On the other hand, the predictions of a movie success, as e.g. in [10], are often oriented on maximizing accuracy. The usage of "black-box" methods, such as neural networks, with the exploring a big number of attributes, such as the movie rating of the reviewers, social media statistics etc., could bring good performance, but sometimes hides the influence of particular predictors. For example, the work [11] uses machine learning methods for the prediction the user ratings and based on a lot of attributes, including the analysis of related tweets.

As it was mentioned before, the main motivation of this work is not to build a "good" predictive model, but to consider the influence of genres and movies' plots directly on the box-office performance as well as its possible evolution over time.

# Data preparation

The IMDb data itself is a set of compressed semi-structured, subject-oriented text files. The example of the text file for the genres is shown in the Table 1.

To reduce amount of data parsing-related work the Python-based tool IMDbPY [13] was used. With the help of the tool, the data was converted and transferred into a SQLite [12] database. The further reverse engineering showed that the database is build with the Entity-Value-Attribute design principle. As a result of this approach all numerical values in the database, such as budget and gross (sales), are represented as free text values (s. Table 2)

It can be seen that the box-office (gross) data consists of the multiple rows, each of them represents the gross amount by particular date.

```
"1-0 til Danmark" (2014)                                History
"1-2-3 Istanbul!" (2009)                                Adventure
"1-2-3 Istanbul!" (2009)                                Comedy
"1-2-3 Moskau!" (2008)                                  Adventure
"1-2-3-los!" (1967)                                     Music
```

Table 1: *The sample content of an IDMb text file*

```
GR: USD 352,114,898 (USA) (3 January 2010)
GR: USD 283,811,000 (USA) (31 December 2009)
GR: USD 212,711,184 (USA) (27 December 2009)
```

Table 2: *The* gross *records in IMDb files*

As USA is the biggest movie producing and consuming country in the world, the author has decided to concentrate on the movies, which were produced (full or partially) in USA. Then the movies were chosen, which have both budget and gross data. As movies' gross data is stored as multi-row free-text, the following transformation logic was applied: with the help of regular expressions the data with the strings, containing "USD" was filtered, then numerical values were extracted and finally the maximum of all movie-related values was taken. This transformation and other data engineering tasks were performed with the help of *R* software and, particularly, its package *RSQLite* [16], used for the SQLite data manipulation.

For the further analysis it was decided to consider the time period, covering the last 20 years, excluding year 2015 (1995-2014).

As a result of data processing and cleansing the R data frame with 3654 rows and 25 columns, which are *title, production year, budget, gross* as well as genre columns of Boolean types. was created. The most of the movies associated with more than one genre and this was the reason, why the genres are not represented as a single column. The list of considered genres is in the Table 3.

## Analysis

### Data Exploration

As it was mentioned above, most of the movies, namely 3085 from 3654, belongs to more than one genre.

On the Figure 1 the total number of movies and number of unprofitable (e.g which gross is less than budget) movies in each genre in the form of bar chart are shown.

On the Figure 2 the mean budget and mean gross are plotted. It can be seen that the highest

```
Action, Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy,
History, Horror, Music, Musical, Mystery, Romance, Sci.Fi, Sport, Thriller, War,
Western
```

Table 3: *Genres*

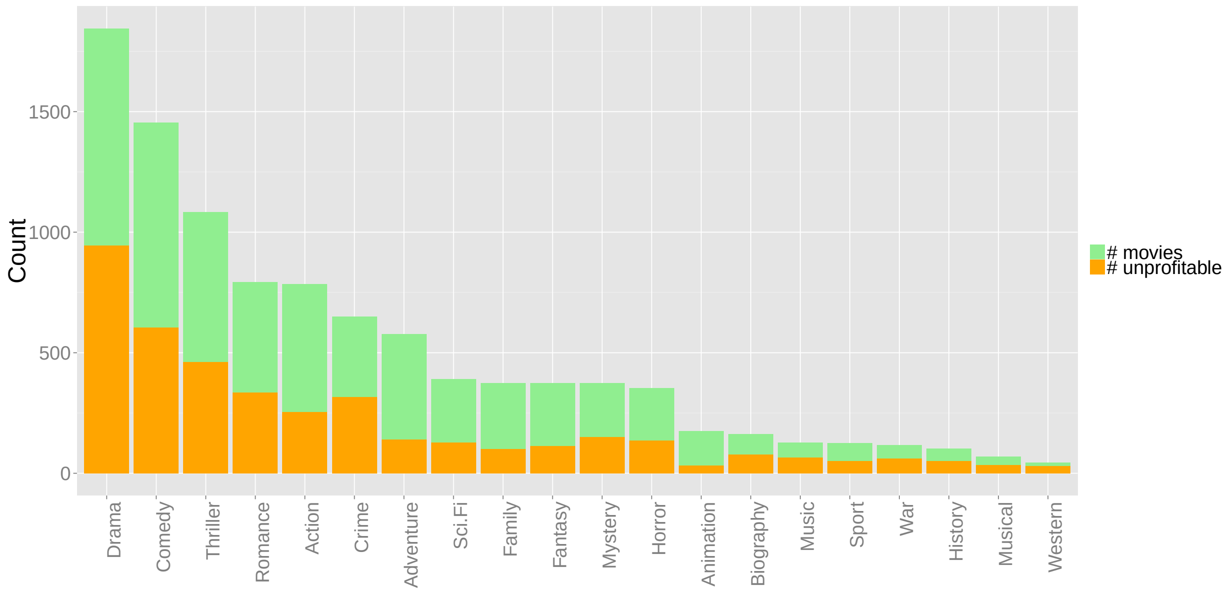

*Figure 1: Count of profitable and non-profitable movies*

average budget and gross values have the movies in the genres *Animation* and *Adventure*. The mean of the gross of genre *Western* is just slightly over the budget mean.

It is worth to mention that *Action* and *Animation* genres have a rather high proportion of unprofitable movies, although the mean (expected) profit for them is relatively high. One can suppose, that a big proportion of the movies in these genres is unprofitable, but a few movies collect a "good" box-office.

On the Figure 3 there is an attempt to bring the break even probability and expected returns pro genre in one chart. One can see, that all genres are generally profitable, but the probability to be profitable for some of genres (e.g. *Western*) is lower than 0.5.

On the Figure 4 the budget distribution, reflecting the proportion of profitable/non-profitable movies for genre *Drama* is shown. It can be seen, that the profitable part is bigger for the movies with the bigger budget and conforms with the assumption made above.

## Linear models

Initially the logistic regression model, predicting the probability of break-even (e.g. the movies, for which the gross exceeds the budget) with the attributes *budget* and *production year*, was decided to be considered. The calculated coefficients and *p-value* can be seen in the Table-4

| | Estimate | Pr(>\|z\|) |
|---:|---:|---:|
| (Intercept) | -2.8E+01 | 3.3E-02 |
| production_year | 1.4E-02 | 3.6E-02 |
| budget | 2.1E-08 | 6.5E-60 |

*Table 4: Coefficients for the logistic regression for break-evens (gross > budget) with the predioctors production year and budget*

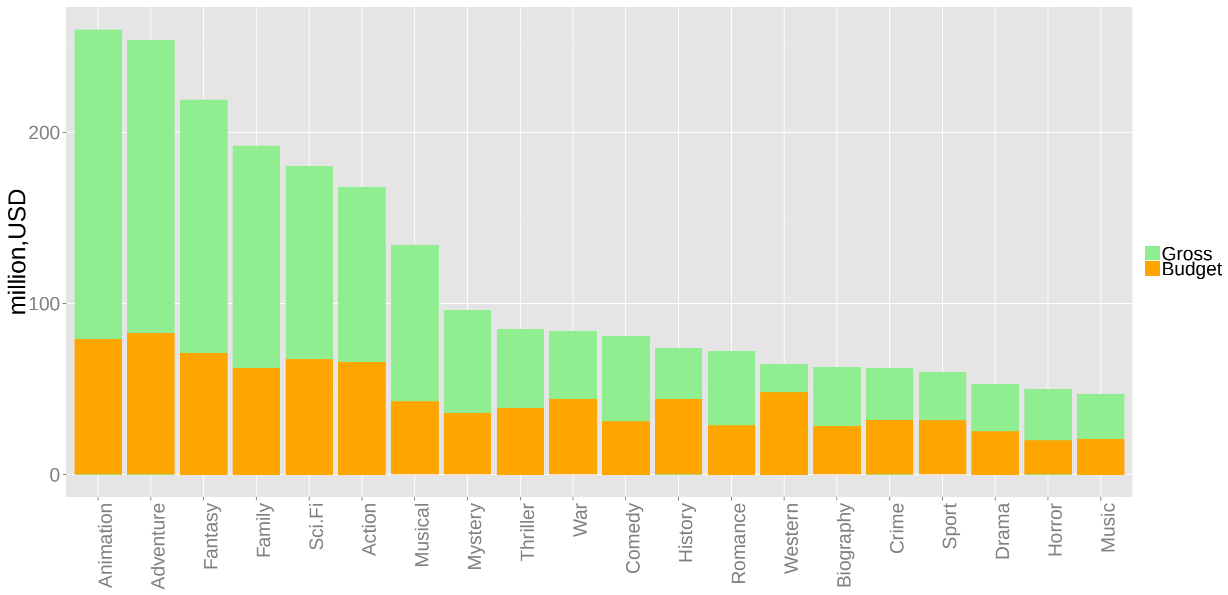

*Figure 2: Mean Gross and Budget pro Genre*

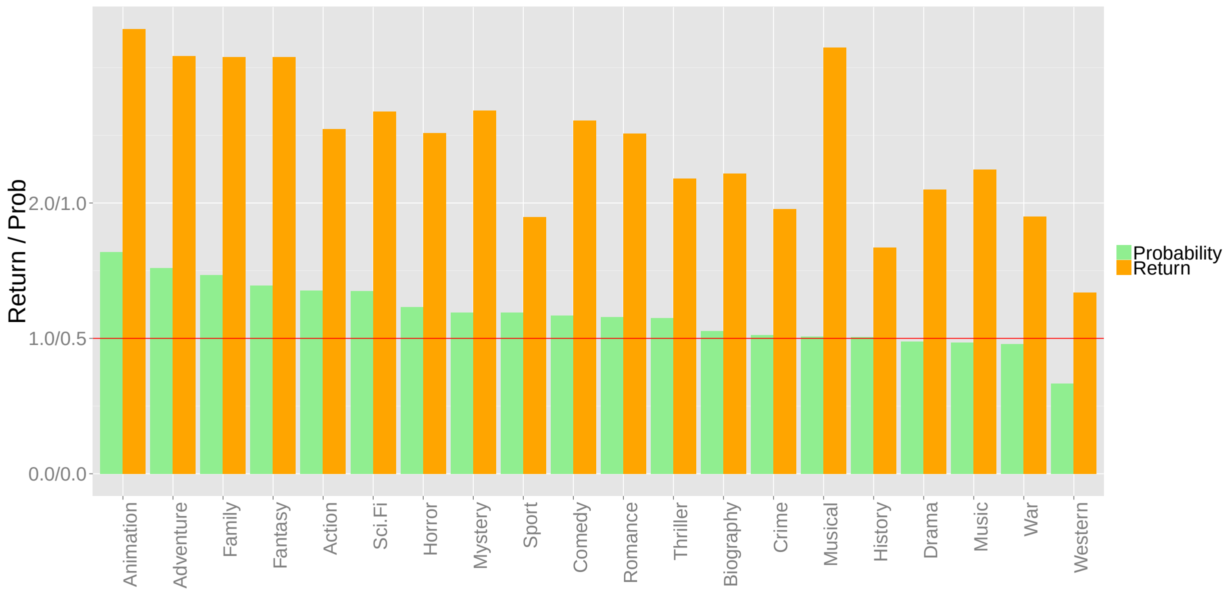

*Figure 3: Probability to be profitable and expexted return pro genre*

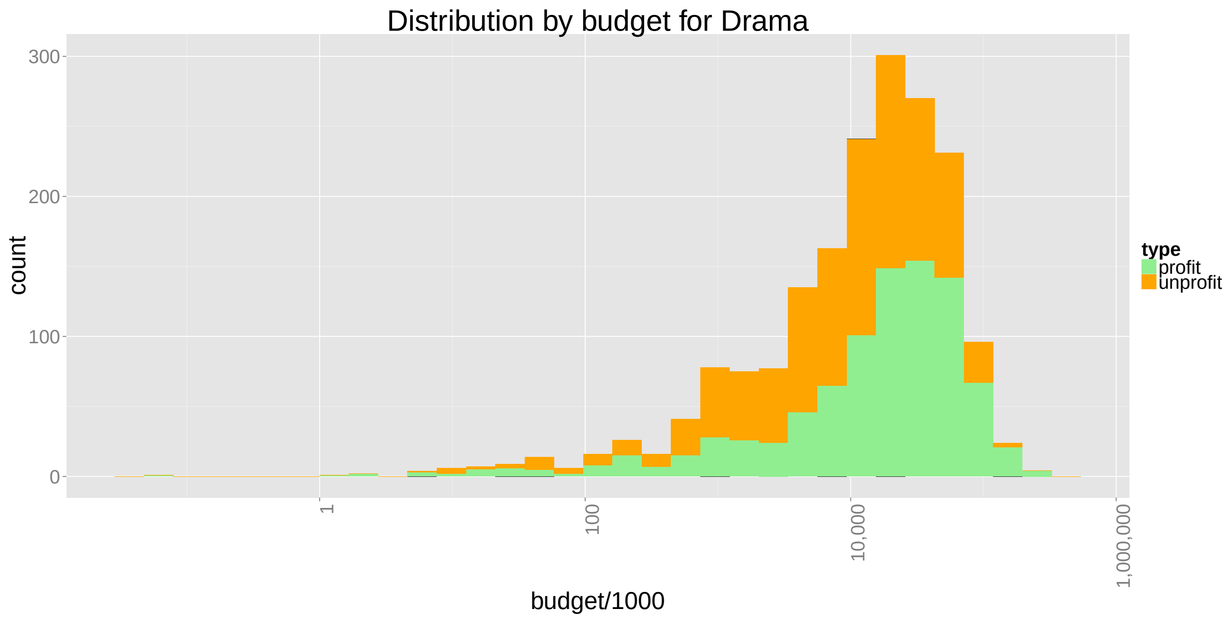

*Figure 4: Probability to be profitable and expexted return pro genre*

It can be seen, that the break-even chance has strong significant dependency on a budget and increases with it.

Then the previous logistic regression model was extended with the genres and their interactions with *budget* and *production year* (to capture possible evolution of genres' influence). After that the backward step-wise predictors' elimination, based on the Bayesian information criterion (BIC), was applied. As a results the list of the most influence predictors was retrieved, that can be seen in the Table 5

| | Estimate | Pr(>\|z\|) |
|---:|---:|---:|
| (Intercept) | -2.5E-01 | 6.3E-04 |
| budget | 2.1E-08 | 2.5E-56 |
| Drama | -3.7E-01 | 5.7E-07 |
| Horror | 4.0E-01 | 1.3E-03 |
| Romance | 2.6E-01 | 2.2E-03 |
| Western | -1.4E+00 | 1.1E-04 |

*Table 5: Coefficients and p-values for the logistic regression (gross > budget) with the predictors production year, budget, genres and genres' interaction with production year and with budget after BIC-based predictors' elimination*

Generally, the bigger budget is the more chance to get break even. It can be seen, that the movies of genres *Horror* and *Romance* have a higher probability to be profitable in compare to the other genres. At the same time the genres *Drama* and *Western* have a higher risk to be unprofitable.

Another metric, that could be interested from the business point of view, is expected return, that is a ratio of gross amount to budget. As the returns' distribution is right-skewed, the log-transformation was applied for the target variable. Similar to the logistic regression the BIC-based backward step-wise predictor elimination algorithm was applied. The results can be seen in the table 6

|  | Estimate | Pr(>\|t\|) |
|---:|---:|---:|
| (Intercept) | -7.5E-01 | 1.0E-22 |
| budget | 2.0E-08 | 1.3E-39 |
| Action | 2.8E-01 | 4.1E-02 |
| Adventure | 4.0E-01 | 1.5E-02 |
| Drama | -3.1E-01 | 1.6E-05 |
| Family | 5.4E-01 | 2.0E-03 |
| Horror | 4.6E-01 | 9.9E-05 |
| Mystery | 3.7E-01 | 1.1E-03 |
| Romance | 3.4E-01 | 3.3E-05 |
| Thriller | -2.7E-01 | 1.3E-03 |
| Western | -9.7E-01 | 8.9E-04 |
| budget:Action | -6.9E-09 | 8.7E-04 |
| budget:Adventure | -6.2E-09 | 3.1E-03 |
| budget:Family | -7.5E-09 | 1.9E-03 |

Table 6: *Coefficients and p-values for the linear regression (log(gross/budget)) with the predictors production year, budget, genres and genres' interaction with production year and with budget after BIC-based predictors' elimination*

The result of linear regression (s. Table 6) mainly conforms with the break even probability analysis. The genres *Romance* and *Horror* are "out-performers" (they have higher probability of break-even as well as above average return), while the genres *Drama* and *Western* can be considered as "under-performers".

## Plot text analysis

It is interesting to analyse the influence of the movie plots alone on the box office performance. Through this analysis one can retrieve some information about the audience preferences.

The detail level of the plots in the database is very different: the number of words varies from 13 to 567 (s. Table 7).

|  | No. of words |
|---:|---:|
| Min. | 13 |
| 1st Qu. | 66 |
| Median | 97 |
| Mean | 106 |
| 3rd Qu. | 127 |
| Max. | 567 |

Table 7: *Statistic of plot texts' word counts*

For the text analysis the author used *R* with the package *tm* [18] (plus package *caret* [17] for the cross validation and parameter optimization) as well as Rainbow program [14]. The last is a powerful command-line text analysis toolkit.

The main idea for the plot text analysis was to classify the plot texts into 2 classes: one class contains the movies with the positive returns (the gross is bigger than budget), while the other contains the rest. As it can be seen above, the genres have significant influence on the box-office performance. As some words in the movies' plots can be genre-specific, it was decided to consider the genres separately (although the most of the movies belong to more than one genre, it would minimize the "genre classification" effect).

The experimenting with *Rainbow* with *Roccio* and *Naive Bayes* classification methods showed, that both have approximately the same performance. The usage of 2-gram of word sequences did not improved the performance too. It was then decided to use further *R* with the *Naive Bayes* for the convenience reason.

As the precision by a simple classification with target profitable/unprofitable was relatively small the following strategy was applied: the movies were chosen from bottom and top 25 percents of the returns (ratio gross/budget) to make the distinction of the "success" and "fail" movies more clearly.

The text was pre-processed by lowercase transformation, stop words removal and stemming.

After experimenting (Naive Bayes, SVM) the Naive Bayes model was chosen. The word presence flags are used as predictors. This "binary" weighting gave better performance in compare with the traditional *tf-idf* weighting. For each of 10 genres, with the most number of movies in the database. the Area Under Curve (AUC or, often, ROC) was calculated, based on 10 cross-validations.

| Genre | ROC | Sens. | Spec. | F1 |
|---|---|---|---|---|
| Fantasy | 0.72 | 0.72 | 0.48 | 0.67 |
| Sci.Fi | 0.68 | 0.70 | 0.55 | 0.67 |
| Family | 0.65 | 0.68 | 0.48 | 0.67 |
| Comedy | 0.64 | 0.68 | 0.55 | 0.67 |
| Thriller | 0.62 | 0.65 | 0.53 | 0.67 |
| Romance | 0.60 | 0.57 | 0.50 | 0.67 |
| Adventure | 0.59 | 0.55 | 0.52 | 0.68 |
| Action | 0.59 | 0.62 | 0.49 | 0.67 |
| Drama | 0.57 | 0.60 | 0.49 | 0.67 |
| Crime | 0.53 | 0.62 | 0.47 | 0.67 |

Table 8: *ROC, sensitivity, specificity and F1 score for some genres. Calculated for the data, containing the top and bottom 25 percents of the returns*

As it can be seen, the AUC value is not spectacular high. For the plot word influence analysis, the ratio of the probability of particular word presence, given that the film is "successful" to probability of the word presence, given that the film is unprofitable, or $\frac{P(word|success)}{P(word|fail)+P(word|success)}$, was taken. The result object of the *R* Naive Bayes model contains the table with these values (see example in the Table 9).

This logic was applied for the genres, which are most "predictable" (high ROC-s). To filter out the seldom words, the words with the relative frequency more than 0.1 (at least 10% of movies plots

| *love* | exists | not.exists |
|---|---|---|
| fail | 0.1702128 | 0.8297872 |
| success | 0.1276596 | 0.8723404 |

Table 9: *Example of probability table for the term love*

have the word) were chosen.

Some word "success" probability numbers one can see in the tables: for *Sci-Fi* in 10 and in 11; for the genre *Fantasy* in 12 and in 13; for *Comedy* in 14 and 15. One should take into attention, that the words are stemmed.

| head | teenag | meanwhil | day | give | place | strang | plan | anoth | armi |
|---|---|---|---|---|---|---|---|---|---|
| 0.92 | 0.90 | 0.87 | 0.83 | 0.83 | 0.83 | 0.83 | 0.82 | 0.82 | 0.80 |

Table 10: *Sci.Fi : the words with the highest probabilities for a movie success*

| murder | crew | woman | young | leav | men | someth | creat | behind | left |
|---|---|---|---|---|---|---|---|---|---|
| 0.17 | 0.30 | 0.30 | 0.32 | 0.33 | 0.33 | 0.33 | 0.33 | 0.36 | 0.36 |

Table 11: *Sci.Fi : the words with the lowest probabilities for a movie success*

| armi | happen | begin | hous | know | never | lord | even | place | follow |
|---|---|---|---|---|---|---|---|---|---|
| 0.82 | 0.82 | 0.81 | 0.80 | 0.80 | 0.80 | 0.79 | 0.75 | 0.75 | 0.73 |

Table 12: *Fantasy : the words with the highest probabilities for a movie success*

Although it is rather hard to interpret explicitly the results, one can speculate about some topics' acceptance. E.g. for the *Sci-Fi* genre, the plot words *murder* and *crew* have low "success" probability.

It is interesting to see how the vocabulary of the plots has been changed over the considered time. To investigate it, the classification of a movies' "success/fail" was performed on all entries (not only top and bottom 25%), but separately for each decade. For the result visualization the word-cloud was chosen, in which the color and its intensity reflects "profitability" of a word, so the more intensity of the red color is the more probability to lose; the more intensity of the green color is the more probability to win. The font size corresponds to the word usage frequency: the bigger the font size is the frequenter the word. The comparison of the two decades for the *Comedy* genre can be seen on the Figures 5 and 6 (the most frequent 100 words are chosen). The interpretation of the charts is not simple, but, for example, one can get a conclusion, that there are more "criminal" or "black" comedies in the last 10 years and they have relative bad acceptance (the words *kidnap*, *dead*, *die*, *murder* have low "success" probability in the second decade chart and are missing in the first decade chart)

Finally the analysis was performed, to determine predictability of the future break evens. For this purpose the rolling historical window of 5 last years movies' data is used to predict the next year movies' break-even probabilities. As a predictive method the random forest [20] was chosen.

For the predictors budget, genres and plot words (without seldom words, sparse factor was defined as 0.95) with the binary weight and for the validation period from 2005-2014 calculated ROC value was about 0.72.

The same rolling historical window analysis for the pure plot text words as predictors (without budget and genres) gave ROC value slightly over 0.58 (Table 17).

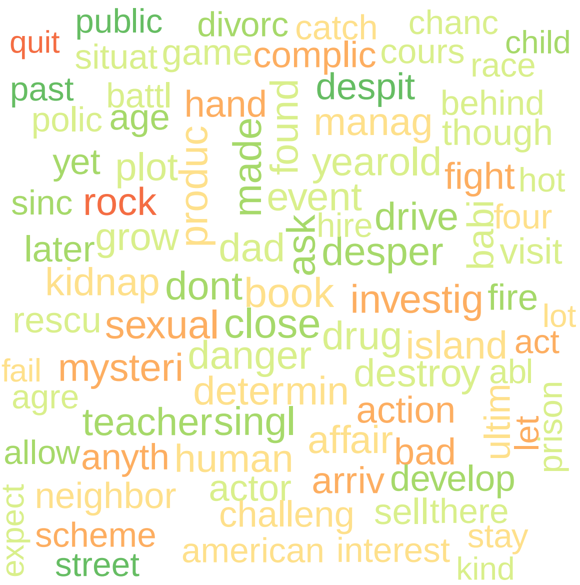

*Figure 5: Comedy, first decade (1995-2004)*

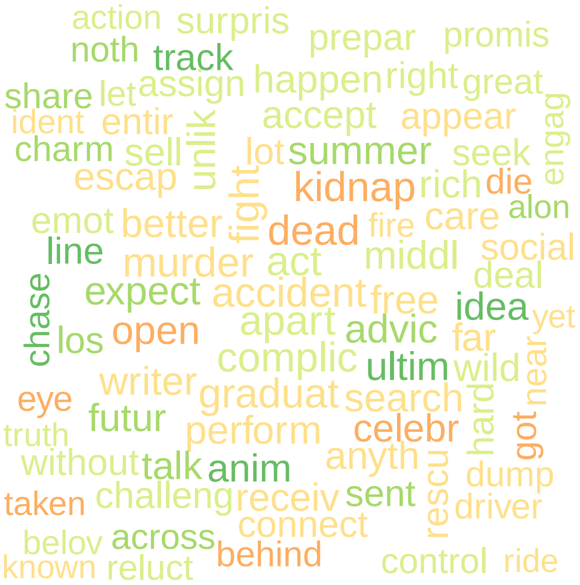

*Figure 6: Comedy, second decade (2005-2014)*

| whose | right | stori | hero | show | best | charact | unlik | york | young |
|---|---|---|---|---|---|---|---|---|---|
| 0.10 | 0.18 | 0.24 | 0.25 | 0.25 | 0.27 | 0.27 | 0.27 | 0.27 | 0.28 |

Table 13: *Fantasy : the words with the lowest probabilities for a movie success*

| person | place | true | begin | feel | look | away | old | parti | learn |
|---|---|---|---|---|---|---|---|---|---|
| 0.78 | 0.71 | 0.70 | 0.69 | 0.68 | 0.68 | 0.67 | 0.67 | 0.67 | 0.65 |

Table 14: *Comedy : the words with the highest probabilities for a movie success*

| comedi | hit | whose | film | hes | star | local | stori | group | young |
|---|---|---|---|---|---|---|---|---|---|
| 0.32 | 0.33 | 0.34 | 0.35 | 0.36 | 0.36 | 0.38 | 0.40 | 0.40 | 0.41 |

Table 15: *Comedy : the words with the lowest probabilities for a movie success*

| ROC | Sens | Spec | F1 |
|---|---|---|---|
| 0.72 | 0.55 | 0.75 | 0.73 |

Table 16: *Break-even probability prediction based on the previous 5-year historical window training. One-year ahead rolling calcualations for the target period 2005-2014*

| ROC | Sens | Spec | F1 |
|---|---|---|---|
| 0.59 | 0.33 | 0.76 | 0.73 |

Table 17: *Break-even probability prediction based on the previous 5-year historical window training. One-year ahead rolling calcualations for the target period 2005-2014. The prediction is based only on the plot text (bag-of-words).*

## Summary

The investigation showed that the budget is the strongest factor for the movies financial success. The general rule is: the bigger budget is the more chances for a movie to have a positive financial result. It can be explained that a big budget allows to get famous actors, directors as well as more publicity.

Some genres, such as *Horror* and *Romance*, were identified as "profitable", that is a chance to get a break even and an expected return are bigger than by other genres; some genres, such as *Drama* and *Western*, show generally worse box office performance in compare with the rest.

The analysis of plot text was performed on a "bag-of-words" with the help of Naive Bayes method for the break even (gross is more than budget) classification. For the better separation of the cases, the data subset, including the top and bottom 25 percentages of the returns, was chosen. It appeared, that the break-even predictability of the plot texts varies from genre to genre.

For the visualization of movie plots and their comparison over different time periods the word-cloud charts were used with the color-coded the conditional probabilities for a movie, containing the word in its plot text, to be profitable.

The prediction random forest model with the 5-years rolling historical window showed above 0.7 ROC performance over the last 10 target years in the database (2005-2014).

The plot texts alone have the slight predictability of the movies' profit chances too: the ROC of the similar rolling window models was about 0.58 over the same time period.

Further steps of analysis could include the building better predictive model for the future returns, investigation of possibility genre assignment, based on the plot texts, as well as detection the words, which have "positive" and "negative" trends by users' perception.

The similar analysis could be naturally used in other fields e.g. for the investigation the perception of item descriptions by target customer groups etc.

# Used Sources

[1] IMDb Data, http://www.imdb.com/interfaces

[2] IMDb Data Ftp Server , ftp://ftp.fu-berlin.de/pub/misc/movies/database/

[3] IMDb Conditions of Use , http://www.imdb.com/conditions

[4] Correlations between user voting data, budget, and box office for films in the Internet Movie Database, Max Wasserman, Satyam Mukherjee, Konner Scott, Xiao Han T. Zeng, Filippo, Radicchi and Luis A. N. Amaral; http://arxiv.org/pdf/1312.3986.pdf

[5] Ensemble of Generative and Discriminative Techniques for Sentiment Analysis Of Movie Reviews, Gregoire Mesnill, Tomas Mikolov & Marc'Aurelio Ranzato, Yoshua Bengio; http://arxiv.org/pdf/1412.5335.pdf

[6] Fast and accurate sentiment classification using an enhanced Naive Bayes model. Vivek Narayanan, Ishan Arora, Arjun Bhatia; http://arxiv.org/ftp/arxiv/papers/1305/1305.6143.pdf

[7] Mining gold from the Internet Movie Database, part 1: decoding user ratings http://blog.moertel.com/posts/2006-01-17-mining-gold-from-the-internet-movie-database-part-1.html

[8] Movie and Actors: Mapping the Internet Movie Database, http://nwb.cns.iu.edu/papers/2007-herr-movieact.pdf

[9] Comparing IMDB Network of Actors to Random Graph Models, Dan Kimball, Eric Herdzik http://www.cs.rpi.edu/~magdon/courses/casp/projects/KimballHerdzik.pdf

[10] Predicting Movie Success Based on IMDB Data, http://www.academia.edu/7763644/Predicting_Movie_Success_Based_on_IMDB_Data

[11] Visual Analytics for the Prediction of Movie Rating and Box Office Performance, http://bib.dbvis.de/uploadedFiles/elassady.pdf

[12] SQLite https://www.sqlite.org/

[13] IMDbPy http://imdbpy.sourceforge.net/

[14] Rainbow program, http://www.cs.cmu.edu/~mccallum/bow/rainbow/

[15] The R Project for Statistical Computing, https://www.r-project.org/

[16] RSQLite: SQLite Interface for R https://cran.r-project.org/web/packages/RSQLite/index.html

[17] Documentation of *caret* package, http://topepo.github.io/caret/index.html

[18] Documentation of *tm* package, https://cran.r-project.org/web/packages/tm/tm.pdf

[19] Documentation of *wordcloud* package, https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf

[20] Documentation of *randomForest* package, https://cran.r-project.org/web/packages/randomForest/index.html

[21] Documentation of *ROCR* package, https://rocr.bioinf.mpi-sb.mpg.de/

[22] Multilingual Information Retrieval: From Research To Practice, Carol Peters, Martin Braschler, Paul Clough, ISBN: 9783642230080