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

Novoseltsev M., under the guidance Dr. Braschler M. and Dr. Cieliebak M.

Introduction

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

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

The sentiment analysis is currently 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 on business value, namely the movie genre and movie plot on the 'book office' performance.

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

Problem description and previous work

Although there is a rather lot of statistical analysis works performed on IMDB data, such as Predicting movie ratings with IMDb data and R [4]. Mining gold from the Internet Movie Database, part 1: decoding user ratings [5] or "Movie and Actors: Mapping the Internet Movie Database" [6], they are mainly focused either on user ratings of the movies or social network analysis of the involved staff (actors, directors etc.).

On the other hand the prediction of movie success (e.g.Predicting Movie Success Based on IMDB Data) [7] explores in addition to the "standard" attributes, such as genres and budget, mainly the movie rating of the reviewers.

The work Visual Analytics for the Prediction of Movie Rating and Box Office Performance [8] uses machine learning methods for the prediction the user ratings and based on a lot of attributes, including the related tweets analysis.

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 movie plotS directly on the box-office performance as well as its evolution over time.

Data preparion

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 below.

```
"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
```

To reduce amount of data parsing-related work the Python-based tool IMDbPY was used. With the help of the tool, the data was converted and transferred into a SQLite database. The further reverse engineering shows that the database is build with the Entity-Value-Attribute design principle. As a result of this approach the all numerical values in the database such as budget and gross statistic are represented as a free text values e.g.:

```
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)
```

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.

The author has decided to concentrate on the movies, where USA as a country took participation. From these movies, the part was chosen, that has both budget and gross data. As a gross data is a multirow free-text, the following logic was used: with the help of regular expressions the data with the strings, containing "USD" was filtered, then numerical values were extracted and after that the maximum of all particular movie-related values was taken.

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, which have Boolean types (most of movies have more than one genre and this is a reason, why the genre are not represented as a single column).

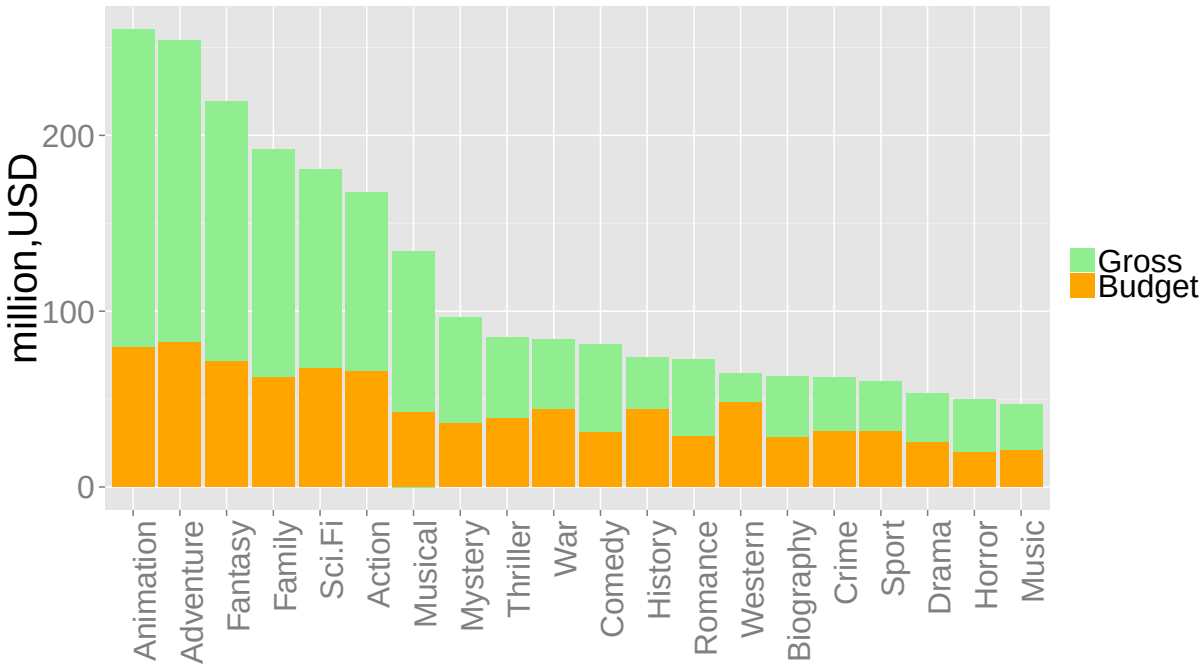
```
##  [1] "Action"    "Adventure" "Animation" "Biography" "Comedy"
##  [6] "Crime"     "Drama"     "Family"    "Fantasy"   "History"
## [11] "Horror"    "Music"     "Musical"   "Mystery"   "Romance"
## [16] "Sci.Fi"    "Sport"     "Thriller"  "War"       "Western"
```

Analysis

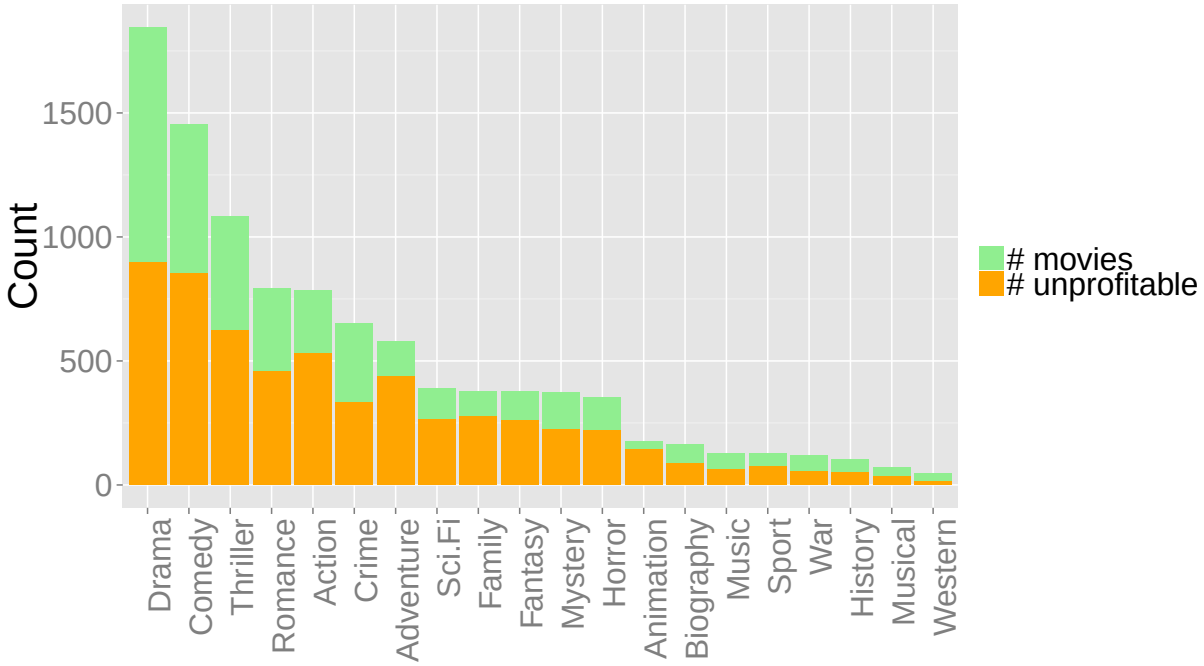
Data Exploration

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

On the figure below the mean budget and mean gross are plotted. It can be seen that the biggest budget have *Animation* and *Adventure* movies have the highest gross and budget. The mean of gross of genre *Western* is just slightly over the budget mean.



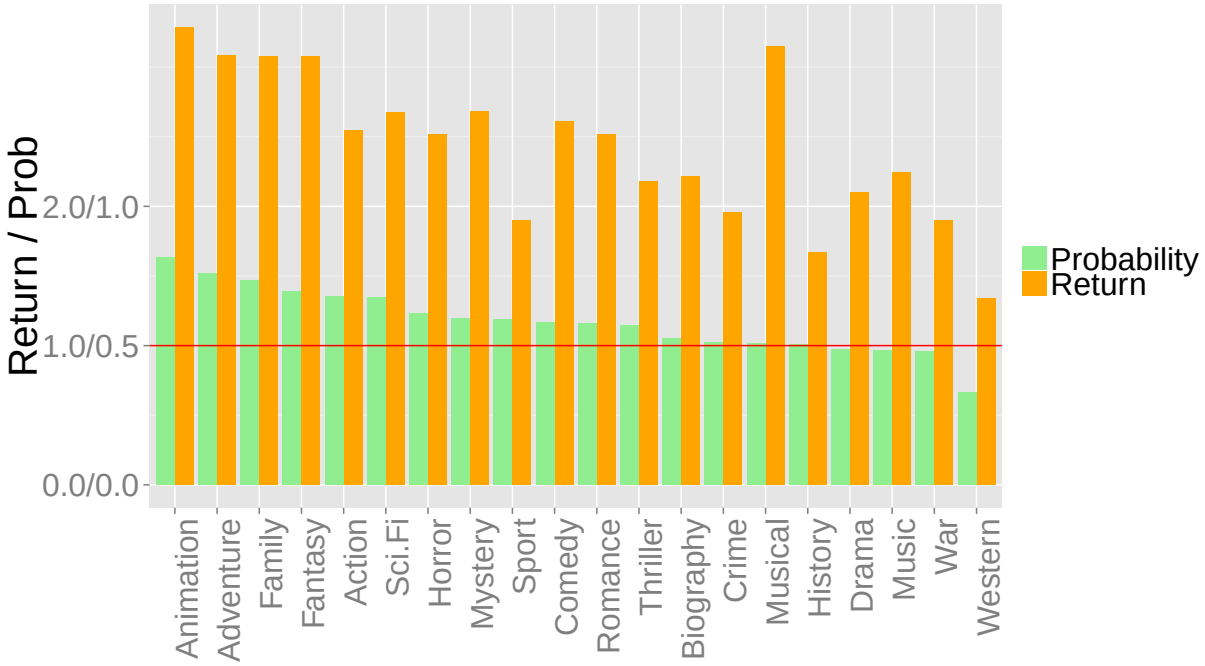
On the figure below the total number of movies and number of unprofitable (e.g which gross is less than budget) movies in each genre is considered.



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 big proportion of these genres is unprofitable, but a few movies collect a "good" box-office.

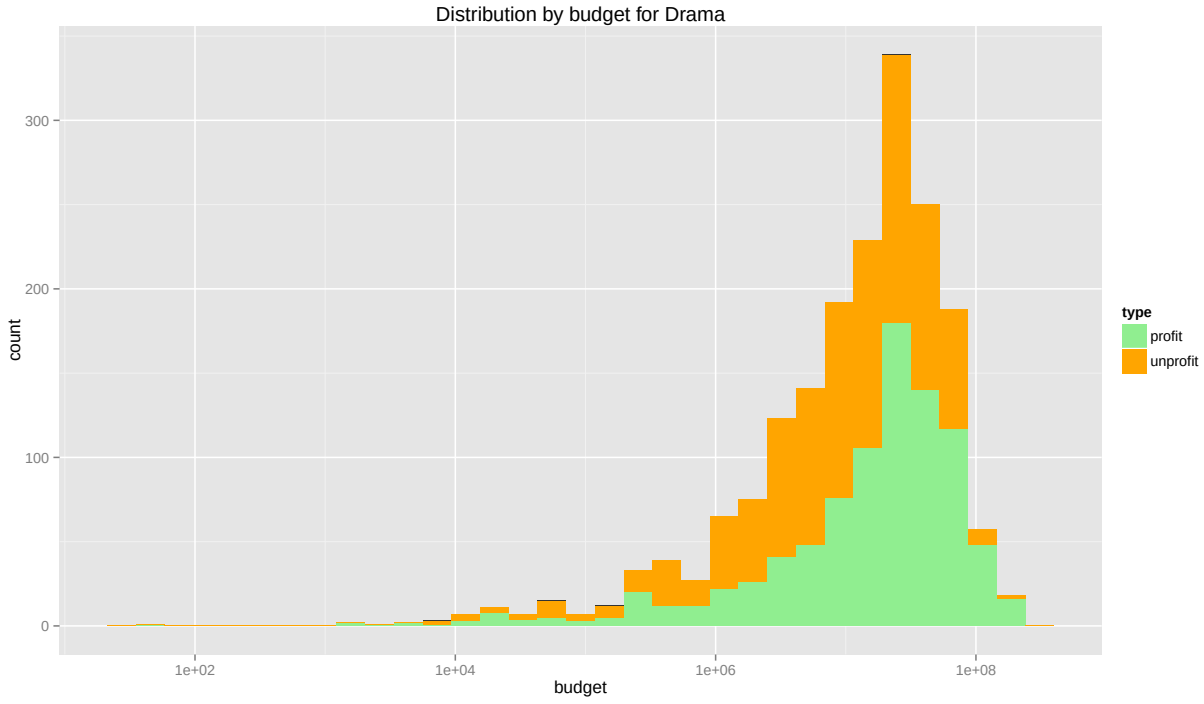
On the next chart there is attempt to bring the break-even and expected returns pro genre in one chart. One can

see that all genres profitable, but the probability to be profitable for some of genres (e.g. *Western*) are lower that



0.5.

On the figure below the budget distribution for *Drama* is shown with the ratio profitable/non-profitable movies. It can be seen that this ratio is bigger for the movies with the bigger budget.



Linear models

Initially the following simple logistic regression model is considered

```
##
## Call:
## glm(formula = I(gross > budget) ~ production_year + budget, family = "binomial",
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6269  -1.1062   0.5523   1.1169   1.3951
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.758e+01  1.294e+01  -2.131   0.0331 *
## production_year  1.357e-02  6.457e-03   2.102   0.0355 *
## budget           2.099e-08  1.286e-09  16.326   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5001.2  on 3653  degrees of freedom
## Residual deviance: 4608.2  on 3651  degrees of freedom
## AIC: 4614.2
##
## Number of Fisher Scoring iterations: 4
```

It can be seen that the break-even chance has strong significant dependency on a budget, that conforms with the observation of the charts.

For the investigation of other influence factors, logistic regression is further used. To analyse the impact of genres and its evolving over time, the next formula is proposed, in which the interaction if production year with the genres was taken into account:

```
## I(gross > budget) ~ production_year + budget + Action + Adventure +
##     Animation + Biography + Comedy + Crime + Drama + Family +
##     Fantasy + History + Horror + Music + Musical + Mystery +
##     Romance + Sci.Fi + Sport + Thriller + War + Western + budget:(Action +
##     Adventure + Animation + Biography + Comedy + Crime + Drama +
##     Family + Fantasy + History + Horror + Music + Musical + Mystery +
##     Romance + Sci.Fi + Sport + Thriller + War + Western) + production_year:(Action +
##     Adventure + Animation + Biography + Comedy + Crime + Drama +
##     Family + Fantasy + History + Horror + Music + Musical + Mystery +
##     Romance + Sci.Fi + Sport + Thriller + War + Western)
```

So in addition to the production year and genres as independent variables, the interaction between production year and each of the genres was taken.

After applying the AIC step-wise algorithm for elimination non-important predictors, the following model is returned:

```
##
```

```
## Call:
## glm(formula = I(gross > budget) ~ budget + Drama + Horror + Romance +
##     Western, family = "binomial", data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6189  -1.0910   0.5446   1.0724   1.9067
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.504e-01  7.331e-02  -3.416 0.000635 ***
## budget       2.084e-08  1.318e-09  15.813  < 2e-16 ***
## DramaTRUE   -3.730e-01  7.457e-02  -5.001 5.69e-07 ***
## HorrorTRUE   3.983e-01  1.239e-01   3.214 0.001309 **
## RomanceTRUE  2.649e-01  8.644e-02   3.064 0.002182 **
## WesternTRUE -1.391e+00  3.600e-01  -3.863 0.000112 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5001.2  on 3653  degrees of freedom
## Residual deviance: 4544.2  on 3648  degrees of freedom
## AIC: 4556.2
##
## Number of Fisher Scoring iterations: 4
```

Generally, the bigger is budget 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 contrary to the genres Drama and Western, which have a higher risk to be unprofitable.

Another metric that could be interested by a potential investor is a expected return, that defined here as a ratio of gross to budget. As a distribution of returns is right-skewed, the log-transformation is applied

```
##
## Call:
## lm(formula = I(log(gross/budget)) ~ budget + Action + Adventure +
##     Drama + Family + Horror + Mystery + Romance + Thriller +
##     Western + budget:Action + budget:Adventure + budget:Family,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3405  -0.8196   0.2304   1.1524   9.3843
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -7.533e-01  7.626e-02  -9.878  < 2e-16 ***
## budget                2.024e-08  1.518e-09  13.329  < 2e-16 ***
## ActionTRUE            2.769e-01  1.354e-01   2.046 0.040863 *
## AdventureTRUE         3.994e-01  1.643e-01   2.431 0.015093 *
## DramaTRUE            -3.075e-01  7.120e-02  -4.319 1.61e-05 ***
## FamilyTRUE            5.397e-01  1.744e-01   3.095 0.001985 **
## HorrorTRUE            4.640e-01  1.190e-01   3.899 9.85e-05 ***
```

```
## MysteryTRUE           3.690e-01  1.127e-01   3.275 0.001067 **
## RomanceTRUE           3.415e-01  8.219e-02   4.155 3.33e-05 ***
## ThrillerTRUE         -2.671e-01  8.281e-02  -3.226 0.001268 **
## WesternTRUE          -9.670e-01  2.907e-01  -3.326 0.000889 ***
## budget:ActionTRUE    -6.870e-09  2.062e-09  -3.331 0.000874 ***
## budget:AdventureTRUE -6.207e-09  2.096e-09  -2.961 0.003084 **
## budget:FamilyTRUE    -7.515e-09  2.420e-09  -3.105 0.001916 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.929 on 3640 degrees of freedom
## Multiple R-squared:  0.1059, Adjusted R-squared:  0.1027
## F-statistic: 33.17 on 13 and 3640 DF,  p-value: < 2.2e-16
```

The result mainly conforms with the break even analysis. For a potential investment the genres *Romance* and *Horror* are recommended (they have higher probability of break-even as well as above average return), the investment in the genres *Drama* and *Western* are rather risky.

Plot text analysis

The plot text contains some information about film content. The detail level is very different: the number of words is varying from 13 to 567.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    13.0    66.0    97.0   106.4   127.0   567.0
```

For the text analysis the author used R with the package *tm* (plus package caret for cross validation) as well as Rainbow program. The latter is a powerful command-line text analysis toolkit.

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

The experimenting with Rainbow with Roccio and Naive Bayes classification methods showed that the both have approximately the same performance. The usage 2-word-grams 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 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" 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. For the predictors word presence flags was chosen, which gave better performance in compare with tf-idf weighting. The 2-word grams did not improve the performance. For each of 10 genres with the most number of movies in the database the Area Under Curve (AUC or, sometimes, ROC) was calculated, based on 10 cross-validations

	ROC	Sens	Spec	F1
Fantasy	0.72	0.72	0.48	0.58
Sci.Fi	0.68	0.7	0.55	0.62
Family	0.65	0.68	0.48	0.57

	ROC	Sens	Spec	F1
Comedy	0.64	0.68	0.55	0.61
Thriller	0.62	0.65	0.53	0.59
Romance	0.6	0.57	0.5	0.53
Adventure	0.59	0.55	0.52	0.54
Action	0.59	0.62	0.49	0.55
Drama	0.57	0.6	0.49	0.54
Crime	0.53	0.62	0.47	0.53

The AUC value (Area Under Curve, here it is named as ROC) is not spectacular high. Fro the word analysis, the ratio of the probability of presence a particular word given that a film is "successful" to probability of the word presence, given that film is unprofitable or P(word | success) / (P(word | fail) + P(word | success)) was taken. The model has the table with these values, e.g.

```
##          var
## grouping     exists not.exists
##   fail    0.1702128  0.8297872
##   success 0.1276596  0.8723404
```

This logic will apply for the genres, which is 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 have the word) were chosen.

Sci.Fi

```
## [1] "Sci.Fi : the highest probability words for success"
##      head    teenag  meanwhil       day      give     place    strang
## 0.9166667 0.9000000 0.8666667 0.8333333 0.8333333 0.8333333 0.8333333
##      plan     anoth      armi
## 0.8235294 0.8181818 0.8000000
## [1] "Sci.Fi : the lowest probability words for success"
##    murder      crew     woman     young      leav       men    someth
## 0.1666667 0.3000000 0.3000000 0.3181818 0.3333333 0.3333333 0.3333333
##     creat    behind      left
## 0.3333333 0.3571429 0.3636364
```

One can interpret the results, that the themes with *murder*,*crew* and *women* are not well accepted.

Fantasy

```
## [1] "Fantasy : the highest probability words for success"
##      armi    happen     begin      hous      know     never      lord
## 0.8181818 0.8181818 0.8095238 0.8000000 0.8000000 0.8000000 0.7857143
##      even     place    follow
## 0.7500000 0.7500000 0.7333333
## [1] "Fantasy : the lowest probability words for success"
##     whose     right     stori      hero      show      best   charact
## 0.1000000 0.1818182 0.2380952 0.2500000 0.2500000 0.2727273 0.2727273
##     unlik      york     young
## 0.2727273 0.2727273 0.2800000
```

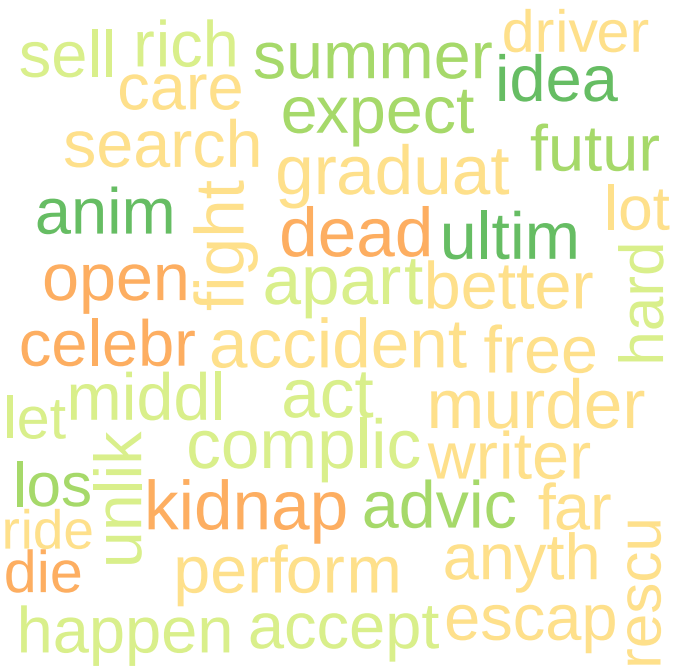
Comedy

```
## [1] "Comedy : the highest probability words for success"
##    person     place      true     begin      feel      look      away
## 0.7750000 0.7111111 0.6976744 0.6865672 0.6829268 0.6805556 0.6744186
##       old     parti     learn
## 0.6666667 0.6666667 0.6491228
## [1] "Comedy : the lowest probability words for success"
##    comedi       hit     whose      film       hes      star     local
## 0.3170732 0.3333333 0.3421053 0.3478261 0.3604651 0.3617021 0.3829787
##     stori     group     young
## 0.4029851 0.4042553 0.4090909
```

For *Comedy* and *Fantasy* the result interpretation is not so simple.

It is interesting to show how the vocabulary of the plots has been changed over time. To investigate it the classification of the decade was used with the similar procedure as for the success/fail plot investigation, but with all entries (not only top and bottom quantiles). For the result visualization the word-cloud was chosen, in which the color and its intensity reflects "profitability" of a word, so the bigger font size the frequenter word; the more intensity of the red color the more probability to lose; the more intensity of the green color the more probability to win. The comparison of the two decades for the *Comedy* genre can be seen below (the most frequent 100 words are chosen).

Comedy, first decade (1995-2004).



Comedy second decade (2005-2014).

Finally the analysis how the future break-evens can be predicted is performed. For this purpose the rolling historical window of 5 last years movies' data is used to predict the next year break-even probabilities. As a predictive method the random forest was chosen.

For the predictors budget, genres and plot words (without seldom words, sparse factor is 0.95) with binary weight and for the validation period from 2005-2014 a ROC value is about 0.7.

```
##         ROC     Sens      Spec      ROCSD    SensSD     SpecSD
## 1 0.7220644 0.550657 0.7513566 0.06598615 0.1138821 0.03597218
```

Summary

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

Some genres as *Horror* and *Romance* were identified as "profitable", that is a chance to be profitable is more than by other genres, some genre, such as *Drama* and *Western* show generally worse book-office performance in compare with the rest.

The analysis of plot text was performed on a "bag-of-words" model with the help of Naive Bayes method for the prediction, whether a movie will be profitable. For the better separation of the cases, the data subset, including the top and bottop 25 percentiles of the "profits", was choosen. It appeared that the break-even predictability of the plot texts varies from genre to genre.

For the visualisation of the movie plots the wordcloud chart was used with the color-coded of the conditional probabilities for a movie, containing the word, to be profitable.

The prediction model with the 5-years sliding historocal window show about 0.7 ROC performance over last 10 years

Further steps of analysis could include the building better predictive model for the future returns, investigation of possibility the automatical genre assignment, based on the plot texts.

The similar analysis could be naturally used in other fields e.g. for the invetigation 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] Predicting movie ratings with IMDb data and R (https://rulesofreason.wordpress.com/2014/03/02/predicting-movie-ratings-with-imdb-data-and-r/)

[5] 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)

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

[7] Predicting Movie Success Based on IMDB Data (http://www.academia.edu/7763644/Predicting_Movie_Success_Based_on_IMDB_Data)

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

[9] Documentation of *caret* Package (http://topepo.github.io/caret/index.html)

[10] Documentation of *tm* Package (https://cran.r-project.org/web/packages/tm/tm.pdf)

[11] Documentation of *wordcloud* Package (https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf)

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