

IMDB Movies Analysis

Social Networks Analysis

**Bokos Evangelos**

**Giannakidi Anna-Maria**

**Koffa Vassiliki**

**Kostopoulos Stavros**

**Plakia Maria**

Table of Contents

[1. Business Purpose & Description 2](#_Toc491638038)

[*Α. Some statistics for the Film Industry about 2016* 3](#_Toc491638039)

[B. Film industry Trends 3](#_Toc491638040)

[C. From the Film Industry to the SNA 4](#_Toc491638041)

[2. Mission 4](#_Toc491638042)

[3. Data Management 5](#_Toc491638043)

[A. Data Set 5](#_Toc491638044)

[B. ETL 5](#_Toc491638045)

[4. Methodology 7](#_Toc491638046)

[A. Logistic Regression 7](#_Toc491638047)

[B. Multilayer Perceptron (MLP 2 layers) 8](#_Toc491638048)

[C. Multilayer Perceptron (MLP 3 layers) 10](#_Toc491638049)

[D1. Text Analytics (Movie Title) 12](#_Toc491638050)

[D2. Text Analytics (Key Words) 13](#_Toc491638051)

[5. Results 15](#_Toc491638052)

[6. Members & Roles 17](#_Toc491638053)

[7.TimePlan 18](#_Toc491638054)

[8. Bibliography 19](#_Toc491638055)

**Table of Figures**

**Figure 1- Global Box Office (US$ Billions)** 4

**Figure 2- Correlation Matrix of the different genres** 7

**Figure 3- Correlation Matrix of the Numerical Variables** 8

**Figure 4- Accuracy Table of the Logistic Regression Model** 9

**Figure 5- Accuracy Plot of the Logistic Regression** 9

**Figure 6-Cross Entropy Loss per Epoch (MLP 2)** 10

**Figure 7- Test & Train Accuracy (MLP 2)** 10

**Figure 8- Final Weights MLP 2 Layers Histogram** 11

**Figure 9- Cross Entropy Loss per Epoch (MLP 3)** 11

**Figure 10- Train & Test Accuracy (MLP 3)** 12

**Figure 11-Final Weights MLP 3 Layers Histogram** 12

**Figure 12-Histogram of the # of Words in Texts** 13

**Figure 13- Avg Training Accuracy over the past 5o iterations** 14

**Figure 14- Avg Test Accuracy over the past 50 iterations** 14

**Figure 15-Histogram of the # of Words in Texts** 15

**Figure 16-Avg Training Accurace over the past 50 Iterations** 15

**Figure 17-Avg Test Accuracy over the past 50 Iterations** 16

# **1. Business Purpose & Description**

## ***Α. Some statistics for the Film Industry about 2016******[[1]](#footnote-2)***

It's not a lie that film industry is one of the top industries worldwide. Distribution and production companies spend millions of money in order their movies to become "Top box office" or "Blockbusters".

In 2016, the global box office for all films released worldwide enlarged 1% from 2015, reaching the incredible number of $38,6 billion. Only the US/Canada market climbed to $11,4 billions. Apart from the movies released, cinema screens also increased -in number- by 8% to nearly 164,000 screens.

Furthermore, the number of frequent moviegoers (people who go to the cinema at least once a month) increased and in 2016 it represents the 48% of all tickets sold in US/Canada, and at the same time more than the 71% of this population (246 millions of viewers) went to the cinema at least once in 2016.

**Figure** 1**- Global Box Office (US$ Billions)**

Source: Theatrical Market Statistics 2016-MPAA

## B. Film industry Trends

As shown in the figure 1 above, film industry both internationally and in US/Canada is constantly evolving and growing. [[2]](#footnote-3)One of the most drastic trend from 2016 is the smart participation of online services and platforms (Netflix, Youtube, Amazon) in the production and distribution of movies and TV series. From low-budget films to huge TV series and movies, use these platforms for distribution, promotion and even production. And how this is interesting for our project? Budget. The budget of these films automatically declines or is spent on other resources. What kind of budget (quality and quantity) possibly makes a blockbuster?

Another interesting tendency is the increased social commentary. Many films –from horror (Get Out) to Dramas (Moonlight) focus on relevant social issues, including diversity acceptance and celebration, highlighting political unrest, and fostering unity across populations.

Finally[[3]](#footnote-4), a trend that seems serious with available data to be explored and analyzed, is the high box office of the R-rated comedies. Amy Schumer emerged as a bona fide star with “Trainwreck,” but most films hoping to ride raunch to box office gold derailed. “Vacation,” “Ted 2” and “[The Night Before](http://variety.com/t/the-night-before/)” left audiences cold, and even well-reviewed “Spy” fell short of previous Melissa McCarthy efforts such as “The Heat” and “Identity Thief.”

## C. From the Film Industry to the SNA

As it became clear in the previous sections, production companies are very interested in knowing what the future brings for their business. Thus, a project like this can accurately provide them information about topics and issues related to the success of their movies.

So, we will try to determine which factors influence the success of a movie, in other words we want to be able to detect the crucial factors that lead a movie to be a blockbuster.

# **2. Mission**

We have in our disposal a dataset of 5.000 movies and we want to predict which features determine the success of a movie. This is a tricky issue due to the vast set of variables that we have inherited from the dataset of IMDB movies in kaggle. So, we are going to use the gross profit given in the dataset and conclude to the variables that put the maximum weight on it and influence it the most. The procedure of predicting how well a movie will perform in theaters is extremely difficult, and often relies on variables such as critic reviews or box office revenues. However, these variables are only known after a movie is released - is it possible to predict whether or not a movie will be good without even watching it? What are the qualities that make a movie good or bad?

The prediction of a movie’s success can be modeled as a classification problem to decide whether a movie should be considered a success or not based on either ROI or profit. The accuracy of a predictive model depends a lot on the extraction and engineering of features (a.k.a., independent variables). When it comes to studying movie success, three types of features have been explored: audience-based, release-based, and movie-based features. Finally, a first approach could be to examine prediction of movie grosses based on information available prior to release.

# **3. Data Management**

## Data Set[[4]](#footnote-5)

We have selected the IMDB 5000 Movie Dataset which contains 28 different variables. This data set has already been processed via a scrapping algorithm so as a column could be added to the dataset. The extra column has to do with ***face recognition*** on the posters of each movie. The idea is to analyze these variables and to develop a neural network that will be able to predict the possible success of a movie. The success will be determined through the IMDB final score and maybe additionally the gross of the movie. We will test the results by the fitting on the data and we will compare the results to other known models. Our final goal is to develop the best model possible for these data.

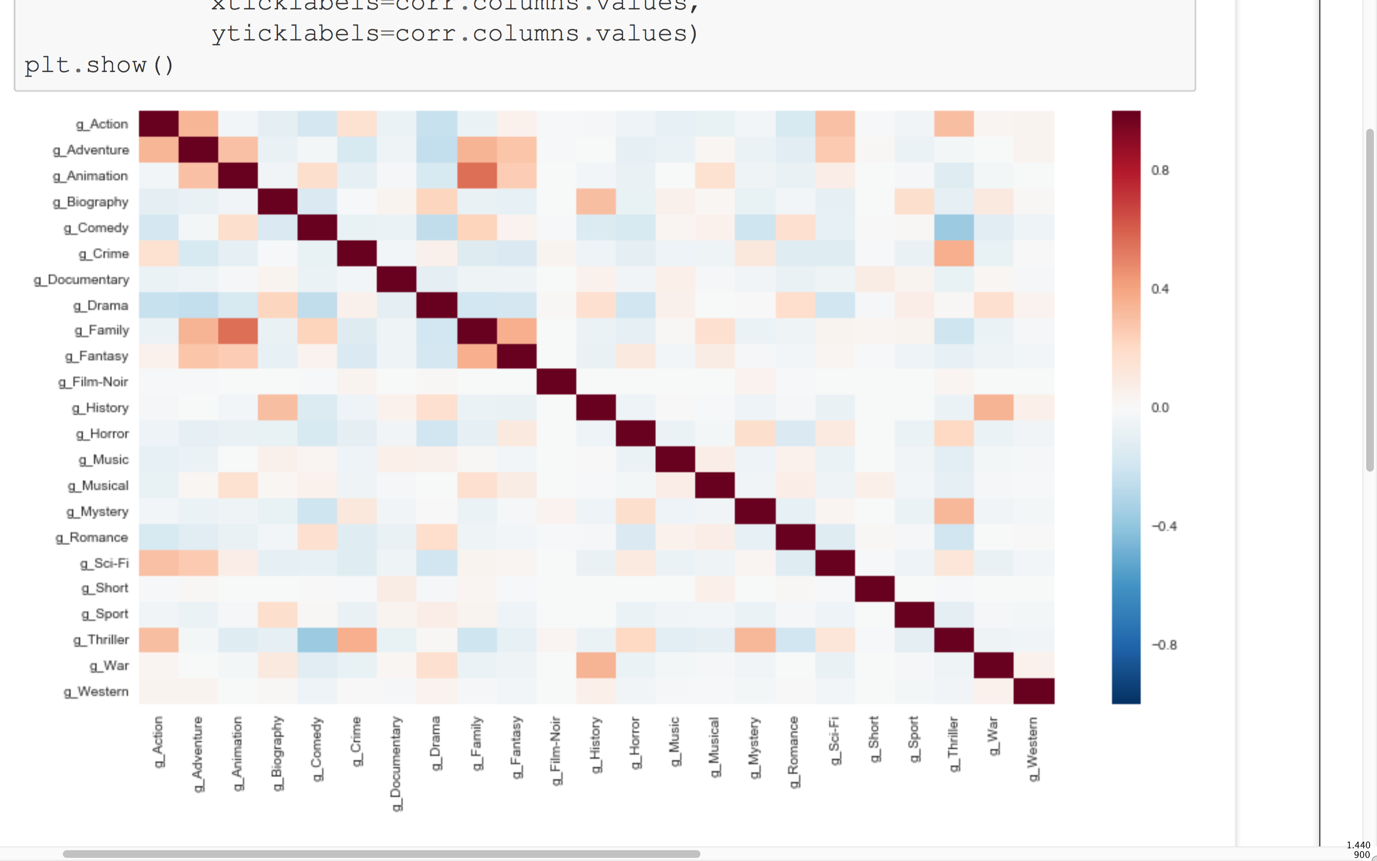
**Table** 1**-Variables**

|  |  |
| --- | --- |
| **Variables** | |
| Movie Title  Color or BW  Director Name  Number of Critics for Reviews  Duration (minutes)  Leading Actor (Name)  Supporting Actor 1 (Name)  Supporting Actor 2 (Name)  Facebook Likes (Director)  Facebook Likes (L. Actor)  Facebook Likes (S. Actor)  Facebook Likes (Total Cast) | Gross  Budget  Genres  Number of voted Users  Facenumber in Poster  Plot Keywords  Movie imdb link  Number of userd for reviews  Language  Country  Content Rating  Aspect Ratio |

## ETL

The first step of the whole process is the "Extract-Transform- Load". The ETL was the same for all methods (which will be analyzed in the Chapter "Methods"), except for some minor differences that will be mentioned below.

After importing the dataset, the first move was to overview the data, by counting the null rows. It was essential to drop the rows with N/As' from the variables "Gross" and "Budget" since those two categories will be used to construct the dependent variable y, (in the Logistic Regression model).Next, dummies need to be created in the "Genre" variable, Then a closer look at the "Genre" variable needs to be taken, because it will be the target variable in the next part where text analytics will be performed. In order to use the genre in our logistic model we create dummies variables, that correspond to 1 and 0, depending if the movie has the specific genre. Note that a movie may belong to more than one genres, which creates some problems in the prediction part, so we need to see if the different genres are related to some point.

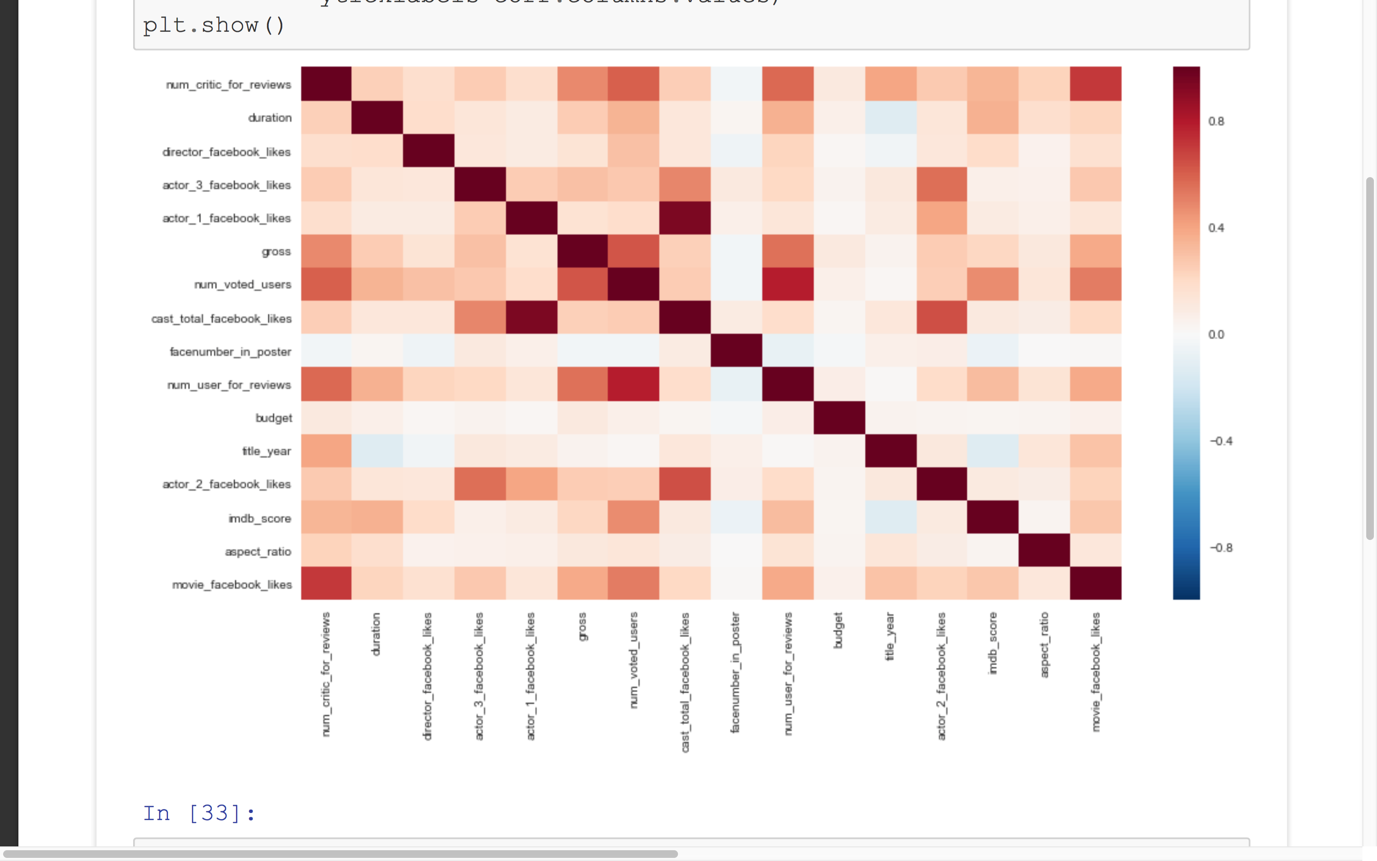
**Figure** 2**- Correlation Matrix of the different genres**

As it is clear from Figure 2, there is not a strong relation between the genres of a movie, with some exceptions: Animation and Family movies seem to have a medium to strong positive relation and Thrillers with Comedies give a mediocre negative relation. Now, it’s time to create the target variable: Is the movie a blockbuster or not? A movie is considered to be a blockbuster if it's earnings are much higher than it's budget (about 70% and more). Thus, we calculate a new variable named "Gross Margin". Gross margin is the total gross of a movie divided by the sum of the total gross and the total budget of a movie. We will use the gross margin variable in order to create a new variable called "Blockbuster". Blockbuster takes the value 0 if the gross margin is below 70% and value equal to 1 if the gross margin is equal or above 70%. After that we create a correlation plot in order to remove highly correlated variables from our model. We also remove gross margin, budget and gross since those variables were used to create the depended variable of the logistic regression. The variables that are excluded from the model- as they are highly correlated in Figure 3- are:

Furthermore, "Gross" and "Budget" are dropped, since "Blockbuster" contains the information of these two columns.

|  |
| --- |
| **Country, Gross, Budget, Leading Actor, Supporting Actor 1 & 2, Number of users for reviews and Facebook likes of the total Cast** |

Finally, we use Stepwise Regression to determine which variables will be in our final model.[[5]](#footnote-6) It's important to clarify that other variables, which contain information with no business value, will remain to the model, because they technically give value to the model (f.e. aspect ratio).

**Figure** 3**- Correlation Matrix of the Numerical Variables**

# 4. Methodology

Now, it's time to move on to the main models. We are going to use 3 methods that later will be compared to each other: Logistic Regression, Multilayer Perceptron with 2 layers (MLP 2) and MLP 3. Additionally, text analytics will be performed, but for a different reason.

## Logistic Regression

Some ETL will be continued in this part, because it belongs only to this method.

After the Stepwise Regression and the Variable selection, we move to the Sets of the model.

We split our data set into two parts. The first is the **Training Set** and the second is the **Testing Set**. The Training set is 80% of the original data set and the Test set is the remaining set. The train set is selected with a random function in order to have a balanced sample.

Right after, tensorflow will be used in order to convert the categorical variables to integers and develop the logistic regression model. The model of the logistic regression is:

|  |
| --- |
| **y = xw + b** |

Where, y: The movie is a blockbuster (or not),

x : the independent variables,

w: weights that will be assigned at each of the independent variables,

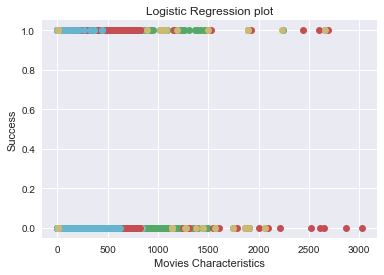
b: the independent constant vector

We would have to use epochs if the model needed to learn something, but since there is no learning need in the logistic regression, no learning curve and vectors will be used.

Therefore, we run the model first with all the train variables and then for all the test set, but because the model doesn't learn anything it doesn't improve and thus both train and test sets produce similar accuracy. The outcomes of the model, that show its accuracy and strength are:

**Figure** 4**- Accuracy Table of the Logistic Regression Model**

|  |  |
| --- | --- |
| **Loss** | 0.693147 |
| **Train Accuracy** | 0.769181 |
| **Test Accuracy** | 0.737113 |

**Figure** 5**- Accuracy Plot of the Logistic Regression**

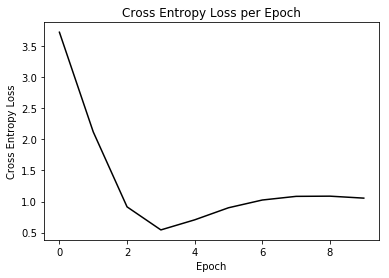
Figures 4 and 5 show the accuracy of the model. Figure 4 informs us that train and test sets give exactly the same accuracy (0,77), which is fine-not perfect-but gives some trustworthiness to the model. Loss, on the other hand is a bit high, but still lower than the accuracy.

In table 5 we see the success of the coefficients. From this plot we can evaluate our model in total. It is not the best however we can determine some patterns among the characteristics of the movies.

## Multilayer Perceptron (MLP 2 layers)

The ETL process is the same as the logistic regression, from the importing to the training & test data. The difference here we have to put layers to our model. The parameters are set into similar variables.

We divide the train and the test set into batches and we have epochs. In those we take each batch and we train the MLP, hoping that the model through the training process will learn and therefore adjust better to the test set. The learning process takes place with the layers we have put in and hidden in between these layers we have weights.

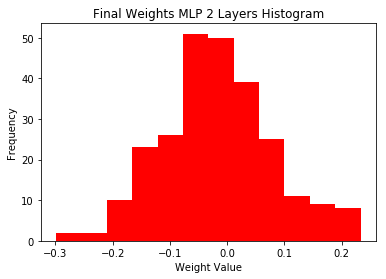
**Figure** 6**-Cross Entropy Loss per Epoch (MLP 2)**

**Figure** 7**- Test & Train Accuracy (MLP 2)**

|  |
| --- |
|  |

Figures 6 & 7 are the same as the table in Figure 4, only now we have plots because our model evolves over time (epochs). With the 2 layer MLP we can detect some significant differences. Cross Entropy Loss declines until it reaches a lowest number (0,54) about the 4th epoch and then it steadily rises. Simultaneously, even though test accuracy remains still in the 0,23 throughout all epochs, train accuracy rises in the second epoch until the fourth, where it remains steady in 0,78.

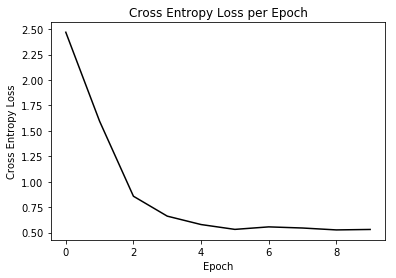
After that we plot the weights to see which ones have the biggest and smallest values and therefore which produce the most impact to our model. So, we plot the frequency of the weights and we observe that most of them are close to zero meaning that they have little impact to our model. None of the weights has absolute value above 0.3

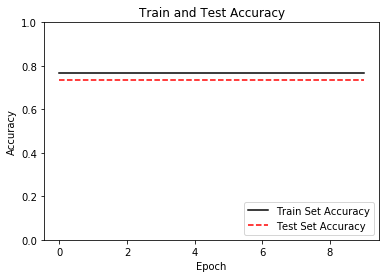
**Figure** 8**- Final Weights MLP 2 Layers Histogram**

## Multilayer Perceptron (MLP 3 layers)

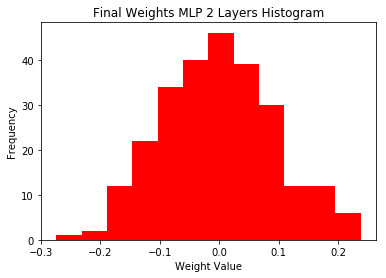
Once again the ETL process is exactly the same as before, but now the layers from two become three.

With three layers however the results do not improve. The cross entropy loss declines with each iteration. However, the train and test accuracy remain impressively steady and close in all iterations, close to 0,8.

**Figure** 9**- Cross Entropy Loss per Epoch (MLP 3)**

**Figure** 10**- Train & Test Accuracy (MLP 3)**

**Figure** 11**-Final Weights MLP 3 Layers Histogram**



## D1. Text Analytics (Movie Title)

Text Analytics will be used for a different purpose, so it won't be compared to the other methods. The idea is to be able to predict the genre of a movie from the title, thus the only variables needed are "Movie Title", "Genre" and "Plot Keywords".

In these 3 variables, the previous ETL process will be performed –removing the nulls values (actually the remaining set now is much bigger) and setting the "Genre" as the target variable. Therefore, we will use multinomial regression with Y representing the different genres. We set as text the movie titles and now we need to normalize the text, put all words in lowercase and remove punctuations, numbers and whitespace. Therefore, we will use multinomial regression with Y representing the different genres. We will create a histogram of the average length of each title and then we will choose the max text word length at 25.

**Figure** 12**-Histogram of the # of Words in Texts**

|  |
| --- |
|  |

From Figure 12 it's clear that there is a positive skewness, the mass of the average length of the titles is concentrated on lower numbers (3 and 4 words).

Next, we will create a train and test set, with 80 and 20 percent ratio as before and then we will develop the multinomial regression.

|  |  |
| --- | --- |
| **Figure** 13**- Avg Training Accuracy over the past 5o iterations**  **Figure** 14**- Avg Test Accuracy over the past 50 iterations** | Training Accuracy Test Accuracy |

From the Figures 13-14 we see that neither the train nor the test accuracy improve over the iterations.

## D2. Text Analytics (Key Words)

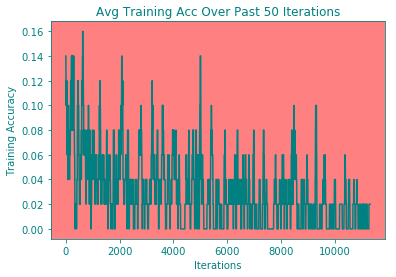
Now, another text analytics will be performed but now we will try to determine the genre of a movie from some plot key words. The ETL process is pretty much the same, but we keep genre and plot key words. Now the target is the different genres, multinomial regression for each genre (y is the genres), but text is the plot key words (not the movie titles).

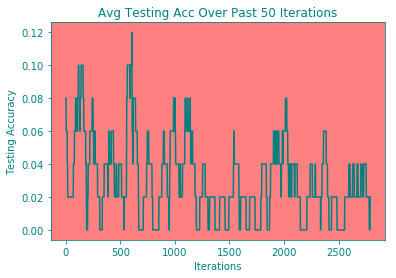
**Figure** 15**-Histogram of the # of Words in Texts**

|  |
| --- |
|  |

Once again, from Figure 15 it's clear that there is a positive skewness, the mass of the average length of the titles is concentrated on lower numbers (2,5 to 3 words).

The results of the test and training sets are very poor in both sets:

**Figure** 16**-Avg Training Accurace over the past 50 Iterations**

**Figure** 17**-Avg Test Accuracy over the past 50 Iterations**

# 5. Results

Time for results. The main results are obvious from the previous chapter, but now we are going to compare the models to each other and end up to the most accurate one.

**Table** 2**- Aggregated Results (1)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Logistic Regression** | **MLP 2 Layers** | **MLP 3 Layers** |
| **Loss**  **Training Acc.**  **Test Acc.** | 0.69  0.77  0.77 |  |  |
| Weights | - |  |  |

According to the loss and accuracy results, clearly the MLP 3 Layers is the best.

**Table** 3**- Aggregated Results (2)**

|  |  |  |
| --- | --- | --- |
| **Text Analytics**  **Accuracy** | **Movie Title** | **Plot Keywords** |
| **Training Set** |  |  |
| **Test Set** |  |  |

In both cases the results and accuracy are very poor. Especially, Text Analytics with Movie Titles was highly time consuming (1-2 hours to run the code in movie titles, 10 minutes for plot keywords).

Apparently, the problem with such small accuracy may be that a movie can have more than one genre and this may cause some confusion to the analysis. Also the data set was very small to lead to accurate results.

# 6. Members & Roles

**Table** 4**- Group Members**

|  |  |
| --- | --- |
| **Name** | **ID** |
| Bokos Evangelos  Giannakidi Anna-Maria  Koffa Vassiliki  Kostopoulos Stavros  Plakia Maria | BAFT1601  BAFT1605  BAFT1614  BAFT1612  BAFT1602 |

**Members Roles and Profiles:**

We have divided the needs for this assignment into three major categories:

**1) Developers and Software engineering**

**Koffa Vassiliki:** Graduate student of Department of Informatics at University of Piraeus with specialization in Software Engineering and Intelligent Systems, granted from the department. Worked as web developer for three operations.

**Plakia Maria:** Graduate student of Department of Management Science and Technology at Athens University of Economics and Business, with specialization in Quantitative methods in Finance. Worked as product manager at an e-Super Market and as Front engineer for a cloud-monitoring site.

**2) Statistical Analysis and Models Implementation**

**Bokos Vaggelis:** Graduate student of Department of Mathematics at National University of Athens. Direction in Applied Mathematics, with specialization in Statistics and Operation Research through Markov Chains and Stochastic methods, granted from the department. Worked as web developer for two operations and, from now on, he is Java Developer in Intracom.

**Giannakidi Anna-Maria:** Graduate student of Department of Mathematics at National University of Athens. Direction in Applied Mathematics, with specialization in Statistics and Operation Research through Markov Chains and Stochastic methods, granted from the department. Worked at National Bank and as a private mathematics tutor.

**3) Business Development and Project Manager**

**Kostopoulos Stavros:** Graduate student of Department of Business Administration. Participant on events of TEDxUniverisy of Piraues. Worked at Sales department in a shipping company and currently is a filmmaker and producer.

# 7.TimePlan

# 8. Bibliography

Brent Lang: Variety.com, "5 Trends Making the Movie Business Lose Sleep", 2015

Theatrical Market Statistics 2016-MPAA

The Los Angeles Film School: 5 emerging trends in the film industry, 31.05.2017

<https://arxiv.org/ftp/arxiv/papers/1301/1301.7363.pdf>

<http://www.sciencedirect.com/science/article/pii/S095741740800496X>

<https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>

<https://www.biz.uiowa.edu/faculty/kangzhao/pub/JMIS_2016.pdf>

<http://www.cvast.tuwien.ac.at/sites/default/files/bakkarbeit/omenitsch.pdf>

<http://cs229.stanford.edu/proj2013/EricsonGrodman-APredictorForMovieSuccess.pdf>

<http://pages.stern.nyu.edu/~jsimonof/movies/movies.pdf>

1. Theatrical Market Statistics 2016-MPAA [↑](#footnote-ref-2)
2. The Los Angeles Film School: 5 emerging trends in the film industry, 31.05.2017 [↑](#footnote-ref-3)
3. Brent Lang: Variety.com, "5 Trends Making the Movie Business Lose Sleep", 2015 [↑](#footnote-ref-4)
4. <https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset> [↑](#footnote-ref-5)
5. NOTE: We set seed important every time we run the code to get the same results [↑](#footnote-ref-6)