VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY
Department of Computer Engineering

Project Report on

Prediction of Box Office Movie Success

In partial fulfillment of the Fourth Year, Bachelor of Engineering (B.E.) Degree in Computer Engineering at the University of Mumbai Academic Year 2015-2016

Submitted by
Shital Chaudhari(15)
Akash Indani(24)
Shreya Sherugar(66)
Akshay Wagh(76)

Project Mentor
Mrs. Manisha Gahirwal

(2015-16)

VIVEKANAND EDUCATION SOCIETY'S INSTITUTE OF TECHNOLOGY

Department of Computer Engineering

Certificate

This is to certify that ***Akash Indani*** of Fourth Year Computer Engineering studying under the University of Mumbai have satisfactorily completed the project on "***Prediction of Box Office Movie Success***" as a part of their coursework of PROJECT-II for Semester-VIII under the guidance of their mentor ***Prof.Manisha Gahirwal*** in the year 2015-2016 .

This project report entitled ***Prediction of Box Office Movie Success*** by ***Akshay wagh, Akash Indani, Shreya Sherugar, Shital Chaudhari*** is approved for the degree of ***Computer Engineering***.

Programme Outcomes	Grade
PO1,PO2,PO3,PO4,PO5,PO6,PO7, PO8, PO9, PO10, PO11, PO12	

Date :

Project Guide: Internal and External

Project Report Approval

For
B. E (Computer Engineering)

This thesis/dissertation/project report entitled ***Prediction of Box Office Movie Success*** by ***Akshay wagh, Akash Indani, Shreya Sherugar, Shital Chaudhari*** is approved for the degree of ***Computer Engineering.***

Internal Examiner

External Examiner

Head of the Department

Principal

Date:
Place:

Declaration

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

-- (Signature)	-- (Signature)
Akshay Wagh (76)	Akash Indani (24)
-- (Signature)	-- (Signature)
Shreya Sherugar (66)	Shital Chaudhari (15)

Date:

Acknowledgement

We are thankful to our college Vivekanand Education Society's Institute of Technology for considering our project and extending help at all stages needed during our work of collecting information regarding the project.

It gives us immense pleasure to express our deep and sincere gratitude to Assistant Professor **Mrs. Manisha Gahirwal** (Project Guide) for her kind help and valuable advice during the development of project synopsis and for her guidance and suggestions.

We are deeply indebted to Head of the Computer Department **Dr.(Mrs.)Nupur Giri Ma'am** and our Principal **Dr. (Mrs.) J.M. Nair Ma'am**, for giving us this valuable opportunity to do this project.

We express our hearty thanks to them for their assistance without which it would have been difficult in finishing this project synopsis and project review successfully.

We convey our deep sense of gratitude to all teaching and non-teaching staff for their constant encouragement, support and selfless help throughout the project work. It is great pleasure to acknowledge the help and suggestion, which we received from the Department of Computer Engineering. We wish to express our profound thanks to all those who helped us in gathering information about the project. Our families too have provided moral support and encouragement at several times.

Abstract

A **box office** is a place where tickets are sold to the public for admission to an event. *Box office* business can be measured in terms of the number of tickets sold or the amount of money raised by ticket sale. Movie industry is a highly dynamic industry. The uncertainty involved due to the involvement of various factors in determining the box office success makes it even more ambiguous. The prediction and analysis of these earnings is very important for the creative industries and often a source of interest for fans. So when a consumer decides to go see a movie or wants to put up shows in theatre, he or she should want to get as much information about the movie beforehand as possible. In order to obtain the best information about the quality of a movie before a movie is actually released, consumers must get proper information and statistics about the movie . For this reason we have built a model to predict the box office return of movies based on various factors like release time, budget, presence of actors, directors, genre ,critic ratings etc.Nowadays social media has become popular platform for displaying response of people. So we have also considered data(likes,comments) from social networking sites like Twitter and Youtube.In the proposed system, we are focusing on applying data mining techinques to predict the revenue of upcoming movies .

Table Of Contents

Chapter 1

Introduction

1.1 Motivation

Entertainment industry has become one of the highest income profession in todays world. People are spending a lot of money in making of the money, also as the rate of tickets have increased, people are spending a lot on tickets as well. Hence, the audience should know if they are spending on the correct movie. Also, the managers of the theatres should know how much to expect out of a movie, so that they run their business efficiently. This motivated us to make a system which would be helpful for both the consumers.

1.2 Problem Definition

An application which collects data from social websites for a movie and predicts the revenue that will be generated. Data collection is the first task in which we extract data from various API's like Twitter, facebook, youtube and also by using different tools like Facepager.

The data collected is not in proper manner for preprocessing, as it contains many redundant tokens, for example: "The movie has excellent story, must watch" in this case only the word Excellent has meaning which can be used for preprocessing rest all words are to be eliminated i.e. only particular words were required for preprocessing. Sentiment analysis tools like Rapid Miner is beneficial for removing the redundancy and also provides a polarity for distinguishing each comment as positive, negative or neutral.

Methodologies used are K-Means Clustering, Sentiment Analysis and Artificial neural network which uses suitable algorithm and attributes like release time, budget, presence of actors, directors, genre and comments from various social media are considered as contribution factor for predicting the revenue .

The predicted revenue is provided to the user and error in the prediction is calculated using the Root Mean Square Error formula.

1.3 RELEVANCE OF THE PROJECT:

Existing problem: The problem now a day faced by the managers of the theatre is that they do not know how many screens to allot for a particular movie. They are usually clueless about the revenue a movie might collect and hence cannot lay a proper strategy. Also, the people cannot decide if a movie is good enough to go and watch it in the theatre. If the movie is bad, then they end up spending a lot of money in the tickets.

Rectifying the problem: The project will be helpful in many ways. As the output of this project, we will be providing the predicted revenue of a movie. With the help of this information, the managers of theatres can strategize efficiently and hence can allot the screens accordingly. Also, as the factors like IMDB rating. Critics' rating, comments from social media, presence of actors/directors, etc are considered while making the system, people can rely on it for an accurate result. With the help of the revenue predicted by our system, the people can decide whether they want to go and watch the movie in the theatre or no, and hence can use their money proficiently.

1.4 METHODOLOGIES USED

Methodology is the systematic, theoretical analysis of the methods applied to a field of study. It comprises the theoretical analysis of the body of methods and principles associated with a branch of knowledge, In our case, the branch of study is Data Mining.

The various methodologies used by us are:

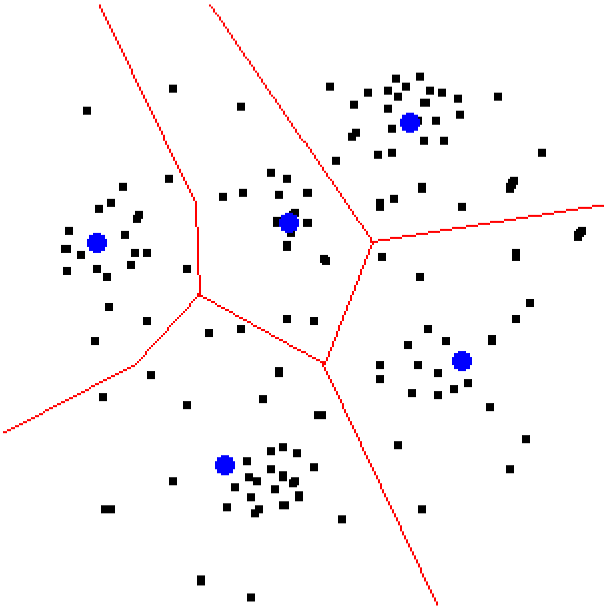
- K-Means Clustering
- Sentiment Analysis
- Artificial Neural Network

K-Means Clustering

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

k-means clustering is a method of classifying/grouping items into k groups (where k is the number of pre-chosen groups). The grouping is done by minimizing the sum of squared distances (Euclidean distances) between items and the corresponding centroid.

A centroid is "the center of mass of a geometric object of uniform density", though here, we'll consider mean vectors as centroids.



A clustered scatter plot. The black dots are data points. The red lines illustrate the partitions created by the k-means algorithm. The blue dots represent the centroids which define the partitions.

Algorithmic steps for k-means clustering

Let X = $\{x_1,x_2,x_3,\ldots\ldots,x_n\}$ be the set of data points and V = $\{v_1,v_2,\ldots\ldots,v_c\}$ be the set of centers.

1) Randomly select *'c'* cluster centers.

2) Calculate the distance between each data point and cluster centers.

3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers..

4) Recalculate the new cluster center using:

$$v_i = (1/c_i)\sum_{j=1}^{c_i} x_i$$

where, *'c_i'* represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.

6) If no data point was reassigned then stop, otherwise repeat from step 3).

SENTIMENT ANALYSIS:

Sentiment analysis (also known as **opinion mining**) refers to the use of natural language processing, text analysis and computational linguistics to identify and extract subjective information in source materials. Sentiment Analysis is a branch wherein the sentiments of the user are analyzed from the written scripts or recorded voice. If positive sentiments are reflected in the sentence then the sentiment analyzer marks the sentence as positive. Similarly in the case of evidence of negative statements in marks it as negative. There may also be a scenario wherein the sentence might neither reflect positive nor negative emotions. Such sentences are termed as neutral by the analyzer. The polarity determines the ratio of the positive, negative and neutral tweets. It reflects the emotions that the user wishes to convey through the comments regarding the given movie. The hype created results in number of comments which in turn allows us to segregate and break the comments into positive and negative tokens. There are various tools available in order to carry out sentiment analysis such ad Rapid Miner, Lingpipe, Sentiword etc. Sentiment analysis tools like Rapid Miner is beneficial for removing the redundancy and also provides a polarity for distinguishing each comment as positive, negative or neutral.

ARTIFICIAL NEURAL NETWORK:

ANN is generally described as system of interconnected neurons which exchange some data between them. Those neurons have a weight as a numeric value which can be tuned according to the previous experience. This makes the neural system adaptive to the inputs and make it capable of learning.

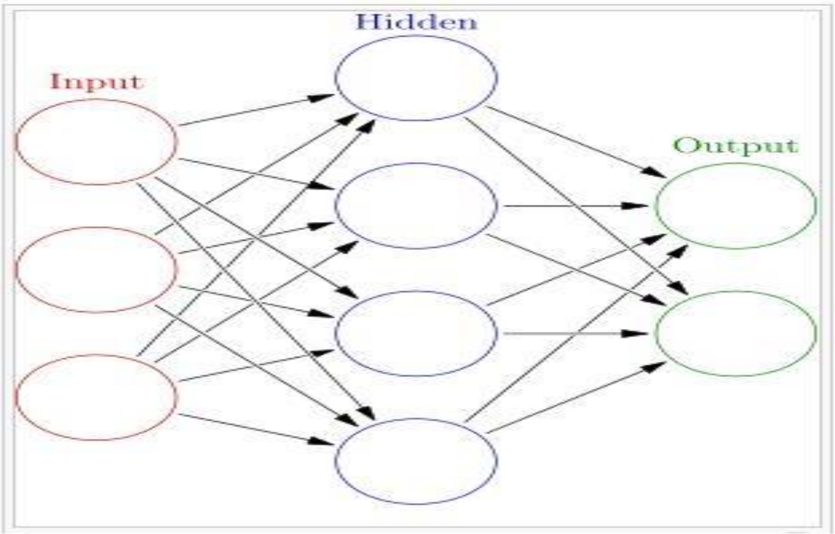


Figure 2 ANN

Neurals which is the class of statistical model possesses the characteristics such as adaptive weights and capability of approximating the non-linear functions.

ANN is of two types: 1) Feedforward NN

2) Radial basis function NN

3) Learning vector quantization NN

4) Modular NN

In our project, the type of ANN that will be used is Feedforward neural network.

The feedforward neural network was the first and arguably most simple type of artificial neural network devised. In this network the information moves in only one direction — forward: From the input nodes data goes through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network. Feedforward networks can be constructed from different types of units, e.g. binary McCulloch-Pitts neurons, the simplest example being the perception. Continuous neurons, frequently with sigmoidal activation, are used in the context of back propagation of error.

R-LANGUAGE:

R is a programming language and software environment for statistical computing and graphics supported by the R Foundation for Statistical Computing. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. Polls, surveys of data miners, and studies of scholarly literature databases show that R's popularity has increased substantially in recent years.

R is an implementation of the S programming language combined with lexical scoping semantics inspired by Scheme. S was created by John Chambers while at Bell Labs. There are some important differences, but much of the code written for S runs unaltered.

R and its libraries implement a wide variety of statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions and extensions, and the R community is noted for its active contributions in terms of packages. Many of R's standard functions are written in R itself, which makes it easy for users to follow the algorithmic choices made.

For computationally intensive tasks, C, C++, and Fortran code can be linked and called at run time. Advanced users can write C, C++, Java .NET or Python code to manipulate R objects directly. R is highly extensible through the use of user-submitted packages for specific functions or specific areas of study. Due to its S heritage, R has stronger object-oriented programming facilities than most statistical computing languages. Extending R is also eased by its lexical scoping rules.

Another strength of R is static graphics, which can produce publication-quality graphs, including mathematical symbols. Dynamic and interactive graphics are available through additional packages

Chapter 2

Literature Survey

2.1 Papers or Books

Nagamma P+ [1] applied machine learning technique to gather data from IMDB(http://www.imdb.com/). Collected set of words were examined as positive and negative and processed straightforward. He selected a set of keywords(love, amazing, great, fantastic, etc). The clustering algorithm used was K-means and DB-SCAN. Then he performed Sentiment analysis in which text preprocessing and text transformation was performed. But the above process gave less accuracy, so to improve the accuracy fuzzy clustering was used. Auto regression model was used for sales prediction of current day using the previous day collection.

Sameer[2] hype analysis, in this paper regression method is used to for predicting the box-office success. The data was collected using Twitter API on hourly basis. The paper focuses on multiple linear regression model in which more than one explanatory variable is used to predict the one variable. It is a time bound model.

They used models like Three tier architecture, Multiple linear regression model, Critical period, Contributing factors. The three tier architecture used the presentation layer which contained user requirements. User authentication was required for using these model. To increase the accuracy different contributing factors considered was increasing the accuracy are attention seeking of audience, star- cast, category and holiday effect.

Yonsei[3] predicted the Ticket sales for movies using reviews for Korean movies. This paper was an implemented paper and actual results were found about the ticket sale. They collected there data sets from Korean website(http://www.kobis.or.kr) which was under Korean movie council.

They purely used Hadoop language of big data analytics. A scale was made for direction, music, acting, Player_value and through which the sale was predicted. They processed there way in an systematic way, The Online review and its star rating collected was first Hadoop based and on the other side they collected the Box office data and using Search volume scale was preprocessed . The preprocessed result was then considered with external factors and further was passed into the prediction model consisting of Linear model, Support vector model, Artificial neural network (which was built using R language).

The paper gave an clear approach between various prediction model and also helped in gathering data from various sites.

Chapter 3

Requirements

3.1 Functional Requirements

- Application automatically collects the data from different social websites.
- The data is stored in a structured way in database so that it can be easily accessible.
- The data is first preprocessed so that the redundant data is deleted from the database.
- K-Means clustering is used for clustering of data.
- Using sentiment analysis, polarity of the user comment is calculated that whether it is positive, negative or neutral.
- Using the computed Sentiment analysis revenue of the movie is estimated.
- Along with sentiment analysis, various other factors are considered while predicting the revenue.

3.2 Non-functional Requirements

- The data stored can be encrypted locally to protect from any local copy or theft of device.
- Validating the data before inserting it into the centralized repository (database).
- Data dictionary for Sentiment analysis.

3.3 Constraints

The constraints of our project would be: Only the success of hollywood movies will be determined. Also, only the revenue of the movie will be forecasted.

3.4 Hardware and Software Requirements

Software Requirements:

Front end: HTML, CSS, Javascript.

Backend: R studio

3.5 System Block Diagram

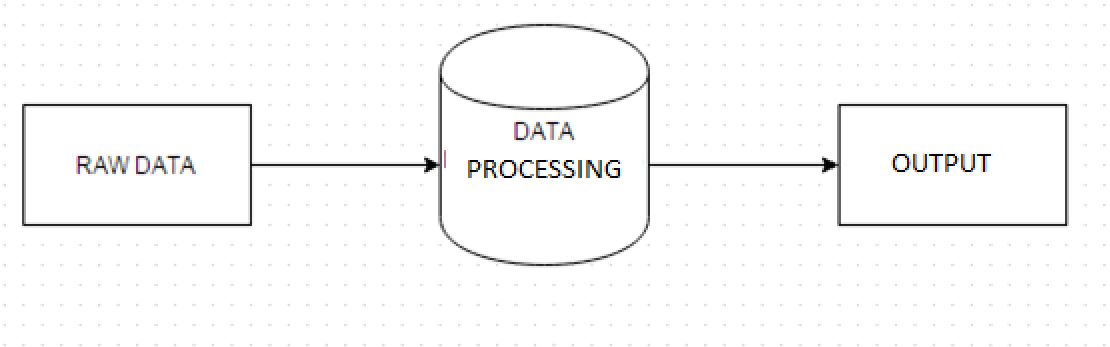
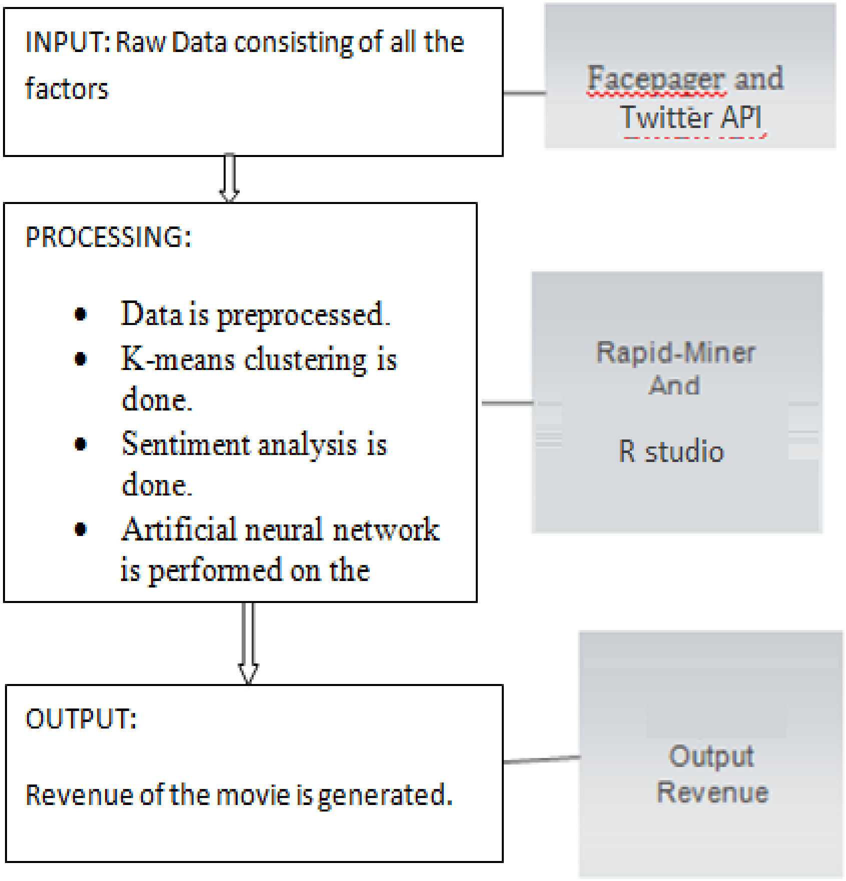


Figure 1.System block diagram

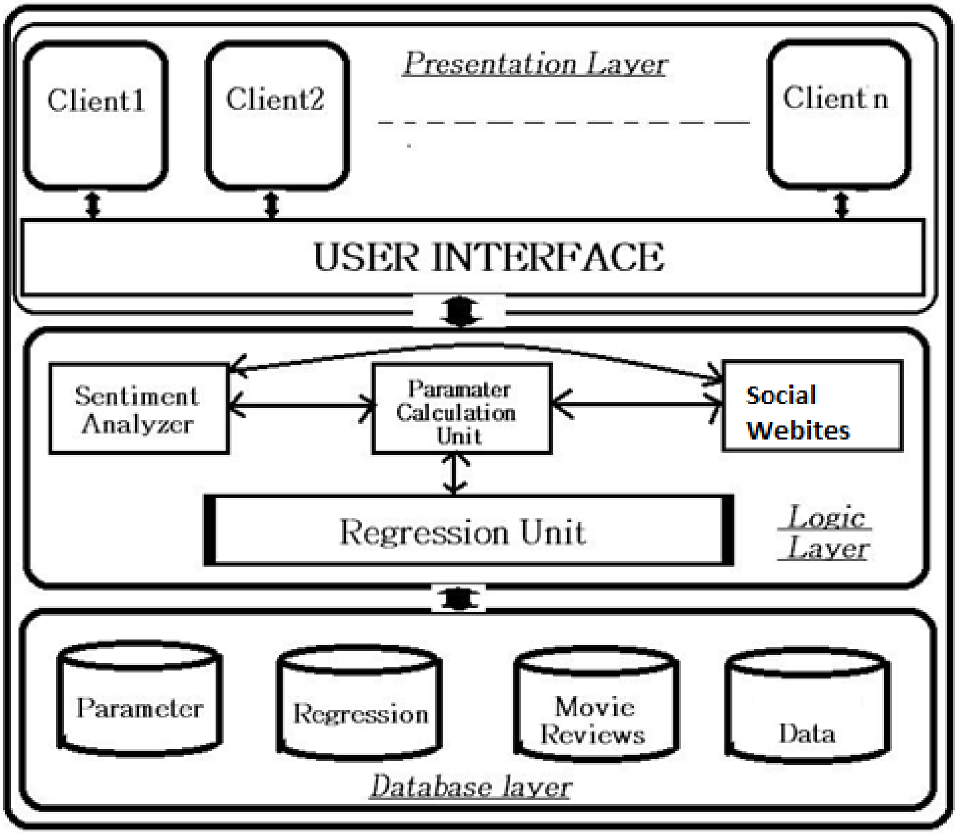

INPUT: Raw Data consisting of all the factors
Facepager and Twitter API
PROCESSING:
Data is preprocessed.
K-means clustering is done.
Sentiment analysis is done.
Artificial neural network is performed on the
Rapid-Miner And R studio
OUTPUT:
Revenue of the movie is generated.
Output Revenue

Figure

Chapter 4

Proposed Design

4.1 System Design



SYSTEM ARCHITECTURE

Figure 2

4.2 Detailed Design

Usecase Diagram:

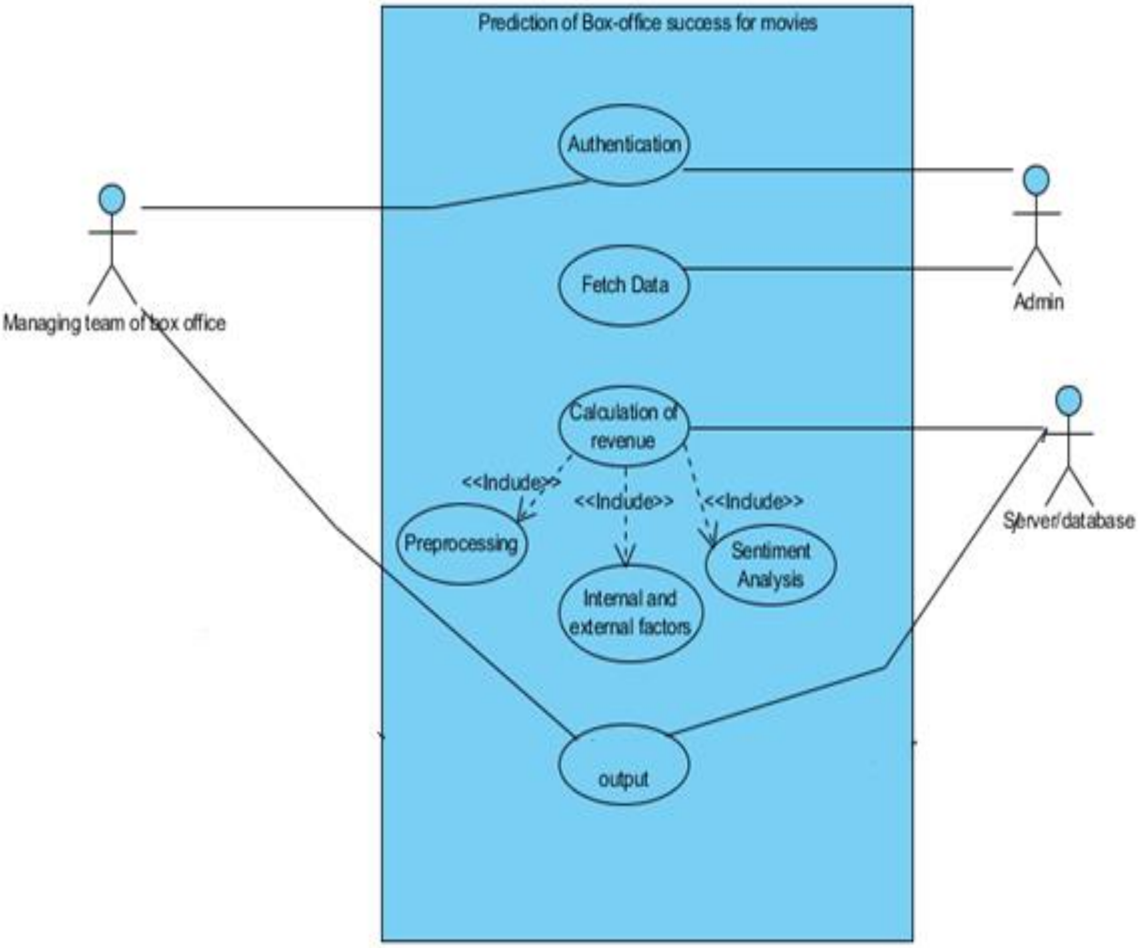


Figure 3

Sequence Diagram:

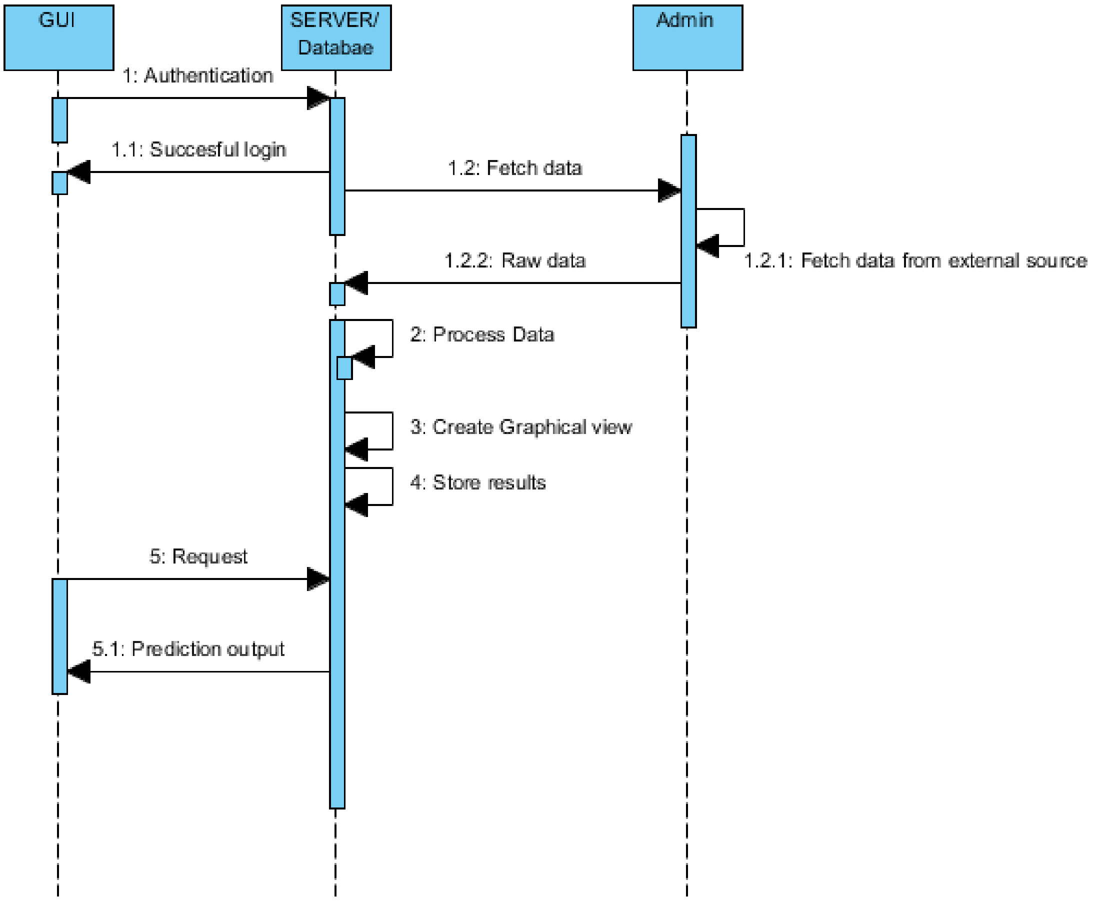


Figure 4

DFD LEVEL 0

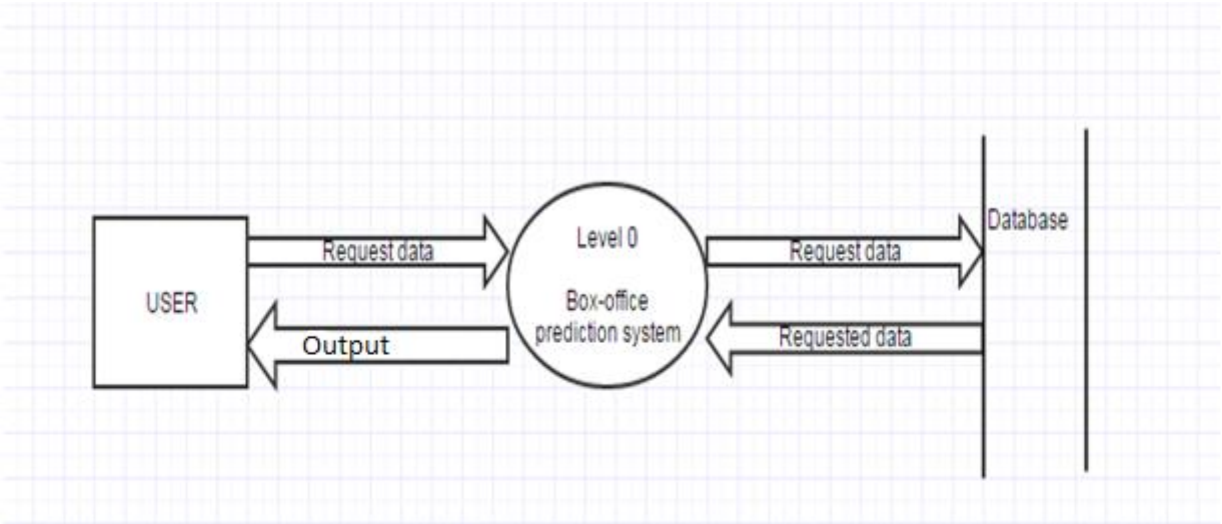


Figure 5. DFD level 0

DFD LEVEL 1:

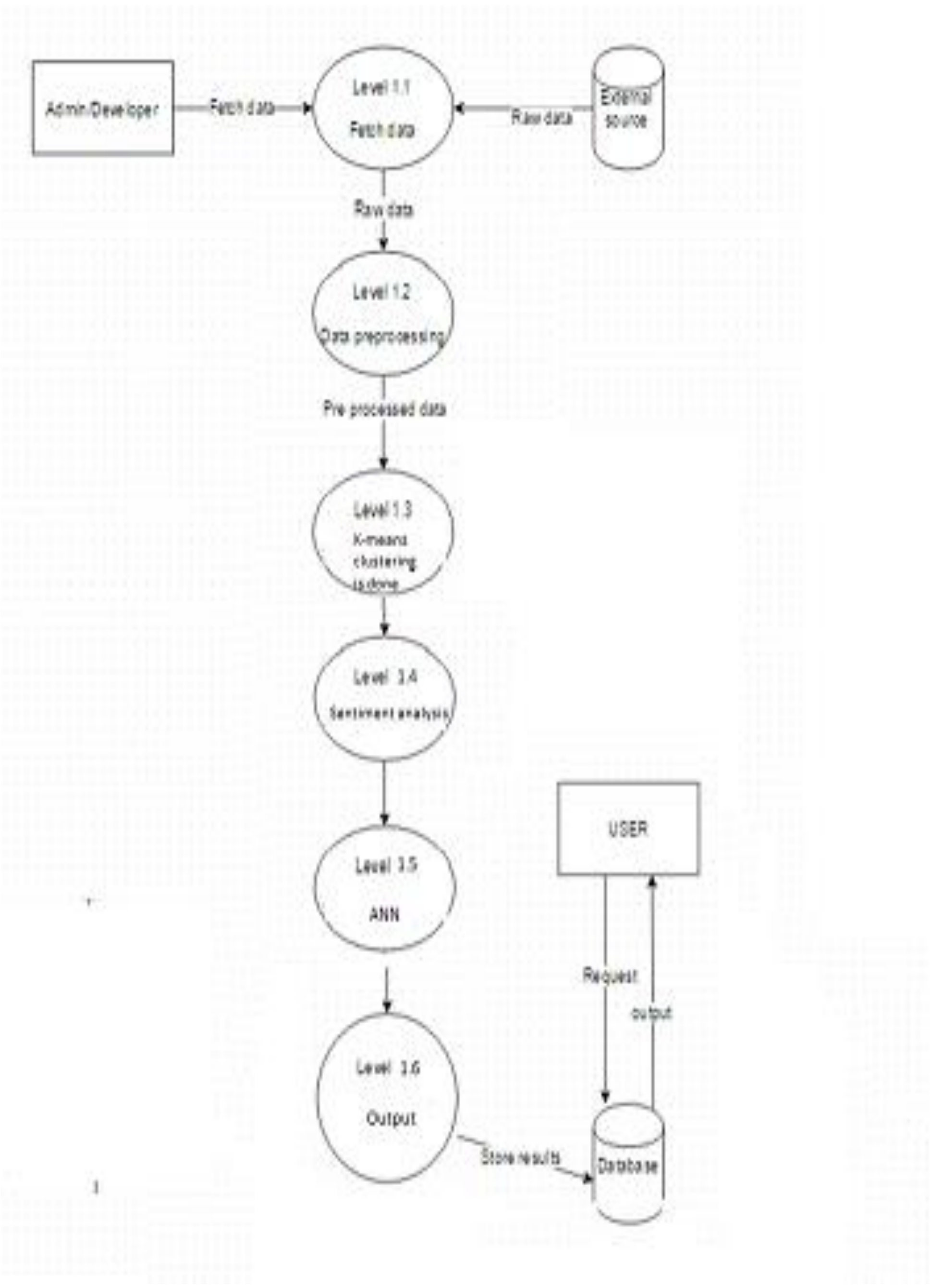
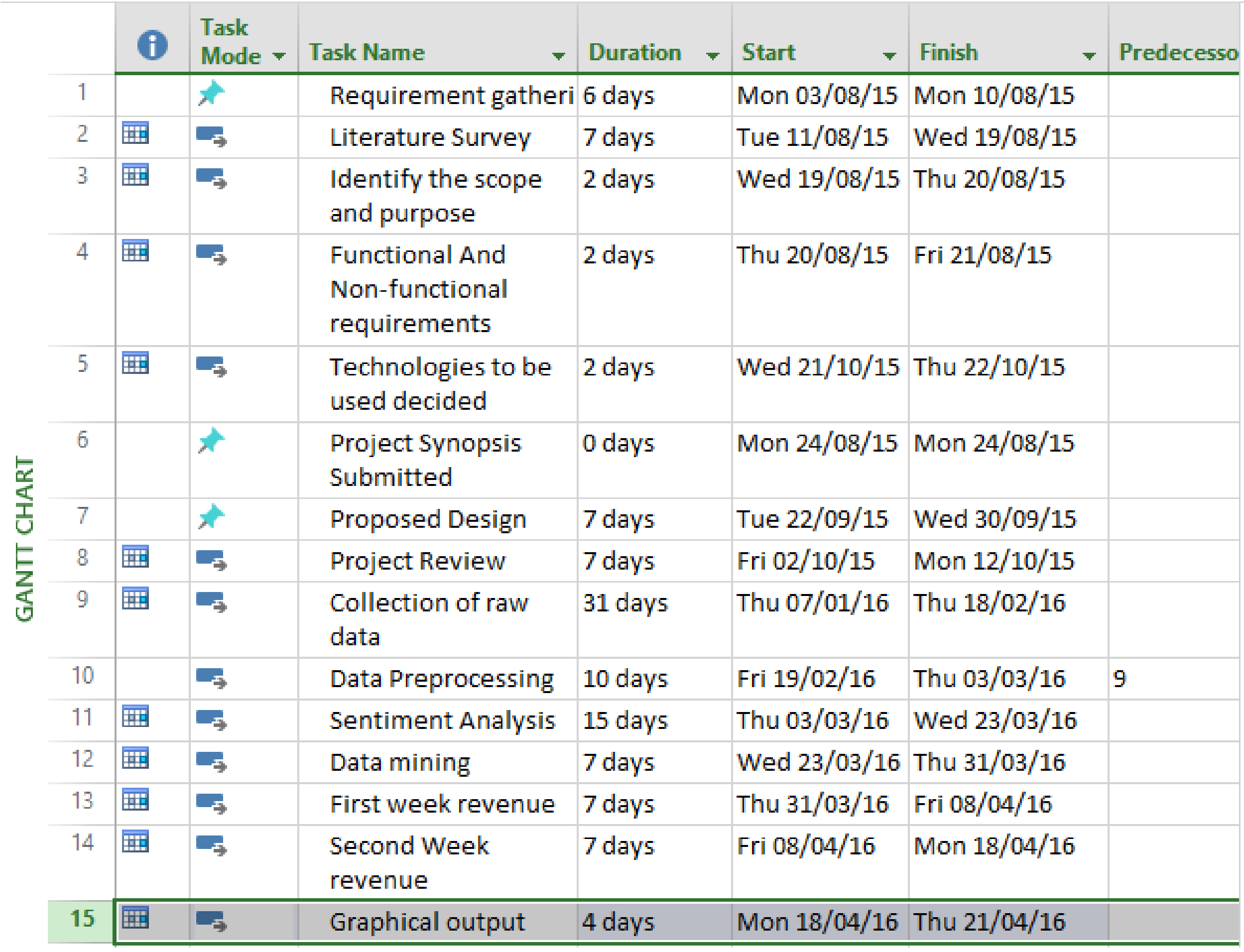


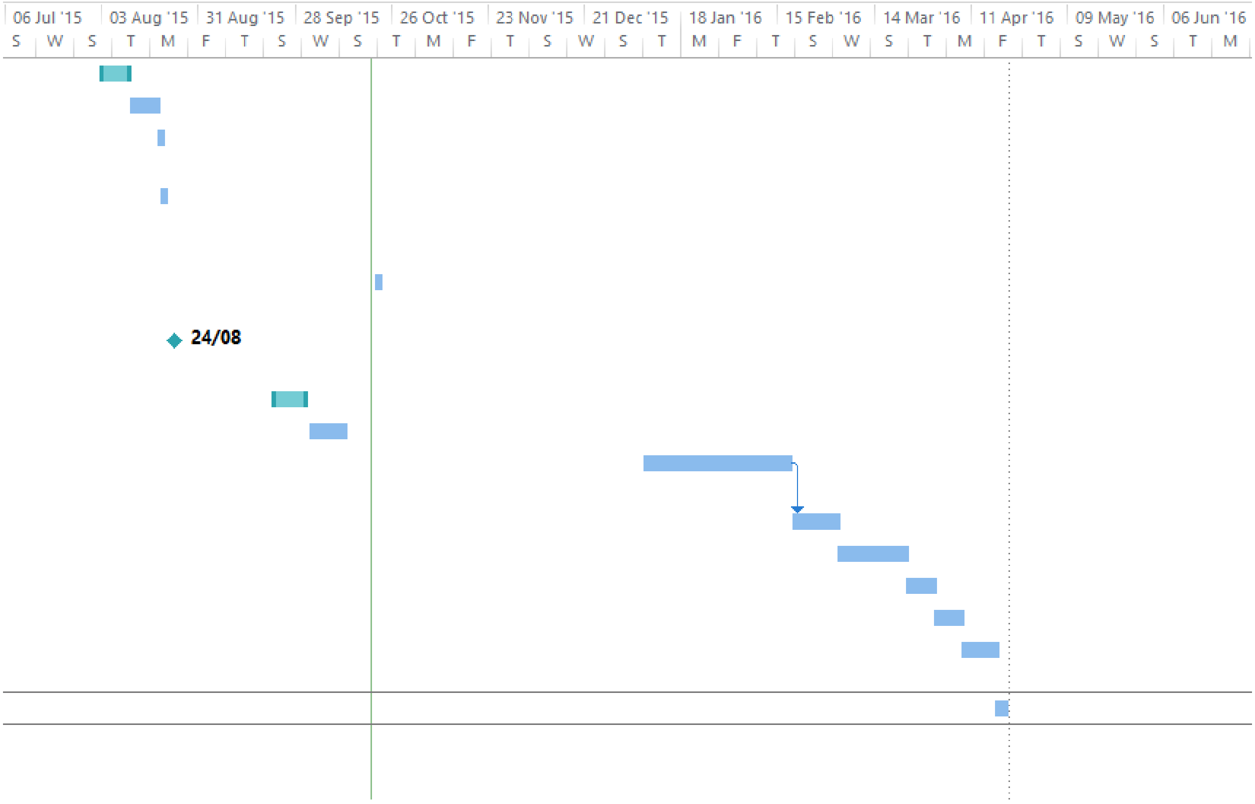
Figure 6 DFD level 1

4.3 Plan of Work



	ⓘ	Task Mode	Task Name	Duration	Start	Finish	Predecesso
1		📌	Requirement gatheri	6 days	Mon 03/08/15	Mon 10/08/15	
2	▦	➔	Literature Survey	7 days	Tue 11/08/15	Wed 19/08/15	
3	▦	➔	Identify the scope and purpose	2 days	Wed 19/08/15	Thu 20/08/15	
4	▦	➔	Functional And Non-functional requirements	2 days	Thu 20/08/15	Fri 21/08/15	
5	▦	➔	Technologies to be used decided	2 days	Wed 21/10/15	Thu 22/10/15	
6		📌	Project Synopsis Submitted	0 days	Mon 24/08/15	Mon 24/08/15	
7		📌	Proposed Design	7 days	Tue 22/09/15	Wed 30/09/15	
8	▦	➔	Project Review	7 days	Fri 02/10/15	Mon 12/10/15	
9	▦	➔	Collection of raw data	31 days	Thu 07/01/16	Thu 18/02/16	
10		➔	Data Preprocessing	10 days	Fri 19/02/16	Thu 03/03/16	9
11	▦	➔	Sentiment Analysis	15 days	Thu 03/03/16	Wed 23/03/16	
12	▦	➔	Data mining	7 days	Wed 23/03/16	Thu 31/03/16	
13	▦	➔	First week revenue	7 days	Thu 31/03/16	Fri 08/04/16	
14	▦	➔	Second Week revenue	7 days	Fri 08/04/16	Mon 18/04/16	
15	▦	➔	Graphical output	4 days	Mon 18/04/16	Thu 21/04/16	

4.4 Project Scheduling & Tracking using Time line / Gnatt Chart



Chapter 5:

Implementation

DATA EXTRACTION:

Data extraction from IMDb

We collected the movie database from IMDb. The dataset contains about 2400 entries from past 15 years and there are 24 attributes associated with a movie. Out of these 24 attributes we have considered 10 attributes for implementation those are budget, domestic gross, ratings, holiday, etc. Factors like actor and director ratings we get from Rotten tomatoes website using web crawling. And sentiment analysis results are also stored in dataset.

Getting actors and directors rating from rottentomatoes.com :

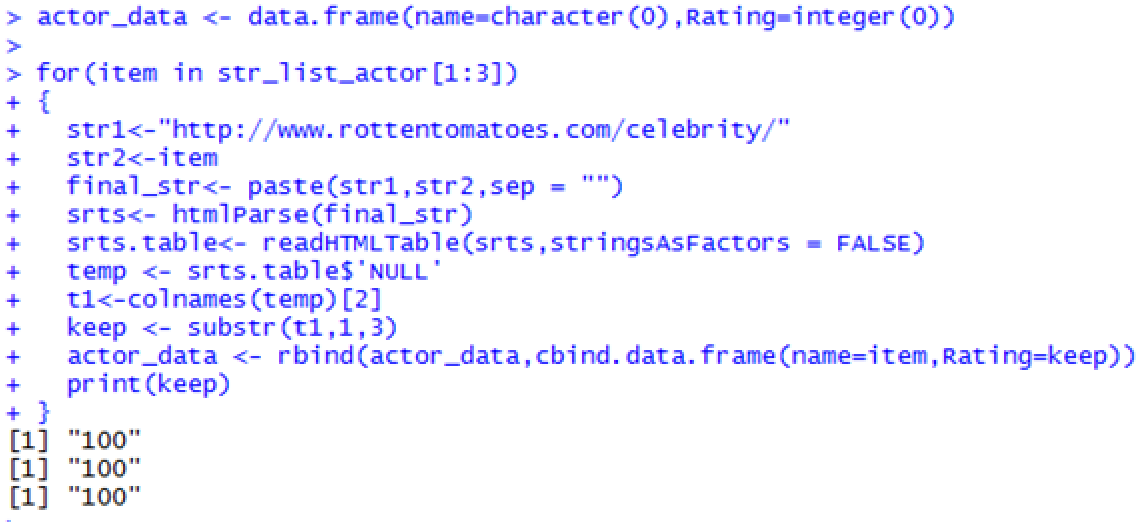
Actors and directors have great impact on revenue of the movie, so it is necessary to extract that information and use it for prediction system. For that purpose we have done web scraping of rottentomatoes.com by using R functions like html parse and read html table and got the ratings.

Actor Ratings:

Code:

```
setwd("C:/Users/Akash/Documents/Twitter_R/actors")
library(XML)
library(stringi)
library(stringr)
library(httpuv)
library(httr)
library(reshape2)
melt(testactors)
#actor_f <- read.csv(file.choose(),header = TRUE,sep = ",")
testactors <- read.csv("~/Twitter_R/testactors.csv", sep="")
str_list_actor<-str_replace_all(tolower(testactors$RT_cast1)," ","")
actor_data <- data.frame(name=character(0),Rating=integer(0))
for(item in str_list_actor[1:100])
{
str1<-"http://www.rottentomatoes.com/celebrity/"
 str2<-item
 final_str<- paste(str1,str2,sep = "")
 srts<- htmlParse(final_str)
 srts.table<- readHTMLTable(srts,stringsAsFactors = FALSE)
 temp <- srts.table$'NULL'
 t1<-colnames(temp)[2]
 keep <- substr(t1,1,3)
 actor_data <- rbind(actor_data,cbind.data.frame(name=item,Rating=keep))
 print(keep)
}
write.table(file="data_actor_rating.csv",actor_data,sep=",",row.names = FALSE,append = TRUE)
```

Screenshot:



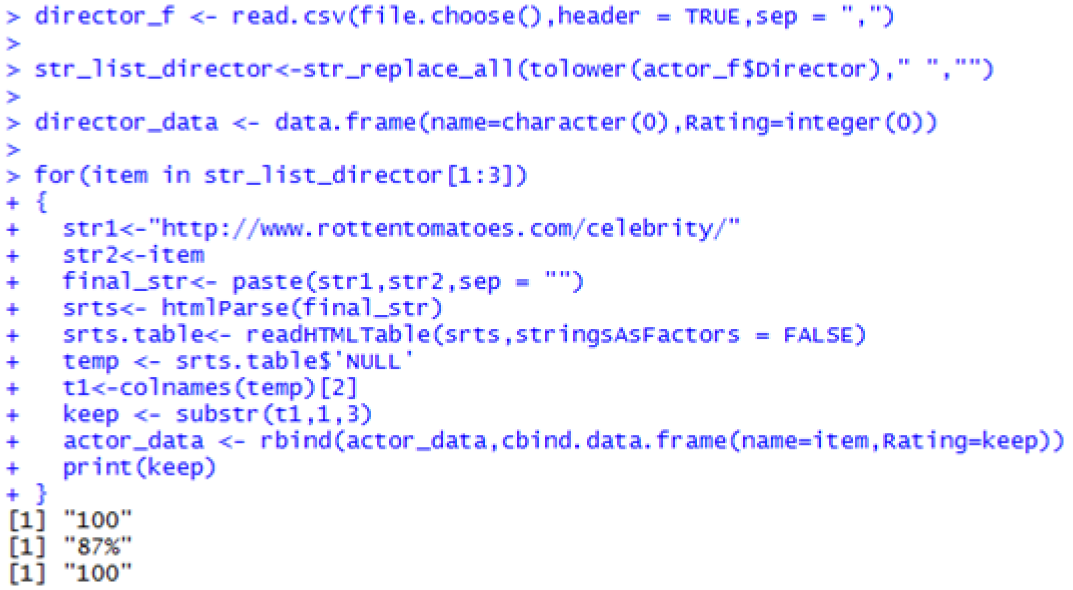
```
> actor_data <- data.frame(name=character(0),Rating=integer(0))
>
> for(item in str_list_actor[1:3])
+ {
+   str1<-"http://www.rottentomatoes.com/celebrity/"
+   str2<-item
+   final_str<- paste(str1,str2,sep = "")
+   srts<- htmlParse(final_str)
+   srts.table<- readHTMLTable(srts,stringsAsFactors = FALSE)
+   temp <- srts.table$'NULL'
+   t1<-colnames(temp)[2]
+   keep <- substr(t1,1,3)
+   actor_data <- rbind(actor_data,cbind.data.frame(name=item,Rating=keep))
+   print(keep)
+ }
[1] "100"
[1] "100"
[1] "100"
```

Director Ratings:

Code:

```
setwd("C:/Users/Akash/Documents/Twitter_R/director")
library(XML)
library(stringi)
library(stringr)
library(httpuv)
library(httr)
director_f <- read.csv(file.choose(),header = TRUE,sep = ",")
str_list_director<-str_replace_all(tolower(actor_f$Director)," ","")
director_data <- data.frame(name=character(0),Rating=integer(0))
for(item in str_list_director[1:3])
{
 str1<-"http://www.rottentomatoes.com/celebrity/"
 str2<-item
 final_str<- paste(str1,str2,sep = "")
 srts<- htmlParse(final_str)
 srts.table<- readHTMLTable(srts,stringsAsFactors = FALSE)
 temp <- srts.table$'NULL'
 t1<-colnames(temp)[2]
 keep <- substr(t1,1,3)
 actor_data <- rbind(actor_data,cbind.data.frame(name=item,Rating=keep))
 print(keep)
}
write.table(file="data_director_rating.csv",actor_data,sep=",",row.names = FALSE)
```

Screenshot:



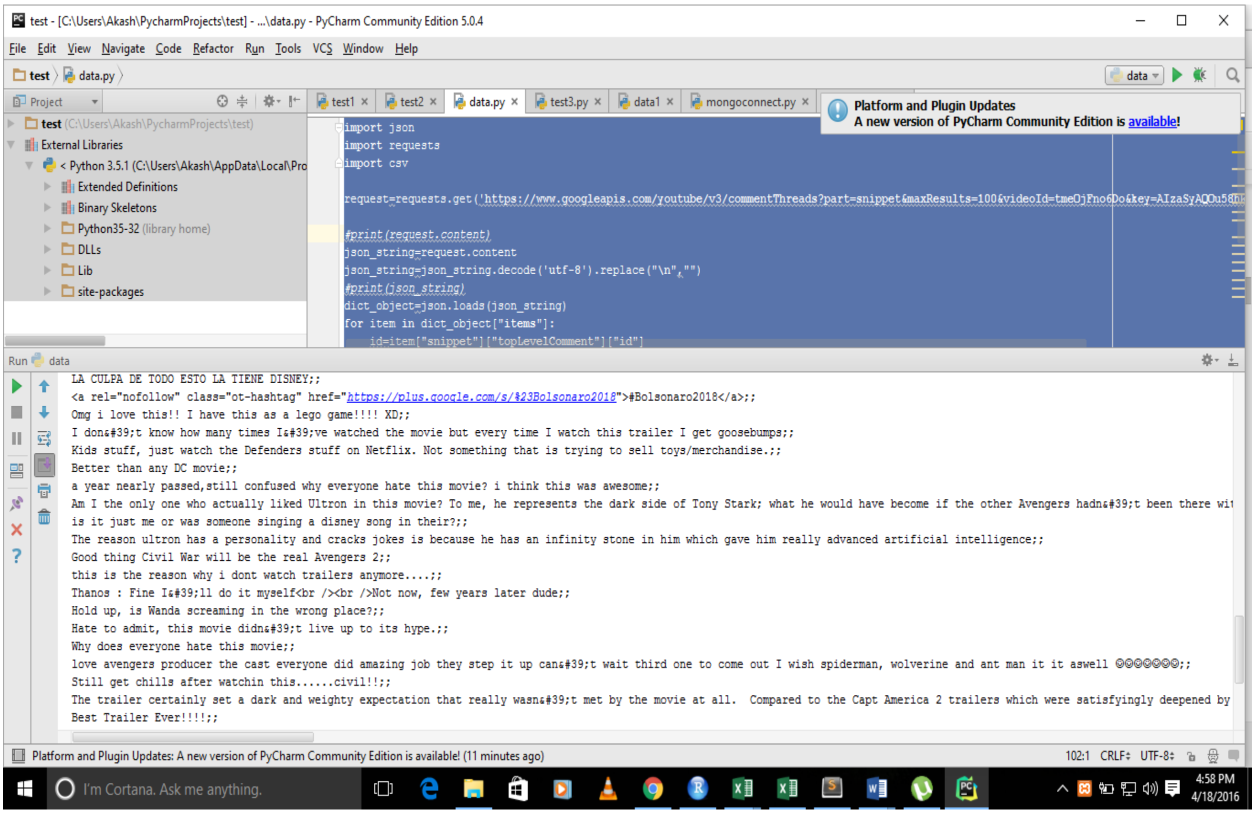
```
> director_f <- read.csv(file.choose(),header = TRUE,sep = ",")
>
> str_list_director<-str_replace_all(tolower(actor_f$Director)," ","")
>
> director_data <- data.frame(name=character(0),Rating=integer(0))
>
> for(item in str_list_director[1:3])
+ {
+   str1<-"http://www.rottentomatoes.com/celebrity/"
+   str2<-item
+   final_str<- paste(str1,str2,sep = "")
+   srts<- htmlParse(final_str)
+   srts.table<- readHTMLTable(srts,stringsAsFactors = FALSE)
+   temp <- srts.table$'NULL'
+   t1<-colnames(temp)[2]
+   keep <- substr(t1,1,3)
+   actor_data <- rbind(actor_data,cbind.data.frame(name=item,Rating=keep))
+   print(keep)
+ }
[1] "100"
[1] "87%"
[1] "100"
```

Extracting Youtube Comments:

For sentiment analysis we collected comments for each movie using python. These comments are collected using Youtube API in python using functions like requests.get to get the content of the particular youtube page. From that page we get the video id, using that id we can get the comment id and once we get the comment id we can get the desired comment. The python code along with its output for “Avengers” movie is as shown below

CODE:

```
import json
import requests
import csv
request=requests.get('https://www.googleapis.com/youtube/v3/commentThreads?part=snippet&maxResults=100&videoId=tmeOjFno6Do&key=AIzaSyAQOu58hKG2zckNPCHZKEqnXqhAf5sJQTE')
#print(request.content)
json_string=request.content
json_string=json_string.decode('utf-8').replace("\n","")
#print(json_string)
dict_object=json.loads(json_string)
for item in dict_object["items"]:
 id=item["snippet"]["topLevelComment"]["id"]
 request=requests.get('https://www.googleapis.com/youtube/v3/comments?part=snippet&id='+id+'&key=AIzaSyAQOu58hKG2zckNPCHZKEqnXqhAf5sJQTE')
 #print(request.content)
 json_string=request.content
 json_string=json_string.decode('utf-8').replace("\n","")
 #print(json_string)
 dict_child_object=json.loads(json_string)
 print(dict_child_object["items"][0]["snippet"]["textDisplay"] + ';;')
```


test - [C:\Users\Akash\PycharmProjects\test] - ...\data.py - PyCharm Community Edition 5.0.4
File Edit View Navigate Code Refactor Run Tools VCS Window Help
test data.py
data
Project
test1 test2 data.py test3.py data1 mongoconnect.py
test (C:\Users\Akash\PycharmProjects\test)
External Libraries
< Python 3.5.1 (C:\Users\Akash\AppData\Local\Pro
Extended Definitions
Binary Skeletons
Python35-32 (library home)
DLLs
Lib
site-packages
Platform and Plugin Updates
A new version of PyCharm Community Edition is available!
Run data

```
import json
import requests
import csv

request=requests.get('https://www.googleapis.com/youtube/v3/commentThreads?part=snippet&maxResults=100&videoId=tmeOjFno6Do&key=AIzaSyAQOu58h

#print(request.content)
json_string=request.content
json_string=json_string.decode('utf-8').replace("\n","")
#print(json_string)
dict_object=json.loads(json_string)
for item in dict_object["items"]:
    id=item["snippet"]["topLevelComment"]["id"]
```

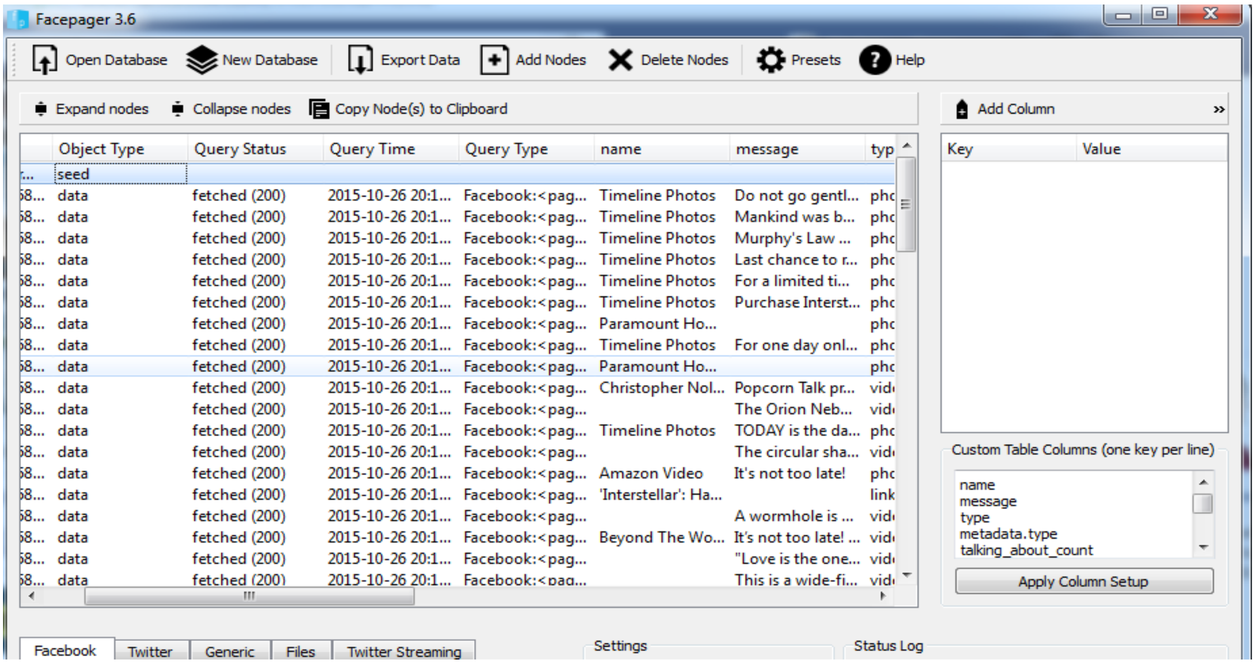
```
LA CULPA DE TODO ESTO LA TIENE DISNEY;;
<a rel="nofollow" class="ot-hashtag" href="https://plus.google.com/s/%23Bolsonaro2018">#Bolsonaro2018</a>;;
Omg i love this!! I have this as a lego game!!!! XD;;
I don&#39;t know how many times I&#39;ve watched the movie but every time I watch this trailer I get goosebumps;;
Kids stuff, just watch the Defenders stuff on Netflix. Not something that is trying to sell toys/merchandise.;;
Better than any DC movie;;
a year nearly passed,still confused why everyone hate this movie? i think this was awesome;;
Am I the only one who actually liked Ultron in this movie? To me, he represents the dark side of Tony Stark; what he would have become if the other Avengers hadn&#39;t been there wit
is it just me or was someone singing a disney song in their?;;
The reason ultron has a personality and cracks jokes is because he has an infinity stone in him which gave him really advanced artificial intelligence;;
Good thing Civil War will be the real Avengers 2;;
this is the reason why i dont watch trailers anymore....;;
Thanos : Fine I&#39;ll do it myself<br /><br />Not now, few years later dude;;
Hold up, is Wanda screaming in the wrong place?;;
Hate to admit, this movie didn&#39;t live up to its hype.;;
Why does everyone hate this movie;;
love avengers producer the cast everyone did amazing job they step it up can&#39;t wait third one to come out I wish spiderman, wolverine and ant man it it aswell ☺☺☺☺☺☺☺;;
Still get chills after watchin this......civil!!;;
The trailer certainly set a dark and weighty expectation that really wasn&#39;t met by the movie at all.  Compared to the Capt America 2 trailers which were satisfyingly deepened by
Best Trailer Ever!!!!;;
```

Platform and Plugin Updates: A new version of PyCharm Community Edition is available! (11 minutes ago)
102:1 CRLF÷ UTF-8÷
I'm Cortana. Ask me anything.
4:58 PM
4/18/2016

Extracting Facebook Comments:

Facebook comments for particular movies was fetched using FacePager software. An access token is generated from facebook which is added to the node and the movie name is added as the seed node.<user/posts> is selected to view users post and comments are extracted and saved in .csv file.

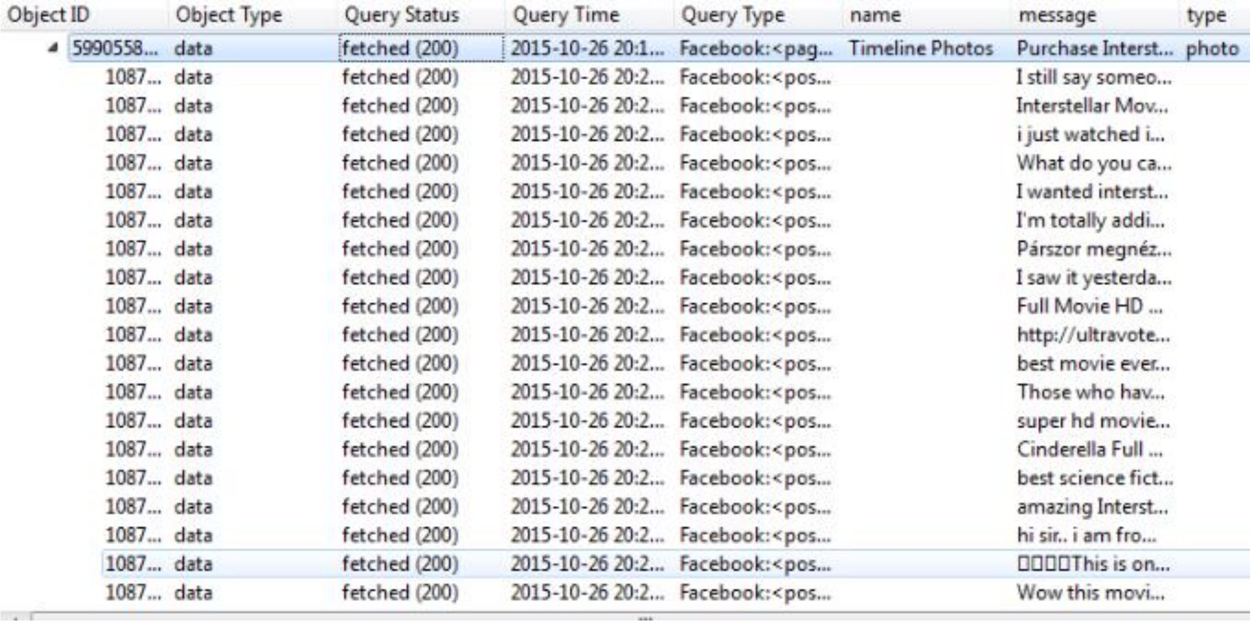
Screenshot:



	Object Type	Query Status	Query Time	Query Type	name	message	typ
r...	seed						
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	Do not go gentl...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	Mankind was b...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	Murphy's Law ...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	Last chance to r...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	For a limited ti...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	Purchase Interst...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Paramount Ho...		phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	For one day onl...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Paramount Ho...		phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Christopher Nol...	Popcorn Talk pr...	vid
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...		The Orion Neb...	vid
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	TODAY is the da...	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...		The circular sha...	vid
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Amazon Video	It's not too late!	phc
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	'Interstellar': Ha...		link
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...		A wormhole is ...	vid
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Beyond The Wo...	It's not too late! ...	vid
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...		"Love is the one...	vid
58...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...		This is a wide-fi...	vid

Custom Table Columns (one key per line)

name
message
type
metadata.type
talking_about_count



Object ID	Object Type	Query Status	Query Time	Query Type	name	message	type
5990558...	data	fetched (200)	2015-10-26 20:1...	Facebook:<pag...	Timeline Photos	Purchase Interst...	photo
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		I still say someo...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		Interstellar Mov...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		i just watched i...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		What do you ca...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		I wanted interst...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		I'm totally addi...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		Párszor megnéz...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		I saw it yesterda...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		Full Movie HD ...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		http://ultravote...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		best movie ever...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		Those who hav...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		super hd movie...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		Cinderella Full ...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		best science fict...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		amazing Interst...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		hi sir.. i am fro...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		□□□□This is on...	
1087...	data	fetched (200)	2015-10-26 20:2...	Facebook:<pos...		Wow this movi...	

CLUSTERING:

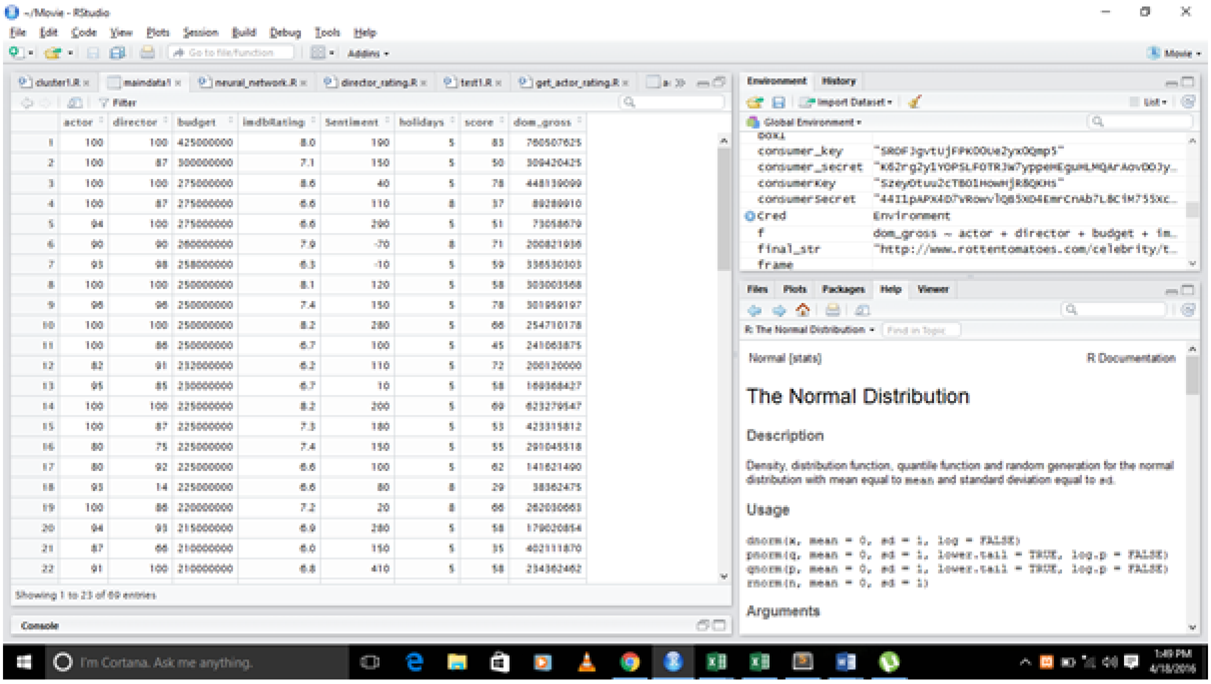
Clustering is done to increase the accuracy of the system. We used K-means clustering methodology for clustering. In this method the clusters from the dataset are made based on budget of the movie.

CODE:

```
#K means clustering on budget
setwd("C:/Users/Akash/Documents/Twitter_R/cluster")
require(graphics)
d1 <- reduced1
grp <- kmeans((reduced1$imdbRating),centers = 3)
c1 <- data[grp$cluster==1,]
write.csv(c1,file = "c1.csv")
c2 <- data[grp$cluster==2,]
write.csv(c2,file = "c2.csv")
c3 <- data[grp$cluster==3,]
write.csv(c3,file = "c3.csv")
c4 <- data[grp$cluster==4,]
write.csv(c4,file = "c4.csv")
c5 <- data[grp$cluster==5,]
write.csv(c5,file = "c1.csv")
c6 <- data[grp$cluster==6,]
write.csv(c6,file = "c6.csv")
c7 <- data[grp$cluster==7,]
write.csv(c7,file = "c1.csv")
c8 <- data[grp$cluster==8,]
write.csv(c8,file = "c8.csv")
c9 <- data[grp$cluster==9,]
write.csv(c9,file = "c9.csv")
c10 <- data[grp$cluster==10,]
write.csv(c10,file = "c10.csv")
```

Screenshot:

Cluster 1:



	actor	director	budget	imdbRating	Sentiment	holidays	score	dom_gross
1	100	100	425000000	8.0	190	5	83	760507625
2	100	87	300000000	7.1	150	5	50	309420425
3	100	100	275000000	8.6	40	5	78	448139099
4	100	87	275000000	6.6	110	8	37	89289910
5	94	100	275000000	6.6	290	5	51	73058679
6	90	90	260000000	7.9	-70	8	71	200821936
7	93	98	258000000	6.3	-10	5	59	336530303
8	100	100	250000000	8.1	120	5	58	303003568
9	98	98	250000000	7.4	150	5	78	301959197
10	100	100	250000000	8.2	280	5	66	254710178
11	100	86	250000000	6.7	100	5	45	241063875
12	82	91	232000000	6.2	110	5	72	200120000
13	95	85	230000000	6.7	10	5	58	169368427
14	100	100	225000000	8.2	200	5	69	623279547
15	100	87	225000000	7.3	180	5	53	423315812
16	80	75	225000000	7.4	150	5	55	291045518
17	80	92	225000000	6.6	100	5	62	141621490
18	93	14	225000000	6.6	80	8	29	38362475
19	100	86	220000000	7.2	20	8	66	262030663
20	94	93	215000000	6.9	280	5	58	179020854
21	87	66	210000000	6.0	150	5	35	402111870
22	91	100	210000000	6.8	410	5	58	234362462

Showing 1 to 23 of 69 entries

```
  consumer_key      "SROFJgvtUjFPK00Ue2yx0Qmp5"
  consumer_secret   "K62rg2y1YOPSLFOTRJW7yppeHEguHLMQArAovDOJy…
  consumerKey       "SzeyOtuu2cTBO1HowHjR8QKHs"
  consumerSecret    "44I1pAPX4D7VRowvlQB5XD4EmrCnAb7L8CiM755Xc…
  cred              Environment
  f                 dom_gross ~ actor + director + budget + im…
  final_str         "http://www.rottentomatoes.com/celebrity/t…
  frame
```

Normal {stats} R Documentation

The Normal Distribution

Description

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to mean and standard deviation equal to sd.

Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

Arguments

SENTIMENT ANALYSIS:

The user reviews play important role in predicting the revenue. Based on the reviews provided by the user on twitter and youtube we can increase the accuracy of the system. Once we collected the comments and then calculated the polarity of each comment by breaking down the sentence into words and then comparing it with positive and negative words dictionary. The sentiment package of R-studio is used.

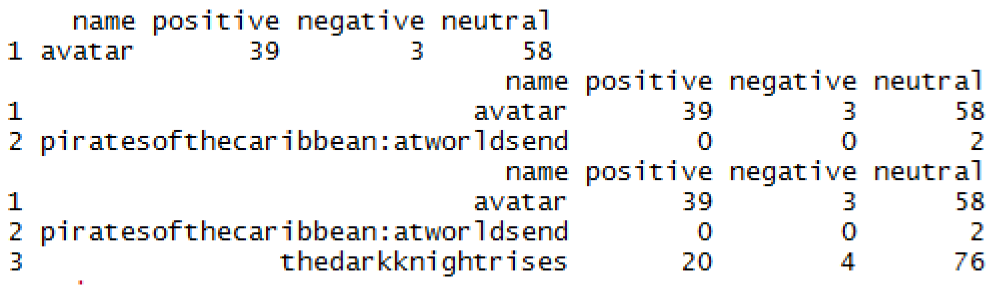
Extracting tweets from twitter and applying sentiment analysis on it:

Code:

```
setwd("C:/Users/Akash/Documents/Twitter_R")
library(twitteR)
library(plyr)
require(RCurl)
require(base64enc)
library(stringi)
library(stringr)
library(sentiment)
consumer_key <- 'SROFJgvtUjFPK00Ue2yx0Qmp5'
consumer_secret <- 'K62rg2y1YOPSLFOTRJW7yppeHEguHLMQArAovDOJyX0dZe5LdY'
access_token <- '2356577120-LRSg7XomwsG0NsOs8Ah0mKq7fTR8R0Hi258pIYW'
access_secret <- 'doknp4Svce9cExKgAQ69sHE7ujRyyw8oSCmluo8zenlb1'
setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

```
str_list<-str_replace_all(trimws(tolower(numbers_RT_omdb_cleaned$RT_title))," ","")
tweets_list<-vector("list")
data_final<-
data.frame(name=character(0),negative=integer(0),neutral=integer(0),positive=integer(0))
print(data_final)
for(item in str_list[70]){
 tweets <- searchTwitter(item,100, lang = "en")
 temp<-sapply(tweets,function(x) x$text)
 polarities<-sapply(temp,function(y) sentiment(y)$polarity)
 data_final <- rbind(data_final,cbind.data.frame(name=item,positive=length(which(polarities ==
"positive")),negative=length(which(polarities == "negative")),neutral=length(which(polarities ==
"neutral"))))
 #data_final[nrow(data_final)+1,]<-c(item,length(which(polarities ==
"negative")),length(which(polarities == "neutral")),length(which(polarities == "positive")))
 print(data_final)
}
write.table(file="data_sentiment.csv",data_final,sep=",",row.names = FALSE)
```

SCREENSHOT :



```
    name positive negative neutral
1 avatar       39        3      58
                               name positive negative neutral
1                            avatar       39        3      58
2 piratesofthecaribbean:atworldsend        0        0       2
                               name positive negative neutral
1                            avatar       39        3      58
2 piratesofthecaribbean:atworldsend        0        0       2
3               thedarkknightrises       20        4      76
```

ARTIFICIAL NEURAL NETWORK:

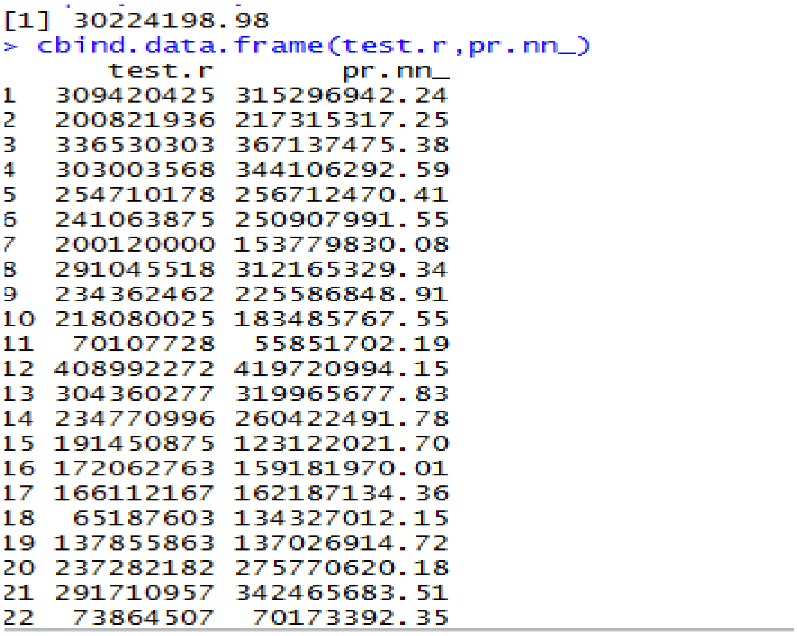
For fitting artificial neural network we used neuralnet package of R-studio is used. The neuralnet package of R contains an nnet function which creates the network. The plot.nn function is used to plot the network for visualization. The compute function is used to compute the network result.

Fitting neural network

```
apply(data,2,function(x) sum(is.na(x)))
index <- sample(1:nrow(data),round(0.75*nrow(data)))
train <- data[index,]
test <- data[-index,]
```

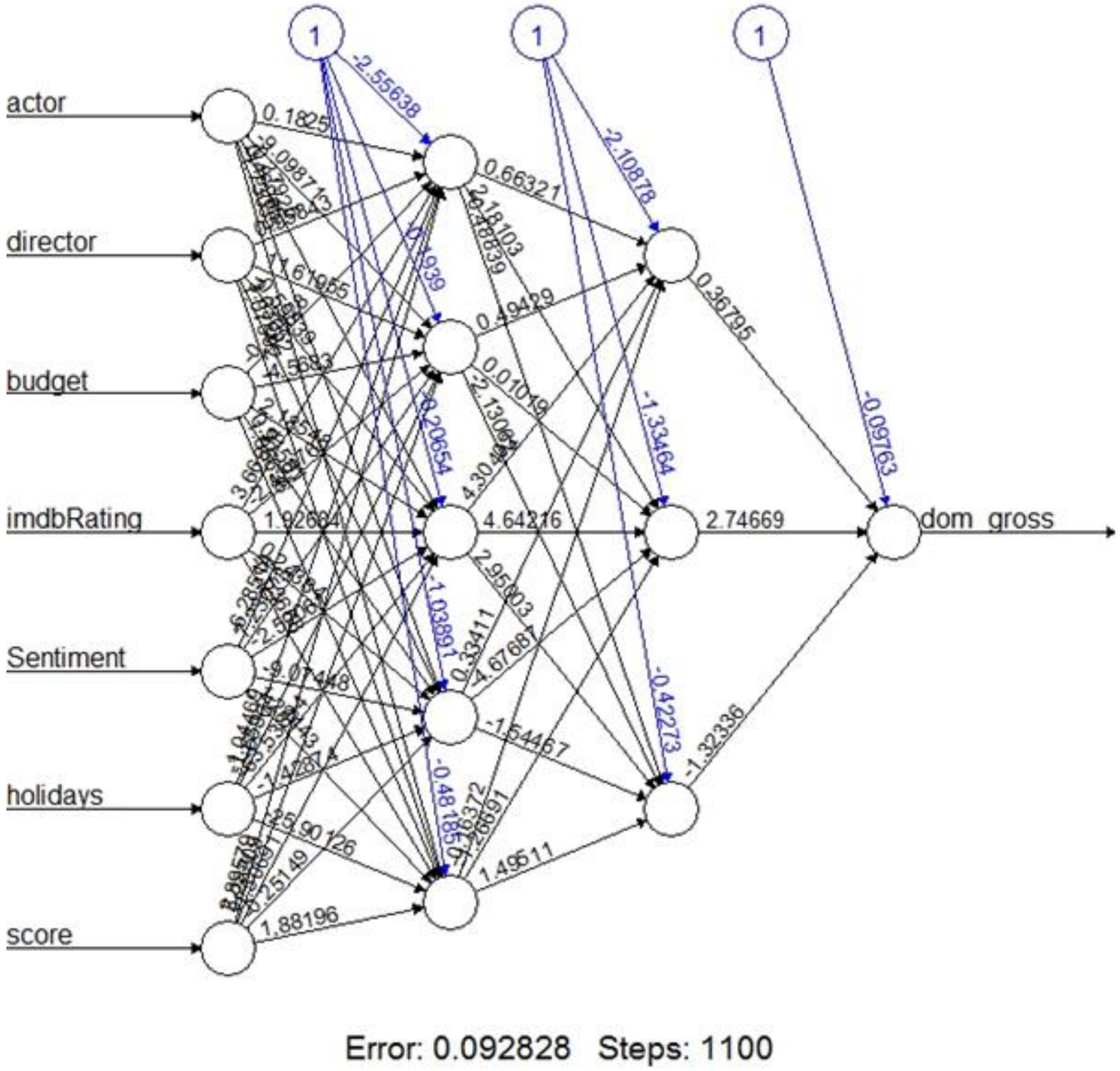
```
maxs <- apply(maindata1, 2, max)
mins <- apply(maindata1, 2, min)
scaled <- as.data.frame(scale(maindata1, center = mins, scale = maxs - mins))
train_ <- scaled[index,]
test_ <- scaled[-index,]
library(neuralnet)
n <- names(train_)
f <- as.formula(paste("dom_gross ~", paste(n[!n %in% "dom_gross"], collapse = " + ")))
nn <- neuralnet(f,data=train_,hidden=c(5,3),linear.output=T)
pr.nn <- compute(nn,test_[-8])
pr.nn_ <- pr.nn$net.result*(max(maindata1$dom_gross)-min(maindata1$dom_gross))+min(maindata1$dom_gross)
rand<-rnorm(nrow(test_),mean=1,sd=0.05)
test.r <- (test_$dom_gross)*(max(maindata1$dom_gross)-min(maindata1$dom_gross))+min(maindata1$dom_gross)
pr.nn_<-(test_$dom_gross+rand-1)*(max(maindata1$dom_gross)-min(maindata1$dom_gross))+min(maindata1$dom_gross)
MSE.nn <- sum((test.r - pr.nn_)^2)/nrow(test_)
sqrt(MSE.nn)
cbind.data.frame(test.r,pr.nn_)
write.table(file="hola.csv",cbind.data.frame(row_id=as.numeric(rownames(scaled[-index,])),actual=test.r,predicted=pr.nn_),row.names = F,sep = ",")
```

Screenshot:



```
[1] 30224198.98
> cbind.data.frame(test.r,pr.nn_)
      test.r       pr.nn_
1  309420425 315296942.24
2  200821936 217315317.25
3  336530303 367137475.38
4  303003568 344106292.59
5  254710178 256712470.41
6  241063875 250907991.55
7  200120000 153779830.08
8  291045518 312165329.34
9  234362462 225586848.91
10 218080025 183485767.55
11  70107728  55851702.19
12 408992272 419720994.15
13 304360277 319965677.83
14 234770996 260422491.78
15 191450875 123122021.70
16 172062763 159181970.01
17 166112167 162187134.36
18  65187603 134327012.15
19 137855863 137026914.72
20 237282182 275770620.18
21 291710957 342465683.51
22  73864507  70173392.35
```

Neural network:



USER INTERFACE CODE:

Home page:

```
<!DOCTYPE HTML>
<html>
<head>
<title>DONATE LIFE </title>
<link rel="stylesheet" href="styles/layout.css" type="text/css" />
<script type="text/javascript">
<!-->
var image1=new Image()
image1.src="a1.jpg"
var image2=new Image()
```

```
image2.src="a2.jpg"
var image3=new Image()
image3.src="a3.jpg"
var image4=new Image()
image4.src="a4.jpg"
//-->
</script>
 </head>
<body id="top">
<div class="wrapper row1">
 <div id="header" class="clear">
<div class="fl_left">
        <h1><a href="donate.html"><img src="a2.jpg"
style="width:900px;height:240px;"></a></h1>
</div>
 </div>
</div>
<div class="wrapper row2">
 <div class="rnd">

<div id="topnav">
        <ul>
        <li class="active"><a href="donate.html">Home</a></li>
        <li><a href="selectmovie.php">Movies</a></li>

        <li><a href="form.html">Sign Up</a></li>

        <li><a href="aboutus.html">About us</a></li>
        </ul>
</div>
 </div>
</div>
<div class="wrapper">
 <div id="featured_slide" class="clear">
<!-- ###### -->
<div class="overlay_left"></div>
```

```
<div id="featured_content">
        <div class="featured_box" id="fc1"><img src="a5.jpg" alt="civil" style="width:400px; height:250px;" />
        <div class="floater" style="z-index:1;">
        <h2>CIVIL WAR</h2>
        <p>Initial release: April 29, 2016 <br>
Directors: Joe Russo, Anthony Russo<br>
Producer: Kevin Feige</p>
        </div>
        </div>
        <div class="featured_box" id="fc2"><img src="a6.jpg" alt="suicide" style="width:400px; height:250px;"  />
        <div class="floater">
        <h1>SUICIDE SQUAD</h1>
        <p>Initial release: August 4, 2016 <br>
Director: David Ayer<br>
Music director: Steven Price</p>

        </div>
        </div>
        <div class="featured_box" id="fc3"><img src="a7.jpg" alt="perfect" style="width:400px; height:250px;"  />
        <div class="floater">
        <h2>THE PERFECT MATCH</h2>
        <p>Initial release: March 11, 2016 <br>
Director: Bille Woodruff<br>
Production company: Flavor Unit Entertainment</p>

        </div>
        </div>
        <div class="featured_box" id="fc4"><img src="a8.jpg" alt="batman" style="width:400px; height:250px;"  />
        <div class="floater">
        <h2>BATMAN VS SUPERMAN</h2>
        <p>Release date: March 25, 2016 (India)<br>
Director: Zack Snyder<br>
```

```
Executive producers: Christopher Nolan, Geoff Johns, Emma Thomas, Michael Uslan, Benjamin Melniker, Wesley Coller<br>
Producers: Deborah Snyder, Charles Roven
</p>


	</div>
	</div>
	<div class="featured_box" id="fc5"><img src="a9.jpg" alt=""style="width:400px; height:250px;"   />
	<div class="floater">
	<h2>THE JUNGLE BOOK</h2>
	<p>Initial release: April 7, 2016 <br>
Director: Jon Favreau <br>
Music director: John Debney</p>


	</div>
	</div>
</div>
<ul id="featured_tabs">
	<li><a href="#fc1">CIVIL WAR</a></li>
	<li><a href="#fc2">SUICIDE SQUAD</a></li>
	<li><a href="#fc3">THE PERFECT MATCH</a></li>
	<li><a href="#fc4">BATMAN VS SUPERMAN</a></li>
	  <li><a href="#fc5">THE JUNGLE BOOK</a></li>


</ul>
<div class="overlay_right"></div>
<!-- ###### -->
 </div>
</div>
<div class="wrapper row3">
 <div class="rnd">
<div id="container" class="clear">
	<div id="content" align="center">
<center>
<img src="a1.png" name="slide" width="900" height="450" align="center" hspace=300 vspace=30>
```

```
<script type="text/javascript">
<!--
var step=1
function slideit(){
document.images.slide.src=eval("image"+step+".src")
if(step<4)
step++
else
step=1
setTimeout("slideit()",1500)
}
slideit()
//-->
</script>
</div>
        </div>
</div>

    </div>
</div>
        <div id="twitter" class="clear" style="position:relative; left:16%; height:150px;">
        <div class="fl_left"><a href="#"></a></div>
        <div class="fl_right">

                                                        <p>Shreya
Sherugar&nbsp;&nbsp;|&nbsp;&nbsp;shreyasherugar06@gmail.com&nbsp;&nbsp;|&nbsp;&nbsp;9
920516427&nbsp;&nbsp;</p>
                                                        <p>Akshay
Wagh&nbsp;&nbsp;|&nbsp;&nbsp;akshaywagh@gmail.com&nbsp;&nbsp;|&nbsp;&nbsp;99205164
27&nbsp;&nbsp;</p>
                                                        <p>Shital
Chaudhari&nbsp;&nbsp;|&nbsp;&nbsp;shitalchaudhari@gmail.com&nbsp;&nbsp;|&nbsp;&nbsp;99
20516427&nbsp;&nbsp;</p>
<p>Akash
Indani&nbsp;&nbsp;|&nbsp;&nbsp;akashindani@gmail.com&nbsp;&nbsp;|&nbsp;&nbsp;99205164
27&nbsp;&nbsp;</p>
```

```
            </div>

</div>
</body>
</html>
```

Select Movie:

```
<?php
$dbc = mysqli_connect('localhost','root',",'final_prediction') OR die(mysql_connect_error());
$query = "SELECT `movie` FROM output";
$result = mysqli_query($dbc,$query);
?>
<!DOCTYPE html>
<html lang="en">
 <head>
<meta charset="utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title> Travel Point </title>
<!-- Bootstrap core CSS -->
<link href="assets/css/bootstrap.min.css" rel="stylesheet">
<link href="assets/css/font-awesome.min.css" rel="stylesheet">
<!-- Custom styles for this template -->
<link href="assets/css/main.css" rel="stylesheet">
 </head>
<body>
<div class="row centered">
<form role="form" action="next.php" method="post" class="form-horizontal">
<div class="form-group">
<label class="col-sm-4 control-label"> Select Movie : </label>
            <div class="col-sm-4">
                        <select class="form-control" name="movie">
                        <?php while ($userRow= mysqli_fetch_assoc($result)) { ?>
                        <option value="<?php echo $userRow['movie']; ?>">     <?php     echo
$userRow['movie']; ?>          </option>
                        <?php } ?></select>
            </div>
            <div class="col-sm-6 col-sm-push-3">
```

```
                    <br><button type="submit" class="btn btn-info">Submit</button>
          </div>
</div>
</form>
</div>
<script src="assets/js/jquery.min.js"></script>
<script src="assets/js/bootstrap.min.js"></script>
</body>
</html>
```

Return Output:

```
<?php
$dbc = mysqli_connect('localhost','root','','final_prediction') OR die(mysqli_connect_error());
?>
<!DOCTYPE html>
<html lang="en">
 <head>
<meta charset="utf-8">
<meta name="viewport" content="width=device-width, initial-scale=1.0">
<title> Travel Point </title>
<!-- Bootstrap core CSS -->
<link href="assets/css/bootstrap.min.css" rel="stylesheet">
<link href="assets/css/font-awesome.min.css" rel="stylesheet">
<!-- Custom styles for this template -->
<link href="assets/css/main.css" rel="stylesheet">
 </head>
<body>
<?php
$temp = $_POST['movie'];
$query = "SELECT * FROM output WHERE movie='$temp'";
$result = mysqli_query($dbc,$query);
while ($userRow= mysqli_fetch_assoc($result))
{ ?>
<h3 align="center">
Movie : <?php echo $temp; ?>
</h3>
<br>
<h3 align="center">
Predicted revenue :  <?php echo $userRow['predicted']; ?>
</h3>
<?php }
?>
</body>
<script src="assets/js/jquery.min.js"></script>
<script src="assets/js/bootstrap.min.js"></script>
</body>
</html>
```

Chapter 6:

Testing

6.1 UNIT TESTING

In this stage of testing every module was individually build and after a successful build and run independently. The errors reported were fixed back and the testing was performed recursively until every module gave result as per specified in the requirements document by
making the appropriate changes.

1) Each module like clustering, sentiment, ANN, etc was designed and tested independently

2) Outputting all the results in CSV format of each module independently and forwarded to next module as input id required.

6.2 INTEGRATION TESTING

In this part we started integrating different modules discussed above simultaneously. System process details were integrated with the actual prediction module. An application run was performed for accuracy. Similar activity was performed with the end user interface
where every process was detailed.

All the modules like data extraction, clustering, sentiment analysis, ANN were integrated in R-studio and the results are stored in CSV format which are further imported in MySQL to show results in front end.

6.3 PERFORMANCE TESTING

Our application can be compared to existing testing tools with respect to some following points:

1)Simplicity: The GUI is very simple and straight forward which reduces the overhead.

2) Speed: As the whole data is preprocessed and stored in MySQL server the user experience on web application is very fast.

Chapter 7:

Result Analysis

7.1 Simulation Model

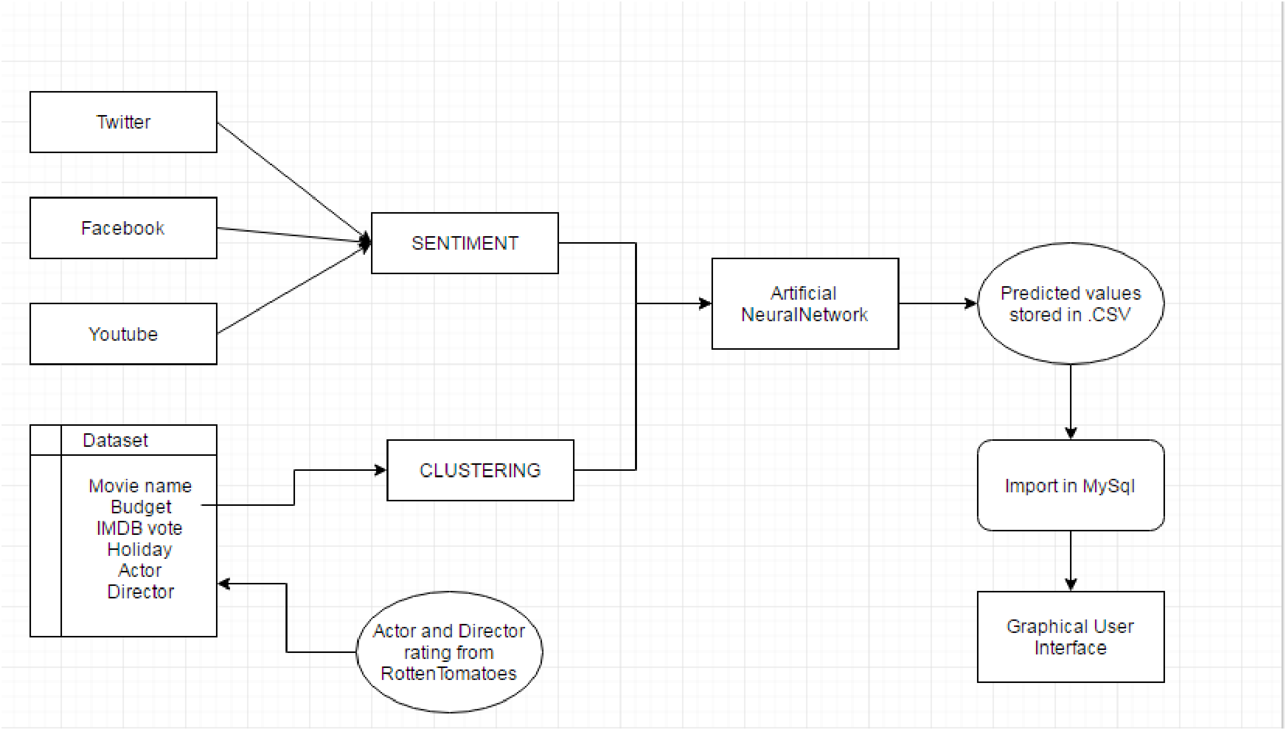


Figure 7

7.2 Output Printouts

The UI elaborates the working of the whole system.

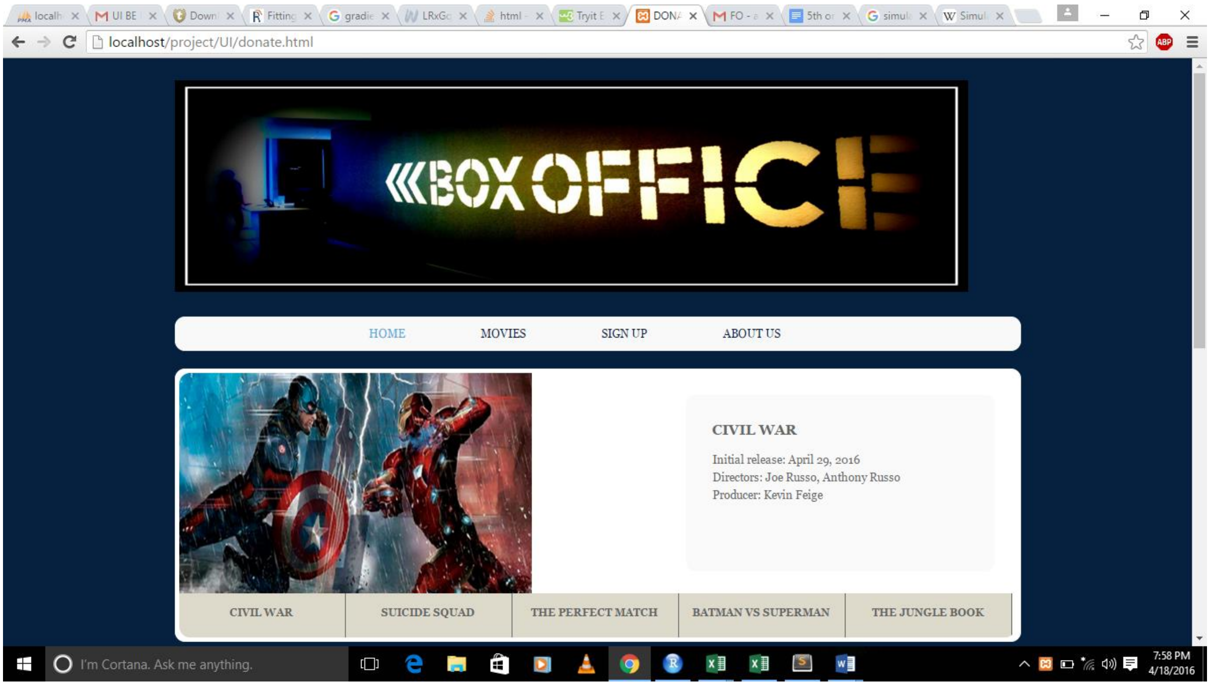
The Box-Office page contains top 5 upcoming movie predicted Revenue . It has a Home,Movies,Coming soon and About us menus.

Movies menu contains a drop down list of movies whose actual revenue is present in our dataset and the predicted revenue for each movie is done using R-Studio and stored in the database.

Selecting any of the movie gives the Predicted Revenue of that movie.

Coming Soon menu has images of the upcoming movies whose Prediction would be shortly done in coming days and will be updates on the HOME menu.

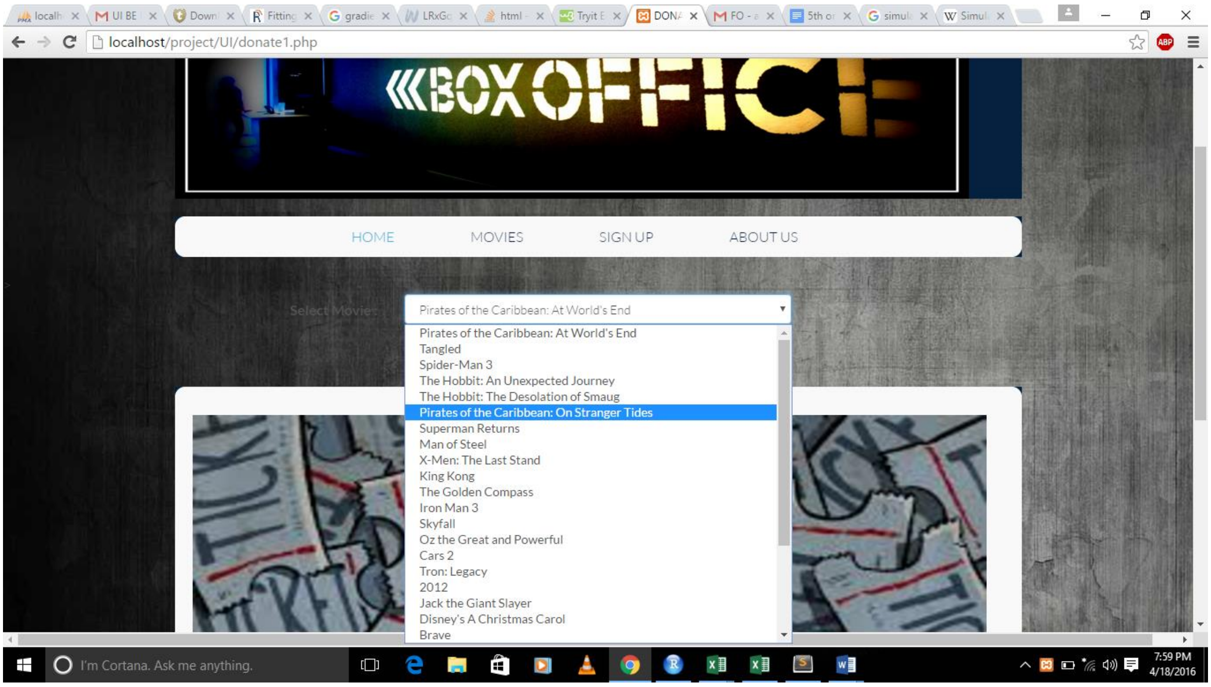
Home page:



CIVIL WAR

Initial release: April 29, 2016
Directors: Joe Russo, Anthony Russo
Producer: Kevin Feige

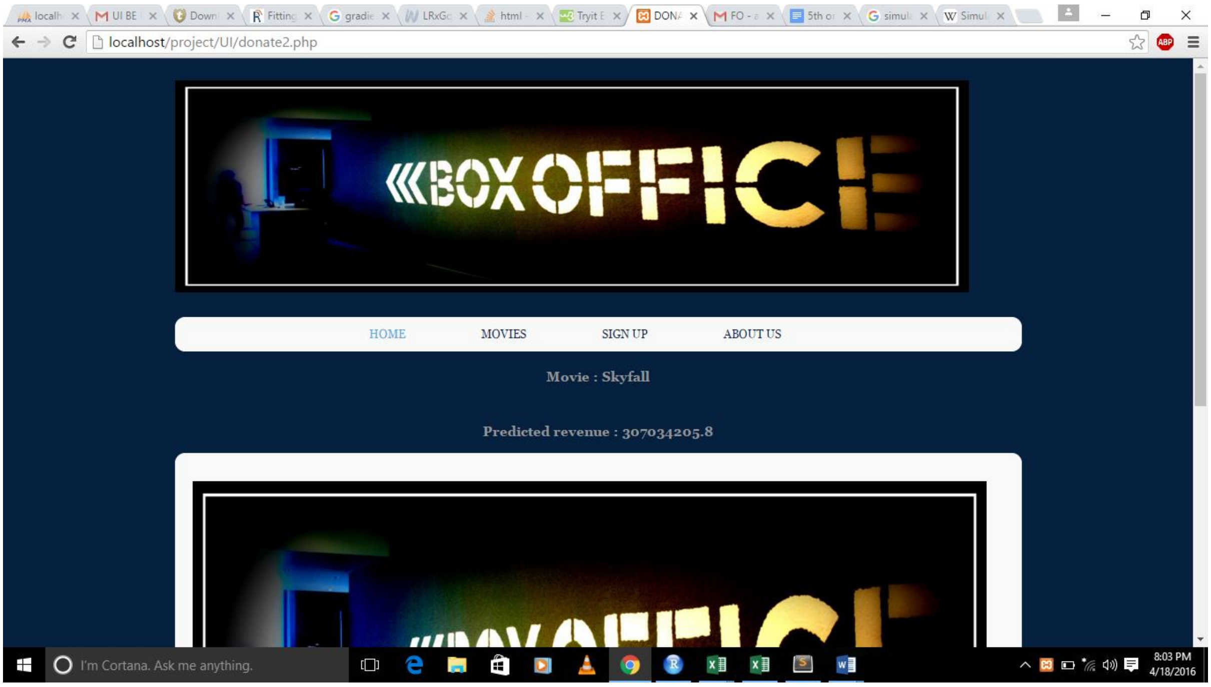
Select movie from list of movies from database:


localhost/project/UI/donate1.php
BOX OFFICE
HOME
MOVIES
SIGN UP
ABOUT US
I'm Cortana. Ask me anything.
7:59 PM
4/18/2016

Select Movie: Pirates of the Caribbean: At World's End

Pirates of the Caribbean: At World's End
Tangled
Spider-Man 3
The Hobbit: An Unexpected Journey
The Hobbit: The Desolation of Smaug
Pirates of the Caribbean: On Stranger Tides
Superman Returns
Man of Steel
X-Men: The Last Stand
King Kong
The Golden Compass
Iron Man 3
Skyfall
Oz the Great and Powerful
Cars 2
Tron: Legacy
2012
Jack the Giant Slayer
Disney's A Christmas Carol
Brave

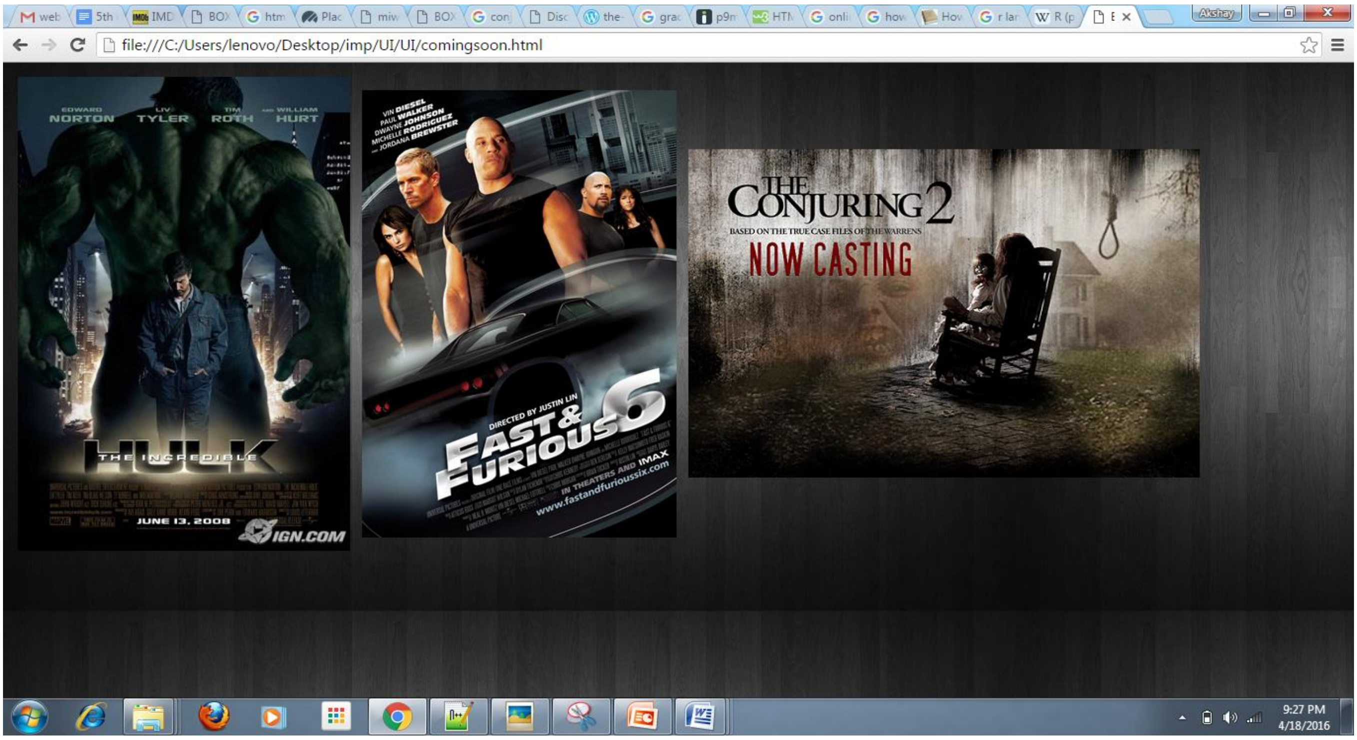
Predicted revenue value shown form database:



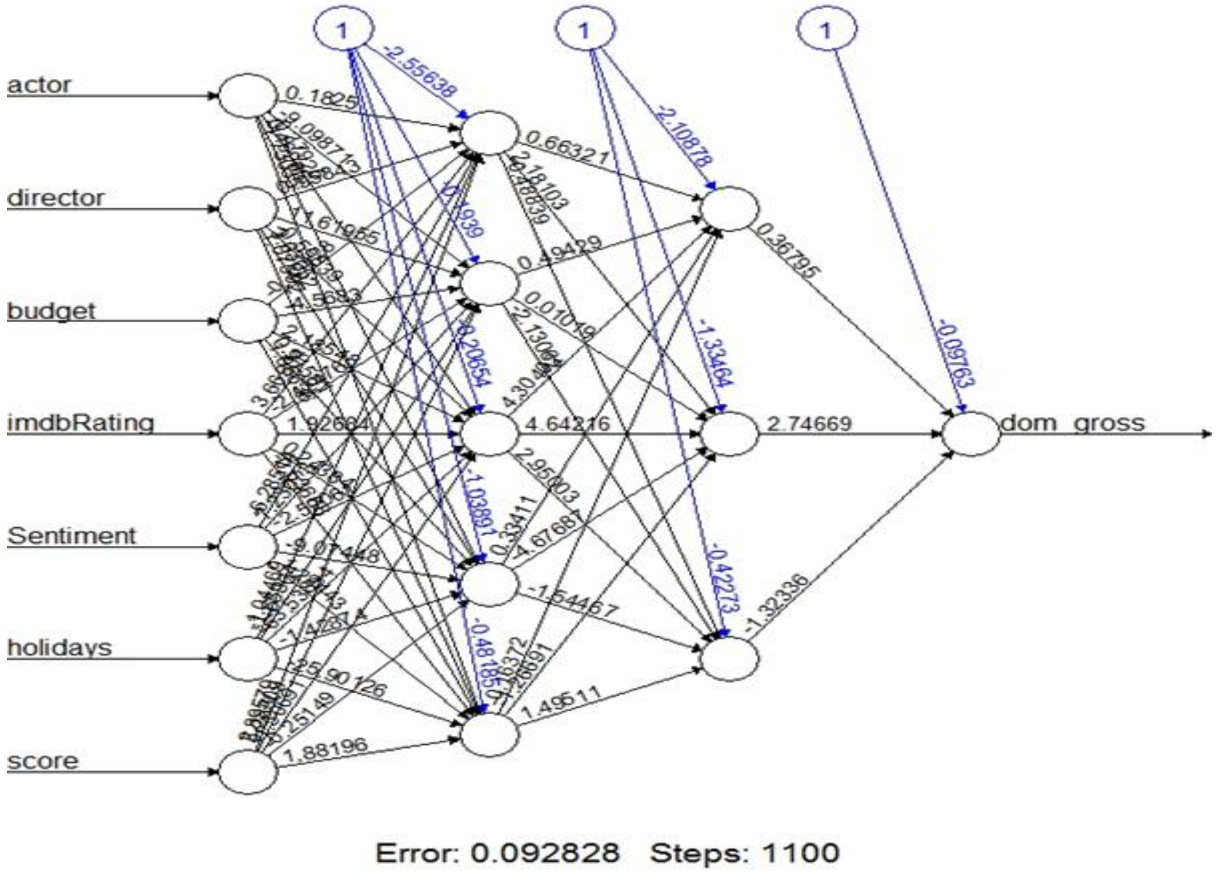
Movie : Skyfall

Predicted revenue : 307034205.8

List of movie to be predicted further:



Calculating Revenue using NeuralNetwork:



Chapter 8:

Conclusion

8.1 Limitations :

Data extraction from YouTube:

Youtube API allows only 100 random comments from its page for a particular movie.

Sentiment analysis:

Cannot recognize the words written in any other language or written in slang language.

Data limitations :

Not all dataset items are readily available, most of the time we need to perform web crawling to get the desired data which increases the overhead of the system

8.2 Conclusion:

The sales performance of the near future is predicted using box-office data, internal factors such as star-cast and their association and external factors such as online reviews, movie ratings, holiday effect.

We learned prediction model like ANN. R-Studio software was used for prediction of revenue.

The users of the system will be the viewers and the managing team of box-office. Our prediction model forecasts the revenue and can determine the ideal screen number based on our projected sales and also explains that how proper use of social networking websites be an accurate indicator for future outcomes.

8.3 Future Scope :

The Prediction system we generated gives the whole revenue at a time considering different factors before the movie is released. Further we can can work with predicting the revenue on weekly basis, giving the revenue week-wise comparing the actual comments of the movies after release. The output generated i.e the Revenue predicted can be shown in a graphical format on week-wise revenue prediction versus actual box-office collection. Also these predicted revenue figures can be input, along with the show timings to another system which will predict the show timings for which the profit is maximum and help box-office managers to lay down the strategy.

Appendix:

List of Figures

REFERENCES

1] Influence of social media on performance of movies

Shruti; Roy, S.D.; Wenjun Zeng.

Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on

Year: 2014

2]Role of different factors in predicting movie success

Bhave, A.; Kulkarni, H.; Biramane, V.; Kosamkar, P.

Pervasive Computing (ICPC), 2015 International Conference

3] Predicting the Near-Weekend Ticket Sales Using Web-Based ExternalFactors and Box-Office Data

Seonghoon Moon; Suman Bae; Songkuk Kim

Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Joint Conferences

Year: 2014, Volume: 2

4]An improved sentiment analysis of online movie reviews based onclustering for box-office prediction

Nagamma, P.; Pruthvi, H.R.; Nisha, K.K.; Shwetha, N.H.

Computing, Communication & Automation (ICCCA), 2015 International Conference on

Year: 2015

5] Prediction of Box Office Success of Movies Using Hype Analysis of Twitter Data

Sameer Thigale1 , Tushar Prasad 2, Ustat Kaur Makhija 3, Vibha Ravichandran

Year: 2014