Project Report:
Movie Recommendation System and Movie Data Visualizer

Vivekananda Adepu
Rohith Kumar Addagalla
Siva Ram Praneeth Vemulapalli

Project-Description:

Collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Applications of collaborative filtering typically involve very large data sets. The basic idea of CFR systems is that, if two users share the same interests in the past, e.g. they liked the same book or the same movie, they will also have similar tastes in the future. If, for example, user A and user B have a similar purchase history and user A recently bought a book that user B has not yet seen, the basic idea is to propose this book to user B.

Collaborative filtering methods have been applied to many kinds of data including: sensing and monitoring data, such as in mineral exploration, environmental sensing over large areas or multiple sensors; financial data, such as financial service institutions that integrate many financial sources; or in electronic commerce and web applications where the focus is on user data, etc. The collaborative filtering approach considers only user preferences and does not consider the features or contents of the items (books or movies) being recommended. In this project, to recommend movies, we are going use a large set of users preferences towards the movies from a publicly available movie rating dataset. The Shiny library allows the user to produce graphs and represent data visually so that the distribution of data based on ratings given to each movie, year of release and genre are being displayed to the user in the form of histograms.

Why R?

R provides a wide variety of statistical and graphical techniques, and is highly extensible. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

Features of R:

Data structure in R:

1. Supports virtually any type of data
2. Numbers, Characters, logicals(TRUE/FALSE)
3. Arrays are of virtually unlimited sizes
4. Simplest data structure: vectors and matrices
5. Lists: can contain mixed type of variables
6. Data frame: rectangular data set.

Vectors in R:

Vectors are used to store the elements of same type

```
> a=c(1,2,3,4) #c() is used to combine the list of elements in the vector
        > a
                [1] 1 2 3 4
        > a[3]
                [1] 3
        > a*2
                [1] 2 4 6 8
```

Lists:

List is a collection of elements. Elements in a list can be of different types. A list can contain numbers, characters, data frames and another list.

```
> id = 1:3
> name=c("Lee","Tom","Rachel")
> age =c(20,40,30)
> x= data.frame(id,name,age)
> lists = list(1,"a",c(2,3,4),x)
> lists
        [[1]]
        [1] 1
        [[2]]
        [1] "a"
        [[3]]
        [1] 2 3 4
        [[4]]
```

id	name	age
1	Lee	20
2	Tom	40
3	Rachel	30

```
> lists[[3]] #if a particular entity is to be read in a list then [[]] is used.
[1] 2 3 4
```

MATRIX:

A matrix stores elements of same type in a two-dimensional format.

```
> MatrixOne= matrix (1:10, nrow=2)
> MatrixOne
        [,1] [,2] [,3] [,4] [,5]
        [1,] 1 3 5 7 9
        [2,] 2 4 6 8 10
```

Various mathematical operations can be done on matrices.

```
> MatrixOne= matrix (1:10, nrow=2)
> MatrixTwo= matrix (11:20, nrow=2)
> MatrixOne

[,1] [,2] [,3] [,4] [,5]
[1,] 1 3 5 7 9
[2,] 2 4 6 8 10
> MatrixTwo
        [,1] [,2] [,3] [,4] [,5]
        [1,] 11 13 15 17 19
        [2,] 12 14 16 18 20
> MatrixOne+MatrixTwo
        [,1] [,2] [,3] [,4] [,5]
        [1,] 12 16 20 24 28
        [2,] 14 18 22 26 30
```

Data frames in R:

```
> id = 1:3
> name=c("Lee","Tom","Rachel")
> age =c(20,40,30)
> x= data.frame(id,name,age)
> x
```

id	name	age
1	Lee	20
2	Tom	40
3	Rachel	30

Data frames help to organize elements on a spreadsheet.

```
> nrow(x)       # determines the number of rows in a dataframe
        [1] 3
> ncol(x)       #determines the number of columns in a dataframe
        [1] 3
> dim(x)        #determines the number rows and columns in a dataframe
        [1] 3 3
> names(x) #returns the number of columns in a dataframe
        [1] "id" "name" "age"
> names(x[3]) #returns the name of a specific column in a data frame
        [1] "age"
```

Abstraction in R:

R has both control and data abstraction. Hence the readability and writability in R is simple. Some of the abstraction concepts are as follows:

Control Abstractions:

R has assignments operators, loops, functions etc. to support control abstractions. R has 3 types of methods to handle loops. Loops use decision making statements to control the loop structures. They are:

- **repeat loop**: Executes a sequence of statements multiple times and abbreviates the code that manages the loop variable.
- **while loop:** Repeats a statement or group of statements while a given condition is true. It tests the condition before executing the loop body.
- **for loop:** Like a while statement, except that it tests the condition at the end of the loop body.

Example program:

```
> x =c(1,2,3,4,5,6,7,8,9,10)
> square=0
> for(i in 1:5)
        { square= x[i]*x[i]
        print(square)
        }
                [1] 1
                [1] 4
                [1] 9
                [1] 16
                [1] 25
```

Data abstractions :

R provides data abstraction of basic, structured, and unit abstraction.
Vectors, arrays, and data frames are few of the examples for data abstraction.

Example programs using vectors:

```
> a=c(1,2,3,4)
> a
        [1] 1 2 3 4
> a[3]
        [1] 3
> a*2
        [1] 2 4 6 8
```

The readability and writability becomes a lot simpler because of the data and control abstraction in R.

SOFTWARE REQUIREMENTS:

TOOLS USED: RStudio
OPERATING SYSTEM: WINDOWS/Unix/MAC
LANGUAGES: R
DATASET: movies.csv, ratings.csv
CODES: server.R, ui.R
WEB BROWSER: Internet explorer

Movie Recommender and Movie Data Analyzer:

Libraries used:

The main libraries that are used in this project are:

1. Shiny

This library is used to create the UI and the components of it.
The methods used from this library are:

- fluidPage
- fluidRow
- titlePanel
- navbarPage
- tabPanel
- plotOutput
- Column etc...

2. Proxy

This library provides an extensible framework for the efficient calculation of auto- and cross-proximities, along with implementations of the most popular ones.
The method used from this library are:

- matrix

3. Recommenderlab

This library provides a research infrastructure to test and develop recommender algorithms including UBCF, IBCF, FunkSVD and association rule-based algorithms.
The methods used from this library are:

- Recommender
- Predict

4. reshape2

This library is used to flexibly restructure and aggregate data using just two functions: melt and 'dcast' (or 'acast').

5. Shinythemes

This library is used for themes with Shiny. Includes several Bootstrap themes from <http://bootswatch.com/>, which are packaged for use with Shiny applications.
The method used from this library are:

- shinythemes

6. ggplot2

This library is a system for 'declaratively' creating graphics, based on "The Grammar of Graphics". You provide the data, tell 'ggplot2' how to map variables to aesthetics, what graphical primitives to use, and it takes care of the details
The methods used from this library are:

- qplot

Application Details:
The application programmed in R allows its user to visualize dataset in the form of histograms that are distributed based on ratings, year-of-release, and genre
It contains the following tabs:

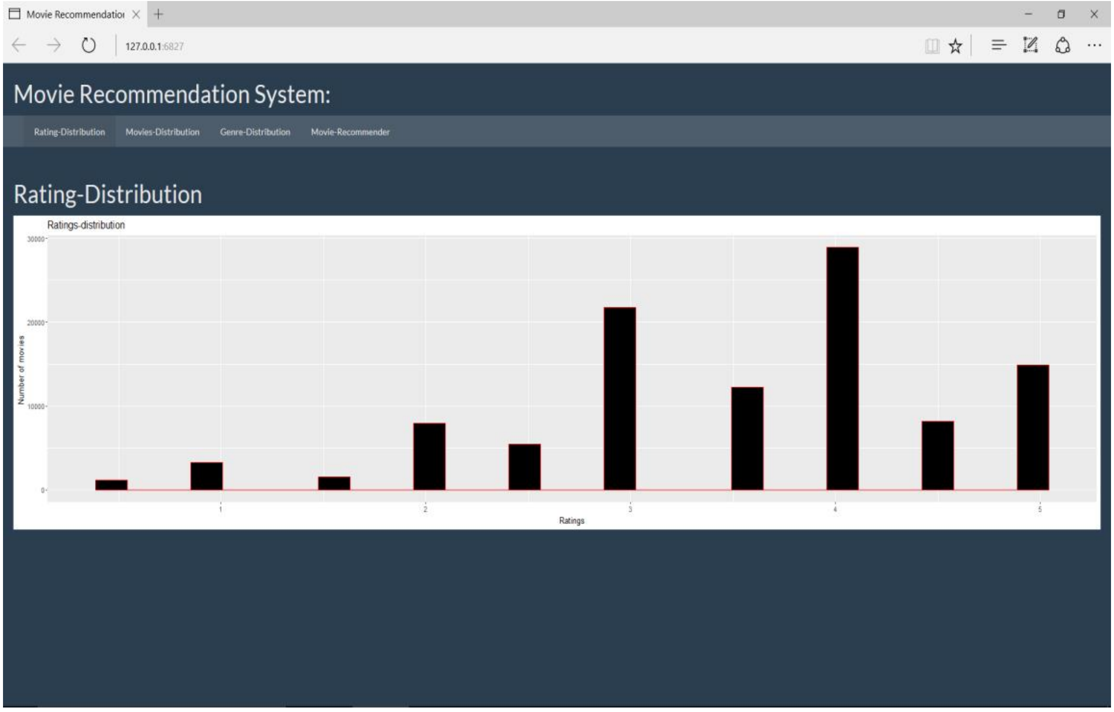
1. Rating-Distribution
2. Movies-Distribution
3. Genre-Distribution
4. Movie-Recommender

The functionalities of the tabs are as follows:

1. **Rating-Distribution:**

On clicking this tab, the user is directed to a page that displays the histogram in which all the movies are distributed based on the ratings where the x-axis consists of ratings ranging from 0-5 and the y-axis consists of number of movies present in the dataset.

It is shown below:



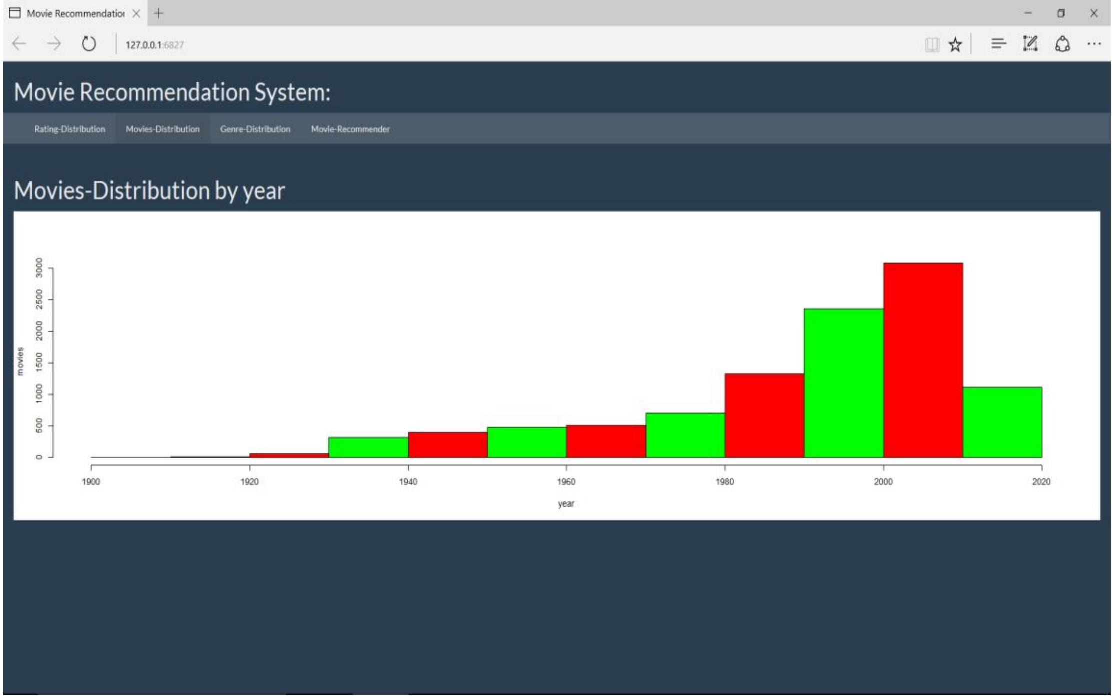
Movie Recommendation System:

Rating-Distribution

2. <u>Movies-Distribution:</u>

On clicking this tab, the user is directed to a page that displays the histogram in which all the movies are distributed based on the <u>year-of-release</u> where the x-axis consists of years ranging from 1990-2020 and the y-axis consists of <u>number-of-movies</u> present in the dataset.

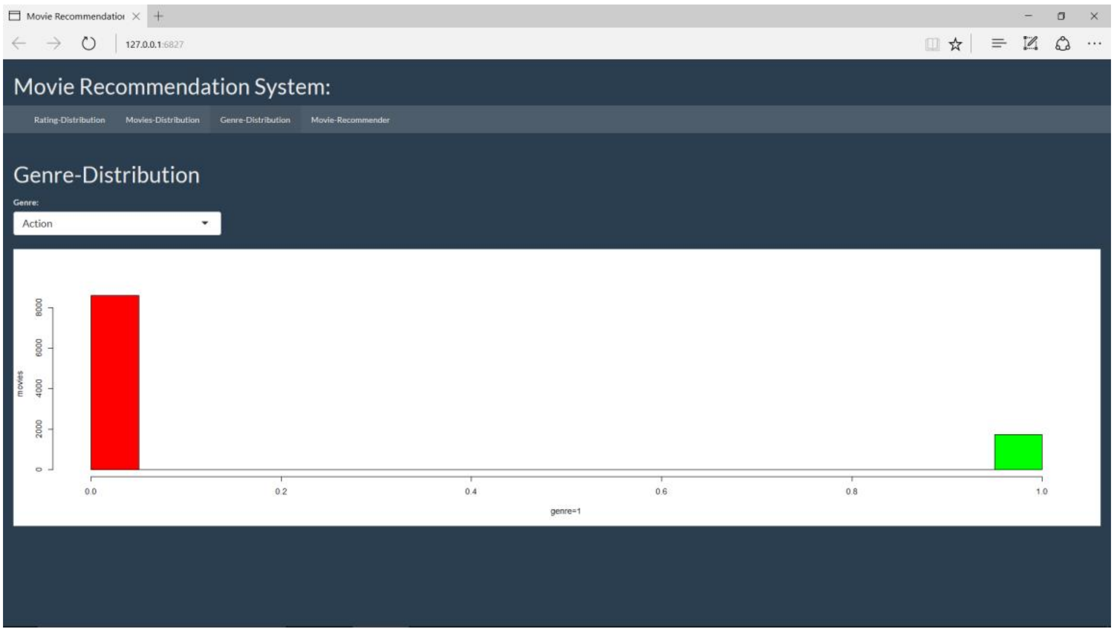
It is shown below:



3. <u>Genre-Distribution:</u>

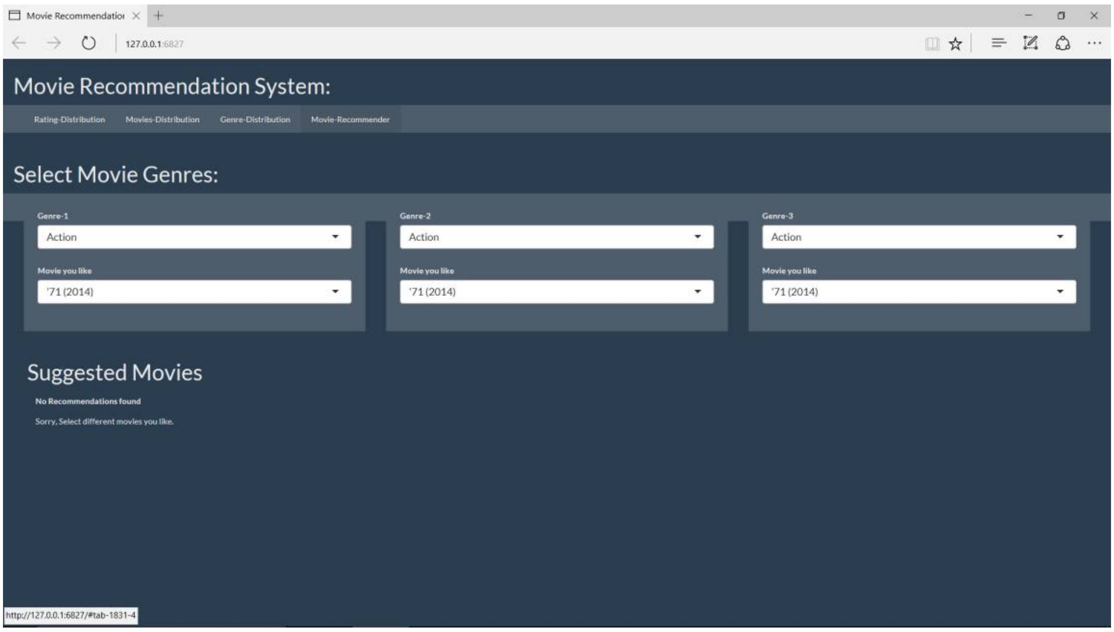
On clicking this tab, the user is directed to a page that displays the histogram in which all the movies are distributed based on the <u>genre</u> where the x-axis consists of genre value which is 1 when the movies belong to the genre that is selected ranging from 0-1 and the y-axis consists of <u>number-of-movies</u> present in the dataset.

It is shown below:


Genre: Action

4. Movie-Recommender:

On clicking this tab, the user is directed to a page that displays three tabs that have genre and the movies are changed, when the genre is selected.
It is shown below:



Movie Recommendation System:

Select Movie Genres:

Genre-1: Action
Movie you like: '71 (2014)

Genre-2: Action
Movie you like: '71 (2014)

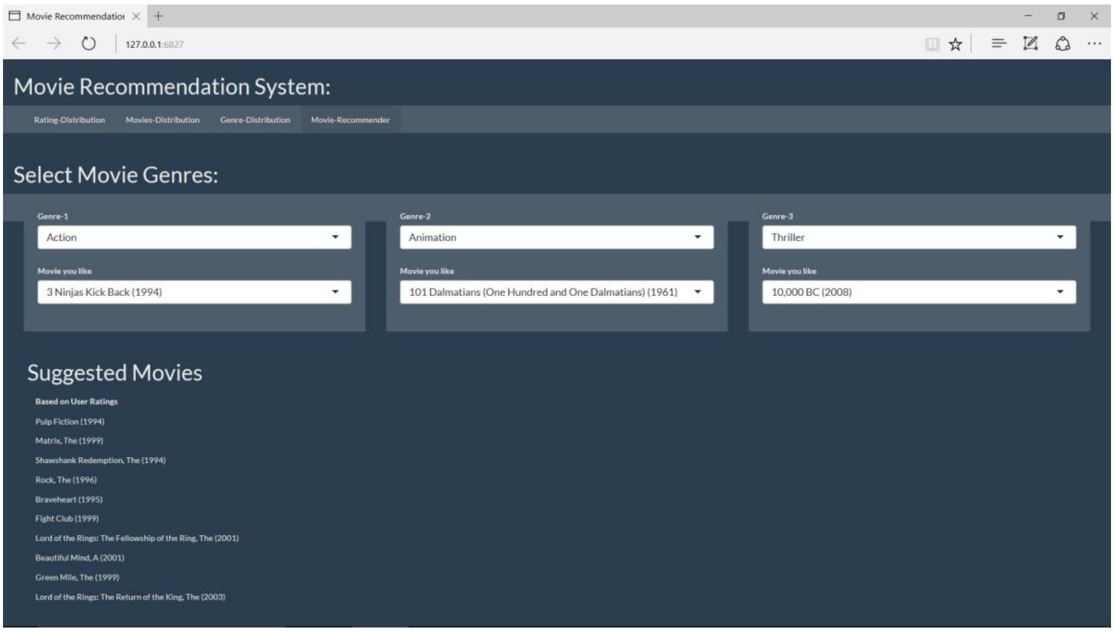
Genre-3: Action
Movie you like: '71 (2014)

Suggested Movies

No Recommendations found

Sorry, Select different movies you like.

After selecting the genre and movie from all the three panels, the application suggests the movie based on "User based collaborative filtering".
It is shown below:



Movie Recommendation System:

Select Movie Genres:

Genre-1: Action
Movie you like: 3 Ninjas Kick Back (1994)

Genre-2: Animation
Movie you like: 101 Dalmatians (One Hundred and One Dalmatians) (1961)

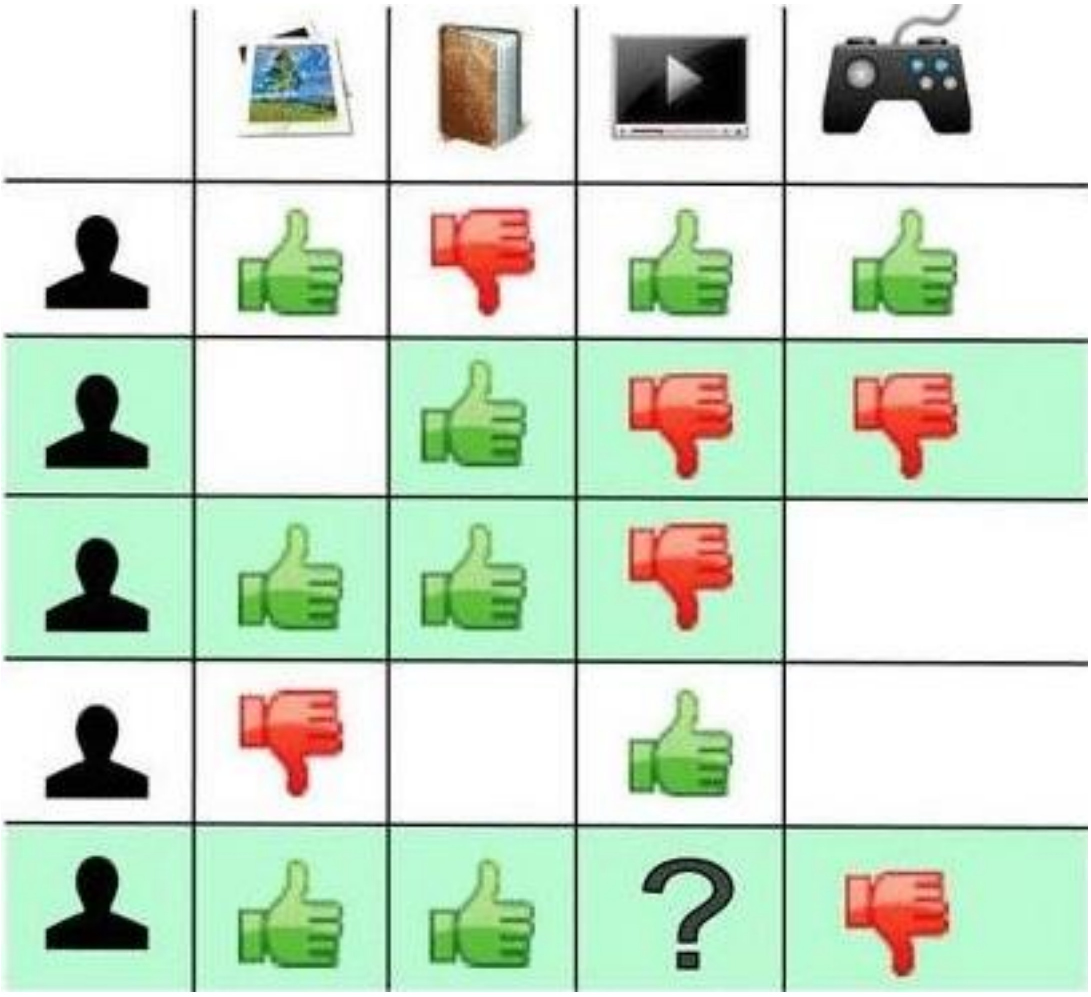
Genre-3: Thriller
Movie you like: 10,000 BC (2008)

Suggested Movies

Based on User Ratings

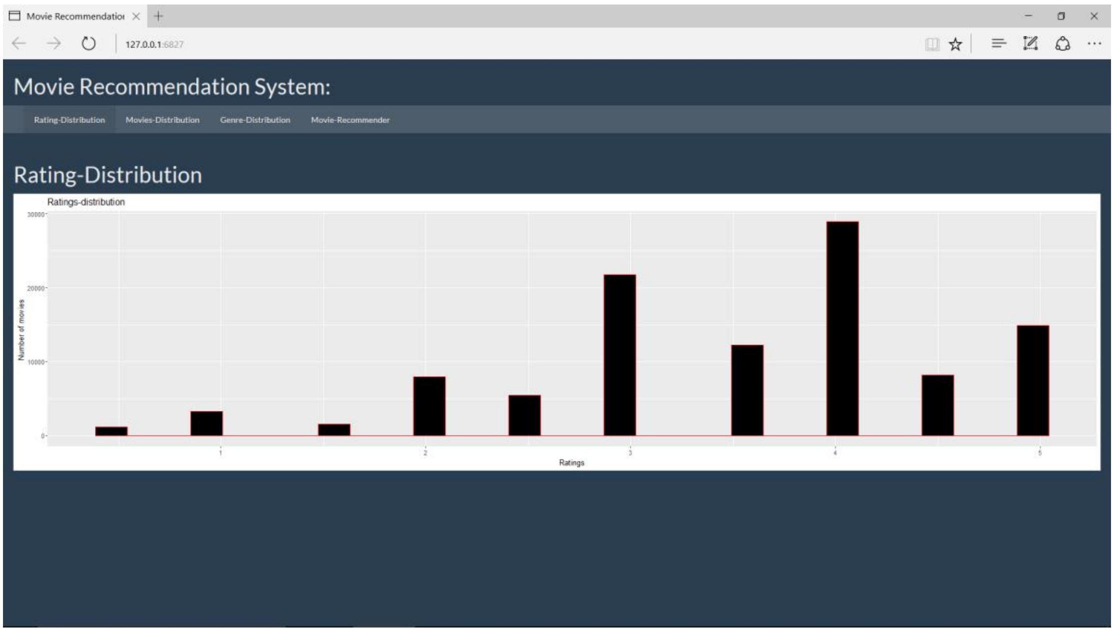
Pulp Fiction (1994)
Matrix, The (1999)
Shawshank Redemption, The (1994)
Rock, The (1996)
Braveheart (1995)
Fight Club (1999)
Lord of the Rings: The Fellowship of the Ring, The (2001)
Beautiful Mind, A (2001)
Green Mile, The (1999)
Lord of the Rings: The Return of the King, The (2003)

<u>User based collaborative filter:</u>
This concept can be explained by the diagram shown below.

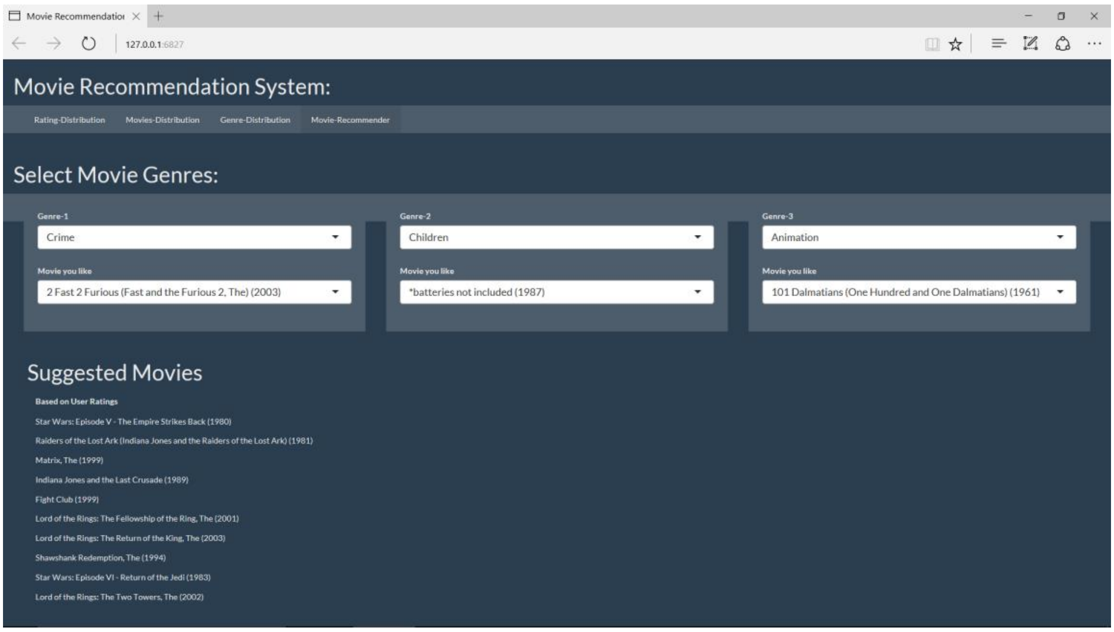


There are five users in the above diagram. The fifth user wants a suggestion about the third object which is represented as”?” in the diagram. The fifth user has liked first and second objects and disliked the fourth object. The fifth user’s interests are compared with the interests of the rest of the users. The user one has liked like the object one but object two has been disliked by the user two which has been liked by the user one. So, the user one’s interests are not considered for any suggestion for the user five. The second user has liked the object two and disliked the object four. User two’s interest are similar to the interests of the user five. So user two’s interests can be suggested to the user five. User three has like the object one and two, which are matching the interests of the user five. So, user three’s interests can also be suggested to the user five. Now the user four has disliked the object one which contradicts the interests of user five. So user four’s interests cannot be suggested to the user five. So, finally the user five’s interests are matching with the interests of the user two and three. User two and three have dislike the object three. So , the user five is most likely to get a suggestion saying that he/she might dislike the object three.

The user interface of our project is shown in the below diagram:



The user can select any genre and any movie. The sample photo is shown below:



Select Movie Genres:

Genre-1: Crime

Movie you like: 2 Fast 2 Furious (Fast and the Furious 2, The) (2003)

Genre-2: Children

Movie you like: *batteries not included (1987)

Genre-3: Animation

Movie you like: 101 Dalmatians (One Hundred and One Dalmatians) (1961)

Suggested Movies

Based on User Ratings

Star Wars: Episode V - The Empire Strikes Back (1980)
Raiders of the Lost Ark (Indiana Jones and the Raiders of the Lost Ark) (1981)
Matrix, The (1999)
Indiana Jones and the Last Crusade (1989)
Fight Club (1999)
Lord of the Rings: The Fellowship of the Ring, The (2001)
Lord of the Rings: The Return of the King, The (2003)
Shawshank Redemption, The (1994)
Star Wars: Episode VI - Return of the Jedi (1983)
Lord of the Rings: The Two Towers, The (2002)

There are three tabs present containing a list of genres and a list of movies. In each tab, the user selects a genre and movie. Now a list of movies will appear under the “Suggested Movies”. This list of movies are liked by the users who liked the above three movies that have been mentioned.

References:

- The Dataset for the movies and user reviews can be downloaded from: http://grouplens.org/datasets/movielens/latest/.
- The User Interface components have been downloaded from: https://shiny.rstudio.com/gallery/.
- R-3.4.0 for Windows (32/64 bit) and its packages have been downloaded from:
  https://cran.r-project.org/bin/windows/base/.
- The installation of RStudio is given in detail in the Readme.pdf.