**Web Scraping from the net and Data Mining**

Objective: To scrape unorganized data from the imdb website, organize it into a structured format and derive insight from it. The insight to be derived is, what was the most popular genre of film in 2017(as per the data scraped here) ?

Note:You can find different insights as per the data you fetch from websites and your business requirement.

Prerequisites: R studio/R, CSS selector gadget for your browser( if you do not know how to inspect CSS)

Code:

#Install the package rvest first

 library(rvest)

url='<https://www.imdb.com/search/title?year=2017&title_type=feature&>'

#Reading the HTML code from the website

webpage=read\_html(url)

#CSS selectors to scrape the ranking section

rank\_data\_html<- html\_nodes(webpage,'.text-primary')

class(rank\_data\_html)

#Convert the ranking data to text

rank\_data=html\_text(rank\_data\_html)

head(rank\_data)

class(rank\_data)

#Convert to numerical format

rank\_data<-as.numeric(rank\_data)

class(rank\_data)

head(rank\_data)

class(rank\_data)

#Using CSS selectors to  scrap the title selector

title\_data\_html<-html\_nodes(webpage,'.lister-item-header a')

class(title\_data\_html)

title\_data<-html\_text(title\_data\_html)

head(title\_data)

class(title\_data)

as.data.frame(title\_data)

#Using CSS selectors to scrap the movie runtime  section

runtime\_data\_html=html\_nodes(webpage,'.text-muted .runtime')

runtime\_data<-html\_text(runtime\_data\_html)

head(runtime\_data)

runtime\_data<-gsub(" min",'',runtime\_data)

as.numeric(runtime\_data)

runtime\_data<-as.data.frame(runtime\_data)

#Using CSS selectors to scrap the genre

genre\_data\_html<-html\_nodes(webpage,'.genre')

genre\_data<-html\_text(genre\_data\_html)

genre\_data

# library(stringr)

# str\_trim(genre\_data)

# genre\_data<-as.data.frame(str\_trim(gsub("\n",'',genre\_data)))

#Data Preprocessing

genre\_data=gsub("\n","",genre\_data)

#Removing excess spaces

genre\_data=gsub(" ","",genre\_data)

#Taking only the first genre of each movie

genre\_data=gsub(",.\*$",'',genre\_data)

#Converting each genre from text to factor

genre\_data=as.factor(genre\_data)

head(genre\_data)

#######combine into a data frame

movies\_df<-data.frame(Rank=rank\_data,Title=title\_data,Genre=genre\_data)

head(movies\_df)

barplot(table(movies\_df$Genre))

#We found the most popular category of film in 2017 i.e. Action