Отчёт по учебной практике

Завгороднего Александра

19 июля 2017

Тема: “Анализ наиболее предпочтительных для просмотра кинопремьер по параметрам на сайте imdb.com”

Ход работы

Первая часть работы посвящена созданию таблицы FAMOUSES с полями имя, место в общемировом рейтинге и баллы для подсчёта коэффициента. Для этого обращаемся к странице “Самые популярные мужчины и женщины”. Люди из этого списка являются режиссёрами или актёрами. Извлекаем необходимые данные и создаём таблицу.

```
library(XML)
library(RCurl)
```

```
## Loading required package: bitops
```

```
library(bitops)
FAMOUSES = data.frame()
j<-1
options(max.print=999999)
for (i in 1:50)
{
  strj <- toString(j)
  page<-paste("http://www.imdb.com/search/name?gender=male,female&ref_=nv_tp_cel_1&start=",strj)

  str<-page
  print(str)
  page<-readLines(str)
```

```
page<-enc2native(page)
page <- iconv(page, from = "UTF-8", to = "CP1251")
numbers = grep('<td class="name">',page)
name = page[numbers+1]
name <- sub('.*/\">','',name)
name <- sub('</a>','',name)

numbers <- grep('<td class="number">',page)
number = page[numbers]
number <- sub('.*\">','',number)
number <- sub('.<.*','',number)

pointcounter = 0
for (k in 1:length(name))
{
  num<-as.integer(number[k])
  if (num<2000)
  {
    pointcounter <- 1
  }
  if (num<1000)
  {
    pointcounter <- 2
  }
  if (num<750)
  {
    pointcounter <- 3
  }
  if (num<500)
  {
    pointcounter <- 4
  }
  if (num<250)
  {
    pointcounter <- 5
  }
  FAMOUSES <- rbind(data.frame(name[k], num, pointcounter),FAMOUSES)
```

```
  }
  j = j + 50
}
print(FAMOUSES)
```

Второй шаг - извлечение данных для основной таблицы PREMIERES, в которой будут поля: название фильма, режиссёр, актёры, возрастное ограничение, рейтинг ожиданий, популярность и дополнительный рейтинг.

```
library(XML)
library(RCurl)
library(bitops)
PREMIERES = data.frame()
page = readLines("http://www.imdb.com/movies-coming-soon/")
page <- enc2native(page)
numbers = grep('<a href="/title',page)
href = page[numbers]
doublicate <- FALSE
j=1



#получение фильмов со второй строки
numbers = grep('<a href="/movies-coming-soon/.*ref_=cs_dt_nx\"',page)
otherpage = page[numbers]

secondpage<-otherpage[1]
secondpage<-sub('.*href=\"','http://www.imdb.com',secondpage)
secondpage<-sub('dt_nx\"','dt_nx',secondpage)
print(secondpage)
secondpage = readLines(secondpage)
secondpage <- enc2native(secondpage)
numbers = grep('<a href="/title',secondpage)
href2 = secondpage[numbers]

#объединение двух страниц
hrefs = character()
for (i in 1:length(href))
```

```
{
  hrefs[j] <- href[i]
  j = j + 1
}
for (i in 1:length(href2))
{
  hrefs[j] <- href2[i]
  j = j + 1
}

j=1

for (i in 1:length(hrefs))
{
  if (doublicate == FALSE)
  {
    str<-hrefs[i]
    str<- sub('.*title','http://www.imdb.com/title', str)
    str<- sub('/?ref_.*.','', str)
    str<-substr(str,1,nchar(str)-1)
    hrefs[i]<-str
    doublicate <- TRUE

    page <- readLines(hrefs[i])
    page <- enc2native(page)
    page <- iconv(page, from = "UTF-8", to = "CP1251")
    #Получение имени
    numbers = grep('<h1 itemprop="name"',page)
    name = page[numbers]
    name <-sub('.*=\"\">','',name)
    name <-sub(' .*','',name)
    print(paste("Название фильма", j, name))
    #Получение рейтинга
    numbers = grep('span itemprop="ratingValue',page)
    rating_of_expectations = page[numbers]
    rating_of_expectations <-sub('.*ratingValue\">','',rating_of_expectations)
    rating_of_expectations <-sub('</span>.*','',rating_of_expectations)
```

```
    print(paste("Рейтинг", j, rating_of_expectations))
    #Получение режиссера
    numbers = grep('span itemprop="director"',page)
    director = page[numbers + 2][1]
    director <-sub('.*itemprop=\"name\">','',director)
    director <-sub('</span>.*','',director)
    print(paste("Режиссер", j, director))
    #Получение актеров
    numbers = grep('<td class="itemprop" itemprop="actor" ',page)
    #print(page[numbers + 2])
    actors <- sub('.*itemprop=\"name\">','',page[numbers + 2])
    actors <- sub('</span>.*','',actors)

    actor1 <- actors[1]
    actor2 <- actors[2]
    actor3 <- actors[3]
    actor4 <- actors[4]
    print(paste("Актер1", j, actor1))
    print(paste("Актер2", j, actor2))
    print(paste("Актер3", j, actor3))
    print(paste("Актер4", j, actor4))
    actor1 <- ifelse(is.na(actor1), "noactor", actor1)
    actor2 <- ifelse(is.na(actor2), "noactor", actor2)
    actor3 <- ifelse(is.na(actor3), "noactor", actor3)
    actor4 <- ifelse(is.na(actor4), "noactor", actor4)


    #возраст
    numbers = grep('itemprop="contentRating"',page)
    age = page[numbers][1]
    age <-sub('.*\">','',age)
    print(paste("Возрастное ограничение", j, age))
    agenum <-0


    age <- ifelse(is.na(age), "G", age)
```

```
    if (age=="PG")
    {
      agenum <-6
    }
    if (age=="PG-13")
    {
      agenum <-12
    }
    if (age=="R")
    {
      agenum <-16
    }
    if (age=="18+")
    {
      agenum <-18
    }
    print(agenum)

    #pop
    numbers = grep('itemprop="ratingCount"',page)
    popularity <- page[numbers]
    popularity <-sub('.*ratingCount\">','',popularity)
    popularity <-sub('</span>.*','',popularity)
    print(paste("Поп", popularity))

    #оценка Metacritic
    numbers = grep('class="metacriticScore', page)
    metacritic = page[numbers + 1]
    metacritic <-sub('<span>','',metacritic)
    metacritic <-sub('</span>','',metacritic)
    print(paste("Оценка Metacritic", j, metacritic))
    if (length(metacritic)==0)
    {
      metacritic = "0"
    }
    if (length(popularity)==0)
    {
```

```
      popularity = "0"
    }
    if (length(rating_of_expectations)==0)
    {
      rating_of_expectations = "0"
    }

    PREMIERES <- rbind(data.frame(Name=name,Director=director,Role1=actor1,Role2=actor2,Role3=actor3,Role4=actor4
,Age=agenum,Rating_of_expectations=rating_of_expectations,Popularity=popularity,Metacritic=metacritic),PREMIERES)

    j=j+1
  }
  else
  {
    doublicate <- FALSE
  }
}
print(PREMIERES)
```

В приведённом выше участке кода строчка:

```
PREMIERES <- rbind(data.frame(Name=name,Director=director,Role1=actor1,Role2=actor2,Role3=actor3,Role4=actor4,Age
=agenum,Rating_of_expectations=rating_of_expectations,Popularity=popularity,Metacritic=metacritic),PREMIERES)
```

отвечает за создание таблицы PREMIERES. Далее идёт анализ фильмов на основе двух таблиц FAMOUSES и PREMIERES. Подсчитывается коэффициент для каждого фильма.

```
library(knitr)
# c(...) - concatnation
coefficient <- c(1:nrow(PREMIERES))
# count k
# nrow - number of rows
for(i in 1:nrow(PREMIERES))
{
  k <- 0
  if (PREMIERES$Age[i] == 18)
```

```
  {
    k <- k - 4
  }
  director <- toString(PREMIERES$Director[i])
  actor1 <- toString(PREMIERES$Role1[i])
  actor2 <- toString(PREMIERES$Role2[i])
  actor3 <- toString(PREMIERES$Role3[i])
  actor4 <- toString(PREMIERES$Role4[i])
  name <- toString(PREMIERES$Name[i])
  for(j in 1:nrow(FAMOUSES))
  {
    famous_name<-toString(FAMOUSES$name.k.[j])
    # adding points for director and roles
    if(director == famous_name || actor1 == famous_name || actor2 == famous_name || actor3 == famous_name || acto
r4 == famous_name)
    {
      k <- k + as.integer(FAMOUSES$pointcounter[j])
      point<-FAMOUSES$pointcounter[j]
      print(point)
    }
  }

  # adding points for rating of expectations
  k <- k + as.numeric(toString(PREMIERES$Rating_of_expectations[i]))
  pop<-PREMIERES$Popularity[i]
  pop <-sub(',','.',pop)
  # adding points for popularity
  k <- k + as.numeric(pop) / 100
  # adding points for metacritic
  k <- k + as.numeric(toString(PREMIERES$Metacritic[i])) / 10
  coefficient[i] <- k
}
```

Последний этап - добавление столбца с коэффициентами в таблицу, сортировка всей таблицы по добавленному столбцу.

```
# adding column coefficient to table PREMIERES
PREMIERES <- cbind(PREMIERES, coefficient)
```

```
print(PREMIERES)
```

```
PREMIERES <- PREMIERES[order(PREMIERES$coefficient, decreasing = TRUE),]
print(PREMIERES)
```

```
##                                     Name             Director
## 15                        Ветреная река       Taylor Sheridan
## 24                    Взрывная блондинка          David Leitch
## 11                      Ingrid Goes West           Matt Spicer
## 4                           Tulip Fever       Justin Chadwick
## 14                  Проклятие Аннабель 2     David F. Sandberg
## 19                          Темная башня         Nikolaj Arcel
## 20                           Sage femme        Martin Provost
## 23 An Inconvenient Sequel: Truth to Power           Bonni Cohen
## 8                  The Hitman's Bodyguard        Patrick Hughes
## 7                            Patti Cake$         Geremy Jasper
## 16                                  Step         Amanda Lipitz
## 9                     The Trip to Spain  Michael Winterbottom
## 3                           Beach Rats         Eliza Hittman
## 22                             Menashe    Joshua Z Weinstein
## 13                      The Glass Castle Destin Daniel Cretton
## 21                         Иллюзия любви         Nicole Garcia
## 12         The Nut Job 2: Nutty by Nature           Cal Brunker
## 10        The Only Living Boy in New York             Marc Webb
## 1                                   Gook           Justin Chon
## 2                                   Gook           Justin Chon
## 5                             Polaroid         Lars Klevberg
## 17                         Midnight Sun           Scott Speer
## 18                              Detroit       Kathryn Bigelow
## 25                       The Emoji Movie          Tony Leondis
## 6                             All Saints           Steve Gomer
##                 Role1                  Role2               Role3
## 15    Elizabeth Olsen           Jon Bernthal       Jeremy Renner
## 24     Sofia Boutella        Charlize Theron        James McAvoy
```

```
## 11       Aubrey Plaza        Elizabeth Olsen  O'Shea Jackson Jr.
## 4     Cara Delevingne        Alicia Vikander         Dane DeHaan
## 14 Alicia Vela-Bailey           Miranda Otto    Stephanie Sigman
## 19   Katheryn Winnick             Idris Elba Matthew McConaughey
## 20  Catherine Deneuve         Catherine Frot     Olivier Gourmet
## 23            Al Gore           Barack Obama     Donald J. Trump
## 8         Gary Oldman          Ryan Reynolds   Samuel L. Jackson
## 7  Danielle Macdonald        Bridget Everett Siddharth Dhananjay
## 16        Paula Dofat          Cori Grainger       Tayla Solomon
## 9        Steve Coogan             Rob Brydon        Marta Barrio
## 3    Harris Dickinson     Madeline Weinstein          Kate Hodge
## 22     Menashe Lustig         Yoel Falkowitz             noactor
## 13        Brie Larson            Naomi Watts     Woody Harrelson
## 21   Marion Cotillard           Louis Garrel     Alex Brendemuhl
## 12        Will Arnett        Katherine Heigl        Maya Rudolph
## 10    Kate Beckinsale           Jeff Bridges      Pierce Brosnan
## 1        Simone Baker            Justin Chon    Curtiss Cook Jr.
## 2        Simone Baker            Justin Chon    Curtiss Cook Jr.
## 5    Kathryn Prescott          Katie Stevens    Madelaine Petsch
## 17       Bella Thorne Patrick Schwarzenegger          Rob Riggle
## 18        John Boyega           Will Poulter         Algee Smith
## 25       James Corden           Maya Rudolph       Steven Wright
## 6          Cara Buono           John Corbett        Barry Corbin
##                 Role4 Age Rating_of_expectations Popularity Metacritic
## 15   Martin Sensmeier  16                    7.5        693         73
## 24     Bill Skarsgard  16                    7.3      1,353         68
## 11      Wyatt Russell  16                    7.0        226         65
## 4     Christoph Waltz  16                    7.2        203          0
## 14    Talitha Bateman  16                    8.0        517         71
## 19 Jackie Earle Haley   0                      0          0          0
## 20   Quentin Dolmaire   0                    7.2        453         72
## 23            noactor   0                    5.0        715         66
## 8         Salma Hayek  16                      0          0          0
## 7      Mamoudou Athie  16                    6.1        467         71
## 16            noactor   6                    7.2         96         90
## 9       Claire Keelan   0                    7.5        259         70
## 3           Neal Huff  16                    6.8        129         79
```

```
## 22            noactor   6                 6.3        123         82
## 13         Sarah Snook  12                   0          0          0
## 21      Brigitte Rouan  18                 6.7      1,951         44
## 12         Jackie Chan   6                   0          0          0
## 10          Debi Mazar  16                   0          0          0
## 1             David So   0                 8.0        112          0
## 2             David So   0                 8.0        112          0
## 5          Javier Botet   0                   0          0          0
## 17      Quinn Shephard   0                   0          0          0
## 18      Jacob Latimore  16                   0          0          0
## 25  Jennifer Coolidge   6                   0          0          0
## 6          David Keith   6                   0          0          0
##    coefficient
## 15    35.73000
## 24    33.11353
## 11    30.76000
## 4     27.23000
## 14    24.27000
## 19    20.00000
## 20    18.93000
## 23    18.75000
## 8     18.00000
## 7     17.87000
## 16    17.16000
## 9     17.09000
## 3     15.99000
## 22    15.73000
## 13    15.00000
## 21    11.11951
## 12    11.00000
## 10    10.00000
## 1      9.12000
## 2      9.12000
## 5      9.00000
## 17     5.00000
## 18     3.00000
```

```
## 25     2.00000
## 6      1.00000
```