 

**MBA 教育管理中心**

电话：029-82669064

传真：029-82668069

管理学院专业学位大作业

课程名称《商务大数据》

专业名称： MBA

班 级： 8146班 第5组

学号/姓名： 赵双喜3118426003

龙佳星3118426048董 璞3117580074

李 娜3118426053常 茵3118426016

谢舒文3118426014李 茜3118426026

赵丽轩3118426004

任课教师： 王 尧

提交时间： 2019年 7 月 8 日

# 目录

目录

[目录 2](#_Toc13505256)

[一、处理Package 2](#_Toc13505257)

[二、数据爬取 3](#_Toc13505258)

[三、处理数据 4](#_Toc13505259)

[四、查看结果 5](#_Toc13505260)

 爬豆瓣电影Top250

# 一、处理Package

#install.packages('xml2')

library("xml2")

#install.packages("rvest")

library("rvest")

#install.packages("dplyr")

library("dplyr")

#install.packages("stringr")

library("stringr")

#install.packages("httr")

library("httr")

# 二、数据爬取

#对爬取页数进行设定并创建数据框。

i=1:10

movie\_inf=data.frame()

#利用for循环封装爬虫代码，进行批量抓取。

for (i in 1:10){

#发现url规律，利用字符串函数进行url拼接并规定编码

web=read\_html(str\_c('https://movie.douban.com/top250?',i),encoding='UTF-8')

#提电影rank

movie\_rank=web %>% html\_nodes('em') %>% html\_text

#提电影名信息

movie\_name=web %>% html\_nodes('.title:nth-child(1)') %>% html\_text

#提电影别名信息

movie\_other\_name=web %>% html\_nodes('.other') %>% html\_text

#movie\_other\_name=str\_replace\_all(movie\_other\_name,' ',' ')

#提电影rating\_num

movie\_rating\_num=web %>% html\_nodes('.rating\_num') %>% html\_text

#提电影inq

movie\_inq=web %>% html\_nodes('.inq') %>% html\_text

#提电影评论人数

movie\_num\_data =web %>% html\_nodes('.rating\_num~ span') %>% html\_text

movie\_num\_data <- str\_match(movie\_num\_data, "[0-9]\*")

movie\_num\_data <- as.numeric(movie\_num\_data)

movie\_num\_data <- movie\_num\_data[!is.na(movie\_num\_data)]

#提取导演信息

movie\_director =web %>% html\_nodes('.bd p:nth-child(1)') %>% html\_text

movie\_director=movie\_director[2:length(movie\_director)]

#创建数据框存储以上信息

movie=data.frame(movie\_rank,movie\_name,movie\_other\_name,movie\_rating\_num,movie\_inq,movie\_num\_data,movie\_director)

movie\_inf=rbind(movie\_inf,movie)

}

#把数据写入CSV文档

write.csv(movie\_inf,file='./movie\_inf8.csv')

a=read.csv('movie\_inf8.csv',encoding="UTF-8")

View(a)

# 三、处理数据

a=read.csv('movie\_inf8.csv')

#处理other name

name= gsub('<U+00A0>/<U+00A0>', "", a$movie\_other\_name,fixed = TRUE)

a$movie\_other\_name=name

#拆分出来导演

director= gsub('<U+00A0>', ";", a$movie\_director,fixed = TRUE)

split\_d<-str\_split(director,";;;")

b<-sapply(split\_d,"[",1)

c<-sapply(split\_d,"[",2)

b=gsub('导演:', "", b,fixed = TRUE)

b=gsub('<U+00F6>m', "", b,fixed = TRUE)

View(b)

a$director\_final=b

f=gsub('...', "割",c,fixed = TRUE)

split\_f<-str\_split(f,"割")

f<-sapply(split\_f,"[",1)

f=gsub('主演:', "", f,fixed = TRUE)

a$actor\_final=f

g<-sapply(split\_f,"[",2)

h=gsub(';/;', "割",g,fixed = TRUE)

split\_h<-str\_split(h,"割")

h1<-sapply(split\_h,"[",1)

h2<-sapply(split\_h,"[",2)

h3<-sapply(split\_h,"[",3)

a$year=h1

a$location=h2

a$type=h3

write.csv(a,file='./movie\_inf8.csv')

a=read.csv('./movie\_inf8.csv')

View(a)

a=a[,-9]

# 四、查看结果





