作业：

1.library(rvest)

url<-'http://sem.buaa.edu.cn/szdw/jsbd.htm'

web<-read\_html(url)

professor\_names<- web %>%

html\_nodes(xpath = '//div[@class="fak-list"]//strong/span/span')%>%

html\_text()

professor\_interest<-web %>%

html\_nodes(xpath = '//div[@class="fak-list"]//p/span/span[1]')%>%

html\_text()

emails<-web %>%

html\_nodes(xpath = '//div[@class="fak-list"]//p/span/span[2]')%>%

html\_text()

link<-web %>%

html\_nodes(xpath = '//div[@class="fak-list"]//td[1]/a')%>%

html\_attr('href')

link<-c(link[1:2],NA,link[3:40])

professor\_information=data.frame(professor\_names,professor\_interest,emails,link)

2.

library(rvest)

movie\_detail<-c()

for(i in 0:9){

p<-as.character(25\*i)

url<-sprintf('https://movie.douban.com/top250?start=%s&filter=',p)

web<-read\_html(url)

movie\_title<-web%>%

html\_nodes(xpath='//div[@class="hd"]/a/span[1]')%>%

html\_text()

movie\_director<-web%>%

html\_nodes(xpath='//div[@class="bd"]/p[1]')%>%

html\_text()

movie\_star<-web%>%

html\_nodes(xpath='//div[@class="star"]/span[2]')%>%

html\_text()

movie\_evaluate<-web%>%

html\_nodes(xpath='//div[@class="star"]/span[4]')%>%

html\_text()

movie\_details=data.frame(movie\_title,movie\_director,movie\_star,movie\_evaluate)

movie\_top250<-c(movie\_detail,movie\_details)

}

3.

library(xml2)

library(rvest)

newdata<-matrix(nrow=250,ncol=9)

for(i in 1:10)

{

ii<-(i-1)\*25

url<-paste("https://movie.douban.com/top250?start=",ii,"&filter=",sep = "")

web<- read\_html(url, encoding = "utf-8")

movie\_title <- web %>%

html\_nodes(xpath = '//div[@class="hd"]/a/span[1]') %>%

html\_text()

new\_movie\_title<-iconv(movie\_title,"UTF-8","gbk")

#鎸囨爣2锛氶摼鎺?

movie\_link <- web %>%

html\_nodes(xpath = '//div[@class="hd"]/a') %>%

html\_attr('href')

movie\_people<-web %>%

html\_nodes(xpath = '//div[@class="bd"]/p[1]/text()[1]') %>%

html\_text()

new\_movie\_people<-iconv(movie\_people,"UTF-8","UTF-8")

new\_movie\_people2<-gsub("([\n ])", "", new\_movie\_people)

new\_movie\_people3<-gsub("([\u00A0])", " ", new\_movie\_people2)

movie\_type<-web %>%

html\_nodes(xpath = '//div[@class="bd"]/p[1]/text()[2]') %>%

html\_text()

new\_movie\_type<-iconv(movie\_type,"UTF-8","UTF-8")

new\_movie\_type2<-gsub("([\n ])", "", new\_movie\_type)

new\_movie\_type3<-gsub("([\u00A0])", " ", new\_movie\_type2)

movie\_rate<-web %>%

html\_nodes(xpath = '//div[@class="star"]/span[@class="rating\_num"]') %>%

html\_text()

movie\_ratenum<-web %>%

html\_nodes(xpath = '//div[@class="star"]/span[4]') %>%

html\_text()

new\_movie\_ratenum<-iconv(movie\_ratenum,"UTF-8","gbk")

newdata[((i-1)\*25+1):((i-1)\*25+25),1]<-new\_movie\_title

newdata[((i-1)\*25+1):((i-1)\*25+25),2]<-movie\_link

newdata[((i-1)\*25+1):((i-1)\*25+25),3]<-new\_movie\_people3

for(j in 1:25)

{

newdata[((i-1)\*25+j),3]<strsplit(new\_movie\_people3[j],split=" ")[[1]][1]

newdata[((i-1)\*25+j),4]<-strsplit(new\_movie\_people3[j],split=" ")[[1]][2]

newdata[((i-1)\*25+j),5]<-strsplit(new\_movie\_type3[j],split=" / ")[[1]][1]

newdata[((i-1)\*25+j),6]<-strsplit(new\_movie\_type3[j],split=" / ")[[1]][2]

newdata[((i-1)\*25+j),7]<-strsplit(new\_movie\_type3[j],split=" / ")[[1]][3]

}

newdata[((i-1)\*25+1):((i-1)\*25+25),8]<-movie\_rate

newdata[((i-1)\*25+1):((i-1)\*25+25),9]<-new\_movie\_ratenum

}

colnames(newdata)<-c

write.csv(newdata,"newdata.csv")