1 Problem 1

1.1 Problem a

An ASCII character takes one bytes, and white space takes no bytes. The tmp1 file contains 1e6 ASCII characters and 1e6 commas, so the file is 2e6 bytes.

tmp2 contains 1e6 ASCII characters, white spaces and 1 ” n”,so it takes 1e6+1 bytes.

tmp3 is a .Rda file. It has 1e6 numbers,eaching occupying 7.6 bytes.

tmp4 is a csv file. Each string contains 18-19 digits,each occupying one bytes. So the file is around 18*1e6 bytes.

Since all numbers in tmp5 has been rounded into twodecimal points, each string contains 5-6 digits. So tmp5 is around 5*1e6 bytes.

1.2 Problem b

The file sisd default set ot compression when saving strings to .Rda form. So generally the file is smaller. Since tmp7 contains only ’a’, the compression level can be higher. So tmp7 is even smaller.

2 Problem 2

2.1 Problem a

In this function, I first get the Google Scholar ID by searching the firstname+lastname in the Google Scholar website. I then use the id to construct the corresponding citication page. I utilize htmlParse to get the citation page. The return is a list consisting of Scholar ID and the corresponding html.

```r
scholar<-function(firstname,lastname){
baseURL<-"http://scholar.google.com/citations?view_op=search_authors&mauthors="
name<-paste0(firstname,"+",lastname)

ending<-"&hl=en&oi=ao"
idURL<-paste0(baseURL,name,ending)

library(RCurl)
library(XML)

idpage<-htmlParse(idURL)

links<-getHTMLLinks(idpage)
index<-grep("user=.*&",links,perl=TRUE)
a<-links[index[1]]
userid<-substr(a,17,28)
citationURL<-paste0("http://scholar.google.com/citations?user=",userid,"&hl=en")
citationpage<-htmlParse(citationURL)
output<-list(userid,citationpage)
return(output)

}
```

2.2 Problem b

In this function, I first get the Google Scholar ID by searching the firstname+lastname in the Google Scholar website. I then use the id to construct the corresponding citication page. I utilize the readHTMLtable to get the citication table and refine it.

```
citationform<-function(firstname,lastname){
baseURL<-"http://scholar.google.com/citations?view_op=search_authors&mauthors="
name<-paste0(firstname,"+",lastname)

ending<-"&hl=en&oi=ao"
idURL<-paste0(baseURL,name,ending)

library(RCurl)
library(XML)

idpage<-htmlParse(idURL)
links<-getHTMLLinks(idpage)
index<-grep("user=.*&",links,perl=TRUE)
a<-links[index[1]]
userid<-substr(a,17,28)
citationURL<-paste0("http://scholar.google.com/citations?user=",userid,"&hl=en")
citationpage<-htmlParse(citationURL)

html<-readLines(citationURL)
tbls<-readHTMLTable(html,header=TRUE,as.data.frame=TRUE,which=2)
library(tidyr)
names(tbls)[1]="Title"
tblsnew<-separate(tbls,col="Title",c("Title","Authors","Journal","location","year"),sep=",",
tblsnew<-unite(tblsnew,Journal,Journal,location,sep=", ",remove=TRUE)
tblsnew<-tblsnew[,-c(4)]
return(head(tblsnew))
}

citationform("Geoffrey","Hinton")

## Loading required package:  bitops
## Warning in readLines(citationURL): ’http://scholar.google.com/citations?user=JicYPdAAAAAJ
##
## Attaching package:  ’tidyr’
## The following object is masked from ’package:RCurl’:
##
##    complete
## Warning:  Too many values at 6 locations:  10, 11, 14, 16, 17, 18
```

```
##                                                                            Title
## 1             Learning representations by back-propagating errorsDE Rumelhart
## 2          Learning internal representations by error-propagationDE Rumelhart
## 3          Learning internal representations by error propagationDE Rumelhart
## 4                                 Parallel distributed processingDE Rumelhart
## 5 Imagenet classification with deep convolutional neural networksA Krizhevsky
## 6                     A fast learning algorithm for deep belief netsGE Hinton
##          Authors
## 1      GE Hinton
## 2      GE Hinton
## 3      GE Hinton
## 4  JL McClelland
## 5    I Sutskever
## 6     S Osindero
##                                                                                          J
## 1                                                               RJ WilliamsNature 323,  5
## 2  RJ WilliamsParallel Distributed Processing: Explorations in the Microstructure of ...,
## 3                                 RJ WilliamsCALIFORNIA UNIV SAN DIEGO LA JOLLA INST FOR,
## 4                                                           PDP Research GroupMIT press 1
## 5                        GE HintonAdvances in neural information processing systems,  109
## 6                                                   YW TehNeural computation 18 (7),  152
##   Cited by Year
## 1   34900* 1986
## 2   27417* 1986
## 3    23094 1985
## 4    18726 1987
## 5    15040 2012
## 6     6618 2006
```

```
citationform("Allen","Craig")
```

```
## Warning in readLines(citationURL): ’http://scholar.google.com/citations?user=lSzjnAwAAAAJ
## Warning:  Too many values at 13 locations:  1, 2, 3, 6, 7, 8, 9,
11, 12, 13, 14, 16, 17
```

```
##
## 1           A global overview of drought and heat-induced tree mortality reveals emerging
## 2 Mechanisms of plant survival and mortality during drought: why do some plants survive w
## 3                                                Regional vegetation die-off in response
## 4                                                    Applied historical ecology: using th
## 5        Drought-induced shift of a foresta<U+0080><U+0093>woodland ecotone: rapid landsc
## 6                                    Ecological restoration of southwestern ponderosa pi
##                                                                 Authors
## 1                                                           AK Macalady
## 2                                                            WT Pockman
## 3                                                               NS Cobb
## 4                                                              CD Allen
## 5  DD BreshearsProceedings of the National Academy of Sciences 95 (25)
## 6                                                              M Savage
##                                                    Journal Cited by Year
## 1                               H Chenchouni,  D Bachelet     2758 2010
## 2                                 CD Allen,  DD Breshears     1581 2008
## 3                                      PM Rich,  KP Price     1434 2005
## 4  JL BetancourtEcological Applications 9 (4),  1189-1206     1247 1999
## 5                                      14839-14842,  1998      912 1998
## 6                                   DA Falk,  KF Suckling      858 2002
```

2.3 Problem c

In this function, I use two if-else to test if the function will fail and give the corresponding notification. The first one is to check if the user provides valid input. The second one is to check if Google Scholar returns a rresult.

```
scholartest1<-function(firstname,lastname){
#The first if-else checks if the user provides valid input
if sum((class(firstname)!="character"),(class(lastname)!="character"))[1]>0
 print("Invalid input")
else
 print("The input is valid")
}

scholartest2<-function(firstname,lastname){
baseURL<-"http://scholar.google.com/citations?view_op=search_authors&mauthors="
name<-paste0(firstname,"+",lastname)

ending<-"&hl=en&oi=ao"
idURL<-paste0(baseURL,name,ending)

library(RCurl)
library(XML)

idpage<-htmlParse(idURL)
links<-getHTMLLinks(idpage)
index<-grep("user=.*&",links,perl=TRUE)

#The second if-else checks if Google Scholar returns a result
if sum(!is.null(index))==0
  print("No scholar matches")
else
  print("We've found the author")
}
scholartest1("Geoffrey","Hinton")
scholartest2("Geoffrey","Hinton")

## Error:  <text>:3:4:
## 2:  #The first if-else checks if the user provides valid input
## 3:  if sum
##        ^
```

The next part use testthat.

```
citationform<-function(firstname,lastname){
baseURL<-"http://scholar.google.com/citations?view_op=search_authors&mauthors="
name<-paste0(firstname,"+",lastname)

ending<-"&hl=en&oi=ao"
idURL<-paste0(baseURL,name,ending)

library(RCurl)
library(XML)

idpage<-htmlParse(idURL)
links<-getHTMLLinks(idpage)
index<-grep("user=.*&",links,perl=TRUE)
a<-links[index[1]]
userid<-substr(a,17,28)
citationURL<-paste0("http://scholar.google.com/citations?user=",userid,"&hl=en")
citationpage<-htmlParse(citationURL)

html<-readLines(citationURL)
tbls<-readHTMLTable(html,header=TRUE,as.data.frame=TRUE,which=2)
library(tidyr)
names(tbls)[1]="Title"
tblsnew<-separate(tbls,col="Title",c("Title","Authors","Journal","location","year"),sep=",",
tblsnew<-unite(tblsnew,Journal,Journal,location,sep=", ",remove=TRUE)
tblsnew<-tblsnew[,-c(4)]
library(testthat)
expect_that(dim(tblsnew)[2],equals(5))
}
```