STAT 6494 Homework 7

Menglei Chen

November 7, 2017

obtain certain fields from the US Patent Database, and store them in a data frame

```
#install.packages("rmarkdown")
library(rmarkdown)
#install.packages("genderdata")
#install.packages("gender")
library(gender)
```

```
## WARNING: Rtools is required to build R packages, but is not currently installed.
##
## Please download and install Rtools 3.3 from http://cran.r-project.org/bin/windows/Rtools/ and then r
```

```
#install.packages("ggmap")
library(ggmap)
```

```
## Loading required package: ggplot2
```

```
library(knitr)

sessionInfo()
```

```
## R version 3.2.2 (2015-08-14)
## Platform: i386-w64-mingw32/i386 (32-bit)
## Running under: Windows 7 (build 7601) Service Pack 1
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] knitr_1.11    ggmap_2.4     ggplot2_1.0.1 gender_0.4.3  rmarkdown_0.7
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.8        htmltools_0.2.6     R6_2.1.0
##  [4] scales_0.2.5        curl_0.9.2          maps_2.3-11
##  [7] assertthat_0.1      grid_3.2.2          stringr_1.0.0
```

```
## [10] RgoogleMaps_1.2.0.7 devtools_1.8.0      httr_1.0.0
## [13] munsell_0.4.2       proto_0.3-10        git2r_0.11.0
## [16] lattice_0.20-33     rversions_1.0.2     jpeg_0.1-8
## [19] DBI_0.3.1           xml2_0.1.1          rstudioapi_0.3.1
## [22] sp_1.1-1            jsonlite_0.9.16     MASS_7.3-43
## [25] plyr_1.8.3          stringi_0.5-5       magrittr_1.5
## [28] png_0.1-7           reshape2_1.4.1      evaluate_0.7.2
## [31] gtable_0.1.2        colorspace_1.2-6    RJSONIO_1.3-0
## [34] yaml_2.1.13         tools_3.2.2         mapproj_1.2-4
## [37] parallel_3.2.2      dplyr_0.4.2         geosphere_1.4-3
## [40] rjson_0.2.15        memoise_0.2.1       Rcpp_0.12.0
```

```
setwd("P:/STAT-6494-Data Management in SAS and R/data/")

#read the Patent numbers
patnums<-read.table("patnums.txt")
names(patnums)<-"patnum"
n<-nrow(patnums)

#define the class of the fields
patn<-as.character(patnums[,1])
title<-rep("",n)
filedate<-rep("",n)
publicdate<-rep("",n)
assign<-rep("",n)
fid<-rep("",n)
inventor<-rep("",n)
nclaim<-rep(0,n)

#scrape the fields for the i-th patent
for(i in 1:n){
  #get the complete full-text of the patent
  url<-paste("http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml
  all<-readLines(url,warn=F)

  #scrape the Title
  l<-grep('<FONT size=\"\\+1\">',all)
  a<-regexpr('>', all[l])[1]
  title[i]<-substr(all[l],a+1,nchar(all[l]))

  #scrape the Filed date
  l<-grep('>Filed:',all)[1]+2
  a<-regexpr('<b>', all[l])[1]
  b<-regexpr('</b>', all[l])[1]
  filedate[i]<-substr(all[l],a+3,b-1)

  #scrape the Publication Date
  l<-grep('<U>Publication Date</U>',all)[1]
  a<-gregexpr('center>', all[l])[[1]][5]
  b<-gregexpr('</TD>', all[l])[[1]][4]
  publicdate[i]<-substr(all[l],a+7,b-1)

  #scrape the Assignee
  l1<-grep('Assignee:</TH>',all)+3
```

```
  l2<-grep('nowrap>Family ID:',all)-4
  assignee<-all[l1:l2]
  m<-length(grep('<BR>',assignee))
  l<-1
  for (j in 1:m){
    a<-regexpr('>', assignee[l])[1]
    b<-regexpr('</B>', assignee[l])[1]
    ch1<-substr(assignee[l],a+1,b-1)
    ch2<-assignee[l+1]
    a<-regexpr('>', assignee[l+2])[1]
    ch3<-substr(assignee[l+2],a+1,a+2)
    assign[i]<-paste(assign[i],ch1,ch2,ch3,")\n",sep="")
    l<-l+4
  }

  #scrape the Family ID
  l<-grep('nowrap>Family ID:',all)+2
  a<-regexpr('>', all[l])[1]
  fid[i]<-substr(all[l],a+1,a+9)

  #scrape the Inventors
  l<-grep('Inventors:</TH>',all)+1
  m<-length(gregexpr('B>',all[l])[[1]])
  for (j in 1:m){
    a<-gregexpr('B>', all[l])[[1]][j]+2
    b<-gregexpr('<', all[l])[[1]][j+1]-1
    ch<-substr(all[l],a,b)
    inventor[i]<-paste(inventor[i],ch,sep="")
  }

  #scrape the Number of claims
  l1<-grep('<b><i>Claims</b></i>',all)
  l2<-grep('<b><i>Description</b></i>',all)
  claims<-all[l1:l2]
  nclaim[i]<-length(grep('<BR><BR>(\\d|\\s\\d)',claims))
}

#save as the fields in a data frame
patent<-as.data.frame(cbind(patn,title,filedate,publicdate,assign,fid,inventor,nclaim),stringsAsFactors
names(patent)<-c("Patent number","Title","Filed date","Publication date","Assignee","Family ID","Invent

#change stored format of dates
patent$`Filed date`<-as.Date(patent$`Filed date`,"%B %d, %Y")
#patent$`Filed date`<-format(patent$`Filed date`,"%Y%m%d")

patent$`Publication date`<-as.Date(patent$`Publication date`,"%B %d, %Y")
#patent$`Publication date`<-format(patent$`Publication date`,"%Y%m%d")

#Parse the inventors data
m<-rep(0,n)
for (i in 1:n){
  m[i]<-length(gregexpr('\\(',patent$Inventors[i])[[1]])
}
```

```
nm<-max(m[1:n])
inventorm<-matrix(rep("",n*nm*10),nrow=n,ncol=nm*10)

for (i in 1:n) {
  ch<-c(-2,gregexpr('\\(|\\)',inventor[i])[[1]])
  j=1
  for (k in seq(1,10*m[i],10)) {
    #name and location
    inventorm[i,k]<-substr(inventor[i],ch[j]+3,ch[j+1]-2)
    inventorm[i,k+1]<-substr(inventor[i],ch[j+1]+1,ch[j+2]-1)
    j<-j+2

    #last name, first name, and gender
    a<-regexpr(';',inventorm[i,k])
    inventorm[i,k+2]<-substr(inventorm[i,k],1,a-1)
    name<-paste(substr(inventorm[i,k],a+2,nchar(inventorm[i,k])),"  ")
    b<-regexpr('\\s',name)
    inventorm[i,k+3]<-substr(name,1,b-1)
    try(inventorm[i,k+4]<-gender(inventorm[i,k+3])$gender,silent=T)



    #city, state, country, longitude, and latitude
    a<-regexpr(',',inventorm[i,k+1])
    inventorm[i,k+5]<-substr(inventorm[i,k+1],1,a-1)
    abb<-substr(inventorm[i,k+1],a+2,nchar(inventorm[i,k+1]))
    if (abb %in% state.abb){
      inventorm[i,k+6]<-state.name[grep(abb,state.abb)]
      inventorm[i,k+7]<-"US"
    } else if(!is.na(abb)){
      inventorm[i,k+6]<-NA
      inventorm[i,k+7]<-abb
    }

    inventorm[i,k+8]<-geocode(inventorm[i,k+1])$lon
    inventorm[i,k+9]<-geocode(inventorm[i,k+1])$lat

  }
}
```

```
## The genderdata package needs to be installed from GitHub.
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=New+York,+NY&sensor=
## The genderdata package needs to be installed from GitHub.
## The genderdata package needs to be installed from GitHub.
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Upton,+MA&sensor=fals
## The genderdata package needs to be installed from GitHub.
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Lexington,+MA&sensor=
## The genderdata package needs to be installed from GitHub.
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Concord,+MA&sensor=fa
## The genderdata package needs to be installed from GitHub.
## Information from URL : http://maps.googleapis.com/maps/api/geocode/json?address=Ithaca,+NY&sensor=fal
## The genderdata package needs to be installed from GitHub.
## The genderdata package needs to be installed from GitHub.
```

```
## The genderdata package needs to be installed from GitHub.
## The genderdata package needs to be installed from GitHub.
```

```
inventors<-as.data.frame(inventorm,stringsAsFactors = F)

#name the new fields
for (i in 1:nm){
  names(inventors)[10*i-9]<-paste("Inventor",i,"_name")
  names(inventors)[10*i-8]<-paste("Inventor",i,"_location")
  names(inventors)[10*i-7]<-paste("Inventor",i,"_Lastname")
  names(inventors)[10*i-6]<-paste("Inventor",i,"_Firstname")
  names(inventors)[10*i-5]<-paste("Inventor",i,"_Gender")

  names(inventors)[10*i-4]<-paste("Inventor",i,"_City")
  names(inventors)[10*i-3]<-paste("Inventor",i,"_State")
  names(inventors)[10*i-2]<-paste("Inventor",i,"_Country")
  names(inventors)[10*i-1]<-paste("Inventor",i,"_Longitude")
  names(inventors)[10*i]<-paste("Inventor",i,"_Latitude")

}

patent<-cbind(patent[1:6],patent[8],inventors)
```

create a table of the number of inventors by gender

```
obs<-c(patnums[1:n,1],"Total")
male<-rep(0,n+1)
female<-rep(0,n+1)
for(i in 1:n){
  for (j in 1:nm){
    male[i][inventors[i,10*j-5]=="male"]<-male[i]+1
    female[i][inventors[i,10*j-5]=="female"]<-female[i]+1
  }
}
male[n+1]<-sum(male[1:n])
female[n+1]<-sum(female[1:n])
gender<-data.frame(patent=obs,male,female,stringsAsFactors = F)

kable(gender,format="latex")
```

patent	male	female
6168946	0	0
7109166	0	0
8202698	0	0
8984396	0	0
Total	0	0