Week1: Parsing Cleaning Transforming ReferencePatent table

```
#Load Necessary packages
library(reshape2)
library(plyr)
library(dbConnect)
```

```
## Loading required package: RMySQL
```

```
## Loading required package: DBI
```

```
## Loading required package: gWidgets
```

```
##
## Attaching package: 'gWidgets'
```

```
## The following object is masked from 'package:plyr':
##
##     id
```

```
#library(ggplot2)
#library(ggmap)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following object is masked from 'package:plyr':
##
##     here
```

```
## The following object is masked from 'package:base':
##
##     date
```

```
library('splitstackshape')
```

```
## Loading required package: data.table
```

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:lubridate':
##
##     hour, isoweek, mday, minute, month, quarter, second, wday,
##     week, yday, year
```

```
## The following objects are masked from 'package:reshape2':
##
##     dcast, melt
```

Parsing Cleaning Transforming ReferencePatent table

Objective:

We are trying to parse the original ReferencePatent table into a final output table, which contains the following attributes: id, authors (original field), title, patentType (e.g. EP, US, GB), PatentNumber (e.g. 0238993), version (e.g. -A2 1), patentDay, PatentMonth, PatentYear, extra_journal (the leftover data that comes after the semi-colon or that is not any of the previously indicated patent data), and then the parsed author names (one per column (e.g. Author 1, Author 2, Author 3).

Get the Reference Patent table from the SQL server

```
#uncomment next line for the real thing
#ReferencePatent<-dbGetQuery(con,'select * from genbank.ReferencePatent')
ReferencePatent<-read.csv('data/ReferencePatent1000.csv',row.names=1)
```

Parse Author Column:

```
#Split Author names of the real set
paste("the original column:")
```

```
## [1] "the original column:"
```

```
head(ReferencePatent$authors,10)
```

```
##  [1] Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##  [2] Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##  [3] Sanchez,F.S., Susan,V.R., Carramolino-Fitera,L. and Ortega,A.P.A.
##  [4] Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##  [5] Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##  [6] Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##  [7] Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##  [8] Warne,S.
##  [9] Brown,J.P., Plowman,G.D., Hellstrom,K.E., Purchio,A.F.,\tPennathur,S., Estin,C.D., Rose,T.M., H
## [10] Brown,J.P., Plowman,G.D., Hellstrom,K.E., Purchio,A.F.,\tPennathur,S., Estin,C.D., Rose,T.M., H
## 273 Levels: . Aberg,B. Aggarwal,B.B. and Lee,S.He. ... Yoshida,M., Sugano,H., Shimizu,F. and Imagawa
```

```
#eliminate "\t"
authorsplit<-gsub("\t", " ", ReferencePatent$authors, fixed=TRUE)
#head(authorsplit)
#eliminate "and"
authorsplit<-gsub(" and ", ", ", authorsplit, fixed=TRUE)
#head(authorsplit)
#split it by ", " because two author names are split by ", "
authorsplit<-strsplit(authorsplit,", ")
#the number of authors the patent with most authors has
paste("the number of authors the patent with most authors has:", max(unlist(lapply(authorsplit, functio
```

```
## [1] "the number of authors the patent with most authors has: 13"
```

```
#The result
paste("the parsed column:")
```

```
## [1] "the parsed column:"
```

```
head(authorsplit)
```

```
## [[1]]
## [1] "Auerswald,E.A." "Schroeder,W."   "Schnabel,E."    "Bruns,W."
## [5] "Reinhardt,G."   "Kotick,M."
##
## [[2]]
## [1] "Auerswald,E.A." "Schroeder,W."   "Schnabel,E."    "Bruns,W."
## [5] "Reinhardt,G."   "Kotick,M."
##
## [[3]]
## [1] "Sanchez,F.S."          "Susan,V.R."            "Carramolino-Fitera,L."
## [4] "Ortega,A.P.A."
##
## [[4]]
## [1] "Auerswald,E.A." "Schroeder,W."   "Schnabel,E."    "Bruns,W."
## [5] "Reinhardt,G."   "Kotick,M."
##
## [[5]]
## [1] "Auerswald,E.A." "Schroeder,W."   "Schnabel,E."    "Bruns,W."
## [5] "Reinhardt,G."   "Kotick,M."
##
## [[6]]
## [1] "Auerswald,E.A." "Schroeder,W."   "Schnabel,E."    "Bruns,W."
## [5] "Reinhardt,G."   "Kotick,M."
```

Transpose the Author column:

For an author cell like

Author
'Kozlov,J.I.' 'Naroditskaya,V.A.'

we want to separate it into two separate cells

Author	Author1	Author2
'Kozlov,J.I.' 'Naroditskaya,V.A.'	Kozlov,J.I.	'Naroditskaya,V.A.'

```
#Create a new dataframe for this task:
pat<-as.data.frame(ReferencePatent$authors)
#Write a for loop to populate the tranposed author columns
for (i in 1:max(unlist(lapply(authorsplit, function(x) length(x))))){
  eval(parse(text = paste0('pat$Author', i, ' <- sapply(authorsplit,function(x) x[i])')))
}

head(pat)
```

```
##                                                    ReferencePatent$authors
## 1 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
## 2 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
## 3               Sanchez,F.S., Susan,V.R., Carramolino-Fitera,L. and Ortega,A.P.A.
```

```
## 4 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
## 5 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
## 6 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##          Author1      Author2               Author3       Author4
## 1 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 2 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 3   Sanchez,F.S.   Susan,V.R. Carramolino-Fitera,L. Ortega,A.P.A.
## 4 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 5 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 6 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
##        Author5   Author6 Author7 Author8 Author9 Author10 Author11
## 1 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 2 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 3         <NA>      <NA>    <NA>    <NA>    <NA>     <NA>     <NA>
## 4 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 5 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 6 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
##   Author12 Author13
## 1     <NA>     <NA>
## 2     <NA>     <NA>
## 3     <NA>     <NA>
## 4     <NA>     <NA>
## 5     <NA>     <NA>
## 6     <NA>     <NA>
```

Parse content out of the journal column:

```
#1. parse patentType out of journal column
head(ReferencePatent$journal,10)
```

```
##  [1] Patent: EP 0238993-A2 1 30-SEP-1987;\tBAYER AG
##  [2] Patent: EP 0238993-A2 2 30-SEP-1987;\tBAYER AG
##  [3] Patent: EP 0240250-A1 1 07-OCT-1987;\tAntibioticos, S.A
##  [4] Patent: EP 0238993-A2 20 30-SEP-1987;\tBAYER AG
##  [5] Patent: EP 0238993-A2 23 30-SEP-1987;\tBAYER AG
##  [6] Patent: EP 0238993-A2 26 30-SEP-1987;\tBAYER AG
##  [7] Patent: EP 0238993-A2 29 30-SEP-1987;\tBAYER AG
##  [8] Patent: GB 2220942-A 1 24-JAN-1990;\tThe Secretary of State of Trade and Industry
##  [9] Patent: GB 2188637-A 1 07-OCT-1987;\tOncogen
## [10] Patent: GB 2188637-A 3 07-OCT-1987;\tOncogen
## 1000 Levels: Patent: EP 0235046-A1 1 02-SEP-1987;\tINSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQUE, Eta
```

```
#sub "\t" and split the journal column by ' '
journal<-strsplit(gsub("\t", " ", ReferencePatent$journal, fixed=TRUE)," ")
head(journal,10)
```

```
## [[1]]
## [1] "Patent:"      "EP"           "0238993-A2"   "1"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[2]]
## [1] "Patent:"      "EP"           "0238993-A2"   "2"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
```

```
## [[3]]
## [1] "Patent:"       "EP"            "0240250-A1"    "1"
## [5] "07-OCT-1987;"  "Antibioticos," "S.A"
##
## [[4]]
## [1] "Patent:"      "EP"           "0238993-A2"   "20"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[5]]
## [1] "Patent:"      "EP"           "0238993-A2"   "23"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[6]]
## [1] "Patent:"      "EP"           "0238993-A2"   "26"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[7]]
## [1] "Patent:"      "EP"           "0238993-A2"   "29"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[8]]
##  [1] "Patent:"      "GB"           "2220942-A"    "1"
##  [5] "24-JAN-1990;" "The"          "Secretary"    "of"
##  [9] "State"        "of"           "Trade"        "and"
## [13] "Industry"
##
## [[9]]
## [1] "Patent:"      "GB"           "2188637-A"    "1"
## [5] "07-OCT-1987;" "Oncogen"
##
## [[10]]
## [1] "Patent:"      "GB"           "2188637-A"    "3"
## [5] "07-OCT-1987;" "Oncogen"
```

```
#the second part of each row of 'journal' list is the patenttype that we need
pat$patentType<-sapply(journal,function(x) x[2])
head(pat$patentType,10)
```

```
##  [1] "EP" "EP" "EP" "EP" "EP" "EP" "EP" "GB" "GB" "GB"
```

```
#2. parse patent number out of journal column
head(journal)
```

```
## [[1]]
## [1] "Patent:"      "EP"           "0238993-A2"   "1"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[2]]
## [1] "Patent:"      "EP"           "0238993-A2"   "2"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[3]]
## [1] "Patent:"       "EP"            "0240250-A1"    "1"
## [5] "07-OCT-1987;"  "Antibioticos," "S.A"
##
```

```
## [[4]]
## [1] "Patent:"      "EP"           "0238993-A2"   "20"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[5]]
## [1] "Patent:"      "EP"           "0238993-A2"   "23"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[6]]
## [1] "Patent:"      "EP"           "0238993-A2"   "26"
## [5] "30-SEP-1987;" "BAYER"        "AG"
```

```
patentNumber<-strsplit(sapply(journal,function(x) x[3]), "-")
pat$patentNumber<-sapply(patentNumber,function(x) x[[1]])
head(pat$patentNumber)
```

```
## [1] "0238993" "0238993" "0240250" "0238993" "0238993" "0238993"
```

```
#3. parse patentVersion (something like -A2 1) out of journal column
head(journal)
```

```
## [[1]]
## [1] "Patent:"      "EP"           "0238993-A2"   "1"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[2]]
## [1] "Patent:"      "EP"           "0238993-A2"   "2"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[3]]
## [1] "Patent:"       "EP"             "0240250-A1"    "1"
## [5] "07-OCT-1987;"  "Antibioticos," "S.A"
##
## [[4]]
## [1] "Patent:"      "EP"           "0238993-A2"   "20"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[5]]
## [1] "Patent:"      "EP"           "0238993-A2"   "23"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[6]]
## [1] "Patent:"      "EP"           "0238993-A2"   "26"
## [5] "30-SEP-1987;" "BAYER"        "AG"
```

```
version1<-sapply(patentNumber,function(x) x[2])
head(version1)
```

```
## [1] "A2" "A2" "A1" "A2" "A2" "A2"
```

```
version2<-sapply(journal,function(x) x[4])
head(version2)
```

```
## [1] "1"  "2"  "1"  "20" "23" "26"
```

```
#mapply is kind of like zip in python
version3<-mapply(c, version1, version2, SIMPLIFY=FALSE)
#head(version3)
```

```
pat$patentVersion<-as.vector(sapply(version3,function(x) paste(x[1],x[2])))
head(pat$patentVersion)
```

```
## [1] "A2 1"  "A2 2"  "A1 1"  "A2 20" "A2 23" "A2 26"
```

```
#4. patentDay, patentmonth, patent year (requires 'lubridate')
head(journal)
```

```
## [[1]]
## [1] "Patent:"      "EP"           "0238993-A2"   "1"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[2]]
## [1] "Patent:"      "EP"           "0238993-A2"   "2"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[3]]
## [1] "Patent:"       "EP"            "0240250-A1"    "1"
## [5] "07-OCT-1987;"  "Antibioticos," "S.A"
##
## [[4]]
## [1] "Patent:"      "EP"           "0238993-A2"   "20"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[5]]
## [1] "Patent:"      "EP"           "0238993-A2"   "23"
## [5] "30-SEP-1987;" "BAYER"        "AG"
##
## [[6]]
## [1] "Patent:"      "EP"           "0238993-A2"   "26"
## [5] "30-SEP-1987;" "BAYER"        "AG"
```

```
fulldate<-dmy(sapply(journal,function(x) x[5]))
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2017c.1.0/
## zoneinfo/America/Los_Angeles'
```

```
pat$patentYear<-year(fulldate)
pat$patentMonth<-month(fulldate)
pat$patentDay<-day(fulldate)
#5. split the journal column on the first occrence of ";\t" to get extra portion of the journal
patentExtra<-as.vector(colsplit(ReferencePatent$journal,";\t",c("a","b"))[2])
pat$patentExtra<-patentExtra$b
head(pat$patentExtra)
```

```
## [1] "BAYER AG"          "BAYER AG"          "Antibioticos, S.A"
## [4] "BAYER AG"          "BAYER AG"          "BAYER AG"
```

```
#Final Results
head(pat)
```

```
##                                                    ReferencePatent$authors
## 1 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
## 2 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
## 3               Sanchez,F.S., Susan,V.R., Carramolino-Fitera,L. and Ortega,A.P.A.
## 4 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
## 5 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
```

```
## 6 Auerswald,E.A., Schroeder,W., Schnabel,E., Bruns,W., Reinhardt,G.\tand Kotick,M.
##          Author1      Author2               Author3       Author4
## 1 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 2 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 3   Sanchez,F.S.   Susan,V.R. Carramolino-Fitera,L. Ortega,A.P.A.
## 4 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 5 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
## 6 Auerswald,E.A. Schroeder,W.           Schnabel,E.      Bruns,W.
##        Author5   Author6 Author7 Author8 Author9 Author10 Author11
## 1 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 2 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 3         <NA>      <NA>    <NA>    <NA>    <NA>     <NA>     <NA>
## 4 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 5 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
## 6 Reinhardt,G. Kotick,M.    <NA>    <NA>    <NA>     <NA>     <NA>
##   Author12 Author13 patentType patentNumber patentVersion patentYear
## 1     <NA>     <NA>         EP      0238993          A2 1       1987
## 2     <NA>     <NA>         EP      0238993          A2 2       1987
## 3     <NA>     <NA>         EP      0240250          A1 1       1987
## 4     <NA>     <NA>         EP      0238993         A2 20       1987
## 5     <NA>     <NA>         EP      0238993         A2 23       1987
## 6     <NA>     <NA>         EP      0238993         A2 26       1987
##   patentMonth patentDay       patentExtra
## 1           9        30          BAYER AG
## 2           9        30          BAYER AG
## 3          10         7 Antibioticos, S.A
## 4           9        30          BAYER AG
## 5           9        30          BAYER AG
## 6           9        30          BAYER AG
```

We are only going to need the

2. Transforming Reference Table:

Parse the genbank.reference table in the same way I did the ReferencePatent Table: keep the id, title, journal, authororginal, author1,2,3,4,pubmed,year,reference

Reference<-dbGetQuery(con,'select * from genbank.Reference') colnames(Reference) colnames(parsedReference) '"