EMIS 7357
Analytics for Decision Support

Assignment # 3
Assessing Cancer Patent Portfolio

By:
APOORVA JAIN

Table of contents

Contents

Introduction

When it comes to providing care for cancer in poor countries, the pharma companies have come in the limelight. Many pharma companies have been denounced as they make the newest cancer drugs costly or they don't support oncology patents in those countries. The information regarding pharmaceutical companies which are working to improve different type of cancer care in poor countries can be found in Medicine Foundation. These high prices of cancer medicines are weakening the ability of any government to impart low-end cancer treatment. All the governments share one common issue that is the eagerness to provide affordable prices for cancer care.

The goal of this project is to assess the cancer patents portfolio using the document term matrixs and cluster analysis (kmeans and hiearchial clustering). Eventually we optimize the number of cluster by using the fviz_nbclust function. We randomly took the companies and the academic organization for our analysis which is shown further in the report.

Questions

Question1:

Explain all the lines of code posted in the R code, looking up functions online if necessary. In particular, look up what the Corpus function does, and explain what the arguments of the DocumentTermMatrix function do.

Solution:

1. **mydata <- read.csv("HW3.csv") :** In this we read our data which is saved in Ms Excel with a file name of HW3.csv
2. **str(mydata) :** Compactly display the internal **str**ucture of an R object.
3. **yt <- strptime(mydata$Filing_Date, format ='%Y%m%d')**
   mydata$year <- as.integer(format(yt,"%Y")) : extracting the year from the date of patent filing using lubridate package.
4. **install.packages("tm")**
   library(tm) : Preparing the Document Term Matrix for clustering loading and installing the TM package for text mining.
5. **mydatatitle <-mydata$Patent_Title :** creating the document term matrix (DTM) on the titles of the patent
6. **ndocs <-length(mydatatitle) :** length sets the length of the vector
7. **minf=ndocs*0.01:** setting the parameters helps keep the DTM of reasonable size and minimum occurrence of words to be used in the DTM.
8. **maxf=ndocs*0.40 :** setting the parameters below helps keep the DTM of reasonable size maximum occurrence of words to be used in the DTM
9. **corpus=Corpus(VectorSource(mydatatitle)):** This function is mainly used for structuring of the data in text mining (TM) package ,putting the input data in a format the main function for DTM can read
10. **dtm=DocumentTermMatrix(corpus,control=list(stopwords=TRUE, wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds = list(global=c(minf,maxf)))) :** main command line to create the DTM with parameters listed in control , in this we have to use the corpus which we defined earlier, define the length of word and boundary of the document term matrix by using minf and maxf function.
11. **inspect(dtm):** this just display the information on the document term matrix
12. **dtmm=as.matrix(dtm)**
    write.csv(dtmm,"cancerdtm.csv")
    View(dtmm): its usually used for the outputting the DTM to a CSV file if we want and viewing the matrix that we saved as CSV file.
13. **FreqMat <- data.frame(ST = colnames(dtmm), Freq = colSums(dtmm))**

y <- as.vector(order(-FreqMat$Freq))
FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20) : We use this to explore the data and finding the most frequent words in the DTM matrix.

14. **bigdata<cbind(mydata$Family_ID,mydata$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm)),make.row.names=TRUE) :** adding the DTMs to the rest of the data frame and outputting that

Question2:

Implement clustering over the DTM matrix. Justify your method: k-means or hierarchical clustering. You must select an appropriate number of clusters and motivate your choice. Also experiment with varying minf and maxf in the R code. What is the size of each cluster? Provide a description (in English) of each cluster using the most commonly used words (you must output the top "n" most frequent words for each cluster, where "n" is a number you choose, using the sample code provided, but in your report you must describe each cluster using full sentences in English instead of just listing the most common words. You might need to research some terms on the Internet.)

Solution:

I will preferably choose K means clustering instead of using hierarchical clustering as K means is implied on all the data sets and K means is a division of objects into clusters such that each object is in exactly one cluster, not several.

The main reason is using number of clusters in K means and Hierarchical clustering which should be realistic according to the data therefore I prefer K means clustering.

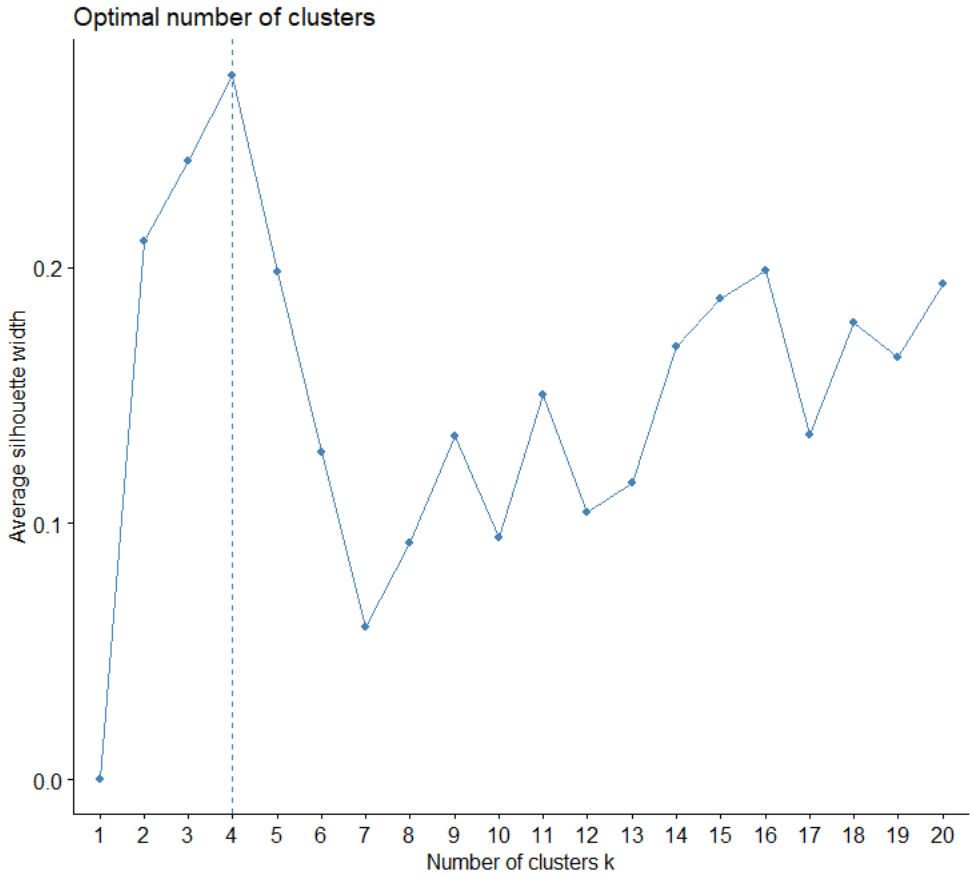
Kmeans clustering :

Below I have assumed number of clusters is equivalent to 14 and its distribution is shown below.

```
> table(kmc.cancer$cluster)

    1     2     3     4     5     6     7     8     9    10    11    12    13    14
11745 19670 23714  8082 11771 13399  9250 101882   720  7501  5615  5940 13018  4552
>
```

Here is the optimal number of clustering when applied on specific data sets of 5000, the output was shown as the number of clusters should be 4 . I used a Fviz_Nbclust function and sillouette method to obtain the optimal number of clusters shown below.

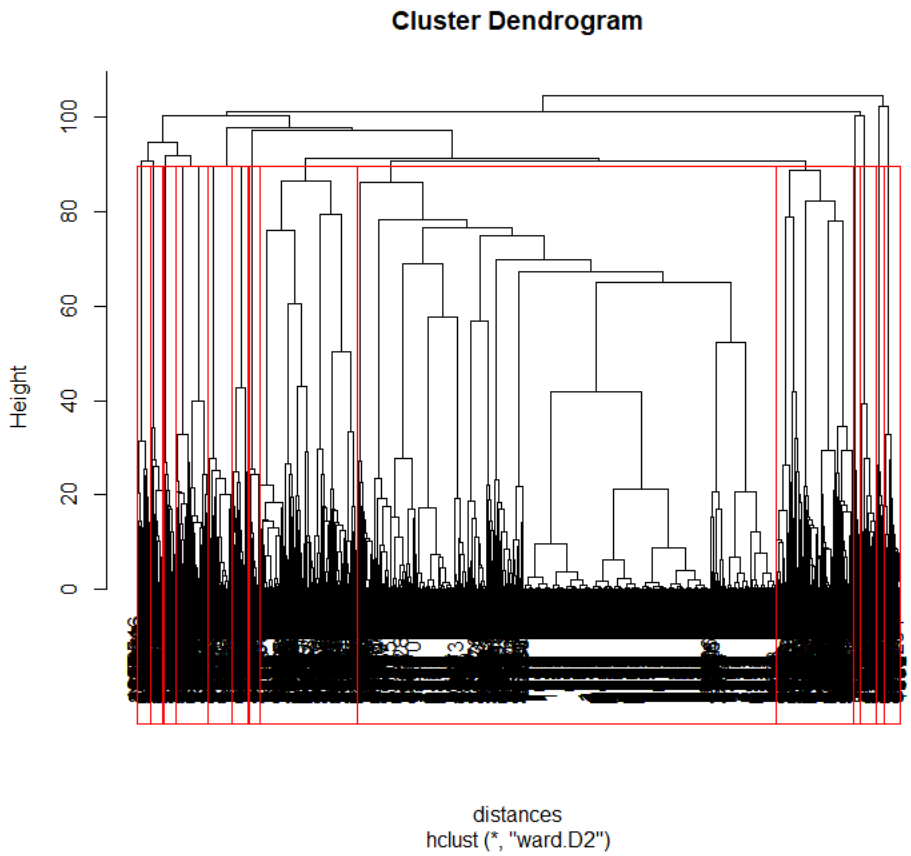


Hierarchical cluster:

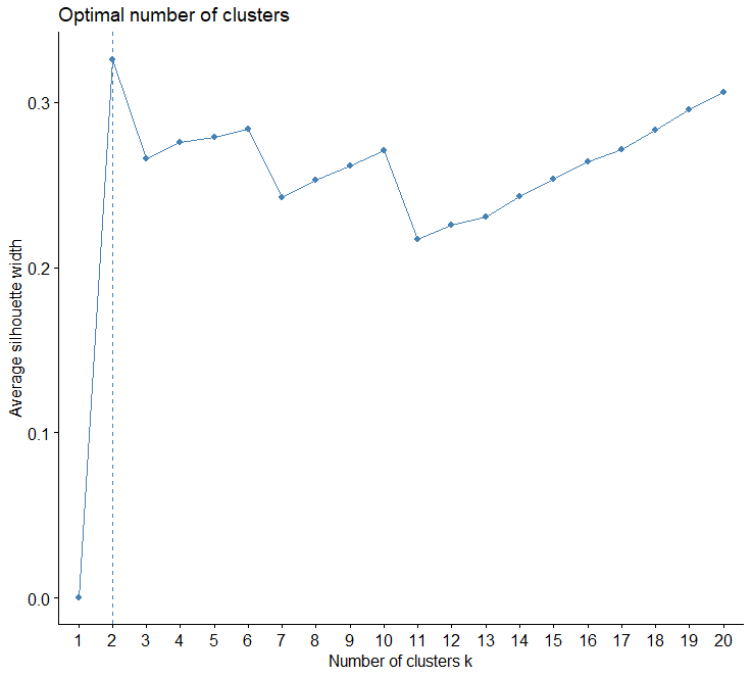
Below I have assumed number of clusters is equivalent to k=14 and its distribution is shown below.

```
   1    2    3    4    5    6    7    8    9   10   11   12   13   14
642 2741  108   80  110  510   73  109   79  215   39  157   90   47
```

We obtain a plot named as Cluster dendrogram which shows the clustering and the red lines shows the number clusters clubbed together based on the similarity. Here I used the ward.D2 method to obtain above dendogram.



Optimal number of clusters by using hierarchical clustering is shown below. Here the optimal number of clusters using silhouette method is 2.



These are the topwords from most to least used words according to number of clusters .

```
> print(topwords)
      [,1]          [,2]            [,3]          [,4]          [,5]           [,6]            [,7]          [,8]
 [1,] "delivery"    "protein"       "diseases"    "inhibitors"  "therapeutic"  "agents"        "cells"       "cell"
 [2,] "cell"        "preparation"   "treating"    "methods"     "compositions" "thereof"       "acid"        "novel"
 [3,] "gene"        "delivery"      "nucleic"     "composition" "therapy"      "substituted"   "protein"     "therapeutic"
 [4,] "nucleic"     "protein"       "acid"        "inhibitors"  "therapeutic"  "novel"         "derivatives" "human"
 [5,] "treating"    "uses"          "substituted" "protein"     "inhibitors"   "treatment"     "thereof"     "novel"
 [6,] "composition" "antibodies"    "thereof"     "protein"     "compounds"    "human"         "therapeutic" "cell"
 [7,] "compounds"   "pharmaceutical" "therapy"    "protein"     "preparation"  "cells"         "composition" "cell"
 [8,] "acid"        "method"        "treating"    "therapeutic" "novel"        "treatment"     "derivatives" "uses"
 [9,] "nucleic"     "using"         "therapeutic" "agents"      "gene"         "uses"          "therapy"     "protein"
[10,] "agents"      "gene"          "thereof"     "antibodies"  "cell"         "novel"         "compounds"   "cells"
[11,] "comprising"  "uses"          "protein"     "compounds"   "cell"         "gene"          "methods"     "human"
[12,] "compounds"   "delivery"      "comprising"  "cells"       "cancer"       "pharmaceutical" "diseases"    "cell"
[13,] "therapeutic" "protein"       "cell"        "agents"      "acid"         "antibodies"    "composition" "using"
[14,] "therapeutic" "pharmaceutical" "gene"      "thereof"     "system"       "cells"         "cell"        "composition"
      [,9]          [,10]         [,11]          [,12]         [,13]          [,14]
 [1,] "antibodies"  "treatment"   "compounds"    "treating"    "using"        "methods"
 [2,] "substituted" "protein"     "compounds"    "derivatives" "kinase"       "inhibitors"
 [3,] "cell"        "human"       "agents"       "antibodies"  "derivatives"  "compounds"
 [4,] "compounds"   "antibodies"  "preparation"  "methods"     "uses"         "thereof"
 [5,] "antibodies"  "human"       "compounds"    "methods"     "derivatives"  "receptor"
 [6,] "gene"        "cells"       "using"        "treating"    "methods"      "cancer"
 [7,] "treatment"   "thereof"     "treating"     "using"       "system"       "method"
 [8,] "preparation" "thereof"     "compounds"    "comprising"  "pharmaceutical" "compositions"
 [9,] "human"       "derivatives" "compounds"    "delivery"    "system"       "novel"
[10,] "using"       "treatment"   "compositions" "therapy"     "methods"      "cancer"
[11,] "method"      "using"       "derivatives"  "nucleic"     "cells"        "acid"
[12,] "thereof"     "using"       "treatment"    "treating"    "compositions" "methods"
[13,] "inhibitors"  "compounds"   "derivatives"  "diseases"    "cancer"       "treatment"
[14,] "therapy"     "treatment"   "using"        "treating"    "method"       "cancer"
>
```

I have explained all the top words used in each of the clusters which are as follows:

Cluster 1: The top word is delivery and the second top word is protein. any of a class of nitrogenous organic compounds that consist of large molecules composed of one or more long chains of amino acids and are an essential part of all living organisms, especially as structural components of body tissues such as muscle, hair, collagen, etc., and as enzymes and antibodies.

Cluster 2: Cell is the smallest structural and functional unit of an organism, typically microscopic and consisting of cytoplasm and a nucleus enclosed in a membrane. Microscopic organisms typically consist of a single cell, which is either eukaryotic or prokaryotic.
Cluster 3: Gene is a unit of heredity which is transferred from a parent to offspring and is held to determine some characteristic of the offspring.
Cluster 4: Nucleic it is any of various complex organic acids (such as DNA or RNA) that are composed of nucleotide chains.
Cluster 5: Treating is medical care given to a patient for an illness or injury.
Cluster 6: Composition but I will prefer Antibodies is a thing that is composed of two or more separate elements; a mixture.
Cluster 7: Compound is a thing that is composed of two or more separate elements; a mixture.
Cluster 8: Acid is a molecule or other entity that can donate a proton or accept an electron pair in reactions.
Cluster 9: Nucleic it is any of various complex organic acids (such as DNA or RNA) that are composed of nucleotide chains.
Cluster 10: Agents is any power, principle or substance capable of producing an effect, whether physical, chemical or biological.
Cluster 11: Comprising
Cluster 12: Compound is a thing that is composed of two or more separate elements; a mixture.
Cluster 13: Therapeutic is the branch of medicine concerned with the treatment of disease and the action of remedial agents.
Cluster 14: Therapeutic is the branch of medicine concerned with the treatment of disease and the action of remedial agents.

Question3:

Repeat 2) by clustering not only on the patent title but also on CPC_Inventive. This involves creating a new DTM matrix for CPC_Inventive and joining it to the previous one by using cbind as in line 48 of the R code. Use this to describe current trends in cancer research. Note that you can find descriptions of the CPC_Inventive codes

Solution:

http://www.patbase.com/linkclass.asp?CCC=CPC&CLASS=C07K14%2F51

KMC cluster analysis:

I have used the K means clustering were number of clusters is equivalent to 8. The distribution of clusters is as shown below.

```
     1      2      3      4      5      6      7      8
138666  22299  13552  18293   1522  13585  11055  17887
> |
```

The top words here are as shown below.

```
> print(topwords)
     [,1]             [,2]          [,3]           [,4]          [,5]           [,6]          [,7]           [,8]
[1,] "acid"           "agents"      "novel"        "protein"     "c07d401/12;"  "human"       "antibodies"   "using"
[2,] "pharmaceutical" "a61k45/06;"  "acid"         "preparation" "composition" "treatment"   "cell"         "cells"
[3,] "treating"       "a61k31/496;" "a61k31/00;"   "a61k39/3955;" "a61k31/506;" "compounds"   "a61k31/5377;" "a61k31/519;"
[4,] "receptor"       "acid"        "compositions" "inhibitors"  "novel"        "derivatives" "human"        "antibodies"
[5,] "c12q1/6883;"    "c07k16/28;"  "c07k14/47"    "a61k39/0011;" "g01n33/5011;" "c07k14/47;"  "c07k16/18;"   "g01n33/574;"
[6,] "acid"           "c12q1/6886;" "method"       "composition" "inhibitors"  "using"       "a61k45/06;"   "derivatives"
[7,] "a61k39/0011;"   "gene"        "cell"         "cells"       "a61k45/06;"   "using"       "therapy"      "c12q1/6886"
[8,] "cell"           "cells"       "thereof"      "compounds"   "comprising"   "using"       "cancer"       "a61k45/06;"
     [,9]          [,10]          [,11]          [,12]         [,13]          [,14]
[1,] "receptor"    "derivatives"  "compounds"    "inhibitors"  "methods"      "make.row.names"
[2,] "thereof"     "treating"     "system"       "using"       "make.row.names" "method"
[3,] "a61k31/337;" "inhibitors"   "treatment"    "methods"     "make.row.names" "a61k45/06;"
[4,] "compounds"   "preparation" "methods"      "uses"        "make.row.names" "thereof"
[5,] "c07k14/4748;" "c12q1/6886;" "g01n33/57484;" "nucleic"   "c07k14/705;"  "make.row.names"
[6,] "compounds"   "methods"     "diseases"     "cancer"      "make.row.names" "treatment"
[7,] "c12q1/6886;" "treating"    "methods"      "method"      "make.row.names" "cancer"
[8,] "treating"    "treatment"   "pharmaceutical" "methods"   "compositions"  "make.row.names"
>
```

Cluster 1: “Acid” is a molecule or other entity that can donate a proton or accept an electron pair in reactions.

Cluster 2: “Pharmaceutical” is the smallest structural and functional unit of an organism, typically microscopic and consisting of cytoplasm and a nucleus enclosed in a membrane. Microscopic organisms typically consist of a single cell, which is either eukaryotic or prokaryotic.

Cluster 3: “Treating” is medical care given to a patient for an illness or injury.

Cluster 4: “Receptor” is an organ or cell able to respond to light, heat, or other external stimulus and transmit a signal to a sensory nerve.

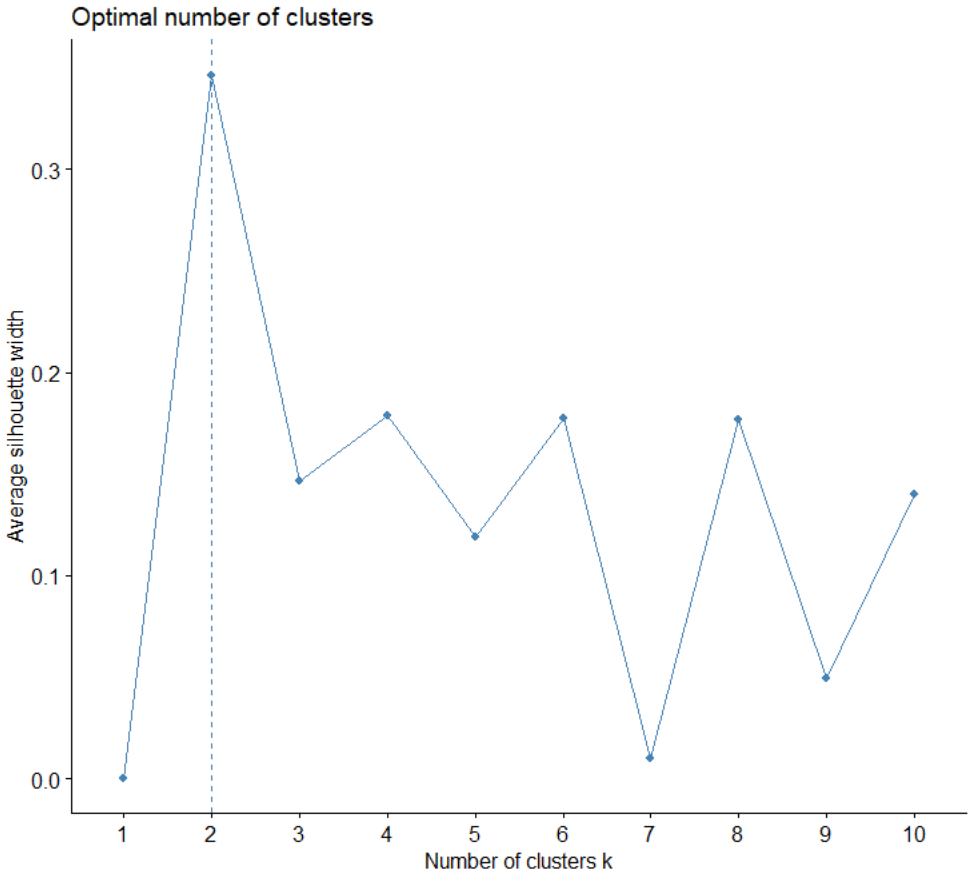
Cluster 5: “`c12q1/6883`” here c12Q means measuring or testing processes involving enzymes, nucleic acids or microorganisms. Though it is for diseases caused by alterations of genetic material.
(Reference: http://www.patbase.com/linkclass.asp?CCC=CPC&CLASS=C07K14%2F51)

Cluster 6: “Acid” is a molecule or other entity that can donate a proton or accept an electron pair in reactions.

Cluster 7: “`A61k39/0011`” here A61K means devices or methods specially adapted for bringing pharmaceutical products into particular physical or administering forms A61J3/00. Though it is a cancer antigen.
(Reference: http://www.patbase.com/linkclass.asp?CCC=CPC&CLASS=C07K14%2F51)

Cluster 8: **“**Cell” is the smallest structural and functional unit of an organism, typically microscopic and consisting of cytoplasm and a nucleus enclosed in a membrane. Microscopic organisms typically consist of a single cell, which is either eukaryotic or prokaryotic.

Optimal number of clusters using silhouette method.



Question4:

for the 5 companies that have the most patent documents with FDA approval (those with data in columns X and beyond). The R code shows how to find the patents that contain the string "NOVARTIS" in the FDA applicant name as an example. Clustering must be on patent title and CPC_Inventive. Note that the number of cluster may change depending on the company considered. You can also add other classification categories if you feel that would be helpful. Which company do you think has the strongest patent portfolio?

Solution:

1. **NOVARTIS**

   Novartis is the company that we used for our analysis , the figure shows the top most word and patent used in that particular data provided by the Novartis company.

Following shows the distribution of clusters :

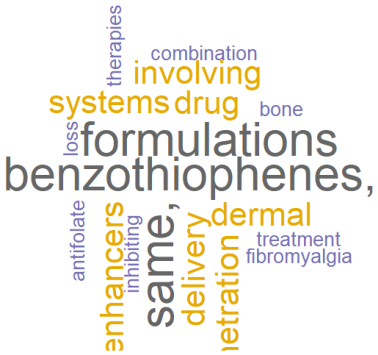
```
 1  2  3  4  5
 7 23  2 43  6
```

These are the topwords from left to right according to the clusters:

```
> print(topwords)
     [,1]           [,2]           [,3]           [,4]           [,5]           [,6]           [,7]           [,8]
[1,] "c07d403/04;"  "c07d409/04;"  "c07d231/48;"  "c07d417/12;" "c07d405/14;" "c07d403/14;" "c07d401/12;"  "c07d401/14;"
[2,] "a61k31/41"    "compositions" "comprising"   "a61k45/06;"  "a61k31/192;" "a61k31/41;"  "a61k31/216;"  "composition"
[3,] "respiratory"  "treatment"    "a61k31/46"    "a61k45/06;"  "a61k9/0075;" "a61k9/145;"  "make.row.names" "a61k31/00;"
[4,] "a61k31/4196;" "a61k31/475;"  "a61k31/502;"  "a61k31/704;" "a61k31/7068;" "c07d413/14;" "treatment"    "a61k31/436;"
[5,] "agonists"     "a61k45/06;"   "a61k31/4709;" "c07c215/80;" "c07c233/43;"  "c07c311/08;" "c07d215/227"   "c07d215/26;"
     [,9]             [,10]
[1,] "c07d403/12;"    "make.row.names"
[2,] "pharmaceutical" "make.row.names"
[3,] "a61k31/46;"     "a61k31/40;"
[4,] "compounds"      "make.row.names"
[5,] "c07d405/12;"    "make.row.names"
>
```

2. LILLY

Lilly is the company that we used for our analysis; the figure shows the top most word and patent used in that particular data provided by the Lilly company.



Following shows the distribution of the cluster:

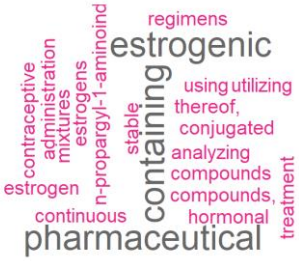
```
1 2
6 2
```

These are the topwords in the cluster:

```
     [,1]              [,2]             [,3]              [,4]             [,5]
[1,] "a61k31/525;.1" "a61k31/714;.1" "c07d333/56"     "c07d333/56.1" "make.row.names"
[2,] "a61k9/122.1"   "a61k9/12;.1"   "a61m35/003;.1" "a61q17/04;.1" "make.row.names"
>
```

3. TEVA

Teva is the company that we used for our analysis; the figure shows the top most word and patent used in that particular data provided by the Teva company.



Following shows the distribution of the cluster:

```
1 2
1 6
```

These are the topwords in the cluster:

```
     [,1]       [,2]             [,3]             [,4]           [,5]
[1,] "mixtures" "pharmaceutical" "make.row.names" "a61k31/565;" "a61k31/566;"
[2,] "hormonal" "regimens"       "treatment"      "utilizing"    "make.row.names"
> }
```

4. ALLOS

ALLOS is the company that we used for our analysis; the figure shows the top most word and patent used in that particular data provided by the ALLOS company.

Following shows the distribution of the cluster:

```
1 2
1 2
```

These are the topwords in the cluster:

```
     [,1]          [,2]             [,3]           [,4]          [,5]
[1,] "c07d475/08;" "make.row.names" "a61k31/555;" "a61k33/24;"  "a61k31/505;"
[2,] "a61k31/505;" "a61k31/555;"    "a61k33/24;"  "c07d475/08;" "make.row.names"
```

5. **MERCK**

MERCK is the company that we used for our analysis; the figure shows the top most word and patent used in that particular data provided by the MERCK company.

Following shows the distribution of the cluster:

```
1  2
2 30
```

These are the topwords in the cluster:

```
     [,1]          [,2]          [,3]             [,4]          [,5]
[1,] "substituted" "a61k9/0019;" "make.row.names" "a61k31/724;" "a61k31/496;"
[2,] "a61k31/44;"  "a61k9/4866;" "treating"       "a61k47/12;"  "make.row.names"
>
```

I think Novartis company has the most strongest patent by looking at the number patents frequency coming in the list of top words in the cluster analysis of the organizations or companies I have assumed.

Useful links:

https://www.novartis.com/our-focus/cancer

http://www.lillyoncology.com/

*Here in all the companies top words don't consider make.row.names.

Question5:

for the 5 academic organizations that have received NIH funding (as measured by the number of patent documents with NIH funding, look at the field "NIH Grant Recipient Organization"). Because academic organizations' names may appear multiple times if multiple inventors were affiliated with it, extract the name of the academic organizations that have received funding most often using the function grep as in the R code for Novartis and then run the clustering.

Solution:

1. **DREXEL UNIVERSITY:**

   Drexel university is the academic organization that we used for our analysis, the figure shows the top most word and patent used in that particular data provided by the academic organization.

Following shows the distribution of the cluster:

```
1 2 3 4 5
2 5 1 2 1
>
```

These are the topwords in the cluster:

```
> print(topwords)
      [,1]          [,2]          [,3]          [,4]          [,5]          [,6]          [,7]          [,8]
[1,] "c12q1/6886"  "a61k31/451;" "a61k31/453;" "a61k31/7088"  "a61k39/3955;" "c07d403/04;" "c07d487/04;" "c07k16/28;"
[2,] "h01141/43;"  "g01n29/036"  "g01n33/53;"  "g01n33/54373;" "a61k38/08"    "c07k7/06;"   "c07k7/08;"   "g01n33/56988;"
[3,] "diagnostic"  "gene"        "identification" "prostate"    "specific"    "treatment"   "useful"      "c12q1/6886;"
[4,] "a61b5/0053;" "a61b5/4244;" "a61b5/4312;" "a61b5/4381"   "a61b5/441;"  "a61b8/085;"  "a61b8/403;"  "a61b8/485;"
[5,] "pre-mrna"    "splice"      "therapeutic" "using"       "a61k48/005;" "c12n15/111;" "c12n15/1138" "c12n15/113;"
      [,9]          [,10]
[1,] "methods"     "make.row.names"
[2,] "piezoelectric" "make.row.names"
[3,] "g01n33/57434" "make.row.names"
[4,] "g01n29/0672;" "make.row.names"
[5,] "c12n15/11;"   "make.row.names"
>
```

2. **BROWN UNIVESITY:**

   Brown university is the academic organization that we used for our analysis , the figure shows the top most word and patent used in that particular data provided by the academic organization.

Following shows the distribution of the cluster:

```
1 2
5 2
```

These are the topwords in the cluster:

```
> print(topwords)
     [,1]       [,2]        [,3]             [,4]          [,5]
[1,] "agents"   "complexes" "organometallic" "therapeutic" "make.row.names"
[2,] "disorders" "group"    "mglur"          "treating"    "make.row.names"
>
```

3. CORNELL UNIVERSITY:

Cornell university is the academic organization that we used for our analysis , the figure shows the top most word and patent used in that particular data provided by the academic organization.

Following shows the distribution of the cluster:

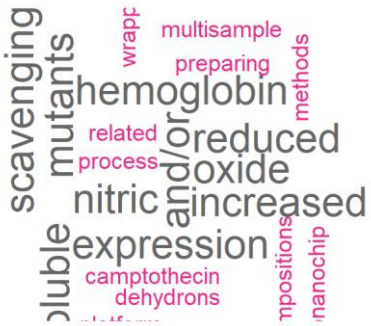
```
 1  2  3
15 65  1
```

These are the topwords in the cluster:

```
     [,1]             [,2]          [,3]          [,4]          [,5]
[1,] "c12q1/6851;"    "c12q1/683;"  "c12q1/6853;" "c12q1/6858;" "c12q1/686;"
[2,] "ligase"         "sequence"    "reaction"    "c12q1/6837;" "make.row.names"
[3,] "make.row.names" "c12q1/6837;" "c12q1/6841;" "c12q1/6855;" "c12q1/6874;"
>
```

4. RICE UNIVERSITY:

Rice university is the academic organization that we used for our analysis the figure shows the top most word and patent used in that particular data provided by the academic organization.



Following shows the distribution of the cluster:

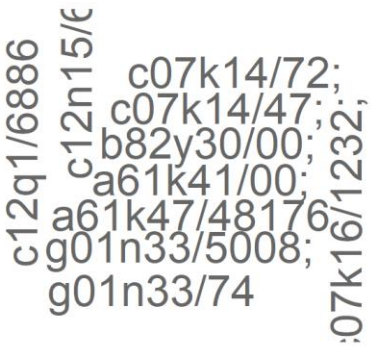
```
1 2
2 3
```

These are the topwords in the cluster:

```
     [,1]          [,2]           [,3]           [,4]          [,5]
[1,] "reduced"     "scavenging"   "soluble"      "c07k14/805" "make.row.names"
[2,] "a61k31/506" "c07d401/04;" "b01l3/5025;" "b01l3/5027" "make.row.names"
>
```

5. **DUKE UNIVERSITY:**

Duke university is the academic organization that we used for our analysis the figure shows the top most word and patent used in that particular data provided by the academic organization.



Following shows the distribution of the cluster:

```
1 2
3 4
```

These are the topwords in the cluster:

```
> print(topwords)
     [,1]          [,2]             [,3]          [,4]           [,5]
[1,] "upregulate" "make.row.names" "pathologic" "nitrosative" "stress"
[2,] "c12q1/6886" "g01n33/74"      "receptor"   "methods"     "make.row.names"
>
```

I think Cornell university has the received most grants by NIH after analyzing the clusters . One factor that can be consider is number of patents coming in the list of top words in the cluster analysis of the academic organizations I have assumed.

*Here in all the academic organizations top words don't consider make.row.names, methods, using, treating these words doesn't count or have that much of importance.

Conclusions

According to the insights developed above, we used cluster analysis to analyse the data and get the desired output that the strongest patent and academic organizations that have received NIH funding.

Strongest Patent company was difficult to choose as there is a close competition between MERCK and NOVARITIS . But while analyzing it Novaritis was opted to be the strongest one.

I think Drexel university and Cornell university has the received most grants by NIH after analyzing the clusters . One factor that can be consider is number of patents coming in the list of top words in the cluster analysis of the academic organizations I have assumed , so the result was Cornell university.

Appendix

1. (Reference: http://www.patbase.com/linkclass.asp?CCC=CPC&CLASS=C07K14%2F51)

```
2. set.seed(31)
install.packages("tm")
library(tm)
data=read.csv("HW3.csv")
str(data)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle=data$Patent_Title
ndocs =length(mydatatitle)
minf=ndocs*0.02
maxf=ndocs*0.50
corpus=Corpus(VectorSource(mydatatitle))
dtm = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
inspect(dtm)
dtmm=as.matrix(dtm)
write.csv(dtmm,"cancerdtm.csv")
View(dtmm)
FreqMat=data.frame(ST = colnames(dtmm), Freq = colSums(dtmm))
y= as.vector(order(-FreqMat$Freq))
data1=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
library(wordcloud)
library(RColorBrewer)
na.omit(data1)
wordcloud(words = data1$ST, freq = data1$Freq, min.freq = 3000,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
bigdata=
cbind(data$Family_ID,data$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm)),make.row.
names=TRUE)
str(bigdata)
write.csv(bigdata,"cancerbigdtm.csv")
# answer 2
#K by kmeans
library(caret)
preproc=preProcess(dtmm)
```

```
cancer=predict(preproc,dtmm)
nc=14
kmc.cancer=kmeans(dtmm, centers = 14)
table(kmc.cancer$cluster)
kmc.cancer$centers

for (i in 1:nc)
{
  clusterdata = subset(as.data.frame(dtmm), kmc.cancer$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)
#used fviz_nbclust silhouette method.
#answer 2 by hierachial
distances = dist(cancer, method="euclidean")
HierClustering = hclust(distances, method="ward.D2")
plot(HierClustering)
clusterGroups = cutree(HierClustering, k = 14)
rect.hclust(HierClustering,k=14, border="red")
table(clusterGroups)
kmax=20
fviz_nbclust(cancer, FUN = hcut, method = "silhouette", k.max=kmax)
#taking number of clusters = 14 that is optimal that i obatined by hierachial cluster analysis

#answer3
library(tm)
data=read.csv("HW3.csv")
str(data)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle1= data$CPC_Inventive
ndocs =length(mydatatitle1)
minf=ndocs*0.01
maxf=ndocs*0.50
corpus=Corpus(VectorSource(mydatatitle1))
dtm1 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
inspect(dtm1)
dtmm1=as.matrix(dtm1)
write.csv(dtmm1,"cancerdtm1.csv")
```

```
View(dtmm1)
FreqMat=data.frame(ST = colnames(dtmm1), Freq = colSums(dtmm1))
y= as.vector(order(-FreqMat$Freq))
data2=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data2$ST, freq = data2$Freq, min.freq = 3000,
          max.words=400, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
bigdata1=
cbind(data$Family_ID,data$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm1)),as.data.fr
ame(as.matrix(dtm)),make.row.names=TRUE)
str(bigdata1)
write.csv(bigdata1,"cancerbigdtm1.csv")
library(caret)
preproc=preProcess(bigdata1)
cancer1=predict(preproc,bigdata1)
nc=8
kmc.cancer1=kmeans(bigdata1[,c(-1,-2)], centers = nc )
table(kmc.cancer1$cluster)
kmc.cancer1$centers
#0.5 marks # used low number of data for the optimal clusters  as fviz was not
#running for large data sets was giving error of memory.
kmax=10
library(cluster)
library(factoextra)
fviz_nbclust(cancer1, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
nb =14
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdata1[,c(-1,-2)], kmc.cancer1$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)

#answer 4
#NOVARTIS
indicesnovartis=grep("NOVARTIS",data$FDA_Applicant)
mydatanova=data[indicesnovartis,]
mydatanovartis=subset(mydatanova,mydatanova$year>=1999)
```

```
str(mydatanovartis)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle2=mydatanovartis$Patent_Title
ndocs =length(mydatatitle2)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle2))
dtm2= DocumentTermMatrix(corpus,control=list(stopwords=TRUE, wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds = list(global=c(minf,maxf))))
dtmm2=as.matrix(dtm2)
mydatatitle2= mydatanovartis$CPC_Inventive
ndocs =length(mydatatitle2)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle2))
dtm3 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE, wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds = list(global=c(minf,maxf))))
dtmm3=as.matrix(dtm3)
write.csv(dtmm3,"cancerdtm1.csv")
FreqMat=data.frame(ST = colnames(dtmm3), Freq = colSums(dtmm3))
y= as.vector(order(-FreqMat$Freq))
data3=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data3$ST, freq = data3$Freq, min.freq = 3000,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(10, "Dark2"))

bigdatanew= cbind(mydatanovartis$Family_ID ,mydatanovartis$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm2)),as.data.frame(as.matrix(dtm3)),make.row.names=TRUE)
str(bigdatanew)
write.csv(bigdatanew,"cancerbigdtmnew.csv")
library(caret)
preproc=preProcess(bigdatanew)
cancer3=predict(preproc,bigdatanew)
nc=5
kmc.cancer3=kmeans(bigdatanew[,c(-1,-2)], centers = nc )
```

```
table(kmc.cancer3$cluster)
kmc.cancer3$centers
kmax=10
fviz_nbclust(bigdatanew, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
nb =10
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew[,c(-1,-2)], kmc.cancer3$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)
#LILLY
indiceslilly=grep("LILLY",data$FDA_Applicant)
mydatalilly=data[indiceslilly,]
mydatalil=subset(mydatalilly,mydatalilly$year>=1999)
str(mydatalil)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle3=mydatalil$Patent_Title
ndocs =length(mydatatitle3)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle3))
dtm4= DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm4=as.matrix(dtm4)
mydatatitle4= mydatalil$CPC_Inventive
ndocs =length(mydatatitle4)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle4))
dtm5 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm5=as.matrix(dtm5)
write.csv(dtmm5,"cancerdtm2.csv")
FreqMat=data.frame(ST = colnames(dtmm4), Freq = colSums(dtmm4))
y= as.vector(order(-FreqMat$Freq))
```

```
data4=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data4$ST, freq = data4$Freq, min.freq = 3000,
          max.words=400, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
bigdatanew2=                                                    cbind(mydatalil$Family_ID
,mydatalil$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm5)),as.data.frame(as.matrix(dt
m4)),make.row.names=TRUE)
str(bigdatanew2)
library(caret)
preproc=preProcess(bigdatanew2)
cancer4=predict(preproc,bigdatanew2)
nc=2
kmc.cancer4=kmeans(bigdatanew2[,c(-1,-2)], centers = nc )
table(kmc.cancer4$cluster)
kmc.cancer4$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew2[,c(-1,-2)], kmc.cancer4$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)

na.omit(cancer4)
kmax=10
fviz_nbclust(cancer4, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
print(topwords)
#TEVA
indicesTEVA=grep("TEVA",data$FDA_Applicant)
mydataTEVA=data[indicesTEVA,]
mydataTEV=subset(mydataTEVA,mydataTEVA$year>=1999)
str(mydataTEV)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle5=mydataTEV$Patent_Title
ndocs =length(mydatatitle5)
minf=ndocs*0.02
```

```
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle5))
dtm6=                    DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm6=as.matrix(dtm6)
mydatatitle6= mydataTEV$CPC_Inventive
ndocs =length(mydatatitle6)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle6))
dtm7          =          DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm7=as.matrix(dtm7)
write.csv(dtmm7,"cancerdtm6.csv")
FreqMat=data.frame(ST = colnames(dtmm7), Freq = colSums(dtmm7))
y= as.vector(order(-FreqMat$Freq))
data5=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data5$ST, freq = data5$Freq, min.freq = 3000,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))

bigdatanew3=                                                cbind(mydataTEV$Family_ID
,mydataTEV$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm6)),as.data.frame(as.matrix(
dtm7)),make.row.names=TRUE)
str(bigdatanew3)
library(caret)
preproc=preProcess(bigdatanew3)
can=predict(preproc,bigdatanew3)
nc=2
kmc.cancer06=kmeans(bigdatanew3[,c(-1,-2)], centers = nc )
table(kmc.cancer06$cluster)
kmc.cancer06$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew3[,c(-1,-2)], kmc.cancer06$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
```

```
  topwords[i,]<- rownames(vv)
}
print(topwords)
kmax=10
fviz_nbclust(can, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
print(topwords)
#ALLOS
indicesALLOS=grep("ALLOS",data$FDA_Applicant)
mydataALLOS=data[indicesALLOS,]
mydataALLO=subset(mydataALLOS,mydataALLOS$year>=1999)
str(mydataALLO)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle8=mydataALLO$Patent_Title
ndocs =length(mydatatitle8)
minf=ndocs*0.02
maxf=ndocs*0.55
corpus=Corpus(VectorSource(mydatatitle8))
dtm8= DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm8=as.matrix(dtm8)
mydatatitle9= mydataALLO$CPC_Inventive
ndocs =length(mydatatitle9)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle9))
dtm9 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm9=as.matrix(dtm9)
write.csv(dtmm7,"cancerdtm6.csv")
FreqMat=data.frame(ST = colnames(dtmm8), Freq = colSums(dtmm8))
y= as.vector(order(-FreqMat$Freq))
data6=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data6$ST, freq = data6$Freq, min.freq = 3000,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
```

```
bigdatanew4= cbind(mydataALLO$Family_ID
,mydataALLO$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm8)),as.data.frame(as.matri
x(dtm9)),make.row.names=TRUE)
str(bigdatanew4)
library(caret)
preproc=preProcess(bigdatanew4)
can=predict(preproc,bigdatanew4)
nc=2
kmc.cancer07=kmeans(bigdatanew4[,c(-1,-2)], centers = nc )
table(kmc.cancer07$cluster)
kmc.cancer07$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew4[,c(-1,-2)], kmc.cancer07$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)

kmax=10
fviz_nbclust(can, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
print(topwords)

#MERCK
indicesDUSA=grep("MERCK",data$FDA_Applicant)
mydataDUSA=data[indicesDUSA,]
mydataDUS=subset(mydataDUSA,mydataALLOS$year>=1999)
str(mydataDUS)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle10=mydataDUS$Patent_Title
ndocs =length(mydatatitle10)
minf=ndocs*0.03
maxf=ndocs*0.50
corpus=Corpus(VectorSource(mydatatitle10))
dtm10= DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm10=as.matrix(dtm10)
```

```
mydatatitle11= mydataDUS$CPC_Inventive
ndocs =length(mydatatitle11)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle11))
dtm11 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm11=as.matrix(dtm11)
write.csv(dtmm07,"cancerdtm8.csv")
FreqMat=data.frame(ST = colnames(dtmm11), Freq = colSums(dtmm11))
y= as.vector(order(-FreqMat$Freq))
data5=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data5$ST, freq = data5$Freq, min.freq = 3000,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
bigdatanew5= cbind(mydataDUS$Family_ID
,mydataDUS$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm10)),as.data.frame(as.matri
x(dtm11)),make.row.names=TRUE)
str(bigdatanew5)
library(caret)
preproc=preProcess(bigdatanew5)
can=predict(preproc,bigdatanew5)
nc=2
kmc.cancer08=kmeans(bigdatanew5[,c(-1,-2)], centers = nc )
table(kmc.cancer08$cluster)
kmc.cancer08$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew5[,c(-1,-2)], kmc.cancer08$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)
na.omit(can)
kmax=10
fviz_nbclust(can, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
```

```
#ANSWER 5
#DREXEL UNIVERSITY
indicesnovartis=grep("DREXEL UNIVERSITY",data$NIH_Grant_Recipient_Organization)
mydataAUBURN=data[indicesnovartis,]
mydataAUB=subset(mydataAUBURN,mydataAUBURN$year>=1999)
str(mydataAUB)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle2=mydataAUB$Patent_Title
ndocs =length(mydatatitle2)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle2))
dtm2= DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm2=as.matrix(dtm2)
mydatatitle2= mydataAUB$CPC_Inventive
ndocs =length(mydatatitle2)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle2))
dtm3 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm3=as.matrix(dtm3)
write.csv(dtmm3,"cancerdtm1.csv")
FreqMat=data.frame(ST = colnames(dtmm2), Freq = colSums(dtmm2))
y= as.vector(order(-FreqMat$Freq))
data3=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data3$ST, freq = data3$Freq, min.freq = 1000,
        max.words=100, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))

bigdatanew= cbind(mydataAUB$Family_ID
,mydataAUB$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm2)),as.data.frame(as.matrix
(dtm3)),make.row.names=TRUE)
str(bigdatanew)
write.csv(bigdatanew,"cancerbigdtmnew.csv")
```

```
library(caret)
preproc=preProcess(bigdatanew)
cancer3=predict(preproc,bigdatanew)
nc=5
kmc.cancer3=kmeans(bigdatanew[,c(-1,-2)], centers = nc )
table(kmc.cancer3$cluster)
kmc.cancer3$centers
kmax=10
fviz_nbclust(bigdatanew, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
nb =10
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew[,c(-1,-2)], kmc.cancer3$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)
kmax=10
fviz_nbclust(can, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)

#BROWN UNIVERSITY
indicesBROWN=grep("BROWN UNIVERSITY",data$NIH_Grant_Recipient_Organization)
mydataBROWN=data[indicesBROWN,]
mydataBRO=subset(mydataBROWN,mydataBROWN$year>=1999)
str(mydataBRO)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle3=mydataBRO$Patent_Title
ndocs =length(mydatatitle3)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle3))
dtm4=                                DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm4=as.matrix(dtm4)
mydatatitle4= mydataBRO$CPC_Inventive
ndocs =length(mydatatitle4)
minf=ndocs*0.01
maxf=ndocs*0.40
```

```
corpus=Corpus(VectorSource(mydatatitle4))
dtm5 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm5=as.matrix(dtm5)
write.csv(dtmm5,"cancerdtm2.csv")
FreqMat=data.frame(ST = colnames(dtmm5), Freq = colSums(dtmm5))
y= as.vector(order(-FreqMat$Freq))
data4=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data4$ST, freq = data4$Freq, min.freq = 3000,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
bigdatanew2= cbind(mydataBRO$Family_ID
,mydataBRO$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm5)),as.data.frame(as.matrix
(dtm4)),make.row.names=TRUE)
str(bigdatanew2)
library(caret)
preproc=preProcess(bigdatanew2)
cancer4=predict(preproc,bigdatanew2)
nc=2
kmc.cancer4=kmeans(bigdatanew2[,c(-1,-2)], centers = nc )
table(kmc.cancer4$cluster)
kmc.cancer4$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew2[,c(-1,-2)], kmc.cancer4$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)
na.omit(cancer4)
kmax=10
fviz_nbclust(cancer4, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
print(topwords)
#CORNELL UNIVERSITY
indicesCORNELL=grep("CORNELL UNIVERSITY",data$NIH_Grant_Recipient_Organization)
mydataCORNELL=data[indicesCORNELL,]
mydataCORN=subset(mydataCORNELL,mydataCORNELL$year>=1999)
```

```
str(mydataCORN)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle5=mydataCORN$Patent_Title
ndocs =length(mydatatitle5)
minf=ndocs*0.02
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle5))
dtm6= DocumentTermMatrix(corpus,control=list(stopwords=TRUE, wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds = list(global=c(minf,maxf))))
dtmm6=as.matrix(dtm6)
mydatatitle6= mydataCORN$CPC_Inventive
ndocs =length(mydatatitle6)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle6))
dtm7 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE, wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds = list(global=c(minf,maxf))))
dtmm7=as.matrix(dtm7)
write.csv(dtmm7,"cancerdtm6.csv")
FreqMat=data.frame(ST = colnames(dtmm6), Freq = colSums(dtmm6))
y= as.vector(order(-FreqMat$Freq))
data5=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data5$ST, freq = data5$Freq, min.freq = 3000,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))

bigdatanew3= cbind(mydataCORN$Family_ID ,mydataCORN$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm6)),as.data.frame(as.matrix(dtm7)),make.row.names=TRUE)
str(bigdatanew3)
library(caret)
preproc=preProcess(bigdatanew3)
can=predict(preproc,bigdatanew3)
nc=3
kmc.cancer06=kmeans(bigdatanew3[,c(-1,-2)], centers = nc )
table(kmc.cancer06$cluster)
```

```
kmc.cancer06$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew3[,c(-1,-2)], kmc.cancer06$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)
kmax=10
fviz_nbclust(can, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
print(topwords)
#RICE UNIVERSITY
indicesRICE=grep("RICE UNIVERSITY",data$NIH_Grant_Recipient_Organization)
mydataRICE=data[indicesRICE,]
mydataRIC=subset(mydataRICE,mydataRICE$year>=1999)
str(mydataRIC)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle8=mydataRIC$Patent_Title
ndocs =length(mydatatitle8)
minf=ndocs*0.02
maxf=ndocs*0.55
corpus=Corpus(VectorSource(mydatatitle8))
dtm8= DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm8=as.matrix(dtm8)
mydatatitle9= mydataRIC$CPC_Inventive
ndocs =length(mydatatitle9)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle9))
dtm9 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm9=as.matrix(dtm9)
write.csv(dtmm7,"cancerdtm6.csv")
FreqMat=data.frame(ST = colnames(dtmm9), Freq = colSums(dtmm9))
y= as.vector(order(-FreqMat$Freq))
```

```
data6=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data6$ST, freq = data6$Freq, min.freq = 3000,
          max.words=400, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))

bigdatanew4=                                                    cbind(mydataRIC$Family_ID
,mydataRIC$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm8)),as.data.frame(as.matrix(
dtm9)),make.row.names=TRUE)
str(bigdatanew4)
library(caret)
preproc=preProcess(bigdatanew4)
can=predict(preproc,bigdatanew4)
nc=2
kmc.cancer07=kmeans(bigdatanew4[,c(-1,-2)], centers = nc )
table(kmc.cancer07$cluster)
kmc.cancer07$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew4[,c(-1,-2)], kmc.cancer07$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
}
print(topwords)

kmax=10
fviz_nbclust(can, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)

#DUKE UNIVERSITY
indicesDUKE=grep("DUKE UNIVERSITY",data$NIH_Grant_Recipient_Organization)
mydataDUKE=data[indicesDUKE,]
mydataDUK=subset(mydataDUKE,mydataDUKE$year>=1999)
str(mydataDUK)
library(tm)
yt=strptime(data$Filing_Date, format ='%Y%m%d')
data$year= as.integer(format(yt,"%Y"))
mydatatitle10=mydataDUK$Patent_Title
ndocs =length(mydatatitle10)
minf=ndocs*0.03
```

```
maxf=ndocs*0.50
corpus=Corpus(VectorSource(mydatatitle10))
dtm10= DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm10=as.matrix(dtm10)
mydatatitle11= mydataDUK$CPC_Inventive
ndocs =length(mydatatitle11)
minf=ndocs*0.01
maxf=ndocs*0.40
corpus=Corpus(VectorSource(mydatatitle11))
dtm11 = DocumentTermMatrix(corpus,control=list(stopwords=TRUE,
wordLengths=c(4,25),removePunctuation = FALSE,removeNumbers = FALSE, bounds =
list(global=c(minf,maxf))))
dtmm11=as.matrix(dtm11)
write.csv(dtmm07,"cancerdtm8.csv")
FreqMat=data.frame(ST = colnames(dtmm10), Freq = colSums(dtmm10))
y= as.vector(order(-FreqMat$Freq))
data5=FreqMat[y,]
head(FreqMat[y,],n=10)
head(FreqMat[y,],n=20)
wordcloud(words = data5$ST, freq = data5$Freq, min.freq = 1,
        max.words=400, random.order=FALSE, rot.per=0.35,
        colors=brewer.pal(8, "Dark2"))
bigdatanew5= cbind(mydataDUK$Family_ID
,mydataDUK$Patent_or_Publication_ID,as.data.frame(as.matrix(dtm10)),as.data.frame(as.matri
x(dtm11)),make.row.names=TRUE)
str(bigdatanew5)
library(caret)
preproc=preProcess(bigdatanew5)
can=predict(preproc,bigdatanew5)
nc=2
kmc.cancer08=kmeans(bigdatanew5[,c(-1,-2)], centers = nc )
table(kmc.cancer08$cluster)
kmc.cancer08$centers
nb =5
topwords <-matrix(0,nrow=nc,ncol=nb)
for (i in 1:nc)
{
  clusterdata = subset(bigdatanew5[,c(-1,-2)], kmc.cancer08$cluster == i)
  vv <- as.data.frame(tail(sort(colMeans(clusterdata)), n=nb))
  topwords[i,]<- rownames(vv)
```

```
}
print(topwords)
kmax=10
fviz_nbclust(can, FUN = kmeans, method = "silhouette", k.max=kmax, nstart=1)
```