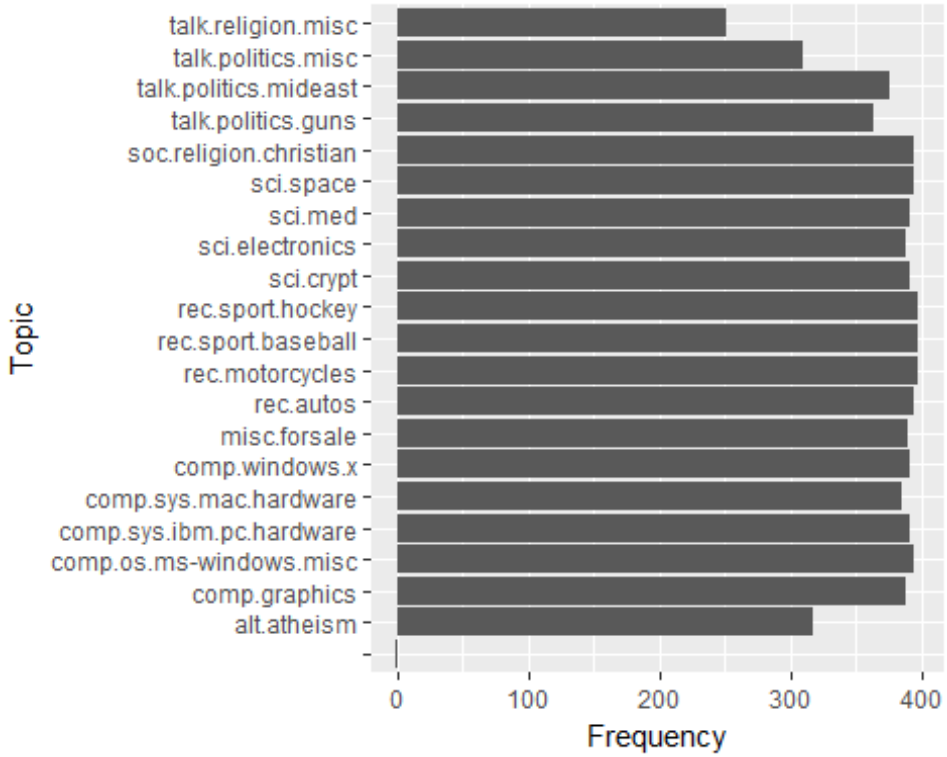
SAI CHARAN TALIPINENI sat122@pitt.edu

homework5

sai charan talipineni

28 March 2017

```
#Task1

1.1

library(plyr)
library(ggplot2)
library(tm)

library(lsa)

library(NMF)

set.seed(123)
ng<-read.csv("D:/semester/2nd sem/DATA_MINING/hw5/Newsgroup.csv",header =
TRUE, stringsAsFactors = FALSE, fill = TRUE)

ggplot(data = ng, aes(x=factor(Topic))) + geom_bar(stat="count") +
scale_x_discrete("Topic") + scale_y_continuous("Frequency") + coord_flip()
```



```
top.topics = sort(table(ng$Topic), decreasing = T)[1:4]
top.topics = names(top.topics)
top.topics
```

```
## [1] "rec.motorcycles"    "rec.sport.baseball" "rec.sport.hockey"
## [4] "rec.autos"
```

1.2

```
doc_idx = which(ng$Topic %in% top.topics)
subdoc = ng[doc_idx,]

corpus = Corpus(VectorSource(subdoc$Content))
corpus
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1587
```

```
corpus = tm_map(corpus, content_transformer(tolower))
inspect(corpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] auto air condit freon articl hvx new cso uiuc edu tspila uxa cso uiuc
edu tim spila romulan write articl apr ntuix ntu mgqlu ntuix ntu max write
work ga solid adsorpt air con system for auto applic thi kind system energi
for regener adsorb exhaust ga interest thi mail email follow thi thread
discuss prospect thi technolog bite thi suppos work tim year ago demonstr
cold air system us air call rovax unit work short come seal technolog todai
## [2] auto air condit freon rovax tobia convex allen tobia write year ago
demonstr cold air system us air call rovax unit work short come seal
technolog todai recal read post back rovax rovac di larger and noisier compet
cheap system dai case bad time system chanc todai that system death row
investor hard come second time jon hacker march beta rom caltech pasadena for
call ibm hacker tumbler ridg caltech edu read comp beta
## [3] auto air condit freon simpl principl porou adsorb zeolit and activ
carbon can adsorb gase evapor adsorb water methanol etc give cool effect heat
ga satur adsorb bed will give gase condens thi form adsorpt refriger cycl
problem that cop low max max phd internet mgqlu ntu divis thermal engin
bitnet mgqlu ntuvax bitnet school mpe nanyang technolog univers phone nanyang
avenu singapor fax
```

```
corpus = tm_map(corpus, removePunctuation)
inspect(corpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
```

```
##
## [1] auto air condit freon articl hvx new cso uiuc edu tspila uxa cso uiuc
edu tim spila romulan write articl apr ntuix ntu mgqlu ntuix ntu max write
work ga solid adsorpt air con system for auto applic thi kind system energi
for regener adsorb exhaust ga interest thi mail email follow thi thread
discuss prospect thi technolog bite thi suppos work tim year ago demonstr
cold air system us air call rovax unit work short come seal technolog todai
## [2] auto air condit freon rovax tobia convex allen tobia write year ago
demonstr cold air system us air call rovax unit work short come seal
technolog todai recal read post back rovax rovac di larger and noisier compet
cheap system dai case bad time system chanc todai that system death row
investor hard come second time jon hacker march beta rom caltech pasadena for
call ibm hacker tumbler ridg caltech edu read comp beta
## [3] auto air condit freon simpl principl porou adsorb zeolit and activ
carbon can adsorb gase evapor adsorb water methanol etc give cool effect heat
ga satur adsorb bed will give gase condens thi form adsorpt refriger cycl
problem that cop low max max phd internet mgqlu ntu divis thermal engin
bitnet mgqlu ntuvax bitnet school mpe nanyang technolog univers phone nanyang
avenu singapor fax

corpus = tm_map(corpus, removeNumbers)
inspect(corpus[1:3])

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] auto air condit freon articl hvx new cso uiuc edu tspila uxa cso uiuc
edu tim spila romulan write articl apr ntuix ntu mgqlu ntuix ntu max write
work ga solid adsorpt air con system for auto applic thi kind system energi
for regener adsorb exhaust ga interest thi mail email follow thi thread
discuss prospect thi technolog bite thi suppos work tim year ago demonstr
cold air system us air call rovax unit work short come seal technolog todai
## [2] auto air condit freon rovax tobia convex allen tobia write year ago
demonstr cold air system us air call rovax unit work short come seal
technolog todai recal read post back rovax rovac di larger and noisier compet
cheap system dai case bad time system chanc todai that system death row
investor hard come second time jon hacker march beta rom caltech pasadena for
call ibm hacker tumbler ridg caltech edu read comp beta
## [3] auto air condit freon simpl principl porou adsorb zeolit and activ
carbon can adsorb gase evapor adsorb water methanol etc give cool effect heat
ga satur adsorb bed will give gase condens thi form adsorpt refriger cycl
problem that cop low max max phd internet mgqlu ntu divis thermal engin
bitnet mgqlu ntuvax bitnet school mpe nanyang technolog univers phone nanyang
avenu singapor fax

corpus = tm_map(corpus, function(x) removeWords(x,
stopwords("english")))
inspect(corpus[1:3])
```

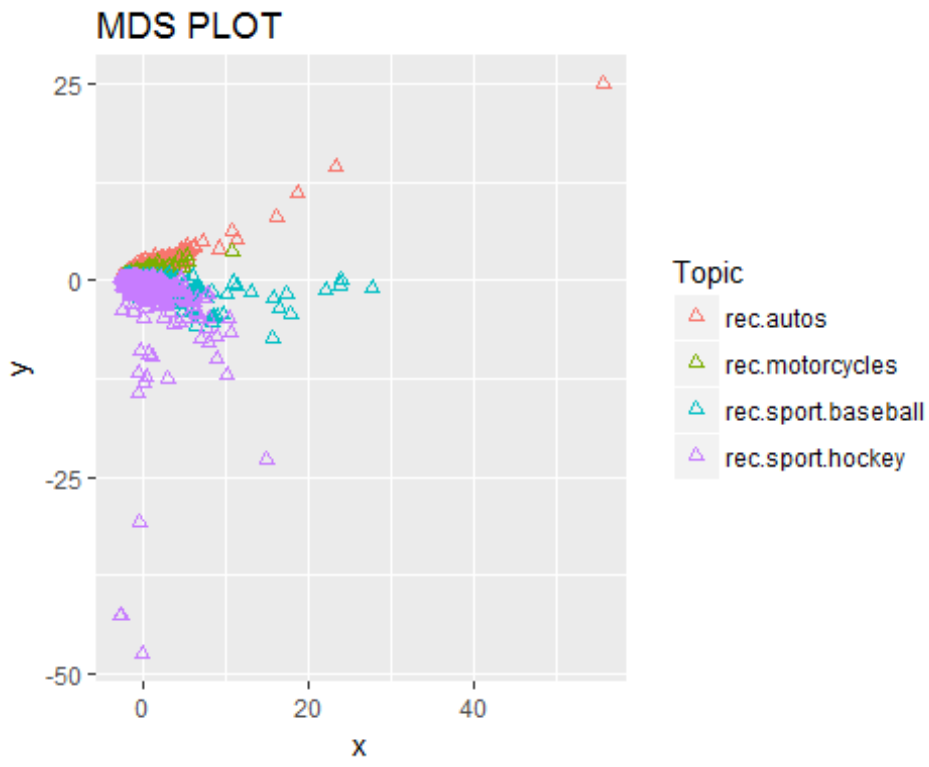
```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] auto air condit freon articl hvx new cso uiuc edu tspila uxa cso uiuc
edu tim spila romulan write articl apr ntuix ntu mgqlu ntuix ntu max write
work ga solid adsorpt air con system  auto applic thi kind system energi
regener adsorb exhaust ga interest thi mail email follow thi thread discuss
prospect thi technolog bite thi suppos work tim year ago demonstr cold air
system us air call rovax unit work short come seal technolog todai
## [2] auto air condit freon rovax tobia convex allen tobia write year ago
demonstr cold air system us air call rovax unit work short come seal
technolog todai recal read post back rovax rovac di larger  noisier compet
cheap system dai case bad time system chanc todai  system death row investor
hard come second time jon hacker march beta rom caltech pasadena  call ibm
hacker tumbler ridg caltech edu read comp beta
## [3] auto air condit freon simpl principl porou adsorb zeolit  activ carbon
can adsorb gase evapor adsorb water methanol etc give cool effect heat ga
satur adsorb bed will give gase condens thi form adsorpt refriger cycl
problem  cop low max max phd internet mgqlu ntu divis thermal engin bitnet
mgqlu ntuvax bitnet school mpe nanyang technolog univers phone nanyang avenu
singapor fax
```

```
corpus = tm_map(corpus, stemDocument, language = "english")
inspect(corpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] auto air condit freon articl hvx new cso uiuc edu tspila uxa cso uiuc
edu tim spila romulan write articl apr ntuix ntu mgqlu ntuix ntu max write
work ga solid adsorpt air con system auto applic thi kind system energi regen
adsorb exhaust ga interest thi mail email follow thi thread discuss prospect
thi technolog bite thi suppo work tim year ago demonstr cold air system us
air call rovax unit work short come seal technolog todai
## [2] auto air condit freon rovax tobia convex allen tobia write year ago
demonstr cold air system us air call rovax unit work short come seal
technolog todai recal read post back rovax rovac di larger noisier compet
cheap system dai case bad time system chanc todai system death row investor
hard come second time jon hacker march beta rom caltech pasadena call ibm
hacker tumbler ridg caltech edu read comp beta
## [3] auto air condit freon simpl principl porou adsorb zeolit activ carbon
can adsorb gase evapor adsorb water methanol etc give cool effect heat ga
satur adsorb bed will give gase conden thi form adsorpt refrig cycl problem
cop low max max phd internet mgqlu ntu divi thermal engin bitnet mgqlu ntuvax
bitnet school mpe nanyang technolog univ phone nanyang avenu singapor fax
```

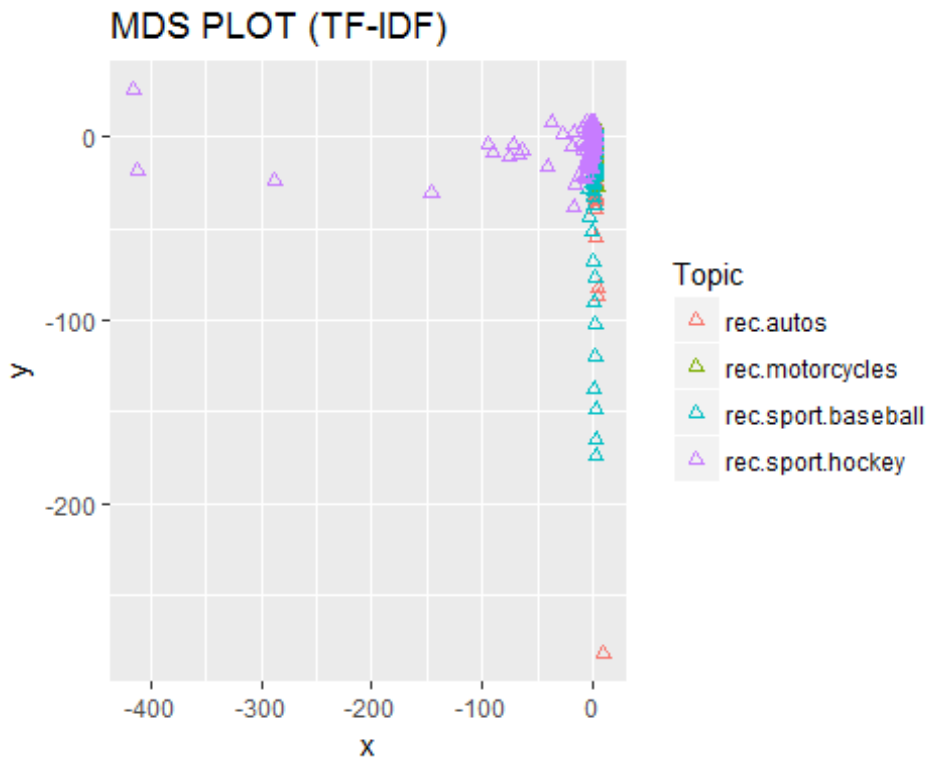
```
corpus = tm_map(corpus, stripWhitespace)
inspect(corpus[1:3])
```

```
## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 3
##
## [1] auto air condit freon articl hvx new cso uiuc edu tspila uxa cso uiuc
edu tim spila romulan write articl apr ntuix ntu mgqlu ntuix ntu max write
work ga solid adsorpt air con system auto applic thi kind system energi regen
adsorb exhaust ga interest thi mail email follow thi thread discuss prospect
thi technolog bite thi suppo work tim year ago demonstr cold air system us
air call rovax unit work short come seal technolog todai
## [2] auto air condit freon rovax tobia convex allen tobia write year ago
demonstr cold air system us air call rovax unit work short come seal
technolog todai recal read post back rovax rovac di larger noisier compet
cheap system dai case bad time system chanc todai system death row investor
hard come second time jon hacker march beta rom caltech pasadena call ibm
hacker tumbler ridg caltech edu read comp beta
## [3] auto air condit freon simpl principl porou adsorb zeolit activ carbon
can adsorb gase evapor adsorb water methanol etc give cool effect heat ga
satur adsorb bed will give gase conden thi form adsorpt refrig cycl problem
cop low max max phd internet mgqlu ntu divi thermal engin bitnet mgqlu ntuvax
bitnet school mpe nanyang technolog univ phone nanyang avenu singapor fax

corpus

## <<SimpleCorpus>>
## Metadata:  corpus specific: 1, document level (indexed): 0
## Content:  documents: 1587

td.mat = TermDocumentMatrix(corpus)
td_freq = findFreqTerms(td.mat, 4)
td.mat_idx = which(row.names(td.mat) %in% td_freq)
td.mat = td.mat[td.mat_idx,]
dim(td.mat)

## [1] 5313 1587

dist.mat = dist(t(as.matrix(td.mat)))
doc.mds = cmdscale(dist.mat)
data = data.frame(x = doc.mds[, 1], y = doc.mds[, 2],
Topic = subdoc$Topic, id = row.names(subdoc))
ggplot(data, aes(x = x , y = y, color = Topic)) + geom_point(shape = 2) +
ggtitle("MDS PLOT")
```

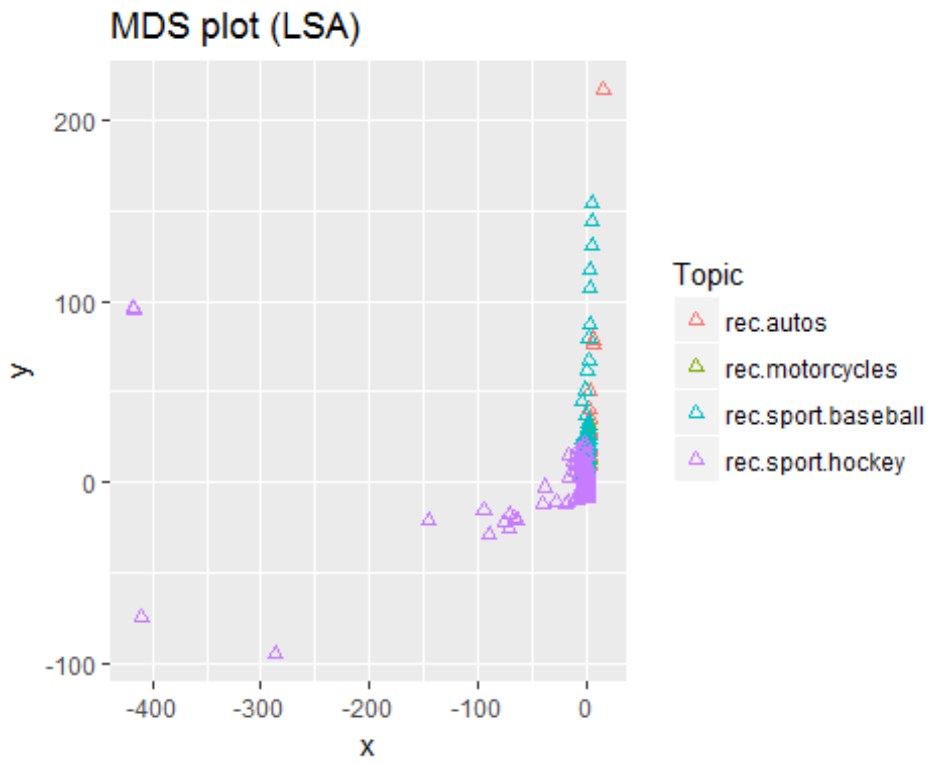


1.3

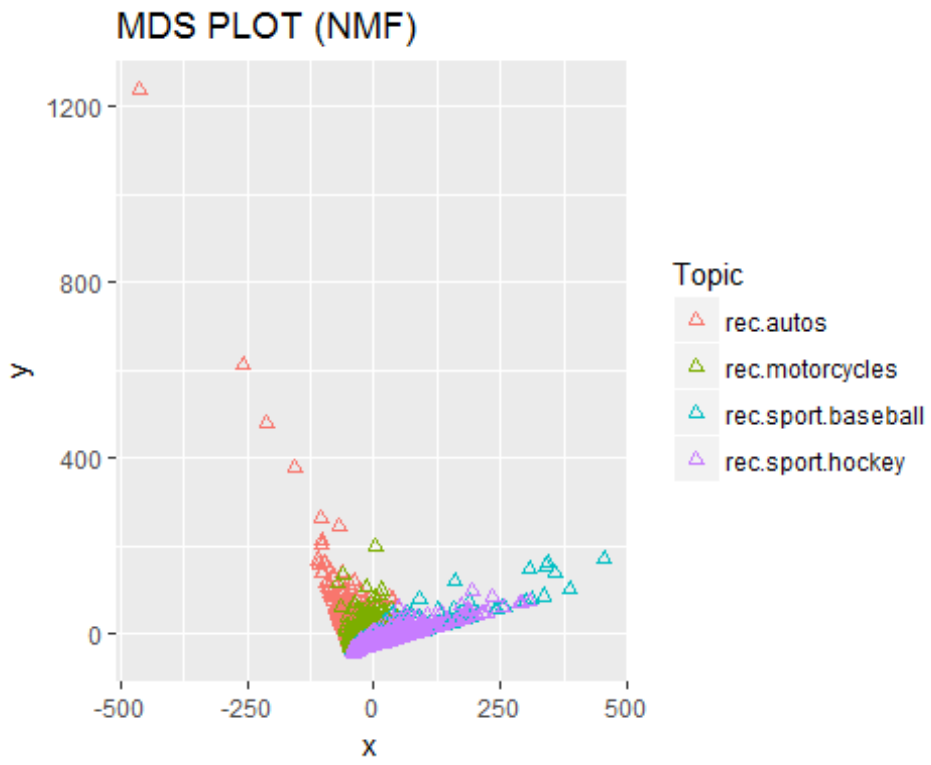
```
td.mat = as.matrix(td.mat)
td.mat.w = lw_tf(td.mat) * gw_idf(td.mat)
dist.mat.w = dist(t(as.matrix(td.mat.w)))
doc.mds.w = cmdscale(dist.mat.w)
data.w = data.frame(x = doc.mds.w[, 1], y = doc.mds.w[, 2], Topic =
subdoc$Topic, id = row.names(subdoc))
ggplot(data.w, aes(x = x, y = y, color = Topic)) + geom_point(shape = 2) +
ggtitle("MDS PLOT (TF-IDF)")
```




```
lsa.space = lsa(td.mat.w, dims = 4)

dist.mat.lsa = dist(t(as.textmatrix(lsa.space)))
doc.mds.lsa = cmdscale(dist.mat.lsa)
dim(doc.mds.lsa)

## [1] 1587    2

data.lsa = data.frame(x = doc.mds.lsa[, 1], y = doc.mds.lsa[, 2],
Topic = subdoc$Topic, id =
row.names(subdoc))
ggplot(data.lsa, aes(x = x, y = y, color = Topic)) + geom_point(shape = 2) +
ggtitle("MDS plot (LSA)")
```




```
nmf_res = nmf(td.mat, 3, "lee")

V_hat = fitted(nmf_res)

W = basis(nmf_res)

H = coef(nmf_res)

dist_mat_nmf = dist(t(H))
doc_mds_nmf = cmdscale(dist_mat_nmf)
data_nmf = data.frame(x = doc_mds_nmf[, 1], y = doc_mds_nmf[, 2],
Topic = subdoc$Topic, id = row.names(subdoc))
ggplot(data_nmf, aes(x = x, y = y, color = Topic)) + geom_point(shape = 2) +
ggtitle("MDS PLOT (NMF)")
```



1.4

```
#From the TF-IDF, we can see the Term importance( = term frequency
(TF)*inverse-document frequency (IDF) )in the documents by Topic.
#From the LSA plot, we can see the similar terms map to similar location in
low dimensional space and a clear low-dimensional space reflects semantic
association
#From the NMF, the data clusters represent the related documents.
```

#Task 2-

2.1

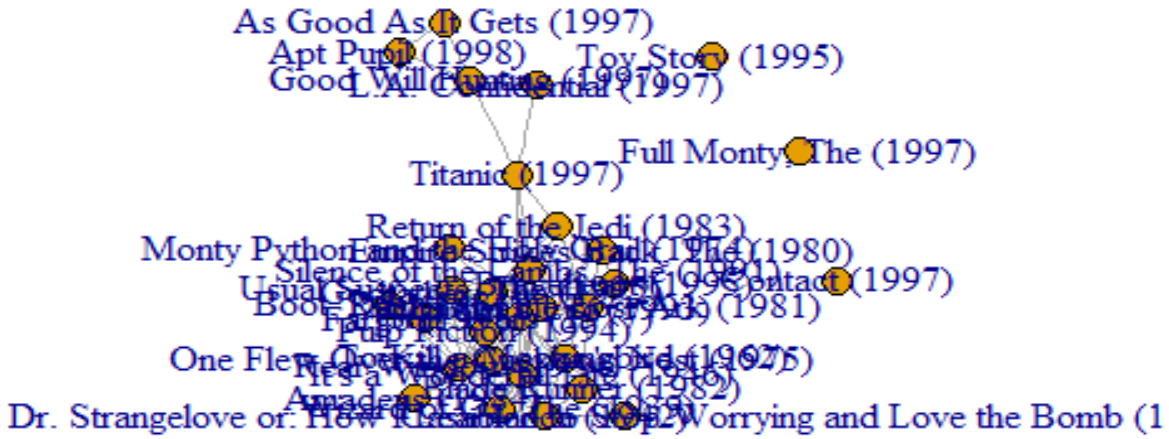
```
set.seed(123)
library(igraph)

udata<-read.table("http://files.grouplens.org/datasets/movielens/ml-
100k/u.data", header = FALSE)
names(udata)<-c("userid","itemid","rating","timestamp")
udata <- udata[which(udata$time >= 890352000),]
udata<-udata[which(udata$rating==5),]
udata$rating<-NULL
udata$timestamp<-NULL
udata$userid<-as.character(udata$userid)
```

```
udata$itemid<-as.character(udata$itemid)
str(sort(unique(udata$userid)))

##  chr [1:206] "100" "107" "11" "111" "112" "116" "119" ...

str(sort(unique(udata$itemid)))

##  chr [1:782] "1" "10" "100" "1005" "1006" "1007" "1009" ...

uitem<-read.delim("http://files.grouplens.org/datasets/movielens/ml-
100k/u.item", sep = "|", header = FALSE)
uitem<-uitem[,c("V1","V2")]
names(uitem)<-c("itemid","movietitle")
uitem$itemid<-as.character(uitem$itemid)
str(sort(unique(uitem$itemid)))

##  chr [1:1682] "1" "10" "100" "1000" "1001" "1002" "1003" ...

umerge<-merge(uitem,udata)
umerge$movietitle<-as.character(umerge$movietitle)
str(sort(unique(umerge$userid)))

##  chr [1:206] "100" "107" "11" "111" "112" "116" "119" ...

str(sort(unique(umerge$itemid)))

##  chr [1:782] "1" "10" "100" "1005" "1006" "1007" "1009" ...

umerge$itemid<-NULL
umerge<-umerge[c(2,1)]

topmovies = sort(table(umerge$movietitle), decreasing = T)
topmovies = rownames(topmovies[1:30])
topmovies[1:10]

##  [1] "Star Wars (1977)"               "Titanic (1997)"
##  [3] "Good Will Hunting (1997)"       "Fargo (1996)"
##  [5] "Schindler's List (1993)"        "Godfather, The (1972)"
##  [7] "L.A. Confidential (1997)"       "As Good As It Gets (1997)"
##  [9] "Raiders of the Lost Ark (1981)" "Rear Window (1954)"

df = subset(umerge, movietitle %in% topmovies)

g = graph.data.frame(df, directed = T)
mat = get.adjacency(g)
mat = as.matrix(mat)
m2 = t(mat) %*% mat
movie.idx = which(colSums(m2) > 0)
mov.matrix = m2[movie.idx, movie.idx]
diag(mov.matrix) = 0
movie.idx = which(colSums(mov.matrix) > 0)
```

```
mov.matrix = mov.matrix[movie.idx, movie.idx]
dim(mov.matrix)

## [1] 30 30

mov.matrix[which(mov.matrix < 10)] = 0

rownames(mov.matrix)[order(colSums(mov.matrix), decreasing = T)][1:10]

##  [1] "Star Wars (1977)"
##  [2] "Godfather, The (1972)"
##  [3] "Schindler's List (1993)"
##  [4] "One Flew Over the Cuckoo's Nest (1975)"
##  [5] "Rear Window (1954)"
##  [6] "Raiders of the Lost Ark (1981)"
##  [7] "Silence of the Lambs, The (1991)"
##  [8] "Pulp Fiction (1994)"
##  [9] "It's a Wonderful Life (1946)"
## [10] "To Kill a Mockingbird (1962)"
```

2.2

```
g = graph.adjacency(mov.matrix, weighted = T, mode = "undirected", diag = F)

plot(g, layout = layout.fruchterman.reingold, vertex.label = V(g)$name)
```



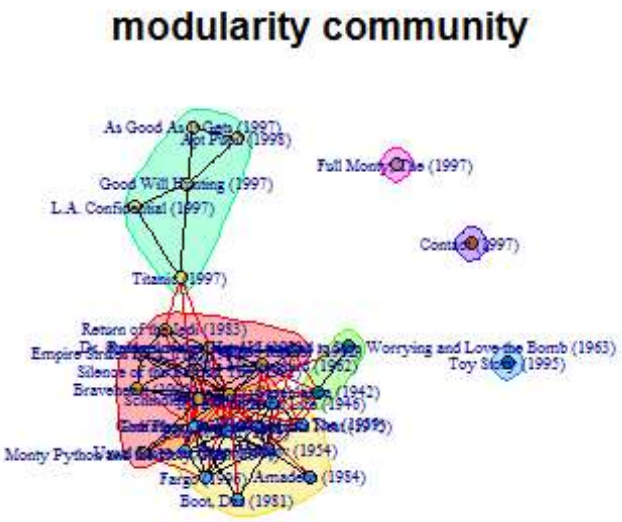
```
plot(g, layout = layout.fruchterman.reingold, vertex.size = 8,
vertex.label.cex = 0.75)
```


```
fc = fastgreedy.community(g)
modularity(fc)

## [1] 0.2296078

membership(fc)

##                                                    Toy Story (1995)
##
5
##                                                    Fargo (1996)
##
2
##                                          Usual Suspects, The (1995)
##
2
##                                               Godfather, The (1972)
##
2
##                                            Wizard of Oz, The (1939)
##
```

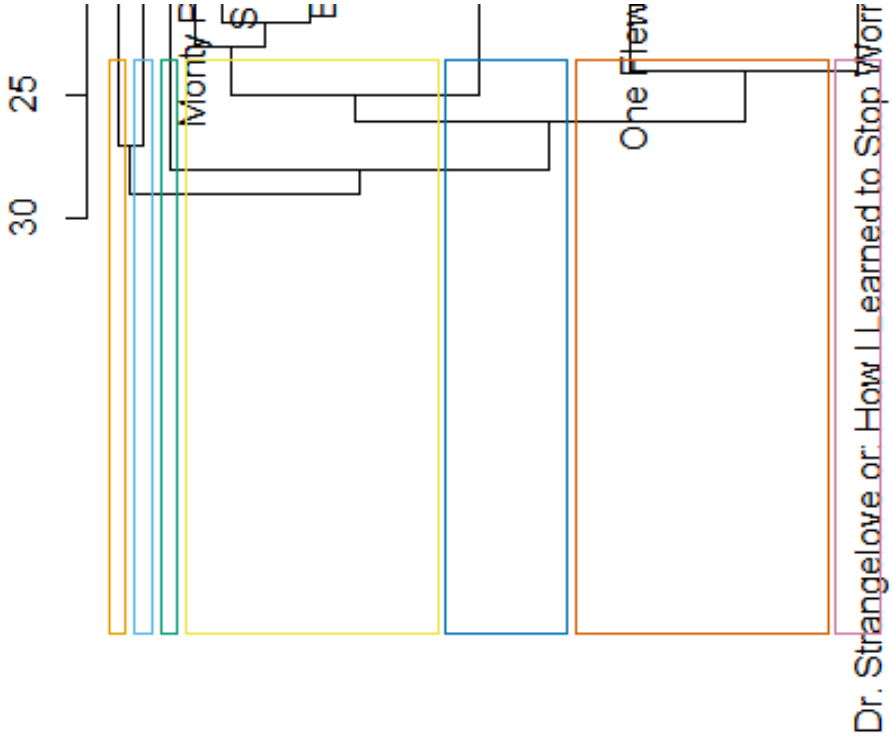
```
2
##                                  Monty Python and the Holy Grail (1974)
##
1
##                                      Empire Strikes Back, The (1980)
##
1
##                                    Raiders of the Lost Ark (1981)
##
1
##                                           Return of the Jedi (1983)
##
1
##                                                Amadeus (1984)
##
2
##                                                 Braveheart (1995)
##
1
##                                                  Contact (1997)
##
6
##                                               Full Monty, The (1997)
##
7
##                                               Good Will Hunting (1997)
##
4
##                                              L.A. Confidential (1997)
##
4
##                                                Titanic (1997)
##
4
##                                           Apt Pupil (1998)
##
4
##                                         As Good As It Gets (1997)
##
4
##                                      Schindler's List (1993)
##
1
##                                   One Flew Over the Cuckoo's Nest (1975)
##
2
##                                  To Kill a Mockingbird (1962)
##
1
## Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb
```

```
(1963)
##
3
##                                                    Casablanca (1942)
##
3
##                                        It's a Wonderful Life (1946)
##
2
##                                                 Star Wars (1977)
##
1
##                                                  Boot, Das (1981)
##
2
##                                                 Pulp Fiction (1994)
##
2
##                                                Rear Window (1954)
##
2
##                                                    Blade Runner (1982)
##
1
##                                         Silence of the Lambs, The (1991)
##
1
```

```
plot(fc, g, main = "modularity community", layout =
layout.fruchterman.reingold,
vertex.size = 8, vertex.label.cex = 0.5)
```


modularity community
As Good As It Gets (1997)
Full Monty, The (1997)
Good Will Hunting (1997)
L.A. Confidential (1997)
Titanic (1997)
Worrying and Love the Bomb (1963)
Toy Story (1995)
Monty Python
Fargo (1996)
Amadeus (1984)
Boot, Das (1981)

```
dendPlot(fc)
```


25
30
One Flew
Dr. Strangelove or: How I Learned to Stop Worr

2.3

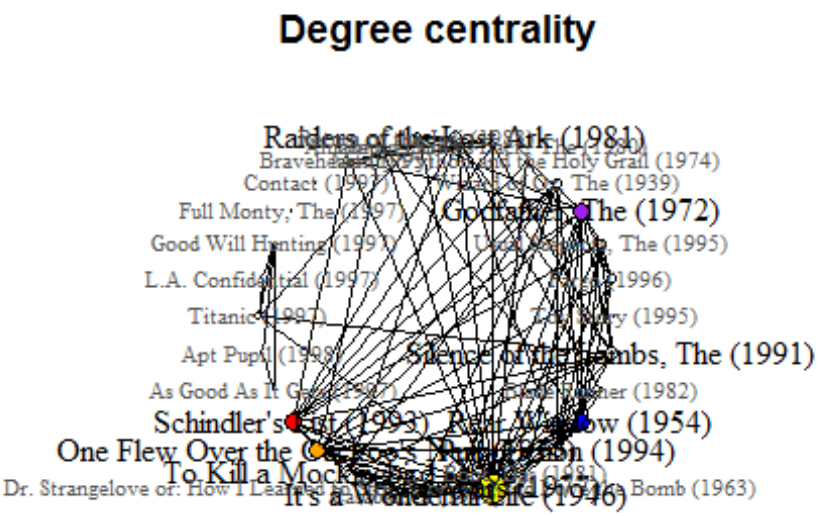
```
deg=degree(g)
deg
```

```
##                                                                 Toy Story
(1995)
##
0
##                                                                     Fargo
(1996)
##
7
##                                                 Usual Suspects, The
(1995)
##
7
##                                                      Godfather, The
(1972)
##
13
##                                                   Wizard of Oz, The
(1939)
##
7
##                                          Monty Python and the Holy Grail
(1974)
##
4
##                                                 Empire Strikes Back, The
(1980)
##
6
##                                                  Raiders of the Lost Ark
(1981)
##
9
##                                                       Return of the Jedi
(1983)
##
5
##                                                                   Amadeus
(1984)
##
3
##                                                               Braveheart
(1995)
##
5
##                                                                   Contact
```

```
(1997)
##
0
##                                                    Full Monty, The
(1997)
##
0
##                                                  Good Will Hunting
(1997)
##
4
##                                                  L.A. Confidential
(1997)
##
2
##                                                           Titanic
(1997)
##
5
##                                                        Apt Pupil
(1998)
##
2
##                                               As Good As It Gets
(1997)
##
2
##                                                 Schindler's List
(1993)
##
12
##                                  One Flew Over the Cuckoo's Nest
(1975)
##
12
##                                             To Kill a Mockingbird
(1962)
##
8
## Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb
(1963)
##
2
##                                                       Casablanca
(1942)
##
8
##                                             It's a Wonderful Life
(1946)
##
```

```
9
##                                                                 Star Wars
(1977)
##
21
##                                                                 Boot, Das
(1981)
##
3
##                                                              Pulp Fiction
(1994)
##
10
##                                                               Rear Window
(1954)
##
12
##                                                              Blade Runner
(1982)
##
4
##                                            Silence of the Lambs, The
(1991)
##
10
```

```
top = order(deg, decreasing=T)[1:10]
top1 = order(deg, decreasing=T)[1:1]
top2 = order(deg, decreasing=T)[2:2]
top3 = order(deg, decreasing=T)[3:3]
top4 = order(deg, decreasing=T)[4:4]
top5 = order(deg, decreasing=T)[5:5]
V(g)$size = abs(deg) * 0.8
V(g)$color = "white"
V(g)$label.color = "gray33"
V(g)$label.cex = 0.66
E(g)$color = "black"
V(g)[top]$label.color = "black"
V(g)[top]$label.cex = 1
V(g)[top1]$color = "yellow"
V(g)[top2]$color = "purple"
V(g)[top3]$color = "red"
V(g)[top4]$color = "orange"
V(g)[top5]$color = "blue"
plot(g, layout = layout.circle)
title("Degree centrality")
```

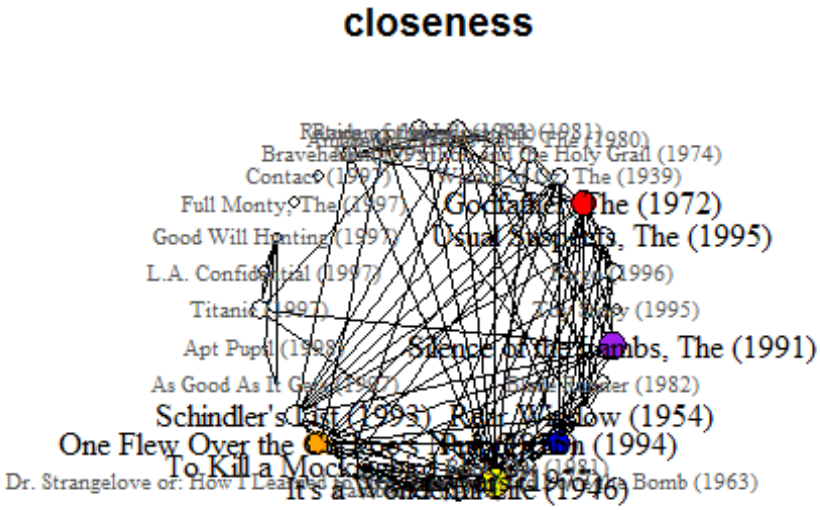



```
clo = closeness(g)
clo

##                                                                 Toy Story
(1995)
##
0.0011494253
##                                                                     Fargo
(1996)
##
0.0014184397
##                                                       Usual Suspects, The
(1995)
##
0.0015673981
##                                                           Godfather, The
(1972)
##
0.0017064846
##                                                         Wizard of Oz, The
(1939)
##
0.0014534884
##                                          Monty Python and the Holy Grail
(1974)
##
0.0014164306
```

```
##                                                Empire Strikes Back, The
(1980)
##
0.0013793103
##                                                 Raiders of the Lost Ark
(1981)
##
0.0014836795
##                                                      Return of the Jedi
(1983)
##
0.0014880952
##                                                                  Amadeus
(1984)
##
0.0013755158
##                                                             Braveheart
(1995)
##
0.0013869626
##                                                                Contact
(1997)
##
0.0011494253
##                                                   Full Monty, The
(1997)
##
0.0011494253
##                                                 Good Will Hunting
(1997)
##
0.0010604454
##                                                 L.A. Confidential
(1997)
##
0.0011160714
##                                                             Titanic
(1997)
##
0.0015082956
##                                                           Apt Pupil
(1998)
##
0.0008264463
##                                                As Good As It Gets
(1997)
##
0.0007936508
##                                                  Schindler's List
(1993)
```

```
##
0.0015923567
##                                      One Flew Over the Cuckoo's Nest
(1975)
##
0.0016666667
##                                                 To Kill a Mockingbird
(1962)
##
0.0015822785
## Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb
(1963)
##
0.0013140604
##                                                              Casablanca
(1942)
##
0.0014471780
##                                                 It's a Wonderful Life
(1946)
##
0.0015552100
##                                                               Star Wars
(1977)
##
0.0018181818
##                                                               Boot, Das
(1981)
##
0.0013513514
##                                                           Pulp Fiction
(1994)
##
0.0016528926
##                                                            Rear Window
(1954)
##
0.0015847861
##                                                           Blade Runner
(1982)
##
0.0013531800
##                                          Silence of the Lambs, The
(1991)
##
0.0017543860

top = order(clo, decreasing=T)[1:10]
top1 = order(clo, decreasing=T)[1:1]
top2 = order(clo, decreasing=T)[2:2]
```

```
top3 = order(clo, decreasing=T)[3:3]
top4 = order(clo, decreasing=T)[4:4]
top5 = order(clo, decreasing=T)[5:5]
V(g)$size = (abs(clo)^2) * 1e+06 * 5
V(g)$color = "white"
V(g)$label.color = "gray33"
V(g)$label.cex = 0.66
V(g)[top]$label.color = "black"
V(g)[top1]$color = "yellow"
V(g)[top2]$color = "purple"
V(g)[top3]$color = "red"
V(g)[top4]$color = "orange"
V(g)[top5]$color = "blue"
V(g)[top]$label.cex = 1
plot(g, layout = layout.circle)
title("closeness")
```




```
bet = betweenness(g)
bet
```

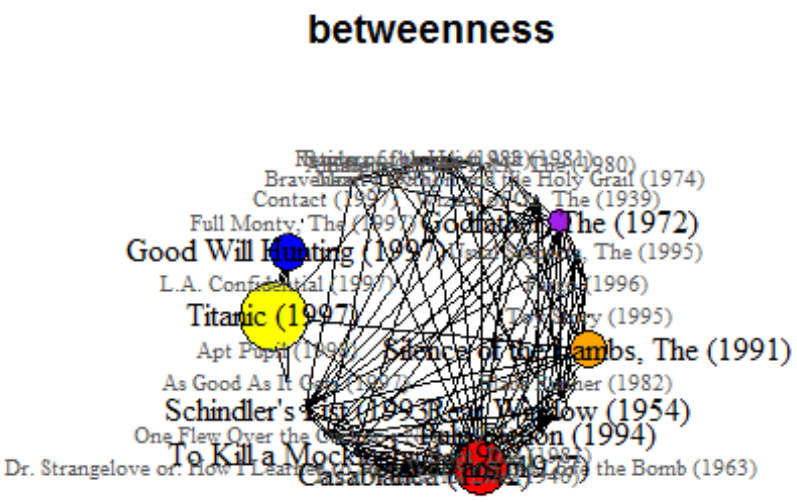
```
##                                                                 Toy Story
(1995)
##
0.000000
##                                                                     Fargo
(1996)
##
```

```
1.000000
##                                                 Usual Suspects, The
(1995)
##
2.666667
##                                                     Godfather, The
(1972)
##
29.333333
##                                                   Wizard of Oz, The
(1939)
##
3.500000
##                                        Monty Python and the Holy Grail
(1974)
##
1.500000
##                                               Empire Strikes Back, The
(1980)
##
1.000000
##                                                Raiders of the Lost Ark
(1981)
##
6.500000
##                                                     Return of the Jedi
(1983)
##
5.000000
##                                                                 Amadeus
(1984)
##
0.000000
##                                                               Braveheart
(1995)
##
2.333333
##                                                                  Contact
(1997)
##
0.000000
##                                                    Full Monty, The
(1997)
##
0.000000
##                                                 Good Will Hunting
(1997)
##
48.000000
##                                                 L.A. Confidential
```

```
(1997)
##
0.000000
##                                                                  Titanic
(1997)
##
89.000000
##                                                              Apt Pupil
(1998)
##
0.000000
##                                                  As Good As It Gets
(1997)
##
0.000000
##                                                    Schindler's List
(1993)
##
11.333333
##                                       One Flew Over the Cuckoo's Nest
(1975)
##
7.333333
##                                               To Kill a Mockingbird
(1962)
##
8.500000
## Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb
(1963)
##
0.000000
##                                                           Casablanca
(1942)
##
10.500000
##                                               It's a Wonderful Life
(1946)
##
1.000000
##                                                            Star Wars
(1977)
##
70.500000
##                                                            Boot, Das
(1981)
##
0.000000
##                                                         Pulp Fiction
(1994)
##
```

```
14.166667
##                                                          Rear Window
(1954)
##
26.666667
##                                                        Blade Runner
(1982)
##
2.333333
##                                          Silence of the Lambs, The
(1991)
##
49.333333
```

```
top = order(bet, decreasing=T)[1:10]
top1 = order(bet, decreasing=T)[1:1]
top2 = order(bet, decreasing=T)[2:2]
top3 = order(bet, decreasing=T)[3:3]
top4 = order(bet, decreasing=T)[4:4]
top5 = order(bet, decreasing=T)[5:5]
V(g)$size = abs(bet) * 0.5
V(g)$color = "white"
V(g)$label.color = "gray33"
V(g)$label.cex = 0.66
V(g)[top]$label.color = "black"
V(g)[top1]$color = "yellow"
V(g)[top2]$color = "red"
V(g)[top3]$color = "orange"
V(g)[top4]$color = "blue"
V(g)[top5]$color = "purple"
V(g)[top]$label.cex = 1
plot(g, layout = layout.circle)
title("betweenness")
```

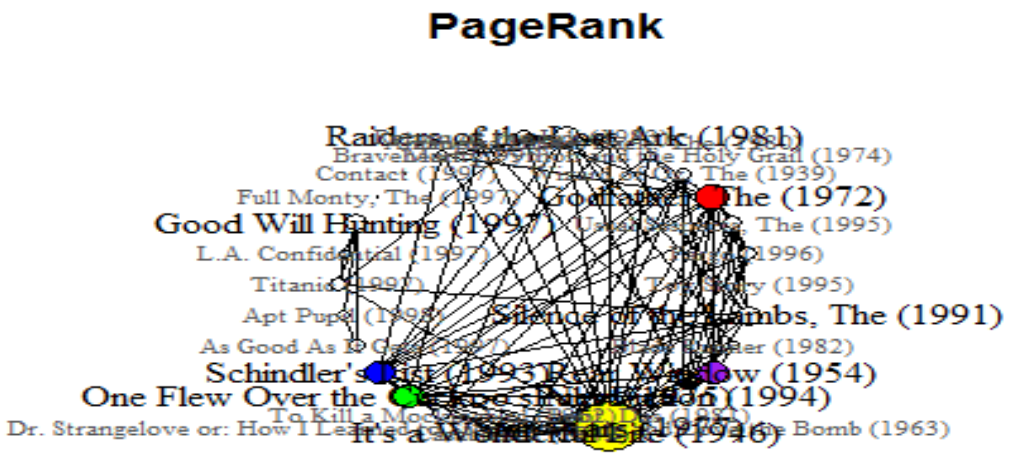



```
pr = page.rank(g)$vector
pr

##                                                                 Toy Story
(1995)
##
0.005464481
##                                                                     Fargo
(1996)
##
0.033035311
##                                                        Usual Suspects, The
(1995)
##
0.031517659
##                                                            Godfather, The
(1972)
##
0.056421551
##                                                          Wizard of Oz, The
(1939)
##
0.030128548
##                                              Monty Python and the Holy Grail
(1974)
##
0.019075277
```

```
##                                                Empire Strikes Back, The
(1980)
##
0.032904550
##                                                 Raiders of the Lost Ark
(1981)
##
0.046831054
##                                                      Return of the Jedi
(1983)
##
0.029391820
##                                                                  Amadeus
(1984)
##
0.015172857
##                                                             Braveheart
(1995)
##
0.024944495
##                                                                Contact
(1997)
##
0.005464481
##                                                  Full Monty, The
(1997)
##
0.005464481
##                                               Good Will Hunting
(1997)
##
0.039742641
##                                               L.A. Confidential
(1997)
##
0.018510141
##                                                             Titanic
(1997)
##
0.034466825
##                                                           Apt Pupil
(1998)
##
0.025941649
##                                                As Good As It Gets
(1997)
##
0.027493757
##                                                  Schindler's List
(1993)
```

```
##
0.054314403
##                                      One Flew Over the Cuckoo's Nest
(1975)
##
0.053750925
##                                                 To Kill a Mockingbird
(1962)
##
0.034442409
## Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb
(1963)
##
0.012250984
##                                                              Casablanca
(1942)
##
0.035808010
##                                                 It's a Wonderful Life
(1946)
##
0.038054782
##                                                               Star Wars
(1977)
##
0.115782423
##                                                               Boot, Das
(1981)
##
0.016481427
##                                                           Pulp Fiction
(1994)
##
0.041328400
##                                                            Rear Window
(1954)
##
0.051755604
##                                                           Blade Runner
(1982)
##
0.021071196
##                                          Silence of the Lambs, The
(1991)
##
0.042987860

top = order(pr, decreasing=T)[1:10]
top1 = order(pr, decreasing=T)[1:1]
top2 = order(pr, decreasing=T)[2:2]
```

```
top3 = order(pr, decreasing=T)[3:3]
top4 = order(pr, decreasing=T)[4:4]
top5 = order(pr, decreasing=T)[5:5]
V(g)$size = abs(pr) * 300
V(g)$color = "white"
V(g)$label.color = "gray33"
V(g)$label.cex = 0.66
V(g)[top]$label.color = "black" ## highlight the top-5 nodes
V(g)[top1]$color = "yellow"
V(g)[top2]$color = "red"
V(g)[top3]$color = "blue"
V(g)[top4]$color = "green"
V(g)[top5]$color = "purple"
V(g)[top]$label.cex = 1
plot(g, layout = layout.circle)
title("PageRank")
```



2.4

#Degree centrality shows the most number of connections for a node. From the figure, the yellow/star wars (1977) one has highest degree centrality.
#Closeness is based on the length of the average shortest path between a node and all nodes in a graph.From the figure, the yellow/star wars (1977) seems to have high closeness.
#Betweenness denotes how many pairs of individuals would have to go through a certain node in order to reach one another. From the figure, Titanic seems to have high betweenness.
#A page's importance is given by the total votes it received and the importance of its voters. Rank(u): importance score of page u. From the figure, Star Wars (1977) has the highest PageRank

#Task 3

3.1

```
set.seed(123)
library(recommenderlab)

library(dplyr)

library(igraph)

book_ratings<-read.csv("D:/semester/2nd sem/DATA_MINING/hw5/BX-CSV-Dump/BX-
Book-Ratings.csv", header = TRUE, sep = ";", stringsAsFactors = FALSE)

books<-read.csv("D:/semester/2nd sem/DATA_MINING/hw5/BX-CSV-Dump/BX-
Books.csv", header = TRUE, sep=";",stringsAsFactors = FALSE)

books$Year.Of.Publication<-as.numeric(books$Year.Of.Publication)

books.m<-merge(book_ratings,books)
books.m<-na.omit(books.m)
books.m<-filter(books.m, Year.Of.Publication>=1998)
books.m<-books.m[,c("ISBN","User.ID","Book.Rating")]

d4 = data.frame(from = books.m$User.ID, to = books.m$ISBN, weight =
books.m$Book.Rating)
g = graph.data.frame(d4)
mat = get.adjacency(g)
mat.w = get.adjacency(g, attr = "weight")
book.idx = which(colSums(mat) >= 10)
user.idx = which(rowSums(mat) >= 10)
rmat = mat.w[user.idx, book.idx]
dim(rmat)

## [1] 1596 1425

m = as.matrix(rmat)
m = as(m, "realRatingMatrix")
dim(m)

## [1] 1596 1425

e = evaluationScheme(m, method = "cross", k=4, given = 5, goodRating = 6)
e

## Evaluation scheme with 5 items given
## Method: 'cross-validation' with 4 run(s).
## Good ratings: >=6.000000
## Data set: 1596 x 1425 rating matrix of class 'realRatingMatrix' with
2274300 ratings.
```

```
r1 = Recommender(getData(e, "train"), "Random")

r2 = Recommender(getData(e, "train"), "Popular")

r3 = Recommender(getData(e, "train"), "UBCF")

r4 = Recommender(getData(e, "train"), "IBCF")

p1 = predict(r1, getData(e, "known"), type="ratings")

p2 = predict(r2, getData(e, "known"), type="ratings")

p3 = predict(r3, getData(e, "known"), type="ratings")

p4 = predict(r4, getData(e, "known"), type="ratings")

error = rbind(
calcPredictionAccuracy(p1, getData(e, "unknown")),
calcPredictionAccuracy(p2, getData(e, "unknown")),
calcPredictionAccuracy(p3, getData(e, "unknown")),
calcPredictionAccuracy(p4, getData(e, "unknown"))
)
rownames(error) = c("Random", "Popular", "UBCF", "IBCF")
t(error)
```

```
##         Random   Popular       UBCF     IBCF
## RMSE 1.0261546 0.5878808 0.59020112  3.37834
## MSE  1.0529933 0.3456039 0.34833736 11.41318
## MAE  0.4814946 0.1023864 0.07817221  1.49569
```

3.2

```
#From the performance table, popular recommendation method has low RMSE, MSE
compared to the others and found to be the best one.
```