Topic Modeling

Qihan Guan

2/28/2021

```
#load data and libraries

library(textir) # to get the data
library(maptpx) # for the topics function
library(fpc)
library(factoextra)
load("congress.RData")
```

1.Fit K-means to the speech text of the members, comprising of the 1000 phrases, for K in 5, 10, 15, 20, 25

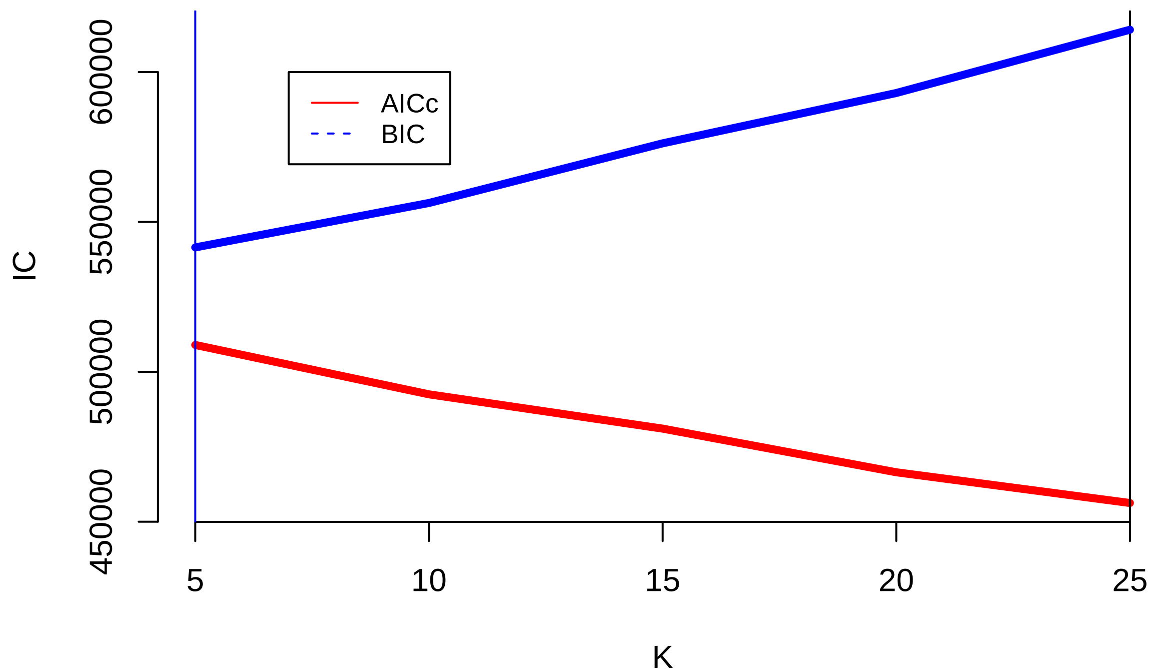
```
fs <- scale(as.matrix( congress109Counts/rowSums(congress109Counts)))

kmfs <- lapply(5*(1:5), function(k) kmeans(fs, k))
```

2.Use AICc and BIC to choose the K. Also use the elbow curve method to identify the most optimal value of K.

```
## get AICc, BIC, and deviance for the output of kmeans
kic <- function(fit, rule=c("A","B","C")){
  df <- length(fit$centers) # K*dim
  #print(df)
  n <- sum(fit$size)
  #print(n)
  D <- fit$tot.withinss # deviance
  rule=match.arg(rule)
  if(rule=="A")
    return(D + 2*df*n/(n-df-1)) #AICc
  else if(rule=="B")
    return(D + log(n)*df)#BIC
  else
    return(D) #Deviance
}
```

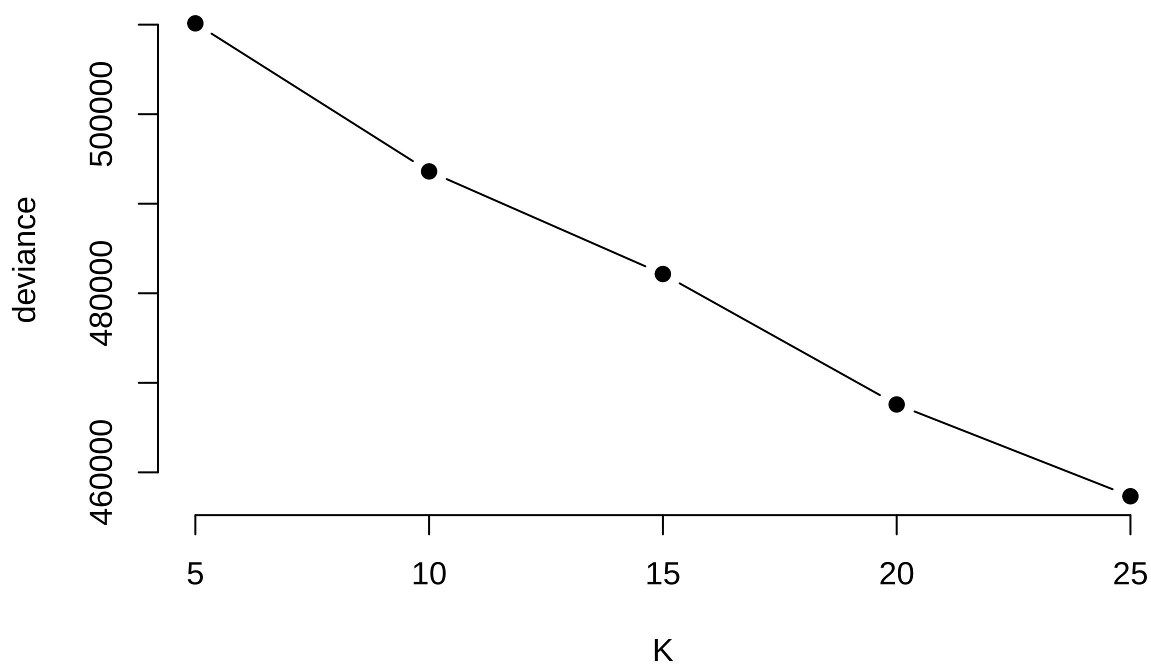
```
## AICc and BIC
km_aicc <- sapply(kmfs, kic, "A")
km_bic <- sapply(kmfs, kic, "B")
```

```r
## Plot IC
plot(5*(1:5), km_aicc, xlab = "K", ylab = "IC", ylim = range(c(km_aicc, km_bic)),
     bty = "n", type = "l", lwd = 4, col = "red")
abline(v=which.min(km_aicc)*5)
lines(5*(1:5), km_bic, col = "blue", lwd = 4)
abline(v=which.min(km_bic)*5, col="blue")
legend(7,600000, legend = c("AICc", "BIC"), col = c("red", "blue"),
       lty=1:2, cex=0.8)
```



IC plot gives contradicting results. AICc decreases as K gets larger while BIC increases as K gets larger. AICc suggests the optimal K is 25, while BIC suggests the optimal K is 5.

```r
##Plot Elbow curve
deviance <- sapply(kmfs, kic, "C")

plot(5*(1:5), deviance,
     type="b", pch = 19, frame = FALSE,
     xlab="K",
     ylab="deviance")
```



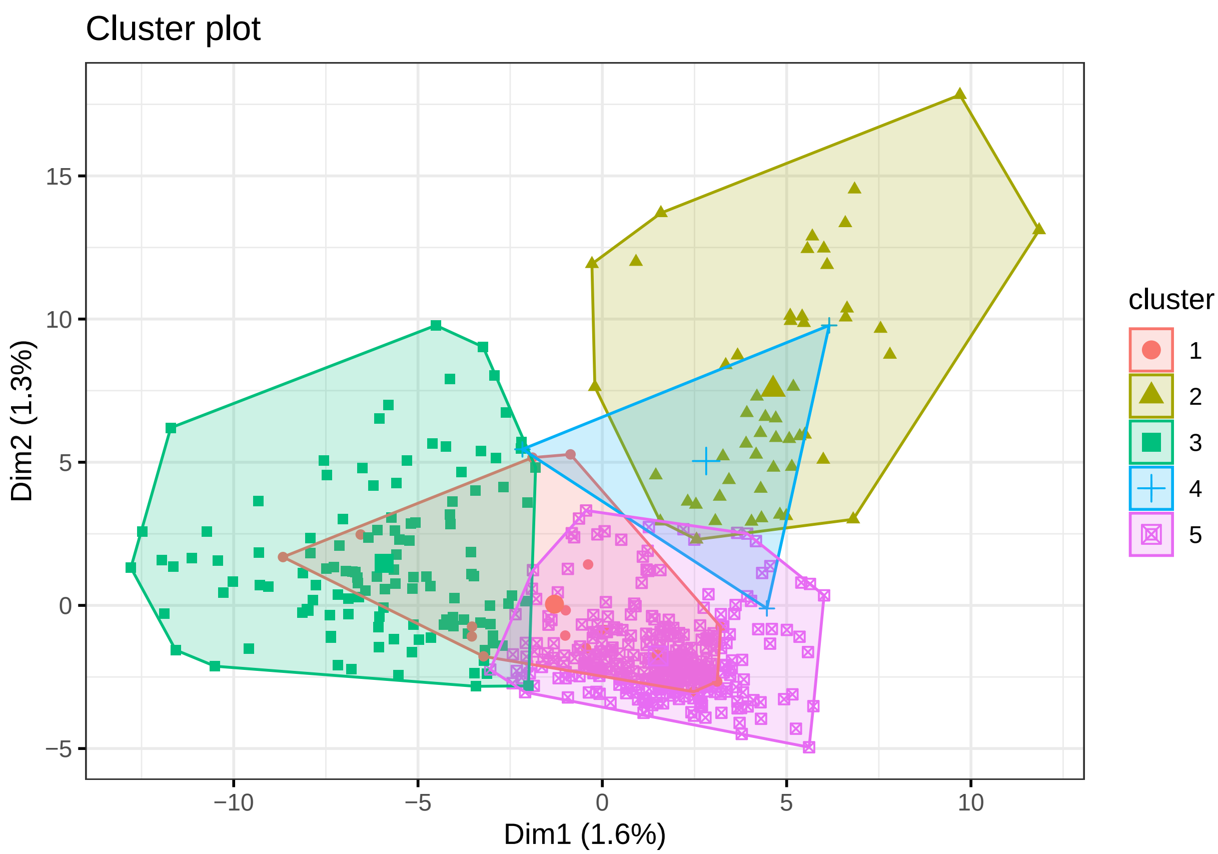
Elbow curve suggests that when K=25, deviance is minimized. Elbow curve yields optimal K=25.

3.Compare the optimal values of K obtained and explain

AICc is the lowest when K=25, while BIC is the lowest when K=5. The elbow curve also suggests that when K=25, deviance is the smallest. AICc aligns with the elbow curves here, while BIC goes the opposite direction. A possible explanation here would be that we have 529 legislators but 1000 phrases. We have small n but very large df. If we print out the n and df for K=5*(1:5), we can see that as K gets larger, the df gets much larger. This would make AICc goes down while BIC goes up. I will use BIC to select the optimal K here as we have really small n here. AICc may overfit. Optimal K = 5.

4.Plot the clusters based on optimal K. I have chosen optimal K=5.

```
kmfs_optimal <- kmfs[[1]] #optimal K=5, the first one

## Use fviz_cluster
fviz_cluster(kmfs_optimal, data=fs, geom='point', ellipse.type="convex",
             ggtheme = theme_bw())
```



5.Interpret the most significant words within that cluster (top 10)

```
print(apply(kmfs_optimal$centers,1,function(c) colnames(fs)[order(-c)[1:10]]))
```

```
##       1                             2
##  [1,] "national.wildlife"           "court.appeal"
##  [2,] "national.wildlife.refuge"    "business.meeting"
##  [3,] "arctic.national.wildlife"    "circuit.court.appeal"
##  [4,] "wildlife.refuge"             "committe.foreign.relation"
##  [5,] "post.traumatic"              "judicial.nomine"
##  [6,] "post.traumatic.stress"       "housing.urban.affair"
##  [7,] "traumatic.stress"            "urban.affair"
##  [8,] "fuel.efficiency"             "court.judge"
##  [9,] "water.act"                   "committe.commerce.science"
## [10,] "traumatic.stress.disorder" "banking.housing.urban"
##       3                             4
##  [1,] "private.account"             "suppli.natural.ga"
##  [2,] "cut.medicaid"                "supply.natural.ga"
##  [3,] "cut.food.stamp"              "ga.natural.ga"
##  [4,] "student.loan"                "natural.ga.natural"
##  [5,] "privatizing.social.security" "able.buy.gun"
##  [6,] "privatize.social.security"   "ga.natural"
##  [7,] "tax.cut.wealthy"             "buy.gun"
```

```
##  [8,] "plan.privatize"              "natural.ga"
##  [9,] "medicaid.cut"                "grand.ole.opry"
## [10,] "cost.war"                    "background.check.system"
##       5
##  [1,] "strong.support"
##  [2,] "urge.support"
##  [3,] "death.tax"
##  [4,] "illegal.immigration"
##  [5,] "private.property"
##  [6,] "repeal.death.tax"
##  [7,] "civil.right.movement"
##  [8,] "embryonic.stem"
##  [9,] "embryonic.stem.cel"
## [10,] "stem.cel"
```

Interpretation: Significant words in cluster 1 seem to focus on environment and humanity, such as wildlife, fuel efficiency, water act, traumatic, etc. Significant words in cluster 2 seem to focus on courts, business and urban. Significant words in cluster 3 seem to focus on finance and social security. Significant words in cluster 4 seem to focus on energy and gun control, such as natural gas supply, buy gun, and background check. Cluster 5 seems to focus on immigration, tax, and civil rights.

6. Fit a topic model for the speech counts.

```r
## Convert matrix
m <- as.simple_triplet_matrix(congress109Counts)

## Choose number of topics
n_topics <- topics(m,K=2:20, tol=10)
```

```
##
## Estimating on a 529 document collection.
## Fit and Bayes Factor Estimation for K = 2 ... 20
## log posterior increase: 961.1, 618.5, 275.3, 231.4, 350.5, 161.7, 63.8, 11.7, 10.3, done.
## log BF( 2 ) = 29905.68
## log posterior increase: 1973.7, 257.2, 163.8, 37.8, 225.3, 77.3, 48.2, 37.4, done.
## log BF( 3 ) = 43982.12
## log posterior increase: 1838.1, 86.8, 88.6, 170.6, 20.7, 28.9, 15.1, done.
## log BF( 4 ) = 51785.98
## log posterior increase: 2953.1, 179.2, 240.2, 66.9, 44.2, 26.4, done.
## log BF( 5 ) = 60418.7
## log posterior increase: 2107.1, 135.3, 43.5, 13.1, done.
## log BF( 6 ) = 65066.89
## log posterior increase: 1905, 79.3, 60.3, 48.6, 48.7, 86.2, 57.4, 48.4, 39.5, 9.6, done.
## log BF( 7 ) = 70427.29
## log posterior increase: 2448.2, 135.2, 15.4, done.
## log BF( 8 ) = 74358.6
## log posterior increase: 1777, 86.7, 120.4, 126.3, 46.7, 12.4, done.
## log BF( 9 ) = 76191.65
## log posterior increase: 1357, 85.6, 253.5, 61.1, 27, done.
## log BF( 10 ) = 79420.89
## log posterior increase: 1394, 42.8, 20, done.
```

```
## log BF( 11 ) = 80317.93
## log posterior increase: 1442.4, 89.3, 45.2, done.
## log BF( 12 ) = 80605.09
## log posterior increase: 1144.4, 65.7, 91.6, 36.5, 32.6, 13.5, done.
## log BF( 13 ) = 79929.44
## log posterior increase: 1159.9, 66.5, 13.5, done.
## log BF( 14 ) = 80622.28
## log posterior increase: 1250.2, 32.8, 30.9, 28.7, 18.6, done.
## log BF( 15 ) = 78236.49
## log posterior increase: 917.1, 37, 20.7, 32.6, 22.8, done.
## log BF( 16 ) = 76057.65
```

```
## Need to choose n that gives biggest BF(n). Results yield n = 14.

## ordering by topic over aggregate lift
summary(n_topics, n=10)
```

```
##
## Top 10 phrases by topic-over-null term lift (and usage %):
##
## [1] 'republic.cypru', 'national.homeownership.month', 'senate.committe.business', 'columbia.river.go
## [2] 'near.retirement.age', 'repeal.death.tax', 'medic.liability.crisi', 'gifted.talented.student', 'p
## [3] 'southeast.texa', 'million.illegal.alien', 'temporary.worker.program', 'amnesty.illegal.alien', 
## [4] 'national.heritage.corridor', 'asian.pacific.american', 'domestic.violence.sexual', 'pacific.ame
## [5] 'united.airline.employe', 'record.budget.deficit', 'private.account', 'security.private.account'
## [6] 'va.health.care', 'troop.bring.home', 'funding.veteran.health', 'bring.troop.home', 'bring.troop
## [7] 'commonly.prescribed.drug', 'hate.crime.legislation', 'change.heart.mind', 'winning.war.iraq', 'I
## [8] 'judicial.confirmation.process', 'judge.alberto.gonzale', 'john.robert', 'fifth.circuit.court', 
## [9] 'indian.art.craft', 'low.cost.reliable', 'ready.mixed.concrete', 'price.natural.ga', 'witness.te
## [10] 'wild.bird', 'arctic.refuge', 'arctic.wildlife.refuge', 'fuel.efficiency', 'drilling.arctic.nat
## [11] 'north.american.fre', 'american.fre.trade', 'central.american.fre', 'buy.american.product', 'tr
## [12] 'pluripotent.stem.cel', 'national.ad.campaign', 'regional.training.cent', 'cel.stem.cel', 'embr
## [13] 'increase.minimum.wage', 'raise.minimum.wage', 'minimum.wage', 'credit.card.issuer', 'northern.
## [14] 'able.buy.gun', 'caliber.sniper.rifle', 'deep.sea.coral', 'assault.weapon', 'defense.intelligen
##
## Log Bayes factor and estimated dispersion, by number of topics:
##
##              2        3        4        5        6        7        8        9
## logBF 29905.68 43982.12 51785.98 60418.70 65066.89 70427.29 74358.60 76191.65
## Disp      4.96     4.27     3.85     3.52     3.32     3.29     3.08     2.93
##             10       11       12       13       14       15       16
## logBF 79420.89 80317.93 80605.09 79929.44 80622.28 78236.49 76057.65
## Disp      2.81     2.69     2.56     2.53     2.47     2.40     2.37
##
## Selected the K = 14 topic model
```

Need to choose n that gives biggest BF(n). Results yield n = 14.

```
## Look at words ordered by simple in-topic prob
print(rownames(n_topics$theta)[order(n_topics$theta[,1],decreasing = TRUE)[1:10]])
```

```
##  [1] "head.start"            "gulf.coast"            "hurricane.katrina"
##  [4] "strong.support"        "appropriation.bil"     "endangered.speci.act"
```

```
##  [7] "low.income"           "medic.malpractice"    "business.owner"
## [10] "million.american"
```

```r
print(rownames(n_topics$theta)[order(n_topics$theta[,2],decreasing = TRUE)[1:10]])
```

```
##  [1] "american.people"        "tax.relief"              "death.tax"
##  [4] "economic.growth"        "finance.committe"        "tax.increase"
##  [7] "budget.committe"        "security.system"         "social.security.system"
## [10] "feder.budget"
```

```r
## Look at party mean
dem <- colMeans(n_topics$omega[congress109Ideology$party=="D",])
rep <- colMeans(n_topics$omega[congress109Ideology$party=="R",])

sort(dem/rep)
```

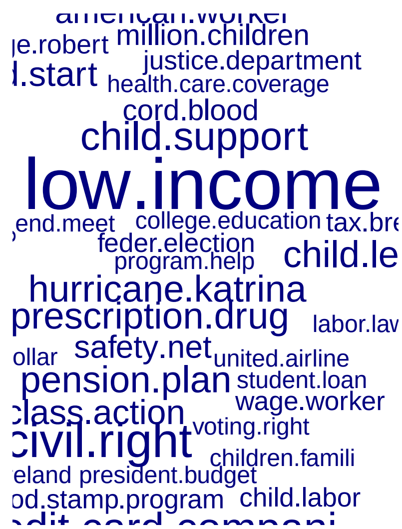
```
##         2         7         3         9        12         8         1        14
## 0.1538006 0.2672923 0.3493013 0.3706401 0.4056197 0.4361000 0.9223347 1.4266409
##        11         6        10         4        13         5
## 1.6764450 2.7208123 2.8105495 2.8620783 3.7659967 8.2887811
```

Topic 2,7,3,9,8,12 are republican while topic 14,11,6,10,4,13,5 are strong democratic.

To further check the validity of our model, we plot some word cloud for strong democratic and republican topics.

```r
## Plot wordcloud
library(wordcloud)
par(mfrow=c(1,2))

## Republican topic
wordcloud(row.names(n_topics$theta),
          freq=n_topics$theta[,2], min.freq=0.004, col="maroon")
## Republican topic
wordcloud(row.names(n_topics$theta),
          freq=n_topics$theta[,3], min.freq=0.004, col="maroon")
```


```r
## Democratic topic
wordcloud(row.names(n_topics$theta),
          freq=n_topics$theta[,5], min.freq=0.004, col="navy")
## Democratic topic
wordcloud(row.names(n_topics$theta),
          freq=n_topics$theta[,13], min.freq=0.004, col="navy")
```



Interpretation

By observing the word clouds from the two parties, we can see that there is a clear difference between the two parties' frequent words. The Republican topics focus on death tax, illegal immigration, etc. The Democratic topics focus on civil right, middle class, low income, etc. These observations fit the ideologies of the corresponding party. In addition, majority of words within each topic share a common theme. Our chosen model makes sense.

7.Connect the unsupervised clusters to partisanship.

```
tapply(congress109Ideology$party, kmfs_optimal$cluster, table)
```

```
## $`1`
##
##  D  I  R
## 11  1  4
##
## $`2`
##
##  D  I  R
##  4  0 48
##
## $`3`
```

```
##
##   D   I   R
## 123   1   0
##
## $`4`
##
## D I R
## 1 0 2
##
## $`5`
##
##   D   I   R
## 103   0 231
```

It appears that cluster 5 is non-partisan because it shows large amount of points from both parties. Cluster 3 is strong democratic. Cluster 2 is strong republican.

To further investigate cluster 5, display top 20 words from cluster 5

```r
colnames(fs)[order(-kmfs_optimal$centers[5,])[1:20]]
```

```
##  [1] "strong.support"       "urge.support"         "death.tax"
##  [4] "illegal.immigration"  "private.property"     "repeal.death.tax"
##  [7] "civil.right.movement" "embryonic.stem"       "embryonic.stem.cel"
## [10] "stem.cel"             "right.movement"       "post.office"
## [13] "cel.research"         "look.forward"         "terri.schiavo"
## [16] "business.owner"       "illegal.immigrant"    "adult.stem"
## [19] "adult.stem.cel"       "property.right"
```

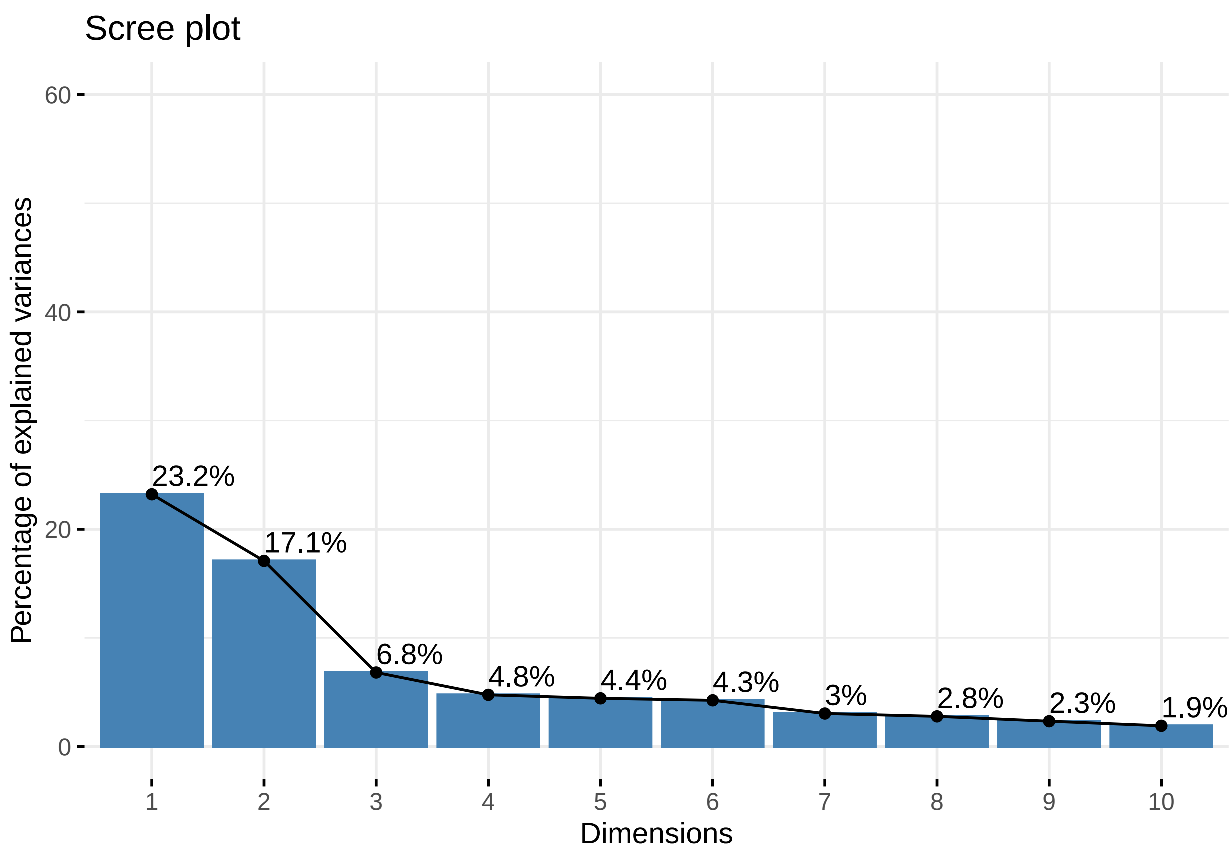
For cluster 5, we can clearly see frequent words from both parties.

8.Fit PCA model to congress counts data.

```r
m.pc <- prcomp(congress109Counts)
m.pca_sum <- summary(m.pc)
```

9.Create a graph that summarizes the percentage of variance explained by the first 10 principle components (scree_plot).

```r
fviz_eig(m.pc, addlabels = TRUE, ylim=c(0,60))
```



From the scree plot, we can see the 'elbow' appears to be at 3.

10.Report results

```r
sum(c(23.2,17.1,6.8,4.8,4.4,4.3,3,2.8,2.3,1.9))
```

```
## [1] 70.6
```

Total proportion of explained variance of the first 10 pc: 70.6%

If we were to eliminate all other components (everything but the first 10), we would eliminate (529-10)=519 dimensions. We would lose (100-70.6)%=29.4% variance.