HW6_Group22

Homework 6

Colley Buwana, Hazna Faiza, Misbah Arshad, Panji Al 'Alam, Sagarika K

```
# Homework: congressional text dataset
library(textir) # to get the data
```

```
Loading required package: distrom
```

```
Loading required package: Matrix
```

```
Loading required package: gamlr
```

```
Loading required package: parallel
```

```
library(maptpx) # for the topics function
```

```
Loading required package: slam
```

```
data(congress109)
```

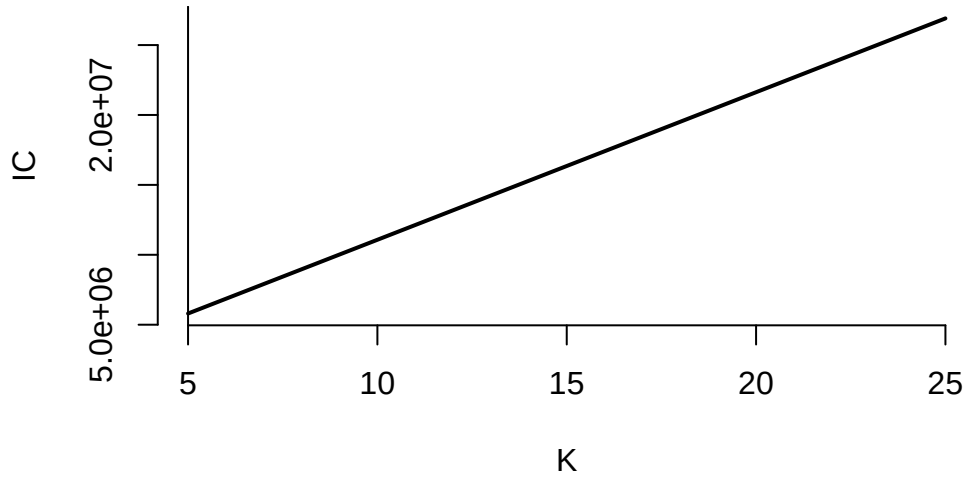
Question 1

Fit K-means to speech text for K in 5,10,15,20,25. Use BIC to choose the K and interpret the selected model.

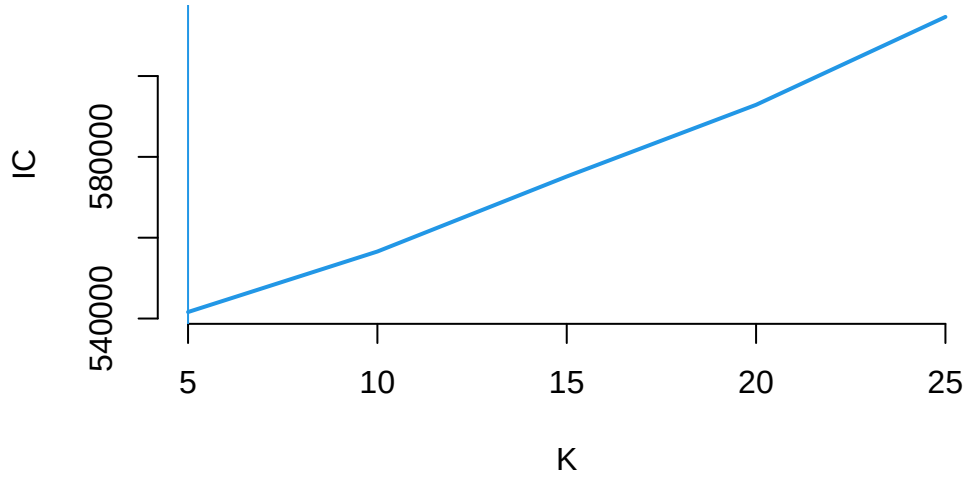
Note: This is different on each person depending on the set.seed. Feel free to add more interpretation!

```r
# [1] fit k-means for k in 5,10,15,20,25.  Use an IC to choose the
# number of clusters and interpret some of the centers.
set.seed(1)

fs <- scale(as.matrix( congress109Counts/rowSums(congress109Counts) ))
kfit <- lapply(5*(1:5), function(k) kmeans(fs,k))

source("kIC.R")

# AICc and BIC
kaicc <- sapply(kfit,kIC)
kbic <- sapply(kfit,kIC,"B")
```

```r
# AICc Plot
plot(5*(1:5), kaicc, xlab="K", ylab="IC",
    bty="n", type="l", lwd=2)
abline(v=which.min(kaicc)*5)
```




```r
# BIC Plot
plot(5*(1:5), kbic, xlab="K", ylab="IC",
    bty="n", type="l", lwd=2, col=4)
abline(v=which.min(kbic)*5,col=4)
```




```
# Based on the AICc and BIC, we used 5 K-means
kmfs <- kfit[[1]]

## Interpretation: we can see the words with cluster centers
## highest above zero (these are in units of standard deviation of f)
print(apply(kmfs$centers,1,function(c) colnames(fs)[order(-c)[1:10]]))
```

```
      1                        2                 3
 [1,] "suppli.natural.ga"      "private.account" "able.buy.gun"
 [2,] "supply.natural.ga"      "tax.cut.wealthy" "buy.gun"
 [3,] "ga.natural.ga"          "cut.medicaid"    "background.check.system"
 [4,] "natural.ga.natural"     "tax.break"       "assault.weapon.ban"
 [5,] "ga.natural"             "child.support"   "assault.weapon"
 [6,] "change.heart.mind"      "cost.war"        "gun.industry"
 [7,] "hate.crime.legislation" "cut.food.stamp"  "gun.violence"
 [8,] "natural.ga"             "medicaid.cut"    "bul.eye"
 [9,] "hate.crime.law"         "student.loan"    "national.rifle.association"
[10,] "grand.ole.opry"         "plan.privatize"  "gun.safety"
      4                             5
 [1,] "oil.food"                    "look.forward"
 [2,] "oil.food.program"            "strong.support"
 [3,] "food.scandal"                "urge.support"
 [4,] "oil.food.scandal"            "death.tax"
 [5,] "food.program"                "illegal.immigrant"
 [6,] "united.nation.reform"        "border.security"
 [7,] "atomic.energy.agency"        "illegal.immigration"
 [8,] "international.atomic.energy" "private.property"
```

```
 [9,] "reform.united.nation"        "pass.bil"
[10,] "un.reform"                   "appropriation.bil"
```

```
kmfs$size
```

```
[1]   3 139   1  15 371
```

Answer

Both AIC and BIC choose k=5 as the optimal number of cluster. The cluster's center revolves around weapon/gun regulation, e.g., able.buy.gun, assault.weapon.ban, gun.industry.

Interpretation:

- Cluster 1 focuses on natural resources, particularly about natural gas. There is also a few mentions of hate crime.
- Cluster 2 focuses on taxation and the use of taxes, particulalry welfare issues.
- Cluster 3 focuses on guns and safety regulation.
- Cluster 4 is about international relations specifically to do with American relations with Iraq. This includes the Oil for Food Program and atomic energy regulation and negotiation.
- Cluster 5 is about support for bills related to the border, specifically on illegal immigration, and also the estate tax (also known as the death tax).

Question 2

Fit a topic model for the speech counts. Use Bayes factors to choose the number of topics, and interpret your chosen model.

```
## [2] topic modelling.
# Convert to slam matrix
set.seed(1)
x <- as.simple_triplet_matrix(congress109Counts)

## Topic modelling:
## Recall: BF is like exp(-BIC), so you choose the biggest BF
tpcs <- topics(x, K=2:25)
```

```
Estimating on a 529 document collection.
Fit and Bayes Factor Estimation for K = 2 ... 25
log posterior increase: 961.1, 618.5, 275.3, 231.4, 350.5, 161.7, 63.8, 11.7, 10.3, 4.3, 2.8
log BF( 2 ) = 30123.15
log posterior increase: 1974.6, 281.6, 131.6, 127.3, 55.2, 82.7, 24.8, 37.1, 6.5, 13.3, 2.2,
log BF( 3 ) = 44142.75
log posterior increase: 1833.9, 174.9, 73, 147, 45.4, 24.5, 10, 37.3, 89.8, 35.1, 18.1, 15.9
log BF( 4 ) = 53865.63
log posterior increase: 2758.2, 80.5, 50.3, 21.2, 21.1, 34.1, 6.5, 5.4, 13.3, 16.7, 8.4, 25.4
log BF( 5 ) = 60318.97
log posterior increase: 2469.9, 39.9, 11.8, 5.8, 7, 6.3, 15.7, 7.9, 72.2, 3.3, 5.1, 2.4, 1.5
log BF( 6 ) = 64330.64
log posterior increase: 1915.2, 75.1, 19.9, 23.4, 59.8, 16.6, 15.5, 52, 82.1, 50.6, 55.6, 26
log BF( 7 ) = 69576.66
log posterior increase: 2035.8, 56.9, 6.8, 27.9, 1.4, 0.6, 7.6, 1.1, 1, 0.4, 0.2, 0.1, done.
log BF( 8 ) = 70825.42
log posterior increase: 1387.8, 131.9, 80.2, 14.9, 6.1, 1.3, 0.4, 0.1, 3.6, done.
log BF( 9 ) = 72622.47
log posterior increase: 1338.8, 62.6, 84.9, 85.4, 115.4, 53, 58.2, 197.9, 55.2, 41.5, 145.6,
log BF( 10 ) = 79285.42
log posterior increase: 1201.7, 47.2, 23.5, 7.5, 9.6, 6.2, 2.9, 3.3, 4.9, 8.2, 2.9, 1.8, 6.9
log BF( 11 ) = 79440.62
log posterior increase: 1141.6, 71.3, 19.3, 16.8, 66.5, 13.9, 9.5, 5.5, 28.8, 19.3, 17.8, 4.7
log BF( 12 ) = 79697.92
log posterior increase: 1186.2, 82.1, 37.1, 10.7, 4.4, 5.3, 0.9, 0.4, 0.3, 0.2, 0.2, 0.1, 1.1
log BF( 13 ) = 78812.7
log posterior increase: 1043.3, 26.9, 18.5, 17.1, 15.8, 3.3, 8.7, 2.6, 3.1, 0.6, 0.1, done.
```

```
log BF( 14 ) = 77383.89
```

```
# It generates 12 topics
```

```
## Interpretation
# ordering by `topic over aggregate' lift:
summary(tpcs, n=5)
```

```
Top 5 phrases by topic-over-null term lift (and usage %):

[1] 'commonly.prescribed.drug', 'medic.liability.insurance', 'medic.liability.crisi', 'death
[2] 'southeast.texa', 'troop.bring.home', 'un.official', 'nunn.lugar.program', 'god.bless.ame
[3] 'national.heritage.corridor', 'asian.pacific.american', 'violence.sexual.assault', 'paci
[4] 'reverse.robin.hood', 'va.health.care', 'passenger.rail.service', 'passenger.rail', 'disa
[5] 'united.airline.employe', 'student.loan.cut', 'security.private.account', 'private.accou
[6] 'near.retirement.age', 'increase.taxe', 'personal.retirement.account', 'gifted.talented.s
[7] 'judge.alberto.gonzale', 'judicial.confirmation.process', 'chief.justice.rehnquist', 'fi
[8] 'low.cost.reliable', 'ready.mixed.concrete', 'indian.art.craft', 'price.natural.ga', 'wit
[9] 'north.american.fre', 'financial.accounting.standard', 'american.fre.trade', 'central.ame
[10] 'change.heart.mind', 'hate.crime.legislation', 'wild.bird', 'republic.cypru', 'hate.crim
[11] 'national.ad.campaign', 'pluripotent.stem.cel', 'regional.training.cent', 'cel.stem.cel
[12] 'able.buy.gun', 'deep.sea.coral', 'buy.gun', 'credit.card.industry', 'caliber.sniper.ri

Log Bayes factor and estimated dispersion, by number of topics:

             2        3        4        5        6        7        8        9
logBF 30123.15 44142.75 53865.63 60318.97 64330.64 69576.66 70825.42 72622.47
Disp      4.96     4.29     3.89     3.58     3.34     3.19     2.99     2.93
            10       11       12       13       14
logBF 79285.42 79440.62 79697.92 78812.70 77383.89
Disp      2.85     2.74     2.66     2.57     2.49

Selected the K = 12 topic model
```

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

```
 [1] "postal.service"    "class.action"      "private.property"
```

```
 [4] "death.tax"         "strong.support"    "american.people"
 [7] "post.office"       "prescription.drug" "property.right"
[10] "hurricane.katrina"
```

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

```
 [1] "american.people"  "iraqi.people"     "saddam.hussein"   "war.iraq"
 [5] "war.terror"       "iraq.afghanistan" "border.security"  "war.terrorism"
 [9] "strong.support"   "god.bless"
```

```
# Look at party mean memberships
DemO <- colMeans(tpcs$omega[congress109Ideology$party=="D",])
RepO <- colMeans(tpcs$omega[congress109Ideology$party=="R",])
sort(DemO/RepO)
```

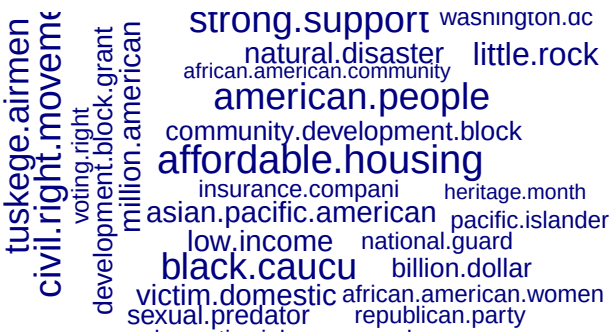
```
        6         8         1         2        11         7         9        10
0.2866818 0.3267723 0.3312262 0.4107537 0.4116513 0.5297540 1.5850736 2.0092146
        3         4        12         5
2.2619136 2.6813479 4.2924384 9.2071912
```

```
set.seed(1)

library(wordcloud)
```

```
Loading required package: RColorBrewer
```

```
# Topic 1 more favored by the Republican
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,1], min.freq=0.004, col="maroon")
```


```
# Topic 3 more favored by the Democratic
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,3], min.freq=0.004, col="navy")
```



Answer

The Republican are indicated by topics: 1, 2, 6, 7, 8, and 11.

The Democratic are indicated by topics: 3, 4, 5, 9, 10, and 12

Bayes factors choose 12 topics in our case, and the top five phrases of each topic is printed below. For instance, the first topic includes medical liability and tax repeal. We can also look at the straight word probabilities; while the first topic is hard to interpret since it contains wide-ranging issues, it is clear that the second topic centers around war and terrorism. The

topics associated with democrats (ratio is higher than one) are 9, 10, 3, 4, 12, 5 and the rest is republican. From Wordles, we can see that the democrats often bring up topics around terrorism while the republican shows more varied interest, such as hurricane and death tax.

Interpretation:

The word clouds tell us the frequency with which democrats and republicans referred to certain topics by increasing the word size proportionally to frequently. This means that when we cluster words, these are usually the words used by the same people and can thus be broken into democrat and republican speaking points for the given time frame they were in of the data.

The maroon colored wordcloud (from Topic 1) is based on the words used more by the Republican party which were: private property, american people, property right, and prescription drug. The navy colored wordcloud (from Topic 3) is based on the words used more by the Democratic party which were: hurricane Katrina and the gulf coast.

Question 3

Connect the unsupervised clusters to partisanship. Tabulate party membership by K-means cluster. Are there any non-partisan topics? Fit topic regressions for each of party and repshare. Compare to regression onto phrase percentages:

```
set.seed(1)

## [3] partisanship
tapply(congress109Ideology$party,kmfs$cluster,table)
```

```
$`1`

D I R
0 0 3

$`2`

  D   I   R
137   2   0

$`3`

D I R
1 0 0

$`4`

 D  I  R
 1  0 14

$`5`

  D   I   R
103   0 268
```

```
colnames(fs)[order(-kmfs$centers[which.max(kmfs$size),])[1:10]]
```

```
 [1] "look.forward"        "strong.support"      "urge.support"
 [4] "death.tax"           "illegal.immigrant"   "border.security"
```

```
 [7] "illegal.immigration" "private.property"    "pass.bil"
[10] "appropriation.bil"
```

```r
## Fit a topic regression
library(gamlr)

gop <- congress109Ideology[,"party"]=="R"

# Logistic regression
partyreg <- gamlr(tpcs$omega, gop,family="binomial")
# odd multipliers for a 0.1 rise in topic weight in doc
print(exp(coef(partyreg)*0.1))
```

```
13 x 1 Matrix of class "dgeMatrix"
              seg100
intercept 1.1489854
1         1.1756840
2         1.1690092
3         0.7205065
4         0.7106229
5         0.1589073
6         2.7613480
7         1.0000000
8         1.3475696
9         0.7515275
10        0.6883877
11        1.1789563
12        0.4573200
```

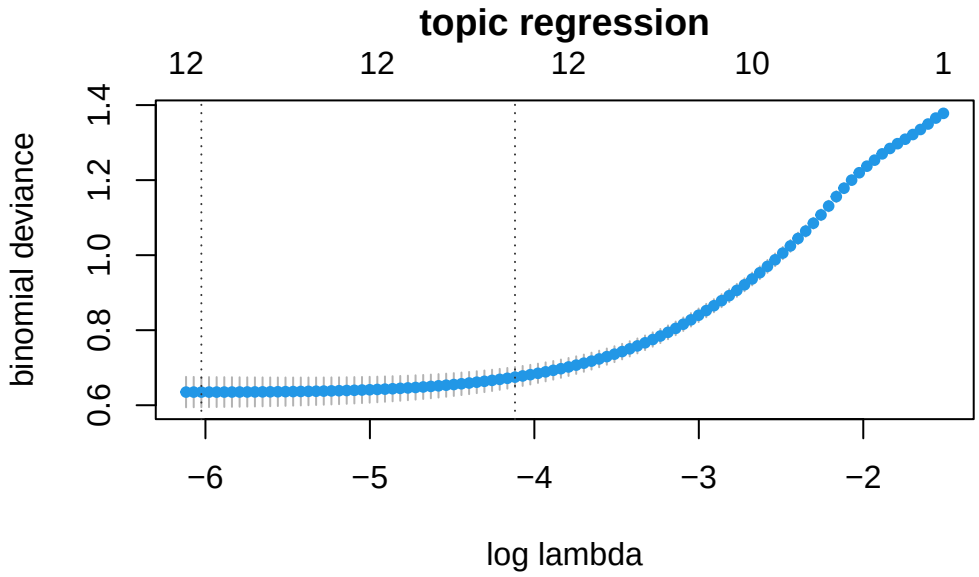
```r
# Linear regression
repreg <- gamlr(tpcs$omega, congress109Ideology[,"repshare"])
# increase in repshare per 0.1 rise in topic in doc
print(coef(repreg)*0.1)
```

```
13 x 1 sparse Matrix of class "dgCMatrix"
                seg100
intercept  0.057635550
1          0.004314108
2          0.001574869
3         -0.025501474
```

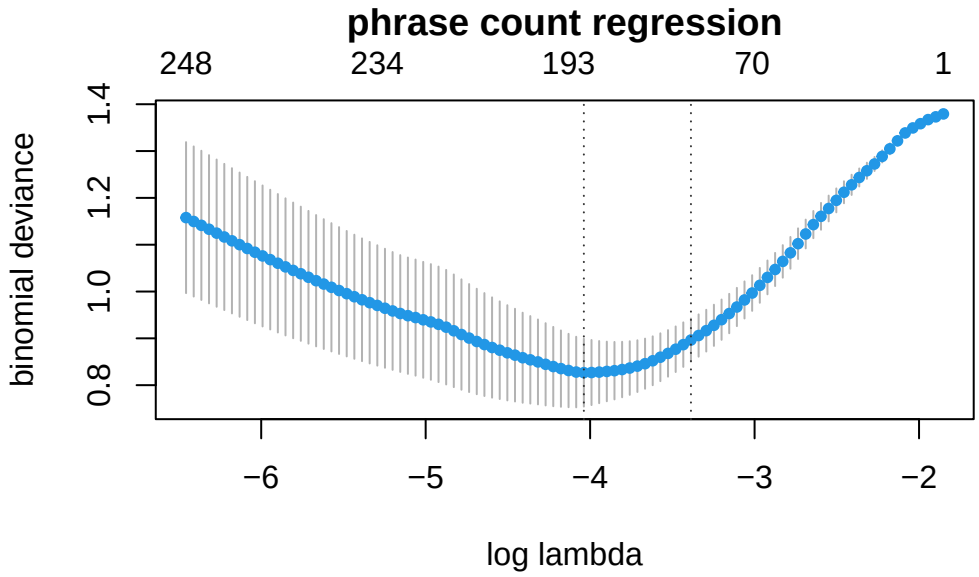
```
4          -0.012536466
5          -0.022275088
6           0.007593223
7           0.002222724
8           0.006946661
9          -0.009867078
10         -0.010989308
11          .
12         -0.021885216
```

```
# Compare to straight regression
regtopics.cv <- cv.gamlr(tpcs$omega, gop, family="binomial")
## give it the word %s as inputs
x <- 100*congress109Counts/rowSums(congress109Counts)
regwords.cv <- cv.gamlr(x, gop, family="binomial")
```

```
plot(regtopics.cv, main="topic regression")
```




```
plot(regwords.cv, main="phrase count regression")
```




```
# Max OOS R^2s
max(1-regtopics.cv$cvm/regtopics.cv$cvm[1])
```

```
[1] 0.5391416
```

```
max(1-regwords.cv$cvm/regwords.cv$cvm[1])
```

```
[1] 0.4005889
```

Answer

There are 10 non-partisan topics:

```
[1] "look.forward"        "strong.support"      "urge.support"        "death.tax"
 [5] "illegal.immigrant"   "border.security"     "illegal.immigration" "private.property"
 [9] "pass.bil"            "appropriation.bil"
```

The topic regression has a higher R2 than the LASSO (53.47% compared to 40.96%, respectively) which means it explains more of the variation in party membership than the LASSO model does. Thus, the topic regression is the better model in this case to pick out party membership based on words.