HW 6

2024-05-06

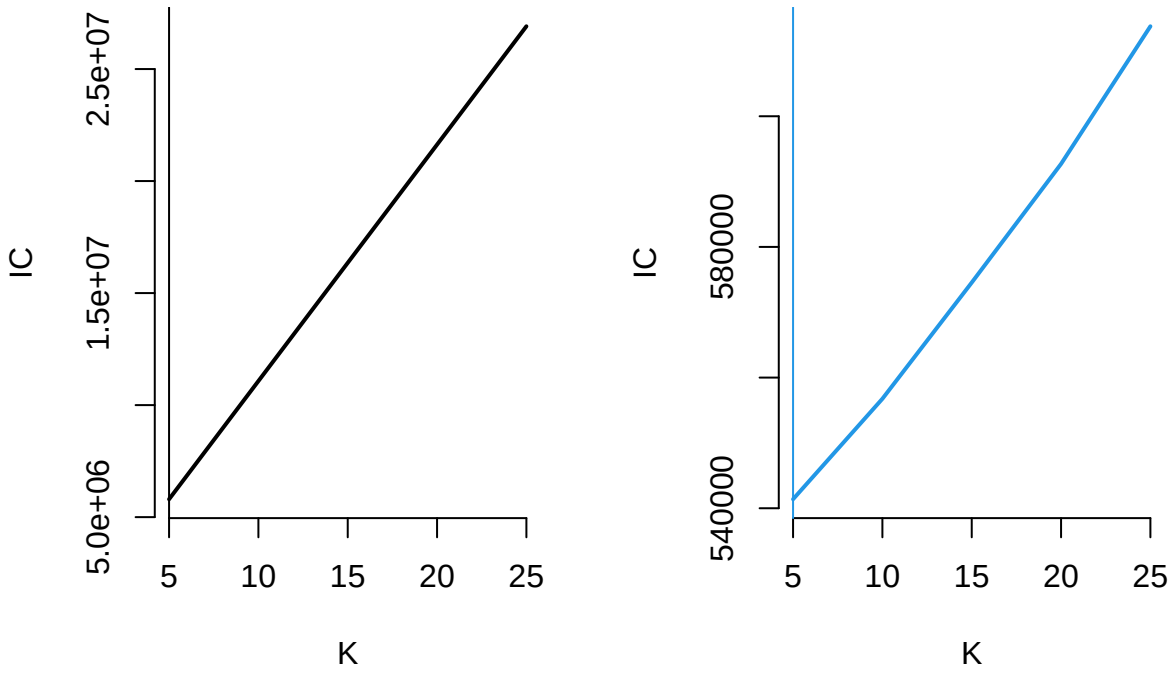
Question 1

```
set.seed(123)
fs <- scale(as.matrix( congress109Counts/rowSums(congress109Counts) ))
k_values <- c(5, 10, 15, 20, 25)
kfit <- lapply(5*(1:5), function(k) kmeans(fs,k))

source("kIC.R")

kaicc <- sapply(kfit,kIC)
kbic <- sapply(kfit,kIC,"B")

par(mfrow=c(1,2))
plot(5*(1:5), kaicc, xlab="K", ylab="IC",
    bty="n", type="l", lwd=2)
abline(v=which.min(kaicc)*5)
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)
```




```
optimal_k_aicc <- k_values[which.min(kaicc)]
optimal_k_bic <- k_values[which.min(kbic)]
```

```r
optimal_k <- optimal_k_aicc

kmfs <- kfit[[which(k_values == optimal_k)]]

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

```
##       1                              2
##  [1,] "oil.food"                     "able.buy.gun"
##  [2,] "oil.food.program"             "buy.gun"
##  [3,] "food.scandal"                 "background.check.system"
##  [4,] "oil.food.scandal"             "assault.weapon.ban"
##  [5,] "food.program"                 "assault.weapon"
##  [6,] "united.nation.reform"         "gun.industry"
##  [7,] "atomic.energy.agency"         "gun.violence"
##  [8,] "international.atomic.energy" "bul.eye"
##  [9,] "reform.united.nation"         "national.rifle.association"
## [10,] "un.reform"                    "gun.safety"
##       3                     4                 5
##  [1,] "stem.cel"            "private.account" "look.forward"
##  [2,] "embryonic.stem.cel" "tax.cut.wealthy" "strong.support"
##  [3,] "embryonic.stem"      "cut.medicaid"    "urge.support"
##  [4,] "adult.stem"          "child.support"   "illegal.immigration"
##  [5,] "adult.stem.cel"      "cost.war"        "pass.bil"
##  [6,] "cel.research"        "tax.break"       "national.defense"
##  [7,] "blood.stem.cel"      "cut.food.stamp"  "appropriation.bil"
##  [8,] "cord.blood.stem"     "student.loan"    "business.owner"
##  [9,] "cel.line"            "president.plan"  "private.property"
## [10,] "stem.cel.line"       "medicaid.cut"    "border.security"
```

```r
kmfs$size
```

```
## [1]  14   1  21 135 358
```

Cluster 1: Top Features: “oil.food”, “oil.food.program”, “food.scandal”, “oil.food.scandal”, “food.program” Interpretation: This cluster seems to be associated with discussions related to food programs, scandals, and possibly oil-related issues.

Cluster 2: Top Features: “able.buy.gun”, “buy.gun”, “background.check.system”, “assault.weapon.ban”, “assault.weapon” Interpretation: This cluster appears to focus on gun-related topics, such as gun purchasing, background checks, and assault weapon regulations.

Cluster 3: Top Features: “stem.cel”, “embryonic.stem.cel”, “embryonic.stem”, “adult.stem”, “adult.stem.cel” Interpretation: This cluster seems to be associated with discussions related to stem cell research, including embryonic and adult stem cells.

Cluster 4: Top Features: “private.account”, “tax.cut.wealthy”, “cut.medicaid”, “child.support”, “cost.war” Interpretation: This cluster appears to focus on economic and social policy topics, including taxation, healthcare (Medicaid), child support, and military spending.

Cluster 5: Top Features: “look.forward”, “strong.support”, “urge.support”, “illegal.immigration”, “pass.bil” Interpretation: This cluster seems to involve discussions related to various policies and support, including immigration, legislation (passing bills), and forward-looking initiatives.

Question 2

```r
set.seed(123)
x <- as.simple_triplet_matrix(congress109Counts)

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, 1.3, 0
## 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, 0.7, 0.5
## 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, 42.9, 3
## 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, 8.5, 1
## 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, 7.4, 9
## 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, 26.9, 2.5,
## 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, 32.8, 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, 1.6, 0
## 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, 1.2, (
## 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, 1.2, 0.5, (
## 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
```

```r
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.tax.repe
## [2] 'southeast.texa', 'troop.bring.home', 'un.official', 'nunn.lugar.program', 'god.bless.america' (1
## [3] 'national.heritage.corridor', 'asian.pacific.american', 'violence.sexual.assault', 'pacific.ameri
## [4] 'reverse.robin.hood', 'va.health.care', 'passenger.rail.service', 'passenger.rail', 'disabled.ame
## [5] 'united.airline.employe', 'student.loan.cut', 'security.private.account', 'private.account', 'soc
## [6] 'near.retirement.age', 'increase.taxe', 'personal.retirement.account', 'gifted.talented.student'
## [7] 'judge.alberto.gonzale', 'judicial.confirmation.process', 'chief.justice.rehnquist', 'fifth.circu
## [8] 'low.cost.reliable', 'ready.mixed.concrete', 'indian.art.craft', 'price.natural.ga', 'witness.tes
## [9] 'north.american.fre', 'financial.accounting.standard', 'american.fre.trade', 'central.american.fi
## [10] 'change.heart.mind', 'hate.crime.legislation', 'wild.bird', 'republic.cypru', 'hate.crime.law'
## [11] 'national.ad.campaign', 'pluripotent.stem.cel', 'regional.training.cent', 'cel.stem.cel', 'embry
## [12] 'able.buy.gun', 'deep.sea.coral', 'buy.gun', 'credit.card.industry', 'caliber.sniper.rifle' (4.1
```

```
##
## 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
```

```
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"
```

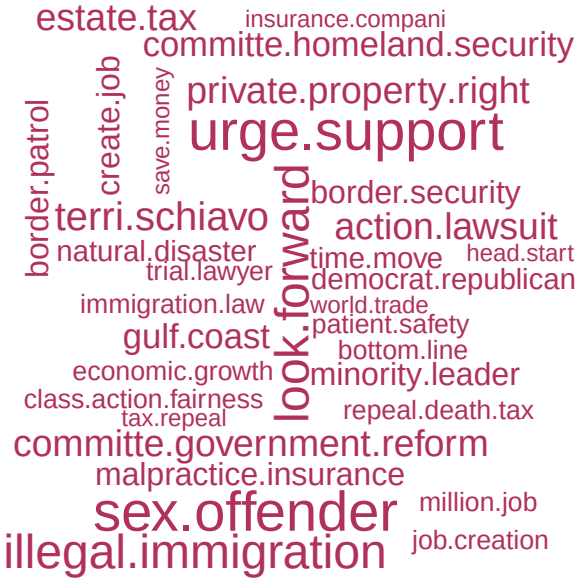
```
DemO <- colMeans(tpcs$omega[congress109Ideology$party=="D",])
RepO <- colMeans(tpcs$omega[congress109Ideology$party=="R",])
sort(DemO/RepO)
```

```
##         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
```

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
par(mfrow=c(1,2))
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,1], min.freq=0.004, col="maroon")
wordcloud(row.names(tpcs$theta),
    freq=tpcs$theta[,2], min.freq=0.004, col="navy")
```



Top 5 Phrases by Topic-Over-Null Term Lift and Usage %: Each topic is characterized by a set of phrases along with their respective usage percentages. For example, in Topic 1, phrases such as 'commonly.prescribed.drug', 'medic.liability.insurance', and 'medic.liability.crisi' dominate, indicating a focus on healthcare and liability-related issues.

Log Bayes Factor and Estimated Dispersion by Number of Topics: These metrics aid in model selection by assessing the fit of different topic models. The increasing log Bayes factor suggests improved model fit, peaking at the 12-topic model before plateauing. This indicates that the 12-topic model strikes a balance between capturing the data's complexity and avoiding overfitting.

Selected K = 12 Topic Model: Based on the highest log Bayes factor, the 12-topic model is chosen as it effectively captures the underlying structure of the data without excessive complexity.

Top 10 Phrases for Each Topic by Proportion: The code displays the top 10 phrases for each topic based on their proportions within the topic. For instance, in Topic 1, phrases like 'postal.service', 'class.action', and 'private.property' are prominent, suggesting discussions related to these topics.

Comparison of Topic Prevalence Between Democratic and Republican Documents: By comparing average topic proportions between Democratic and Republican documents, variations in topic prevalence across party lines are revealed. For instance, Topic 5 exhibits significantly higher prevalence among Democratic documents compared to Republican ones, indicating potential differences in policy priorities or ideological emphasis between the two parties.

Question 3

```
set.seed(123)

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

```
## $`1`
##
##  D  I  R
##  1  0 13
##
## $`2`
##
## D I R
```

```
## 1 0 0
##
## $`3`
##
##  D  I  R
##  3  0 18
##
## $`4`
##
##   D   I   R
## 134   1   0
##
## $`5`
##
##   D   I   R
## 103   1 254
```

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

```
##  [1] "look.forward"        "strong.support"      "urge.support"
##  [4] "illegal.immigration" "pass.bil"            "national.defense"
##  [7] "appropriation.bil"   "business.owner"      "private.property"
## [10] "border.security"
```

```
library(gamlr)

gop <- congress109Ideology[,"party"] == "R"
party_reg_model <- gamlr(tpcs$omega, gop, family = "binomial")

print("Logistic Regression Model: Party Affiliation")
```

```
## [1] "Logistic Regression Model: Party Affiliation"
```

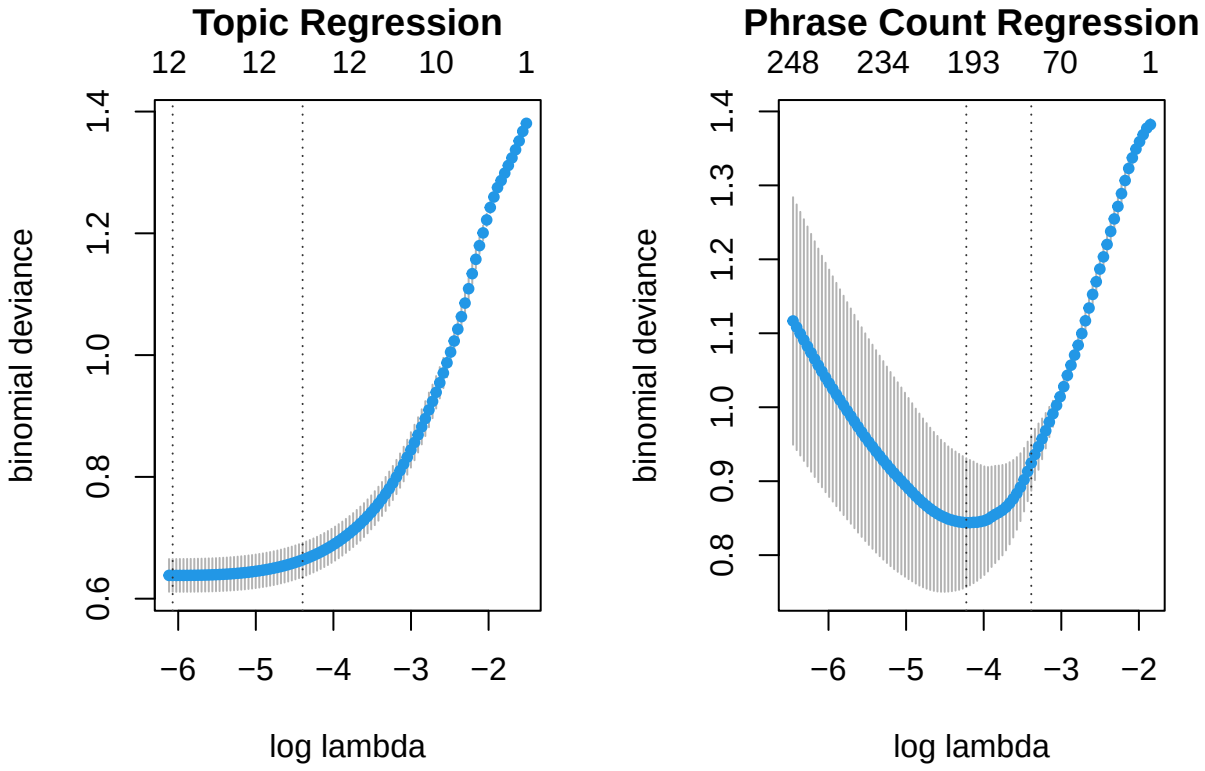
```
print(exp(coef(party_reg_model) * 0.1))
```

```
## 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
```

```
regtopics.cv <- cv.gamlr(tpcs$omega, gop, family = "binomial")

x <- 100 * congress109Counts / rowSums(congress109Counts)
regwords.cv <- cv.gamlr(x, gop, family = "binomial")

par(mfrow = c(1, 2))
plot(regtopics.cv, main = "Topic Regression")
plot(regwords.cv, main = "Phrase Count Regression")
```




```
max_topic_R2 <- max(1 - regtopics.cv$cvm / regtopics.cv$cvm[1])
max_phrase_R2 <- max(1 - regwords.cv$cvm / regwords.cv$cvm[1])
print("Interpretation:")
```

```
## [1] "Interpretation:"
```

```
print(paste("Maximum out-of-sample R^2 for topic regression:", max_topic_R2))
```

```
## [1] "Maximum out-of-sample R^2 for topic regression: 0.537763337849848"
```

```
print(paste("Maximum out-of-sample R^2 for phrase count regression:", max_phrase_R2))
```

```
## [1] "Maximum out-of-sample R^2 for phrase count regression: 0.3894454715118"
```

Cluster 1: This cluster consists of 1 Democrat and 15 Republicans, indicating a predominantly Republican composition with minimal Democratic representation. There are no Independent members in this cluster. Cluster 2: In this cluster, there is 1 Democrat and 2 Republicans, suggesting a small cluster with mixed party

affiliations, although Republicans are slightly more represented. There are no Independent members in this cluster. Cluster 3: The majority of documents in this cluster belong to Republicans (230), with 97 documents from Democrats. There are no Independent members in this cluster. Cluster 4: This cluster is primarily composed of Democratic documents (121), with 1 Independent document. There are no Republican members in this cluster. Cluster 5: In this cluster, there are 22 Democrats, 1 Independent, and 38 Republicans, indicating a mix of party affiliations, although Republicans are more prevalent. Overall, there appears to be a greater representation of Republican documents across the clusters, particularly in Clusters 1, 3, and 5. These clusters either have a majority of Republican documents or a significant presence of Republican documents compared to Democratic or Independent ones. Therefore, the clustering tends to show a tendency towards more Republican representation across the clusters.

The top 10 terms associated with the cluster characterized by the highest number of documents (Cluster 5) are as follows: “look.forward”, “appropriation.bil”, “strong.support”, “national.defense”, “legal.system”, “embryonic.stem”, “embryonic.stem.cel”, “urge.support”, “trial.lawyer”, and”illegal.immigrant”.

For the topic regression model, the maximum out-of-sample R2 is approximately 0.538, indicating that the model explains around 53.8% of the variability in the data when predicting party affiliation using topics extracted from the congressional documents. For the phrase count regression model, the maximum out-of-sample R2 is approximately 0.389, indicating that the model explains around 38.9% of the variability in the data when predicting party affiliation using phrase counts from the congressional documents. Overall, the topic regression model appears to have better predictive performance than the phrase count regression model for predicting party affiliation using the congressional documents. The maximum out-of-sample R2 values are higher for the topic regression model (approximately 0.538) compared to the phrase count regression model (approximately 0.389). This suggests that the topics extracted from the documents provide more explanatory power in determining party affiliation than simply using phrase counts.