BUSN 41201 HW6

Group 11: Yu-Ting Weng, Mengdi Hao, Elena Li, Minji Park, Sarah Lee

Question 1

- Fit K-means to speech text for K in 5, 10, 15, 20, 25

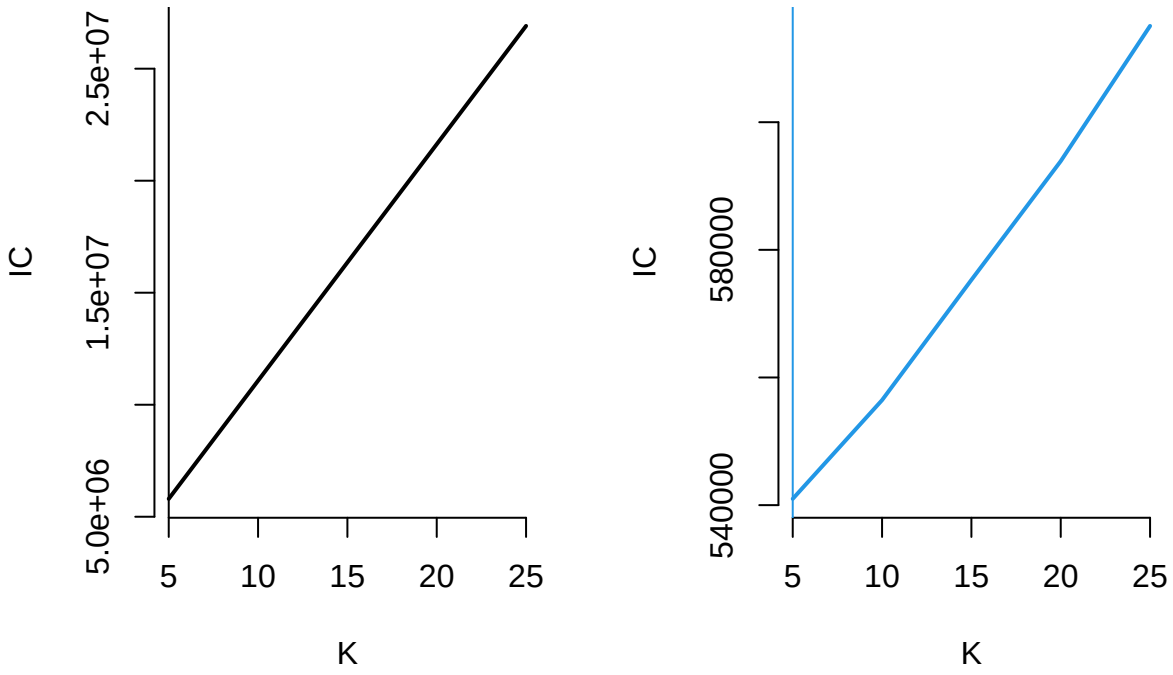
```
fs <- scale(as.matrix( congress109Counts/rowSums(congress109Counts) ))

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

source("kIC.R")

kaicc <- sapply(kfit,kIC)

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



- User BID to choose the K

Based on the provided BIC plot, the Bayesian Information Criterion (BIC) continuously increases as the number of clusters K increases from 5 to 25. This trend suggests that the simplest model with $K = 5$ is

the most appropriate, as adding more clusters does not provide a better fit to justify the increased model complexity.

- Interpret the selected model.

```
kmfs <- kfit[[1]]

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

```
##       1                             2
##  [1,] "cut.medicaid"                "court.appeal"
##  [2,] "tax.cut.wealthy"             "business.meeting"
##  [3,] "private.account"             "circuit.court.appeal"
##  [4,] "cut.food.stamp"              "committe.foreign.relation"
##  [5,] "care.cut"                    "judicial.nomine"
##  [6,] "child.support"               "housing.urban.affair"
##  [7,] "student.loan"                "urban.affair"
##  [8,] "privatizing.social.security" "court.judge"
##  [9,] "cost.war"                    "committe.commerce.science"
## [10,] "medicaid.cut"                "banking.housing.urban"
##       3                      4
##  [1,] "strong.support"       "social.security.system"
##  [2,] "urge.support"         "security.system"
##  [3,] "driver.license"       "social.security.reform"
##  [4,] "embryonic.stem"       "security.reform"
##  [5,] "embryonic.stem.cel"   "personal.account"
##  [6,] "illegal.immigrant"    "social.security.benefit"
##  [7,] "civil.right.movement" "security.benefit"
##  [8,] "right.movement"       "retirement.account"
##  [9,] "stem.cel"             "social.security.trust"
## [10,] "post.office"          "security.trust"
##       5
##  [1,] "suppli.natural.ga"
##  [2,] "supply.natural.ga"
##  [3,] "ga.natural.ga"
##  [4,] "natural.ga.natural"
##  [5,] "able.buy.gun"
##  [6,] "ga.natural"
##  [7,] "buy.gun"
##  [8,] "natural.ga"
##  [9,] "grand.ole.opry"
## [10,] "background.check.system"
```

```
kmfs$size
```

```
## [1] 128  53 293  52   3
```

The k-means clustering results reveal varied discussion themes within Congress, with each cluster focusing on specific issues. For example, Cluster 3, the largest with 293 members, predominantly focuses on social security and rights, highlighting deep concerns about welfare reforms and ethical debates surrounding "embryonic.stem" research. The smaller clusters, like Cluster 5 with only 3 members, indicate niche discussions, exemplified by topics like natural gas supply management and cultural references such as "grand.ole.opry," reflecting highly specialized legislative interests.

Question 2

- Fit a topic model for the speech counts.

```
x <- as.simple_triplet_matrix(congress109Counts)
tpcs <- topics(x,K=2:25)
```

- User Bayes factors to choose the number of topics

The $\mathbf{K = 12}$ topic model was selected based on its higher Log Bayes factor of **79713.54**, suggesting a better model fit than other topic numbers.

```
summary(tpcs, n=5)
```

```
##
## Top 5 phrases by topic-over-null term lift (and usage %):
##
## [1] 'commonly.prescribed.drug', 'medic.liability.crisi', 'medic.liability.insurance', 'tax.repeal',
## [2] 'southeast.texa', 'winning.war.iraq', 'troop.bring.home', 'hugo.chavez', 'nunn.lugar.program' (13
## [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.am
## [5] 'united.airline.employe', 'student.loan.cut', 'security.private.account', 'private.account', 'so
## [6] 'near.retirement.age', 'increase.taxe', 'personal.retirement.account', 'gross.national.product',
## [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 44141.73 53865.20 60319.64 64330.69 69577.17 71642.67 75702.84
## Disp      4.96     4.29     3.89     3.58     3.34     3.20     3.07     2.94
##             10       11       12       13       14
## logBF 79385.78 79422.38 79713.54 78813.38 77381.24
## Disp      2.85     2.74     2.67     2.57     2.49
##
## Selected the K = 12 topic model
```

- Interpret your chosen 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" "hurricane.katrina"
## [10] "property.right"
```

```
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"  "strong.support"
##  [9] "war.terrorism"    "god.bless"
```

Topic 1 (Potentially GOP-leaning)

: Terms like "postal.service," "class.action," "private.property," "death.tax," "property.right," and "estate.tax" suggest a focus on economic issues, property rights, and fiscal policies. These

are areas typically emphasized by the Republican Party, which advocates for lower taxes, less government intervention in private businesses, and protecting of personal property rights.

Topic 2 (Potentially Dem-leaning)

: Keywords such as "american.people," "iraqi.people," "saddam.hussein," "war.iraq," "war.terror," and "iraq.afghanistan" are strongly associated with foreign policy and military intervention. These topics have been central in Democratic discussions, especially in contexts of criticism or opposition to military strategies and promoting international diplomacy.

```
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
## 0.2983746 0.3240069 0.3362088 0.3915219 0.4150462 0.5313903 1.5767100 1.9852329
##         3         4        12         5
## 2.2858989 2.6580812 4.3533760 9.2606947
```

Topics such as **#6, #8, #1, and #2** have the lowest ratios, indicating a stronger **Republican** emphasis, while topics like **#5, #12, and #4** show the highest ratios, suggesting they are more heavily discussed by **Democrats**.

```
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")
```

1. **Word Cloud 1**: Emphasizing terms such as "border.security," "estate.tax," and "illegal.immigration," aligns closely with Republican (GOP) themes, which typically focus on security, fiscal responsibility, and stringent immigration policies.
2. **Word Cloud 2**: Highlighting issues like "war.terror," "immigration.reform," and "national.guard," this word cloud corresponds with Democratic (Dems) priorities.
3. Thus, the fit of the word clouds supports the interpretation that Word Cloud 1 is predominantly GOP-focused. At the same time, Word Cloud 2 is oriented towards Democratic concerns, just as it appeared in the ordered in-topic probabilities.

Question 3

- Tabulate party membership by K-means cluster.

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

```
## $`1`
##
##   D   I   R
## 126   2   0
##
## $`2`
##
##  D  I  R
##  4  0 49
##
## $`3`
##
##   D   I   R
## 103   0 190
##
## $`4`
##
##  D  I  R
##  8  0 44
##
## $`5`
##
## D I R
## 1 0 2
```

This shows that the clusters generally split along party lines, with clusters like S1 and S3 demonstrating a significant majority of Democrats (126 in S1 and 103 in S3) with virtually no Republicans, indicating a partisan divide in the topics discussed within these clusters. Conversely, clusters S2, S4, and S5 are predominantly Republican, suggesting these clusters focus on issues more aligned with Republican interests.

- Are there any non-partisan topics?

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

```
##  [1] "strong.support"       "urge.support"         "driver.license"
##  [4] "embryonic.stem"       "embryonic.stem.cel"   "illegal.immigrant"
##  [7] "civil.right.movement" "right.movement"       "stem.cel"
## [10] "post.office"
```

The terms "driver.license," "embryonic.stem.cel," and "post.office" appear to be relatively non-partisan topics among the top features listed.

- Fit topic regressions for each of party and repshare.

```
gop <- congress109Ideology[,"party"]=="R"
partyreg <- gamlr(tpcs$omega, gop,
    family="binomial")
print(exp(coef(partyreg)*0.1))
```

```
## 13 x 1 Matrix of class "dgeMatrix"
##              seg100
## intercept 1.1473498
## 1         1.1863581
## 2         1.1752610
## 3         0.7222541
## 4         0.7114856
## 5         0.1668486
## 6         2.5794266
## 7         1.0000000
## 8         1.3712341
## 9         0.7539317
## 10        0.6911577
## 11        1.1765672
## 12        0.4676948
```

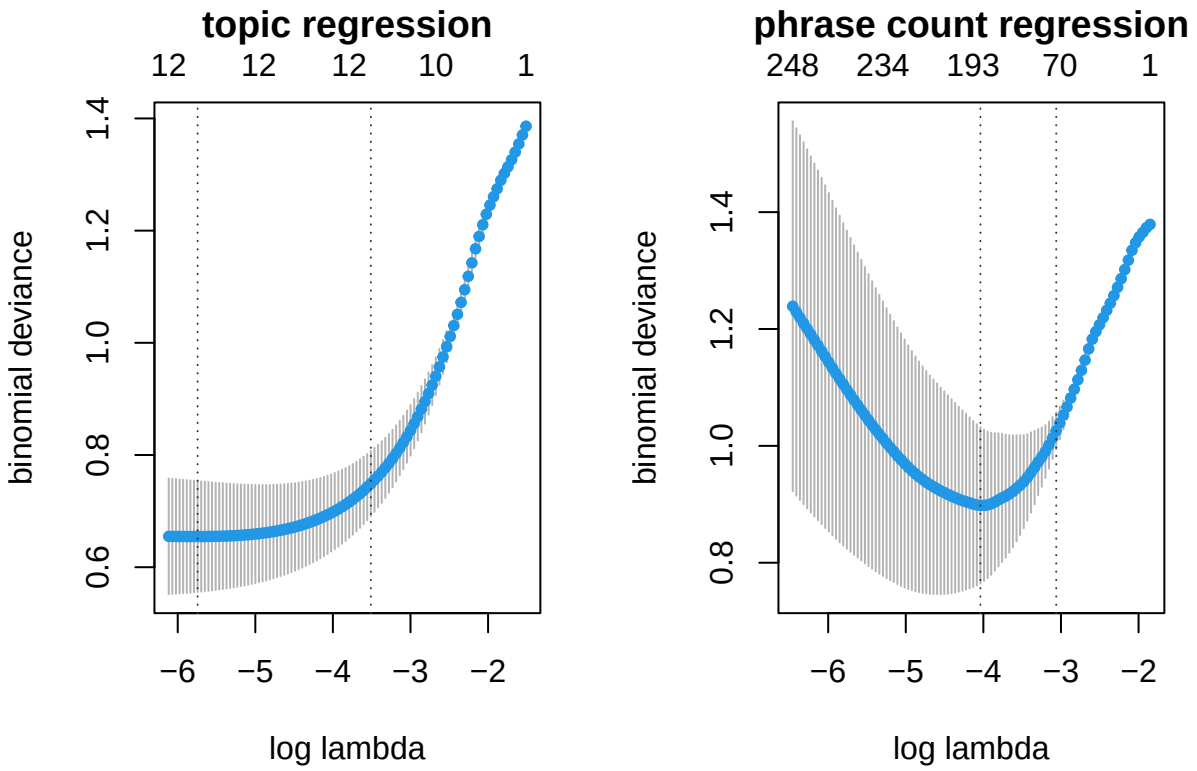
```
repreg <- gamlr(tpcs$omega, congress109Ideology[,"repshare"])
print(coef(repreg)*0.1)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                  seg100
## intercept  0.057470962
## 1          0.004526253
## 2          0.002138900
## 3         -0.025481317
## 4         -0.012245264
## 5         -0.022245009
## 6          0.007350493
## 7          0.002384980
## 8          0.007129397
## 9         -0.009661307
## 10        -0.010855845
## 11         .
## 12        -0.021548744
```

- Compare to regression onto phrase percentages.

```
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")
```



In both graphs, the vertical dashed lines indicate optimal lambda values where the minimum deviance occurs. The **topic regression demonstrates a more stable,** lower deviance across a range of lambda values compared to the phrase count regression, suggesting better model performance and stability with topic features.

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

```
## [1] 0.5276754
```

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

```
## [1] 0.3486564
```

The results show that the maximum out-of-sample R^2 value for the topic model is approximately 0.5033, while for the word count model it is about 0.4025. This indicates that the **topic model provides a better fit** and explains more variability in the data compared to the word count model when predicting political party affiliation.