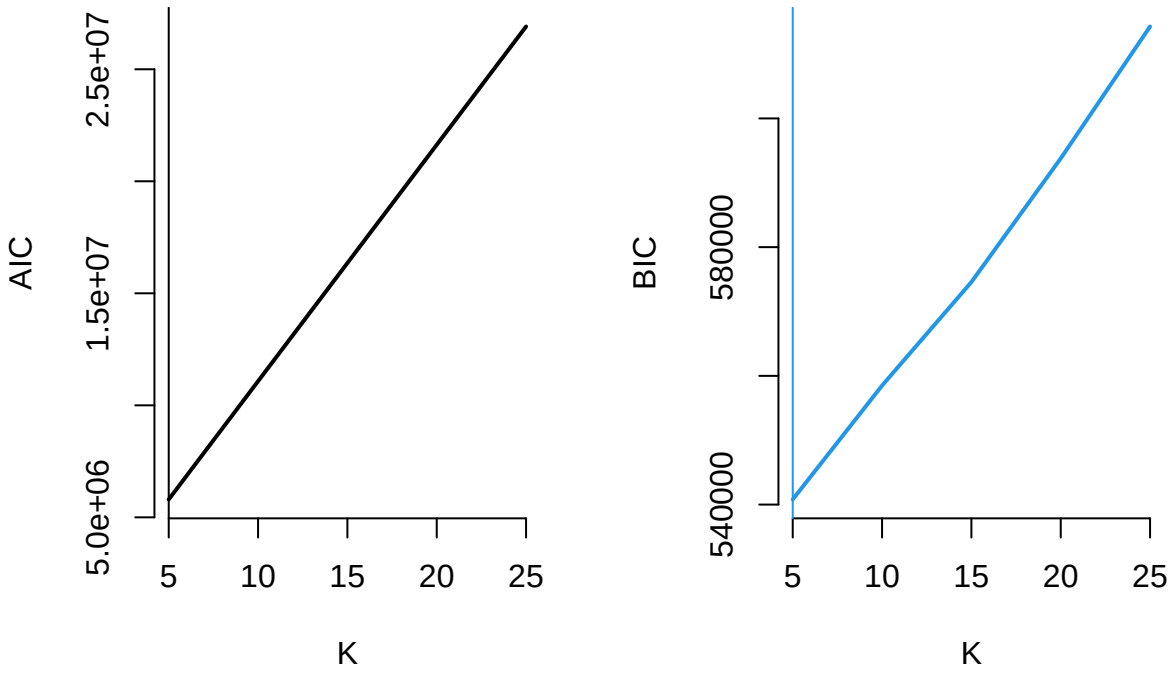
HW6_Mengdi

Mengdi Hao

2024-05-04

Q1. 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.

According to the plot, using BIC, K=5 is chosen. In these 5 clusters, each cluster can be interpreted based on the prominent terms shown below: Cluster 1: Topics related to American heritage and possibly related cultural or historical discussions. Cluster 2: Involves terms like "oil.food", "atomic.energy.agency", suggesting discussions related to energy policies or international relations. Cluster 3: Focused on fiscal policies like tax cuts, medicaid, and social security reforms. Cluster 4: Covers topics related to immigration and taxation issues. Cluster 5: Seems to deal with judicial and urban affairs, reflecting debates on urban development or judiciary matters.




```
##       1                          2
##  [1,] "oil.food"                 "private.account"
##  [2,] "oil.food.program"         "tax.cut.wealthy"
```

```
##  [3,] "atomic.energy.agency"         "cut.medicaid"
##  [4,] "international.atomic.energy"  "cut.food.stamp"
##  [5,] "united.nation.reform"         "child.support"
##  [6,] "food.program"                 "privatizing.social.security"
##  [7,] "reform.united.nation"         "care.cut"
##  [8,] "food.scandal"                 "student.loan"
##  [9,] "oil.food.scandal"             "cost.war"
## [10,] "international.peace.security" "plan.privatize"
##       3                             4
##  [1,] "suppli.natural.ga"           "court.appeal"
##  [2,] "supply.natural.ga"           "business.meeting"
##  [3,] "ga.natural.ga"               "circuit.court.appeal"
##  [4,] "natural.ga.natural"          "court.judge"
##  [5,] "able.buy.gun"                "committe.foreign.relation"
##  [6,] "ga.natural"                  "judicial.nomine"
##  [7,] "buy.gun"                     "housing.urban.affair"
##  [8,] "natural.ga"                  "urban.affair"
##  [9,] "grand.ole.opry"              "committe.commerce.science"
## [10,] "background.check.system"     "banking.housing.urban"
##       5
##  [1,] "strong.support"
##  [2,] "urge.support"
##  [3,] "death.tax"
##  [4,] "illegal.immigration"
##  [5,] "private.property"
##  [6,] "business.owner"
##  [7,] "repeal.death.tax"
##  [8,] "illegal.immigrant"
##  [9,] "pass.bil"
## [10,] "look.forward"
```

```
## [1]  13 125   3  53 335
```

Q2. Fit a topic model for the speech counts. Use Bayes factors to choose the number of topics, and interpret your chosen model.

```
x <- as.simple_triplet_matrix(congress109Counts)
tpcs <- topics(x,K=2:25) # it chooses 11 topics
```

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

```
##
## Top 5 phrases by topic-over-null term lift (and usage %):
##
## [1] 'near.earth.object', 'southeast.texa', 'flag.protection.amendment', 'million.illegal.alien', 'bl
## [2] 'national.heritage.corridor', 'columbia.river.gorge', 'asian.pacific.american', 'little.rock.nin
## [3] 'near.retirement.age', 'medic.liability.crisi', 'personal.retirement.account', 'commonly.prescri
## [4] 'united.airline.employe', 'record.budget.deficit', 'private.account', 'spending.cut.bil', 'priva
## [5] 'global.gag.rule', 'post.traumatic', 'post.traumatic.stress', 'traumatic.stress', 'stress.disord
## [6] 'low.cost.reliable', 'ready.mixed.concrete', 'grand.ole.opry', 'indian.affair', 'witness.testify
## [7] 'judicial.confirmation.process', 'protect.minority.right', 'fifth.circuit.court', 'chief.justice
## [8] 'republic.cypru', 'change.heart.mind', 'wild.bird', 'hate.crime.legislation', 'hate.crime.law' (
## [9] 'american.fre.trade', 'central.american.fre', 'north.american.fre', 'financial.accounting.standa
```

```
## [10] 'able.buy.gun', 'buy.gun', 'increase.minimum.wage', 'assault.weapon', 'credit.card.industry' (4
## [11] 'regional.training.cent', 'national.ad.campaign', 'pluripotent.stem.cel', 'cel.stem.cel', 'embry
##
## Log Bayes factor and estimated dispersion, by number of topics:
##
##              2        3        4        5        6        7        8        9
## logBF 30123.15 44142.75 53865.28 60318.95 64329.85 69583.92 74082.51 76340.36
## Disp      4.96     4.29     3.89     3.58     3.34     3.19     3.06     2.91
##             10       11       12       13       14       15
## logBF 77276.03 80322.45 79383.62 79438.45 77344.55 76508.86
## Disp      2.81     2.75     2.62     2.58     2.49     2.43
##
## Selected the K = 11 topic model
```

A topic model with K=11 is selected using Bayes factors.

Each topic's top 5 phrases, chosen by "topic-over-null term lift" (a ratio comparing the frequency of a term in a topic to its frequency in the corpus), provide a clear snapshot of what each topic is about:

- Topic 1: Dominated by phrases like 'near.earth.object', 'southeast.texa', and 'flag.protection.amendment'. This topic seems to focus on regional issues, immigration, and possibly national pride or security matters.
- Topic 2: Features 'national.heritage.corridor', 'columbia.river.gorge', and 'asian.pacific.american'. This could relate to cultural heritage and conservation areas, indicating discussions on environmental and cultural preservation.
- Topic 3: Includes terms such as 'near.retirement.age', 'medic.liability.crisi', suggesting a focus on healthcare, retirement, and economic issues related to seniors.
- Others: Subsequent topics cover various areas such as airline industry issues, fiscal policies, mental health issues, legal matters, and foreign policy.

These topics show a diverse range of discussions, reflecting different focal points in congressional speeches. The percentages in parentheses (usage %) indicate how much of the total discussion these topics represent, helping to understand their relative importance in the dataset.

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

```
##  [1] "american.people"     "postal.service"      "strong.support"
##  [4] "illegal.alien"       "private.property"    "illegal.immigration"
##  [7] "saddam.hussein"      "border.security"     "driver.license"
## [10] "post.office"
```

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

```
##  [1] "african.american"   "civil.right"        "domestic.violence"
##  [4] "head.start"         "rosa.park"          "hurricane.katrina"
##  [7] "gulf.coast"         "strong.support"     "affordable.housing"
## [10] "violence.women"
```

For the selected 11-topic model, further insights can be gathered by examining the top terms in certain topics:

Top Terms in Topic 1: Terms like 'american.people', 'illegal.alien', and 'border.security' suggest a focus on immigration and national security. Top Terms in Topic 2: Includes 'african.american', 'civil.right', indicating discussions centered around civil rights and possibly historical or cultural recognition.

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

```
##         3         1         6        11         7         9         8         2
## 0.2219662 0.2719377 0.2958398 0.4226948 0.5901655 1.6192768 1.6953960 2.2347976
```

```
##         5        10         4
## 2.6329351 4.1152138 8.2306288
```

The ratio of topic weights between parties (DemO/RepO) sorted:

Lower values indicate topics more frequently discussed by Republicans, while higher values suggest topics favored by Democrats. 3,1,6,11,7 are republican, and 9,8,2,5,10,4 are democratic.

```
library(wordcloud)
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")
```

The word clouds for Topics 1 (GOP-leaning) and 2 (Dem-leaning) visualize these differences, using maroon for GOP and navy for Dems. This visually reinforces the textual analysis, showing the distinct language and thematic concerns of each party.

Q3. 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.

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

```
## $`1`
##
##  D  I  R
##  1  0 12
##
## $`2`
##
##   D   I   R
## 124   1   0
##
## $`3`
##
## D I R
## 1 0 2
##
## $`4`
##
##  D  I  R
##  4  0 49
##
## $`5`
##
##   D   I   R
## 112   1 222
```

Based on the tabulation:

Clusters potentially indicating non-partisan topics:

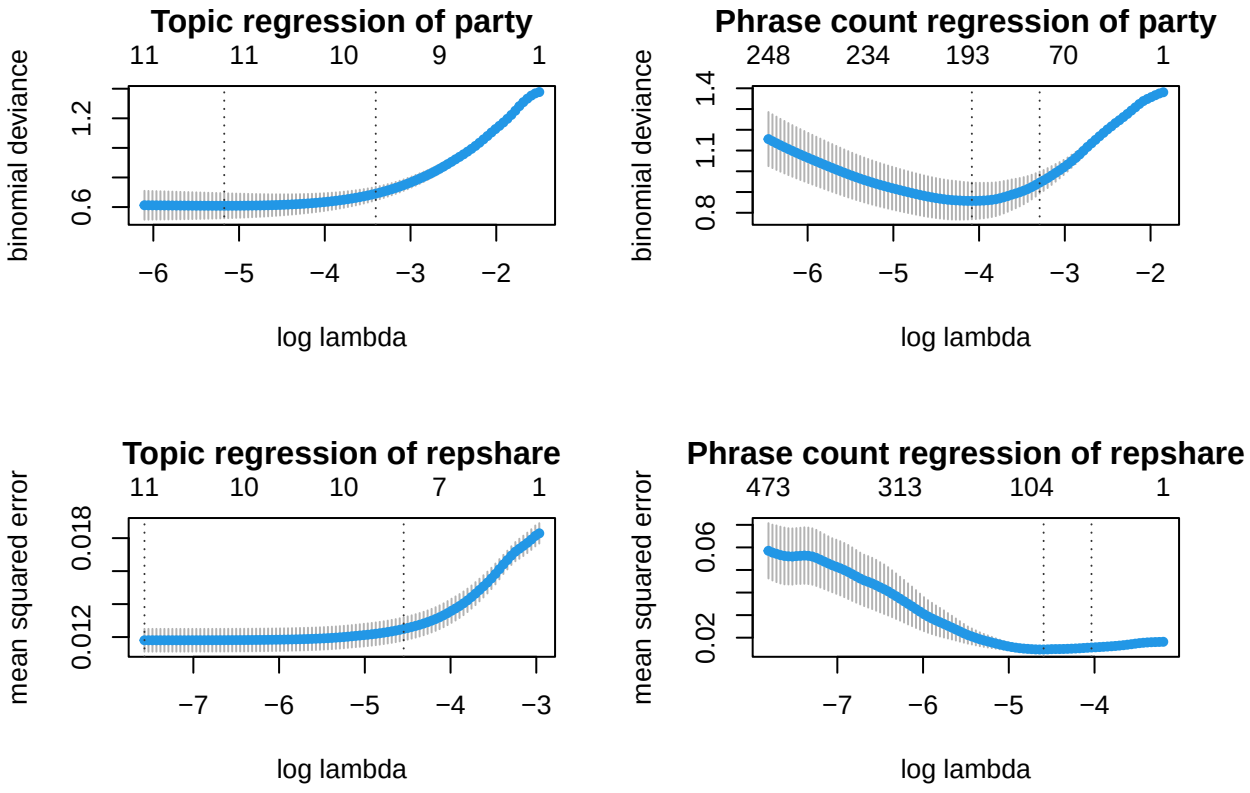
- Cluster 1: Although small in number, the balance between Democrats and Republicans might suggest a topic of common interest. However, due to the very small size of this cluster, it's hard to definitively categorize it as non-partisan without more context on the actual content.
- Cluster 4: This is the largest cluster and includes a significant representation from both Democrats and Republicans. While it is not evenly split, the presence of a large number of members from both parties suggests that the topic might address broad issues impacting many, or it might encapsulate a major policy area where both parties have strong interests even if their specific positions differ.

Partisan Clusters:

- Clusters 2, 3, and 5 show strong partisan leanings and are likely centered around issues that are particularly important to one party.

```
library(gamlr)

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

## give it the word %s as inputs
```

```
x <- 100*congress109Counts/rowSums(congress109Counts)

# 1: Party on topic weights
partytopics.cv <- cv.gamlr(tpcs$omega, gop, family="binomial")

# 2: Party on phrase percentage
partywords.cv <- cv.gamlr(x, gop, family="binomial")

# 3: Repshare on topic weights
reptopics.cv <- cv.gamlr(tpcs$omega, congress109Ideology[,"repshare"])

# 4: Repshare on phrase percentage
repwords.cv <- cv.gamlr(x, congress109Ideology[,"repshare"])

par(mfrow=c(2,2))
plot(partytopics.cv, main="Topic regression of party")
plot(partywords.cv, main="Phrase count regression of party")
plot(reptopics.cv, main="Topic regression of repshare")
plot(repwords.cv, main="Phrase count regression of repshare")
```




```
# Calculate and compare max OOS R2 to evaluate model performance

# Party regressions
max(1-partytopics.cv$cvm/partytopics.cv$cvm[1])
```

```
## [1] 0.5580512
```

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

```
## [1] 0.3792173
```

```r
# Repshare regressions
max(1-reptopics.cv$cvm/reptopics.cv$cvm[1])
```

```
## [1] 0.3545104
```

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

```
## [1] 0.1871234
```

- **Model Choice**: Topic-based models consistently outperform phrase count-based models in both logistic and linear regression settings according to the OOS R2 results. This suggests that the abstracted features captured by topics are more predictive of the outcomes than direct phrase counts, possibly due to better generalization and reduction of over-fitting through dimensional reduction.
- **Performance**: Topic regression not only explains more variance but also likely provides models that generalize better, as indicated by lower deviance and error metrics from the plots.
- **Application**: For practical applications in predicting political party alignment or share based on speech content, using topics as predictors seems to be more effective than using direct phrase counts.

In summary, topics derived from congressional speeches provide a more robust set of features for modeling purposes related to party affiliation and repshare than simply counting phrases. This could be due to topics capturing underlying thematic structures that are more informative and less sparse than individual phrase counts.