Chad Germany, Xinhao Liu
Stat 480: Big Data Analytics
November 18, 2023

Homework 5

We will use congress109 data in package textir. It counts for 1,000 phrases used by each of 529 members of the 109th US congress.

Question 1. Fit K-means to speech text for K in 5,10,15,20,25. Use an IC to choose the K and interpret the selected model.

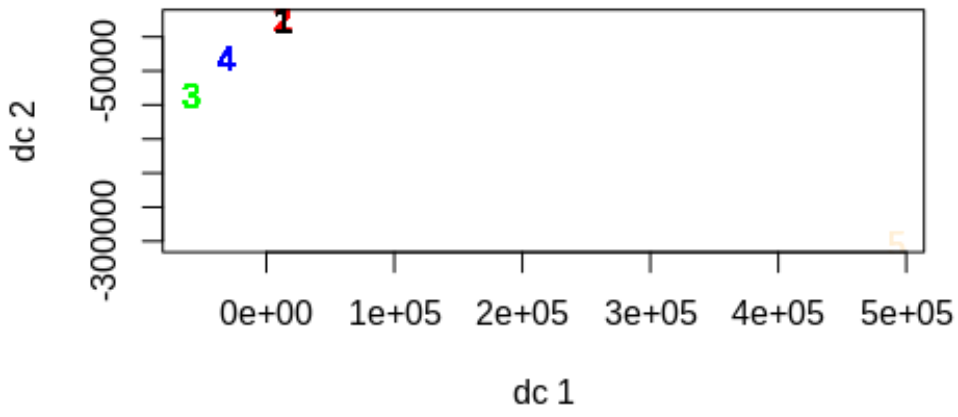

FIGURE 1. Graph of data in 5 different groups.

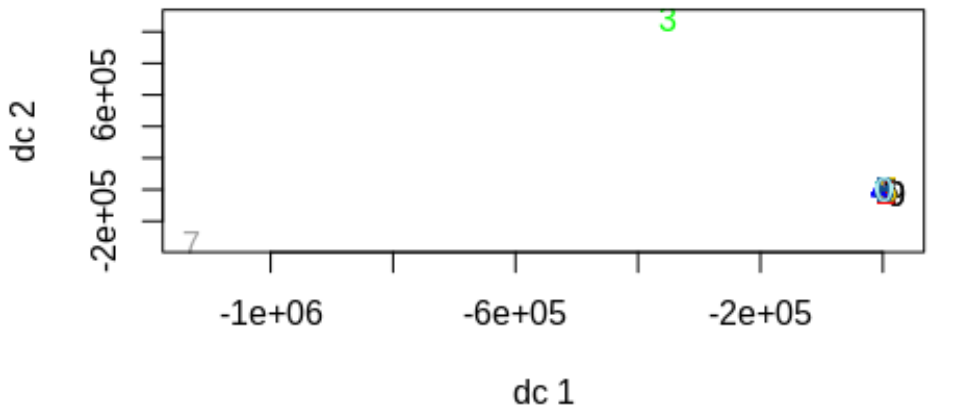

FIGURE 2. Graph of data in 10 different groups.

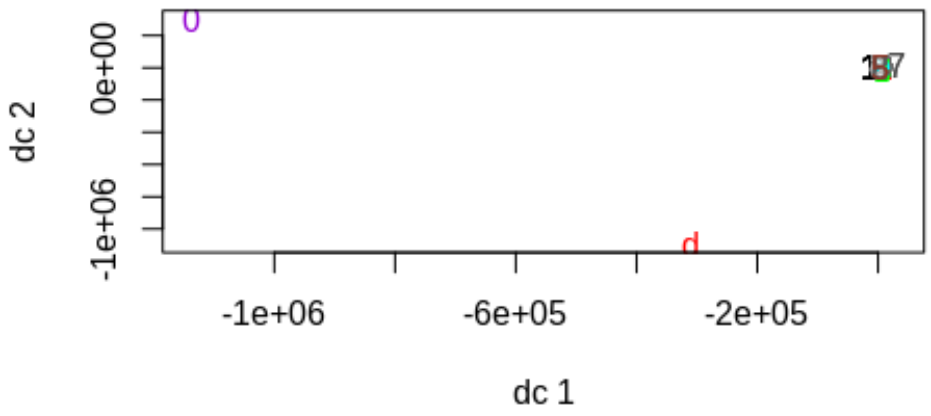

FIGURE 3. Graph of data in 15 different groups.

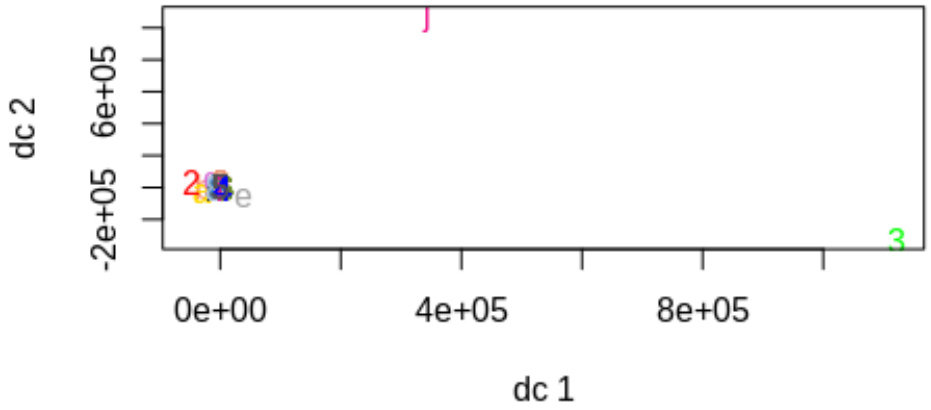

FIGURE 4. Graph of data in 20 different groups.

Question 2. Fit a topic model for the speech counts. Use Bayes factors to choose the number of topics, and interpret your chosen model.

If you supply a vector of topic sizes, it uses a Bayes factor to choose the correct K. (Bayes Factor is like exp(-BIC), so you choose the biggest BF). The bayes factor is largest for k=10 as shown in figure 7 so we choose 10 groups.

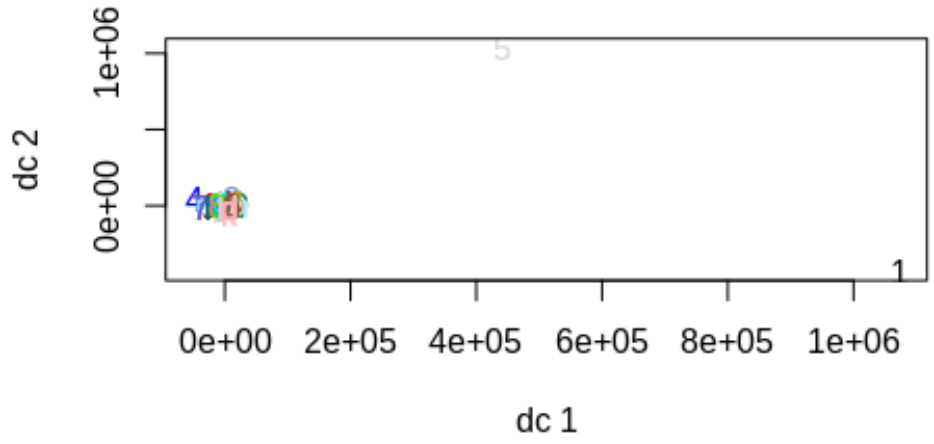

FIGURE 5. Graph of data in 25 different groups.

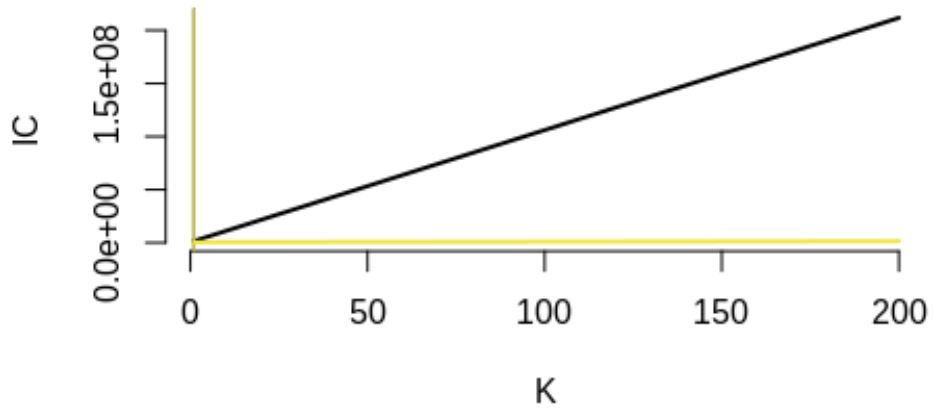

FIGURE 6. This IC chart shows the minimum BIC where the line is which is at K=1.

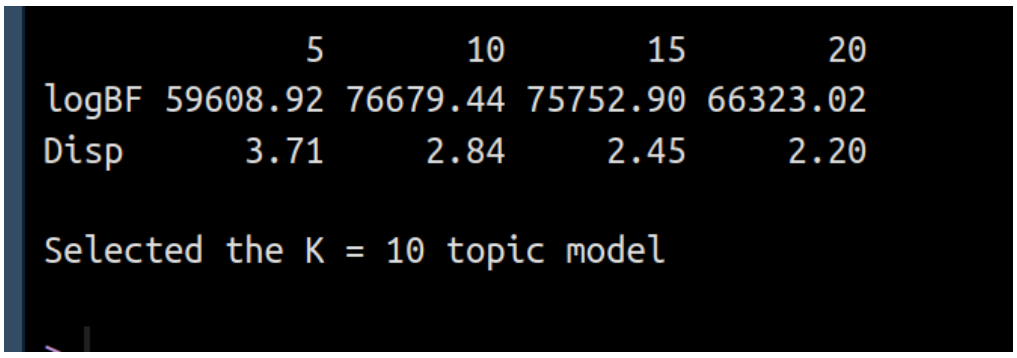

FIGURE 7. The bayes factor for different k values. The model has selected k=10

FIGURE 8. Word cloud 1 Rosa Parks is the largest word.

FIGURE 9. Word cloud 2 illegal immigration is the biggest probability .

```
#question2 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)


tpc <- topics(x,K=10)
```

FIGURE 10. Word cloud 3 tax relief has the largest probability.

FIGURE 11. Word cloud 4 nuclear weapon has the largest probability.

```
8
9
10 dim(tpc$theta)
11 colSums(tpc$theta)
12
13 dim(tpc$omega)
14 rowSums(tpc$omega)
15
16 ## choosing the number of topics
17 ## If you supply a vector of topic sizes, it uses a Bayes factor to choose
18 ## (BF is like exp(-BIC), so you choose the bigggest BF)
19 ## the algo stops if BF drops twice in a row
20
21 tpcs <- topics(x,K=5*(1:6), verb=10) # it chooses 10 topics
22 summary(tpcs)
```

FIGURE 12. Word cloud 5 most words are the same size .

FIGURE 13. Word cloud 6 food stamp looks like the most probable .

```
# summary prints the top `n' words for each topic,
# under ordering by `topic over aggregate' lift:
#    the topic word prob over marginal word prob.

summary(tpcs, n=10)

library(wordcloud)
par(mfrow=c(1,2))

wordcloud(row.names(tpcs$theta),
          freq=tpcs$theta[,5], min.freq=0.004, col="maroon")

wordcloud(row.names(tpcs$theta),
          freq=tpcs$theta[,2], min.freq=0.004, col="navy")

```

FIGURE 14. Word cloud 7 date majority vote is the most probable .

FIGURE 15. Word cloud 8 credit is the most probable.

```
38
39 #question 3  Connect the unsupervised clusters to partisanship.
40 # tabulate party membership by K-means cluster. Are there any non-partisan topics?
41 #  I fit topic regressions for each of party and repshare. Compare to regression onto
       phrase percentages:
42 #  x<-100*congress109Counts/rowSums(congress109Counts)
43
44 library(gamlr)
45
46
47
48
49 party_109 <- congress109Ideology[,"party"]
50 repshare_109 <- congress109Ideology[,"repshare"]
51 x <- 100*congress109Counts/rowSums(congress109Counts)
```

FIGURE 16. Word cloud 9 embryonic stem cells is the most probable.

FIGURE 17. Word cloud 10 trade deficit is the most probable.

```
52
53 party_109_R = congress109Ideology[,"party"][congress109Ideology[,"party"]=="R"]
54
55 party.cv <- cv.gamlr((tpcs$theta[1:length(party_109_R)]), factor(party_109_R))
56 party.cv <- cv.gamlr(drop(tpcs$theta), drop(party_109_R))
57
58
59 summary(party.cv)
60
61 o<-order(as.numeric(abs(coef(party.cv))),decreasing=T)
62 (coef(party.cv)[o[1:10],])
63
64
65 drop(coef(party.cv))*.1
66 party_per.cv <- cv.gamlr(x, party_109)
```

```
67 o<-order(as.numeric(abs(coef(party_per.cv))),decreasing=T)
```

LISTING 1. R source code implemented for the homework.

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

In figure 8 we see phases like black caucus and rosa parks at high probablity to I would associate this group with democrats as black people tend to vote more overwhelmingly for democrats. Looking at figure 9 we see that it has phrases that we would consider partisan. Republicans are more likely to say these phrases. One group that seems nonpartisan is in figure 12. This shows many words that are the same size and the phrases themselves seems related to money but not necessarily partisan. When we fit a topic regression onto each party (dem and republican). There was no relationship with democrat and republican but was a relationship with the share of constituency that voted for bush as shown in the figure 18.

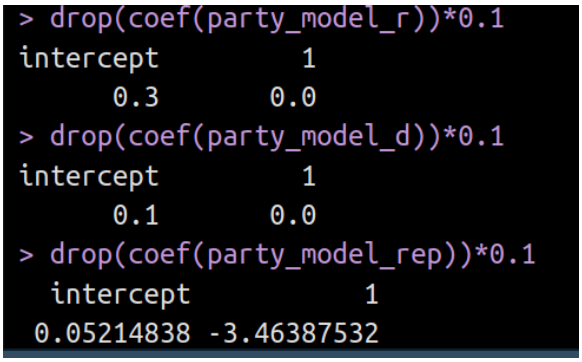

FIGURE 18. No relationship for dems and republicans but a relation for repshare.