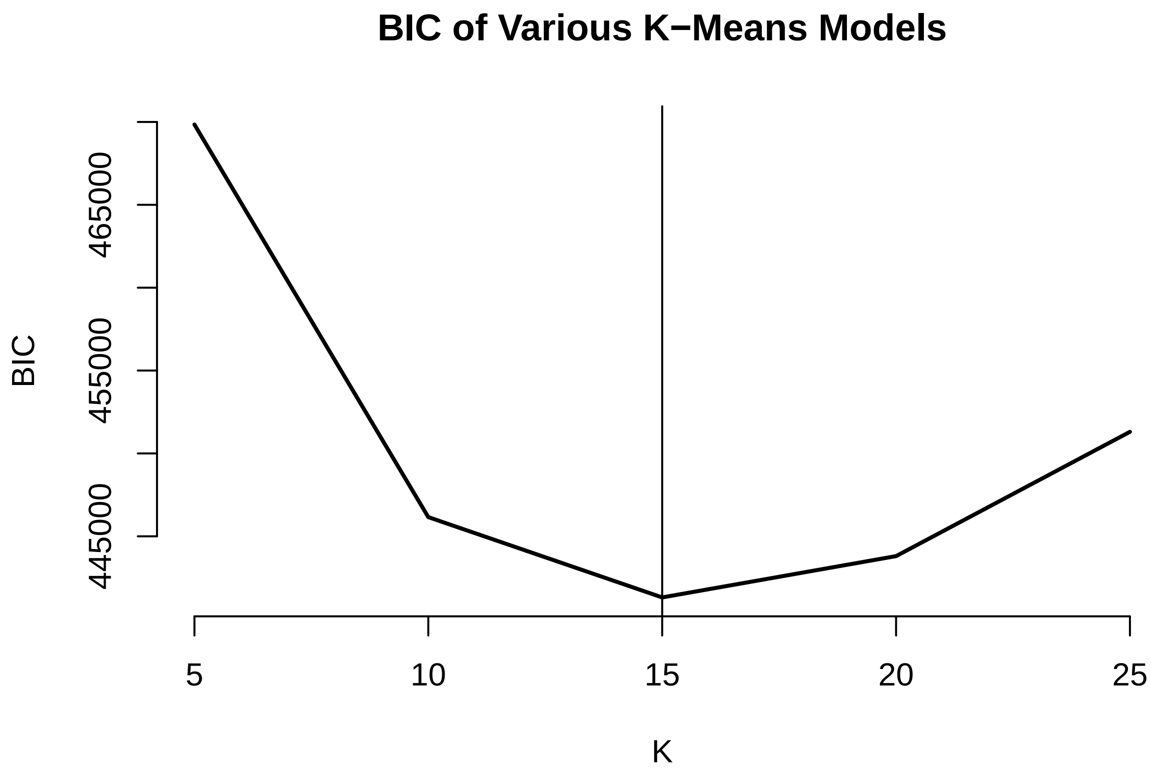
Congressional Speech Topics

(Student: Vinh Luong - 442069)

1. Unsupervised Topic Clustering

We fit a number of K-Means clustering models to the phrase counts data, with various K = 5, 10, 15, 20, 25:



From the above plot, we select K = 15 corresponding lowest BIC. Let's look at the 10 most common phrases in the 15 clusters:

```
##       1                             2
##  [1,] "fuel.efficiency"             "boy.girl.club"
##  [2,] "invasion.iraq"               "republican.senator"
##  [3,] "bankruptcy.court"            "check.bal"
##  [4,] "bush.white.house"            "senate.republican"
##  [5,] "medic.bil"                   "protect.minority.right"
##  [6,] "secretary.colin.powel"       "minority.right"
##  [7,] "president.bush.plan"         "politic.party"
##  [8,] "reduce.dependence.foreign" "confer.civil.right"
##  [9,] "social.security.president" "leadership.confer.civil"
## [10,] "president.cheney"            "system.check.bal"
##       3                           4
##  [1,] "northern.ireland"          "arab.oil.embargo"
##  [2,] "children.live.poverty"     "pluripotent.stem.cel"
##  [3,] "increase.minimum.wage"     "percent.growth.rate"
##  [4,] "basic.right"               "percent.growth"
##  [5,] "live.poverty"              "stem.cel.line"
##  [6,] "minority.women.owned"      "cel.line"
##  [7,] "minority.owned.business" "oil.production"
##  [8,] "tax.break.wealthy"       "oil.field"
##  [9,] "american.living.poverty" "produce.stem.cel"
## [10,] "public.college"          "stem.cel"
##       5                       6
##  [1,] "buy.american.product"  "international.labor.organization"
##  [2,] "world.poorest.people"  "american.force.radio"
##  [3,] "central.american.fre"  "voting.system"
##  [4,] "american.fre.trade"    "feder.election"
##  [5,] "central.american"      "people.disabiliti"
##  [6,] "trade.policy"          "family.medic.leave"
##  [7,] "world.poorest"         "child.labor"
##  [8,] "trade.agreement"       "food.stamp.program"
##  [9,] "president.bush.signed" "disease.control.prevention"
## [10,] "religiou.leader"       "election.reform"
##       7                           8
##  [1,] "cut.funding"               "committe.commerce.science"
##  [2,] "program.help"              "global.war"
##  [3,] "additional.funding"        "growth.job"
##  [4,] "middle.class.american"     "illegal.immigration"
##  [5,] "billion.dollar.tax"        "business.meeting"
##  [6,] "american.pacific.islander" "housing.urban.affair"
##  [7,] "tax.break"                 "urban.affair"
##  [8,] "asian.american.pacific"    "assistant.secretary"
##  [9,] "cut.program"               "banking.housing.urban"
## [10,] "budget.cut"                "feder.spending"
##       9                    10
##  [1,] "tax.cut.spending"   "violent.sexual.predator"
##  [2,] "additional.tax.cut" "sole.source.contract"
##  [3,] "record.deficit"     "integrated.oil.compani"
##  [4,] "cut.spending"       "democrat.white.house"
##  [5,] "spending.cut"       "tax.haven"
##  [6,] "war.cost"           "warren.buffett"
##  [7,] "offset.tax.cut"     "home.heating.fuel"
##  [8,] "cost.billion"       "corporation.public.broadcasting"
##  [9,] "president.tax.cut"  "public.television"
## [10,] "president.budget"   "public.broadcasting"
##       11                          12
##  [1,] "summa.cum.laude"           "final.minute"
##  [2,] "justice.priscilla.owen"    "legislative.session"
##  [3,] "supreme.court.united"      "date.time"
##  [4,] "roe.wade"                  "democratic.leader"
##  [5,] "president.judicial.nomine" "president.business"
##  [6,] "justice.janice.roger"      "third.time"
##  [7,] "law.review"                "busy.week"
##  [8,] "abortion.clinic"           "democrat.leader"
##  [9,] "justice.supreme.court"     "vote.judicial.nomine"
## [10,] "majority.vote"             "president.pro.tempore"
##       13                        14
##  [1,] "military.operation.iraq" "outing.cia.agent"
##  [2,] "military.operation"      "cia.agent"
##  [3,] "iraq.president"          "social.security.plan"
##  [4,] "committe.hold.hearing"   "private.account"
##  [5,] "troop.iraq"              "start.talking"
##  [6,] "iraq.policy"             "issue.facing.american"
##  [7,] "war.iraq"                "privatization.social.security"
##  [8,] "iraq.war"                "security.plan"
##  [9,] "bunker.buster"           "prescription.drug.bil"
## [10,] "cost.war"                "drug.bil"
##       15
##  [1,] "speaker.table"
##  [2,] "malpractice.insurance.rate"
##  [3,] "national.homeownership.month"
##  [4,] "ending.september"
##  [5,] "columbia.river.gorge"
##  [6,] "united.postal.service"
##  [7,] "national.heritage.corridor"
##  [8,] "commonly.prescribed.drug"
##  [9,] "wild.bird"
## [10,] "able.buy.gun"
```

From the above, we can see a few key topics emerging from the congressional speeches:

- General Law/Policy-Making on energy, heritage preservation, drug, etc.
- Minorities, Poverty and Wealth Gap
- The Iraq War
- Economy, Business, Growth, Jobs
- Tax, Government Spending and Budget Decifit
- Law and Justice
- Energy and Scientific Research
- Foreign Affairs, Defense and Immigration
- Foreign Trade

2. Topic Model

We now fit a number of topic models over the phrase counts, trying K = 5, 10, 15, 20, 25 topics:

The selected model by Bayes factor corresponds to K = 15 topics, with the following top-10 phrases:

```
##
## Top 10 phrases by topic-over-null term lift (and usage %):
##
## [1] 'commonly.prescribed.drug', 'southeast.texa', 'million.illegal.alien', 'amnesty.illegal.alien', 'postal
## [2] 'reverse.robin.hood', 'passenger.rail.service', 'rail.service', 'chemic.plant.security', 'passenger.rai
## [3] 'near.retirement.age', 'gifted.talented.student', 'tax.relief.package', 'increase.taxe', 'personal.reti
## [4] 'near.earth.object', 'winning.war.iraq', 'oil.food.program', 'oil.food', 'troop.bring.home', 'bring.tro
## [5] 'united.airline.employe', 'private.account', 'security.private.account', 'outing.cia.agent', 'record.bu
## [6] 'asian.pacific.american', 'american.heritage.month', 'pacific.american.heritage', 'asian.pacific', 'lit
## [7] 'republic.cypru', 'senate.committe.business', 'driver.education', 'national.flood.insurance', 'flood.in
## [8] 'low.cost.reliable', 'ready.mixed.concrete', 'price.natural.ga', 'witness.testify', 'suppli.natural.ga'
## [9] 'judicial.confirmation.process', 'judge.alberto.gonzale', 'vote.judicial.nomine', 'justice.janice.roger
## [10] 'north.american.fre', 'financial.accounting.standard', 'american.fre.trade', 'central.american.fre', '
## [11] 'va.health.care', 'global.gag.rule', 'disabled.american.veteran', 'health.care.budget', 'gag.rule', 'f
## [12] 'pluripotent.stem.cel', 'national.ad.campaign', 'cel.stem.cel', 'embryonic.stem', 'embryonic.stem.cel'
## [13] 'change.heart.mind', 'hate.crime.legislation', 'wild.bird', 'hate.crime.law', 'drilling.arctic.nationa
## [14] 'able.buy.gun', 'deep.sea.coral', 'buy.gun', 'credit.card.industry', 'caliber.sniper.rifle', 'assault.
## [15] 'violence.sexual.assault', 'domestic.violence.sexual', 'indian.art.craft', 'victim.domestic', 'victim.
##
## Log Bayes factor and estimated dispersion, by number of topics:
##
##              5       10       15       20       25
## logBF 56333.75 75922.34 76812.63 67187.48 53329.73
## Disp      3.67     2.89     2.49     2.23     2.04
##
## Selected the K = 15 topic model
```

Topics that emerge are similar to those discovered by K-Means:

- General Law-/Policy-Making

- Immigration Reform
- Gun Control
- Foreign Trade
- Science
- Law and Justice

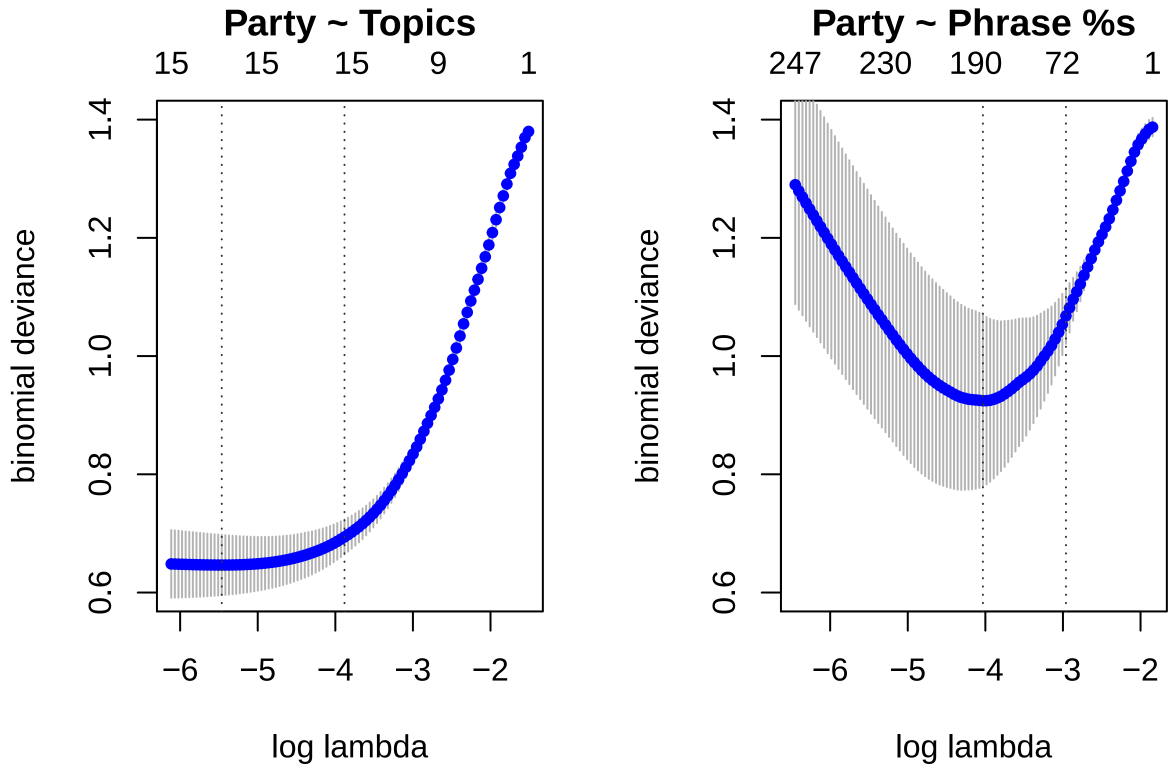
3. Relationship between Topics and Partisanships

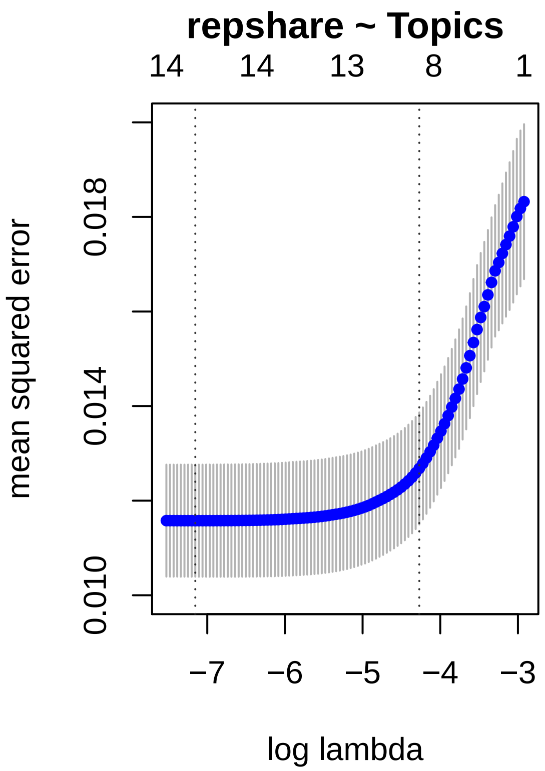
Let's now take a look at the topics by partisanship:

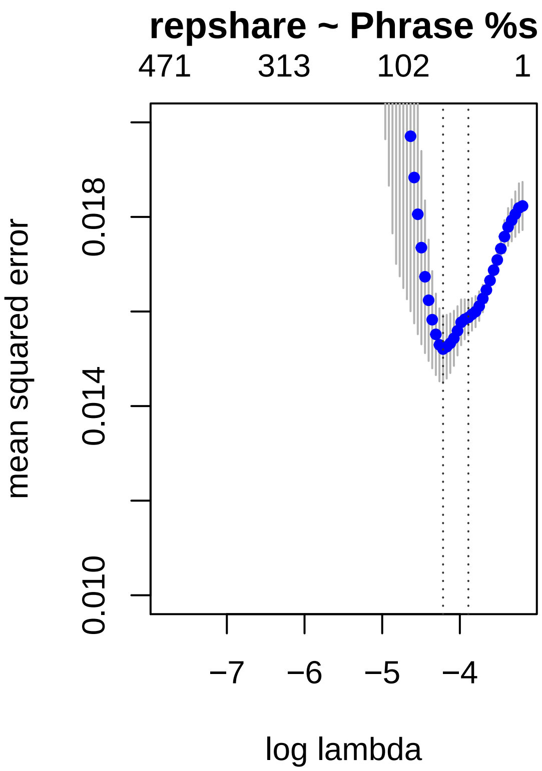
```
##
##        1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
##    D   1   1   1   0   1   2  40   0   2   1   0   0   7   2 184
##    R   0   0   0   1   0   0   0  34   0   0   6   1   0   0 243
```

From this table, we can see that Democrats dominate the topic government spending on various welfare programs (supposedly to boost those programs), whereas Republicans dominate the topic of economy / business / jobs. There is one topic which both major parties talked about, which seems to be related to general law-making.

Next, we can try regression partisanships and *repshare* onto the topic weights ω from the topic model, and then compare these models with regressions of partisanships and *repshare* onto the percentage of phrases used by the representatives:







We can see that both topic regressions perform better in terms of goodness of fit than the regressions on the relative use of individual word phrases. Hence summarizing the 1,000 word phrases into a small collection of topics is a good dimensionality reduction exercise.