# Climate Gentrification: Topic Modeling

Alex Clippinger, Halina Do-Linh, Desik Somasundaram, Alex Vand

## Assignment:

Use the data you plan to use for your final project:

**Prepare the data so that it can be analyzed in the topicmodels package**

**Run three more models and select the overall best value for k (the number of topics) - include some justification for your selection: theory, FindTopicsNumber() optimization metrics, interpretability, LDAvis**

### Import Nexis Uni Data

Our data for the final project uses search results from Nexis Uni for the search term "climate gentrification". This table shows the distribution of the type of results.

**Theory: Climate gentrification is a relatively new topic and may not have many subtopics because it is both new and already very specific. As a matter of fact, we believe it's likely to be considered a subtopic under environmental justice. Nonetheless, we would like to explore the the topic modeling related to our data to see any patterns that may emerge.**

```
my_files <- list.files(pattern = ".docx", path = here("data"),
                       full.names = TRUE, recursive = TRUE, ignore.case = TRUE)

cg_data <- lnt_read(my_files) #Object of class 'LNT output'
```

```
## Creating LNToutput from 8 files...

##  ...files loaded [0.81 secs]

##  ...articles split [1.05 secs]

##  ...lengths extracted [1.06 secs]

##  ...headlines extracted [1.06 secs]

##  ...newspapers extracted [1.06 secs]

##  ...dates extracted [1.12 secs]

##  ...authors extracted [1.13 secs]

##  ...sections extracted [1.13 secs]
```

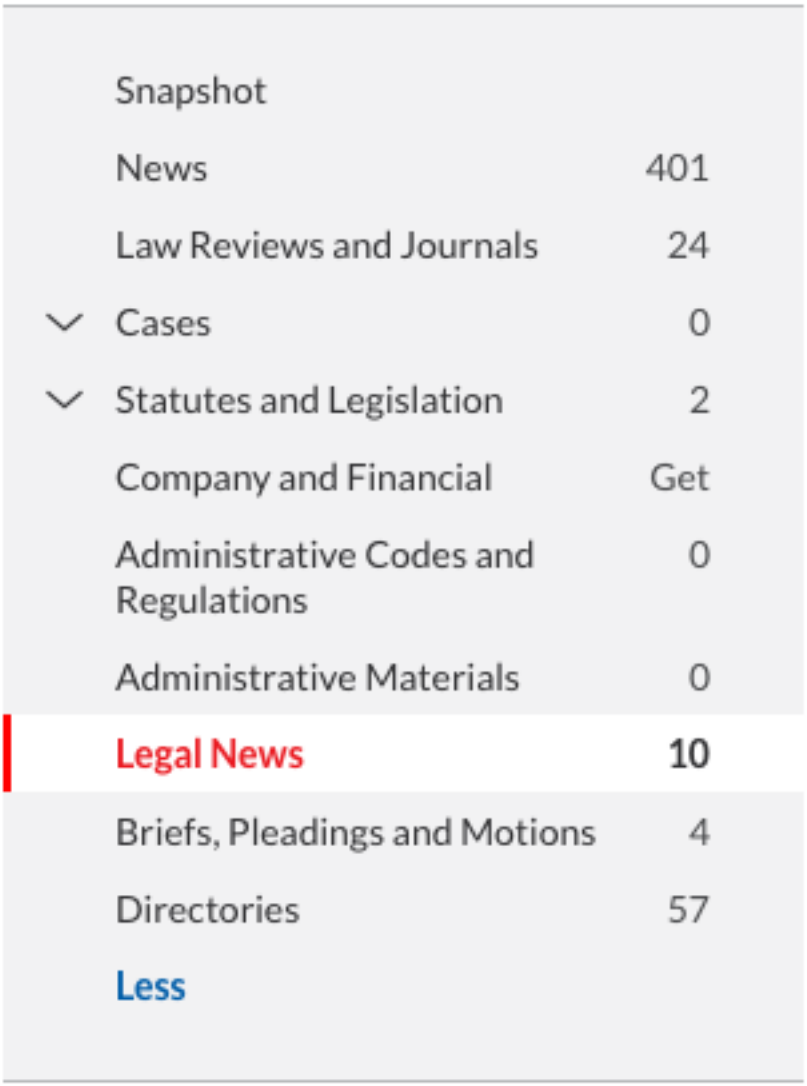



Figure 1: Nexis Uni Results

```
##  ...editions extracted [1.13 secs]

##  ...dates converted [1.14 secs]

##  ...metadata extracted [1.14 secs]

##  ...article texts extracted [1.14 secs]

##  ...superfluous whitespace removed [1.29 secs]

## Elapsed time: 1.29 secs
```

```
cg_meta_df <- cg_data@meta
cg_articles_df <- cg_data@articles
cg_paragraphs_df <- cg_data@paragraphs

cg_data2<- data_frame(element_id = seq(1:length(cg_meta_df$Headline)),
                          Date = cg_meta_df$Date,
                          Headline = cg_meta_df$Headline)
```

## Clean the corpus

```
cg_corp <- corpus(x = cg_articles_df, text_field = "Article")
```

```
cg_corp.stats <- summary(cg_corp)
head(cg_corp.stats, n = 25)
```

```
##      Text Types Tokens Sentences ID
## 1   text1   235    413        14  1
## 2   text2   429    957        24  2
## 3   text3   429    957        24  3
## 4   text4   430    963        24  4
## 5   text5   430    963        24  5
## 6   text6   430    963        24  6
## 7   text7   551   1344        43  7
## 8   text8   562   1110        50  8
## 9   text9   562   1110        50  9
## 10 text10   562   1110        50 10
## 11 text11   319    608        23 11
## 12 text12   871   2187        94 12
## 13 text13   296    608        27 13
## 14 text14   622   1541        59 14
## 15 text15   587   1654        67 15
## 16 text16  1009   2308        65 16
## 17 text17  1022   2345        65 17
## 18 text18   294    544        27 18
## 19 text19   296    546        27 19
## 20 text20   361    776        31 20
## 21 text21   294    544        27 21
```

```
## 22 text22   672   1680        67 22
## 23 text23   736   1871        82 23
## 24 text24  1146   2684        87 24
## 25 text25  1364   3831       152 25
```

```
toks <- tokens(cg_corp, remove_punct = TRUE, remove_numbers = TRUE)
#I added some project-specific stop words here
add_stops <- c(stopwords("en"), "like", "just", "say", "year")
toks1 <- tokens_select(toks, pattern = add_stops, selection = "remove")
```

## Convert to a document-feature matrix

```
dfm_comm<- dfm(toks1, tolower = TRUE)
dfm <- dfm_wordstem(dfm_comm)
dfm <- dfm_trim(dfm, min_docfreq = 2) #remove terms only appearing in one doc (min_termfreq = 10)

print(head(dfm))
```

```
## Document-feature matrix of: 6 documents, 13,104 features (97.91% sparse) and 1 docvar.
##        features
## docs    new york kansa citi miami denver mantra locat alway relev
##   text1   2    2     1    8     2      2      1     3     1     1
##   text2   1    0     0    9     9      0      0     0     0     0
##   text3   1    0     0    9     9      0      0     0     0     0
##   text4   1    0     0    9     9      0      0     0     0     0
##   text5   1    0     0    9     9      0      0     0     0     0
##   text6   1    0     0    9     9      0      0     0     0     0
## [ reached max_nfeat ... 13,094 more features ]
```

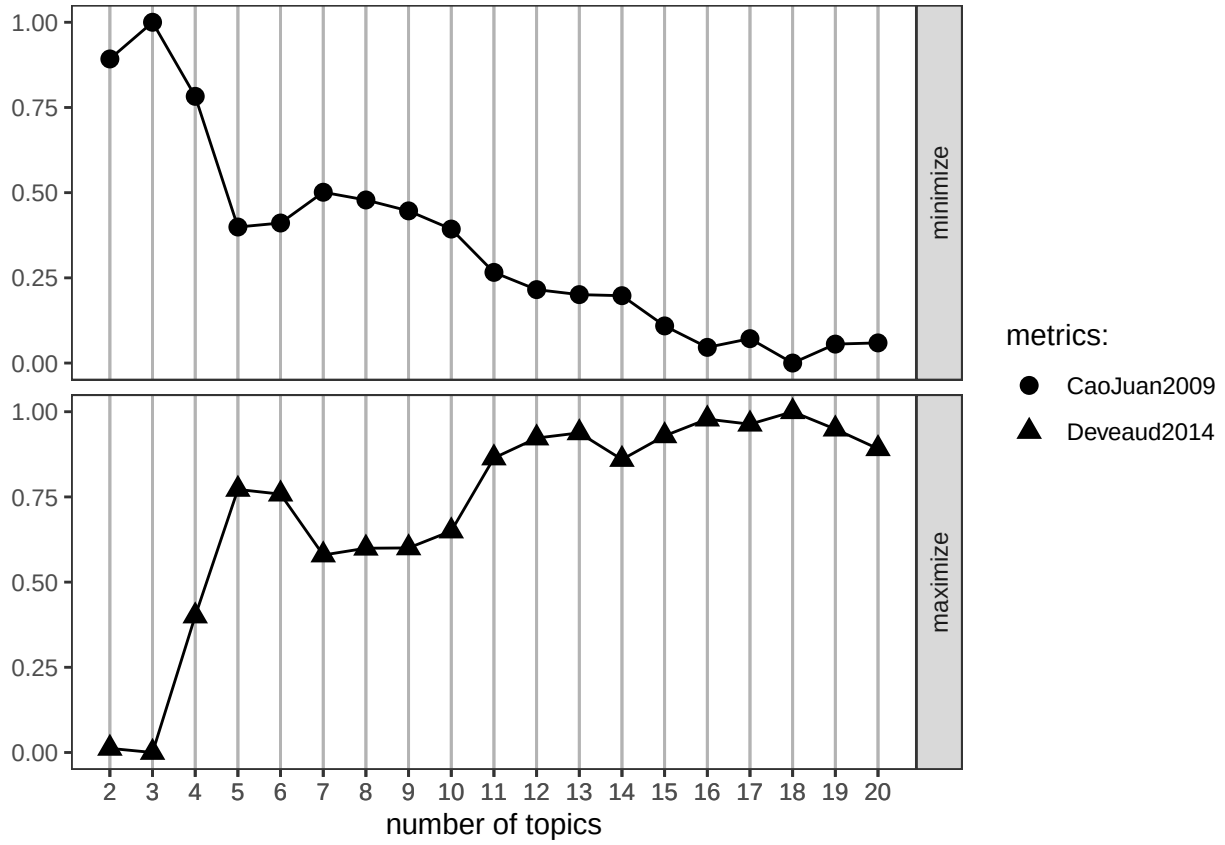
```
#remove rows (docs) with all zeros
sel_idx <- slam::row_sums(dfm) > 0
dfm <- dfm[sel_idx, ]
```

### Optimization for k

```
result <- FindTopicsNumber(
  dfm,
  topics = seq(from = 2, to = 20, by = 1),
  metrics = c("CaoJuan2009",  "Deveaud2014"),
  method = "Gibbs",
  control = list(seed = 77),
  verbose = TRUE
)
```

```
## fit models... done.
## calculate metrics:
##   CaoJuan2009... done.
##   Deveaud2014... done.
```

```
FindTopicsNumber_plot(result)
```




**FindTopicsNumber: 4, 7, 12** k=5: 75%/30% k=7: 55%/50% k=12: 90%/25%

**We ran 3 models based on the number of topics provided by the optimization metrics. We think that k=5, k=7 and k=12 are good values to test for the number of topics according to the results from the CauJuan2009 and Devaud2014 metrics. In this case, we do recognize that k=18 may also seem like a good number to test but we opted for k=5 instead because of our prior knowledge that climate gentrification does not have that many subtopics.**

**Topic models for k=5, k=7 and k=12**

```
k <- 5

topicModel_k5 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = 25))
```

```
## K = 5; V = 13104; M = 440
## Sampling 500 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
```

```
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

```
#nTerms(dfm_comm)

tmResult_5 <- posterior(topicModel_k5)
attributes(tmResult_5)
```

```
## $names
## [1] "terms"  "topics"
```

```
#nTerms(dfm_comm)
beta_5 <- tmResult_5$terms   # get beta from results
dim(beta_5)                # K distributions over nTerms(DTM) terms# lengthOfVocab
```

```
## [1]     5 13104
```

```
terms(topicModel_k5, 10)
```

```
##       Topic 1     Topic 2     Topic 3 Topic 4   Topic 5
##  [1,] "climat"    "communiti" "said"  "florida" "climat"
##  [2,] "energi"    "hous"      "miami" "work"    "chang"
##  [3,] "program"   "develop"   "peopl" "de"      "flood"
##  [4,] "feder"     "use"       "hous"  "miami"   "will"
##  [5,] "act"       "resid"     "citi"  "black"   "risk"
##  [6,] "communiti" "green"     "home"  "art"     "area"
##  [7,] "nation"    "govern"    "go"    "said"    "rise"
##  [8,] "fund"      "plan"      "will"  "trump"   "properti"
##  [9,] "congress"  "can"       "rise"  "will"    "citi"
## [10,] "build"     "citi"      "get"   "new"     "disast"
```

```
k <- 7

topicModel_k7 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = 25))
```

```
## K = 7; V = 13104; M = 440
## Sampling 500 iterations!
## Iteration 25 ...
```

```
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

```
#nTerms(dfm_comm)

tmResult_7 <- posterior(topicModel_k7)
attributes(tmResult_7)
```

```
## $names
## [1] "terms"  "topics"
```

```
#nTerms(dfm_comm)
beta_7 <- tmResult_7$terms   # get beta from results
dim(beta_7)                # K distributions over nTerms(DTM) terms# lengthOfVocab
```

```
## [1]     7 13104
```

```
terms(topicModel_k7, 10)
```

```
##       Topic 1     Topic 2        Topic 3      Topic 4   Topic 5   Topic 6
##  [1,] "energi"    "communiti"    "climat"     "miami"   "peopl"   "work"
##  [2,] "climat"    "hous"         "chang"      "said"    "know"    "de"
##  [3,] "program"   "develop"      "risk"       "flood"   "go"      "florida"
##  [4,] "feder"     "citi"         "citi"       "citi"    "one"     "miami"
##  [5,] "act"       "resid"        "adapt"      "rise"    "get"     "said"
##  [6,] "congress"  "green"        "heat"       "climat"  "now"     "art"
##  [7,] "communiti" "urban"        "will"       "sea"     "think"   "black"
##  [8,] "nation"    "plan"         "die"        "home"    "can"     "via"
##  [9,] "state"     "neighborhood" "impact"     "florida" "will"    "$"
## [10,] "fund"      "social"       "temperatur" "hous"    "percent" "trump"
##       Topic 7
##  [1,] "climat"
##  [2,] "will"
##  [3,] "properti"
```

```
##  [4,] "chang"
##  [5,] "disast"
##  [6,] "govern"
##  [7,] "state"
##  [8,] "cost"
##  [9,] "can"
## [10,] "may"
```

```
k <- 12

topicModel_k12 <- LDA(dfm, 12, method="Gibbs", control=list(iter = 500, verbose = 25))
```

```
## K = 12; V = 13104; M = 440
## Sampling 500 iterations!
## Iteration 25 ...
## Iteration 50 ...
## Iteration 75 ...
## Iteration 100 ...
## Iteration 125 ...
## Iteration 150 ...
## Iteration 175 ...
## Iteration 200 ...
## Iteration 225 ...
## Iteration 250 ...
## Iteration 275 ...
## Iteration 300 ...
## Iteration 325 ...
## Iteration 350 ...
## Iteration 375 ...
## Iteration 400 ...
## Iteration 425 ...
## Iteration 450 ...
## Iteration 475 ...
## Iteration 500 ...
## Gibbs sampling completed!
```

```
tmResult_12 <- posterior(topicModel_k12)
terms(topicModel_k12, 10)
```

```
##       Topic 1     Topic 2   Topic 3  Topic 4       Topic 5    Topic 6
##  [1,] "hous"      "work"    "climat" "climat"      "hous"     "miami"
##  [2,] "know"      "florida" "chang"  "chang"       "flood"    "citi"
##  [3,] "peopl"     "said"    "citi"   "will"        "home"     "rise"
##  [4,] "go"        "art"     "risk"   "peopl"       "said"     "sea"
##  [5,] "can"       "black"   "heat"   "migrat"      "afford"   "florida"
##  [6,] "communiti" "via"     "adapt"  "world"       "peopl"    "said"
##  [7,] "fair"      "artist"  "also"   "global"      "new"      "level"
##  [8,] "one"       "$"       "level"  "environment" "villag"   "littl"
##  [9,] "think"     "trump"   "rise"   "water"       "properti" "elev"
## [10,] "get"       "will"    "fund"   "futur"       "risk"     "haiti"
##       Topic 7   Topic 8 Topic 9     Topic 10    Topic 11        Topic 12
##  [1,] "percent" "de"    "energi"    "develop"   "communiti"     "disast"
##  [2,] "$"       "die"   "climat"    "communiti" "green"         "cost"
```

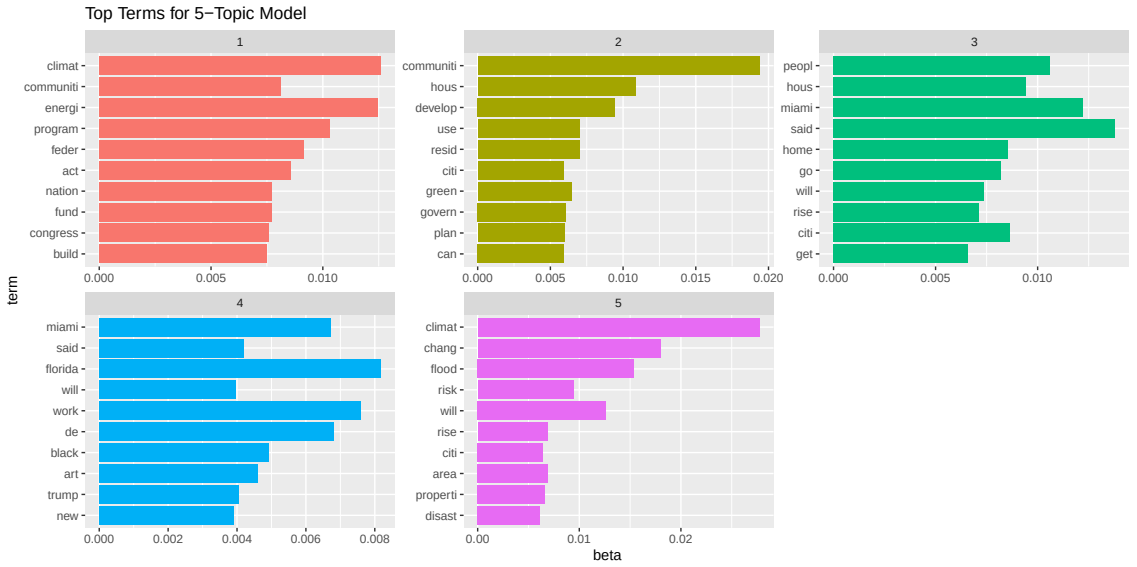
```
##  [3,] "report"  "la"    "program"   "properti"  "urban"         "flood"
##  [4,] "busi"    "miami" "act"       "law"       "social"        "properti"
##  [5,] "will"    "der"   "congress"  "land"      "environment"   "polici"
##  [6,] "new"     "en"    "feder"     "use"       "health"        "market"
##  [7,] "price"   "des"   "build"     "plan"      "gentrif"       "adjust"
##  [8,] "say"     "und"   "communiti" "right"     "neighborhood"  "failur"
##  [9,] "compani" "den"   "nation"    "govern"    "resid"         "communiti"
## [10,] "peopl"   "van"   "state"     "local"     "infrastructur" "feder"
```

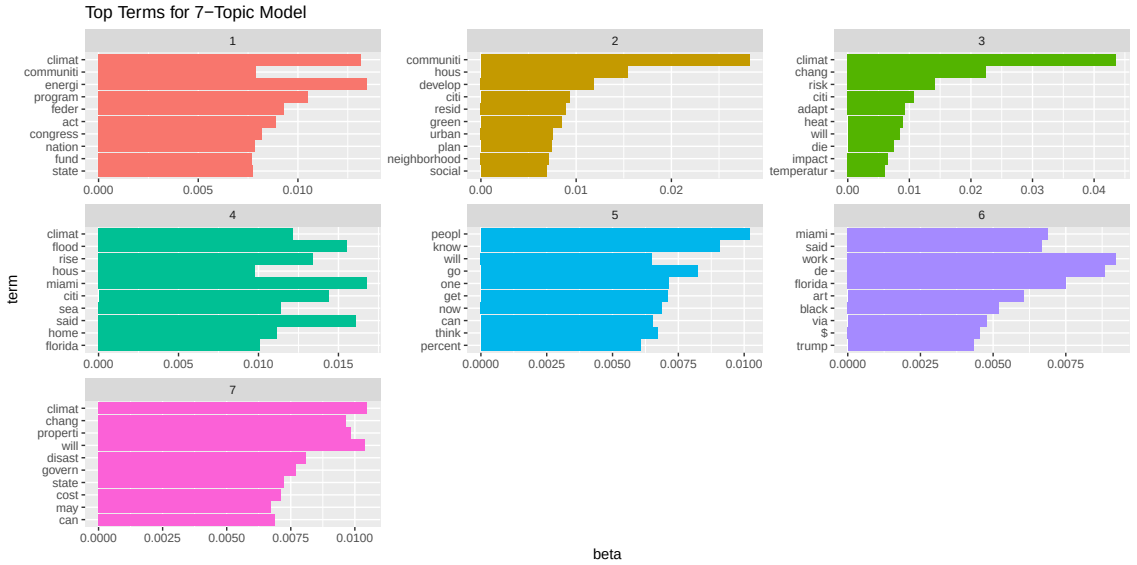
```
theta_12 <- tmResult_12$topics
beta_12 <- tmResult_12$terms
vocab <- (colnames(beta_12))
```

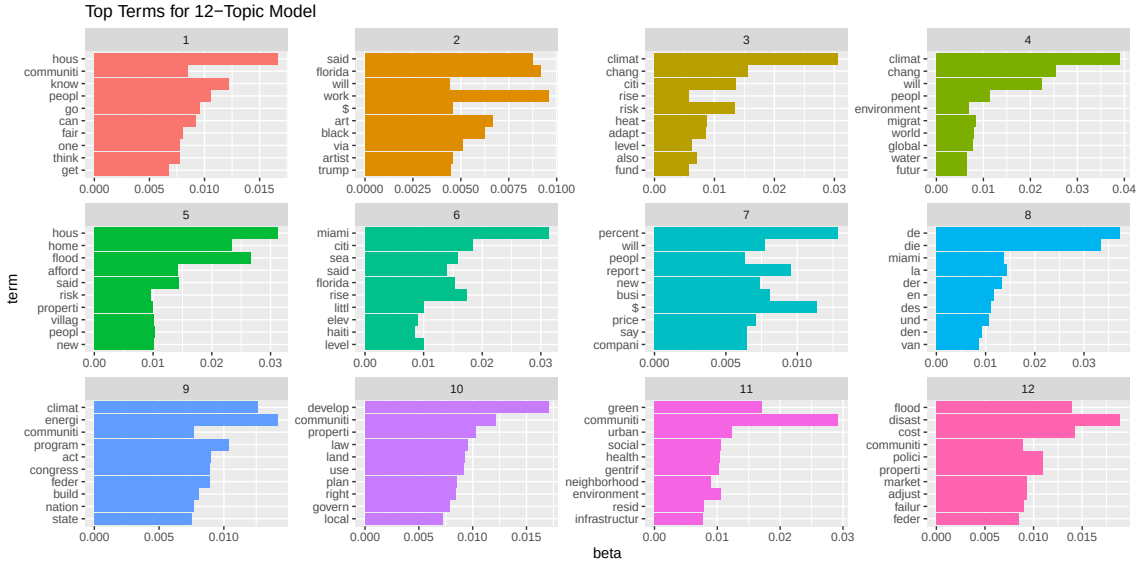
```
comment_topics_5 <- tidy(topicModel_k5, matrix = "beta")

comment_topics_7 <- tidy(topicModel_k7, matrix = "beta")


comment_topics_12 <- tidy(topicModel_k12, matrix = "beta")

top_terms_5 <- comment_topics_5 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)


top_terms_7 <- comment_topics_7 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)


top_terms_12 <- comment_topics_12 %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)
```

```
top_terms_5_plot <- top_terms_5 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  labs(title="Top Terms for 5-Topic Model")

top_terms_7_plot <- top_terms_7 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
```

```
  coord_flip() +
  labs(title="Top Terms for 7-Topic Model")


top_terms_12_plot <- top_terms_12 %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip()+
  labs(title="Top Terms for 12-Topic Model")


top_terms_5_plot / top_terms_7_plot / top_terms_12_plot
```


Top Terms for 5−Topic Model
1
climat
communiti
energi
program
feder
act
nation
fund
congress
build
0.000
0.005
0.010
2
communiti
hous
develop
use
resid
citi
green
govern
plan
can
0.000
0.005
0.010
0.015
0.020
3
peopl
hous
miami
said
home
go
will
rise
citi
get
0.000
0.005
0.010
4
miami
said
florida
will
work
de
black
art
trump
new
0.000
0.002
0.004
0.006
0.008
5
climat
chang
flood
risk
will
rise
citi
area
properti
disast
0.00
0.01
0.02
term
beta


Top Terms for 7−Topic Model
1
climat
communiti
energi
program
feder
act
congress
nation
fund
state
0.000
0.005
0.010
2
communiti
hous
develop
citi
resid
green
urban
plan
neighborhood
social
0.00
0.01
0.02
3
climat
chang
risk
citi
adapt
heat
will
die
impact
temperatur
0.00
0.01
0.02
0.03
0.04
4
climat
flood
rise
hous
miami
citi
sea
said
home
florida
0.000
0.005
0.010
0.015
5
peopl
know
will
go
one
get
now
can
think
percent
0.0000
0.0025
0.0050
0.0075
0.0100
6
miami
said
work
de
florida
art
black
via
$
trump
0.0000
0.0025
0.0050
0.0075
7
climat
chang
properti
will
disast
govern
state
cost
may
can
0.0000
0.0025
0.0050
0.0075
0.0100
term
beta


Top Terms for 12−Topic Model
1
hous
communiti
know
peopl
go
can
fair
one
think
get
0.000
0.005
0.010
0.015
2
said
florida
will
work
$
art
black
via
artist
trump
0.0000
0.0025
0.0050
0.0075
0.0100
3
climat
chang
citi
rise
risk
heat
adapt
level
also
fund
0.00
0.01
0.02
0.03
4
climat
chang
will
peopl
environment
migrat
world
global
water
futur
0.00
0.01
0.02
0.03
0.04
5
hous
home
flood
afford
said
risk
properti
villag
peopl
new
0.00
0.01
0.02
0.03
6
miami
citi
sea
said
florida
rise
littl
elev
haiti
level
0.00
0.01
0.02
0.03
7
percent
will
peopl
report
new
busi
$
price
say
compani
0.000
0.005
0.010
8
de
die
miami
la
der
en
des
und
den
van
0.00
0.01
0.02
0.03
9
climat
energi
communiti
program
act
congress
feder
build
nation
state
0.000
0.005
0.010
10
develop
communiti
properti
law
land
use
plan
right
govern
local
0.000
0.005
0.010
0.015
11
green
communiti
urban
social
health
gentrif
neighborhood
environment
resid
infrastructur
0.00
0.01
0.02
0.03
12
flood
disast
cost
communiti
polici
properti
market
adjust
failur
feder
0.000
0.005
0.010
0.015
term
beta

```
top5termsPerTopic_5 <- terms(topicModel_k5, 5)
topicNames_5 <- apply(top5termsPerTopic_5, 2, paste, collapse=" ")
topicNames_5
```

```
##                             Topic 1                             Topic 2
##  "climat energi program feder act" "communiti hous develop use resid"
##                             Topic 3                             Topic 4
##       "said miami peopl hous citi"      "florida work de miami black"
##                             Topic 5
##     "climat chang flood will risk"
```

```
top5termsPerTopic_7 <- terms(topicModel_k7, 5)
topicNames_7 <- apply(top5termsPerTopic_7, 2, paste, collapse=" ")
topicNames_7
```

```
##                              Topic 1                              Topic 2
##   "energi climat program feder act" "communiti hous develop citi resid"
##                              Topic 3                              Topic 4
##      "climat chang risk citi adapt"        "miami said flood citi rise"
##                              Topic 5                              Topic 6
##             "peopl know go one get"        "work de florida miami said"
##                              Topic 7
## "climat will properti chang disast"
```

```
top5termsPerTopic_12 <- terms(topicModel_k12, 5)
topicNames_12 <- apply(top5termsPerTopic_12, 2, paste, collapse=" ")
topicNames_12
```

```
##                                    Topic 1
##                   "hous know peopl go can"
##                                    Topic 2
##              "work florida said art black"
##                                    Topic 3
##              "climat chang citi risk heat"
##                                    Topic 4
##           "climat chang will peopl migrat"
##                                    Topic 5
##              "hous flood home said afford"
##                                    Topic 6
##              "miami citi rise sea florida"
##                                    Topic 7
##               "percent $ report busi will"
##                                    Topic 8
##                      "de die la miami der"
##                                    Topic 9
##       "energi climat program act congress"
##                                   Topic 10
##      "develop communiti properti law land"
##                                   Topic 11
## "communiti green urban social environment"
##                                   Topic 12
##        "disast cost flood properti polici"
```

```
library(LDAvis)
library("tsne")
svd_tsne <- function(x) tsne(svd(x)$u)
json <- createJSON(
  phi = tmResult_5$terms,
  theta = tmResult_5$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="", ylab="")
)
serVis(json)
```

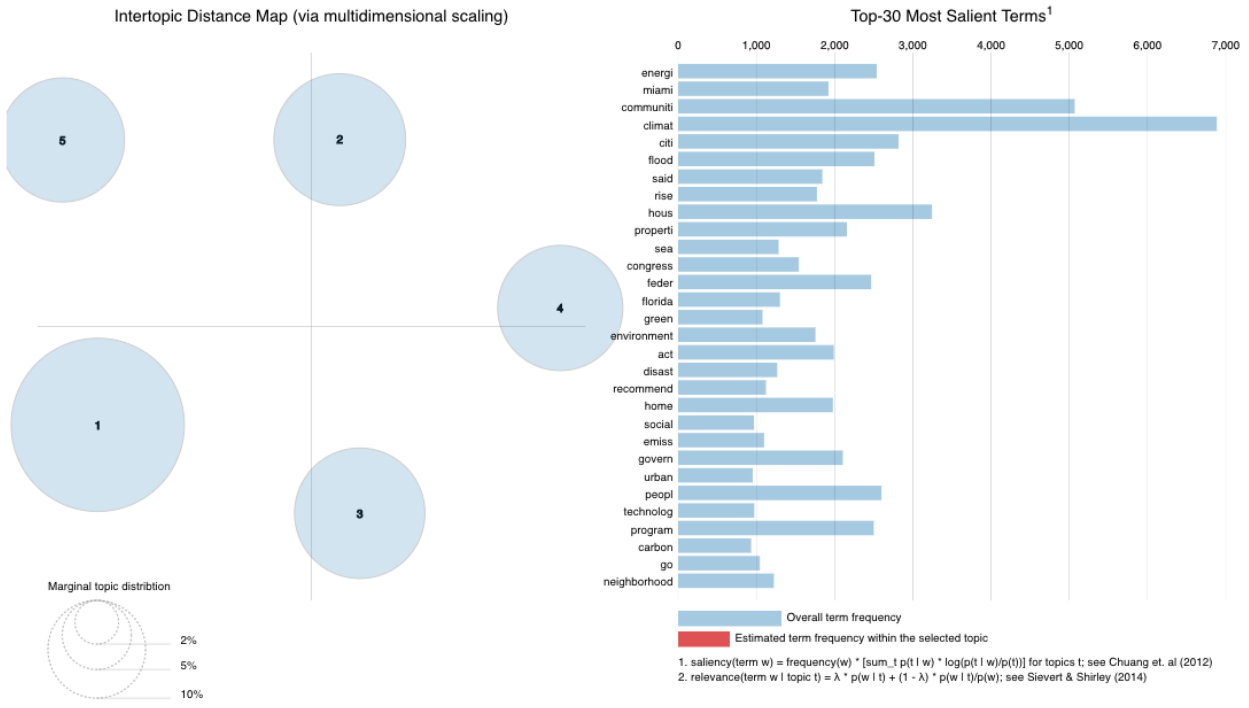



Figure 2: Topic Modelling Intertopic Distance Map for k=5

```
library(LDAvis)
library("tsne")
svd_tsne <- function(x) tsne(svd(x)$u)
json <- createJSON(
  phi = tmResult_7$terms,
  theta = tmResult_7$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="", ylab="")
)
serVis(json)
```

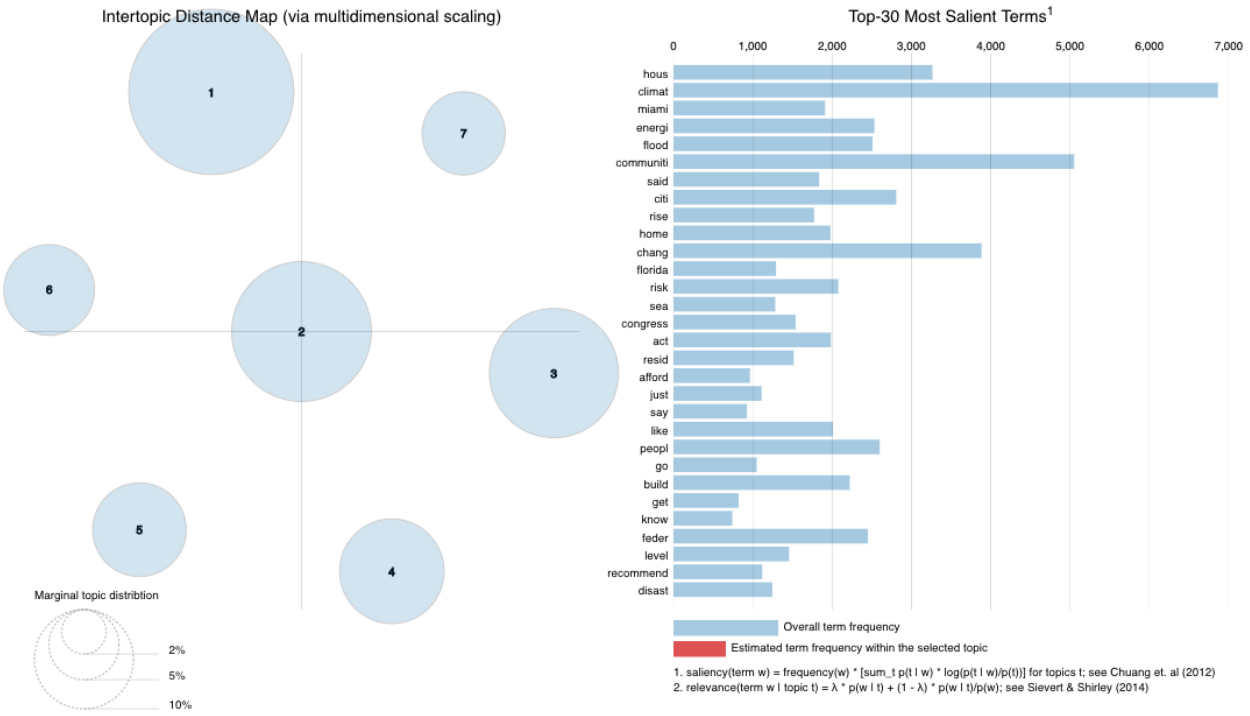
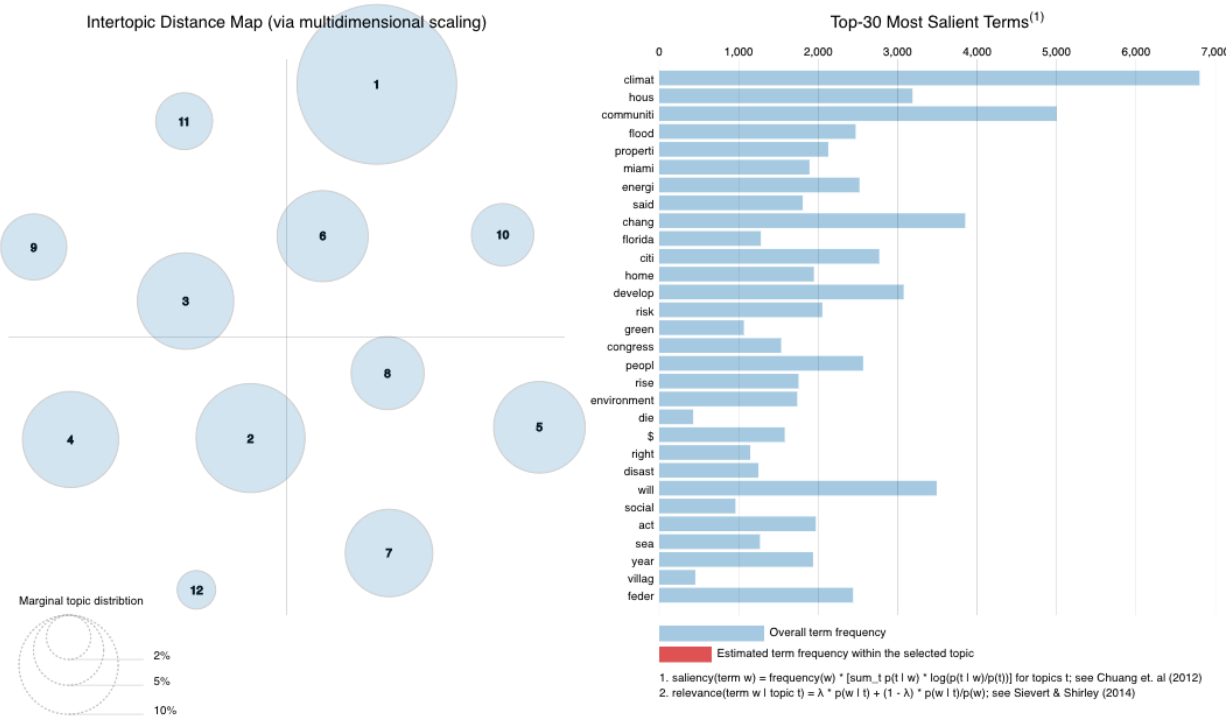



Figure 3: Topic Modelling Intertopic Distance Map for k=7

```
library(LDAvis)
library("tsne")
svd_tsne <- function(x) tsne(svd(x)$u)
json <- createJSON(
  phi = tmResult_12$terms,
  theta = tmResult_12$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="", ylab="")
)
serVis(json)
```




The interopic distance maps show that there's not much overlap either of the three topic models. Given our limited dataset, there were alot of small topics in k=12 model that may not be important to the overarching themes of climate gentrification We found the interpretation for 12 was not useful because it parsed out topics more than necessary (i.e. flood and rise were in two different topics). The k=5 model has a good spread in the intertopic distance map but it only scored 75/30% on the optimization metrics. The k=7 model has good spread in the intertopic distance map as well but there seems to be one very large topic in the center of it all. Since the k=7 model scored nearly 55%/50% on the optimization metric and did not add much value through the additional 2 topics, we believe that the optimal number of topics is 5. However, as mentioned earlier, the newness and specificity of our search term does not make it well suited for topic modelling analysis.