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 [0.84 secs]

##  ...articles split [1.03 secs]

##  ...lengths extracted [1.04 secs]

##  ...headlines extracted [1.04 secs]

##  ...newspapers extracted [1.05 secs]

##  ...dates extracted [1.10 secs]

##  ...authors extracted [1.11 secs]

##  ...sections extracted [1.11 secs]
```

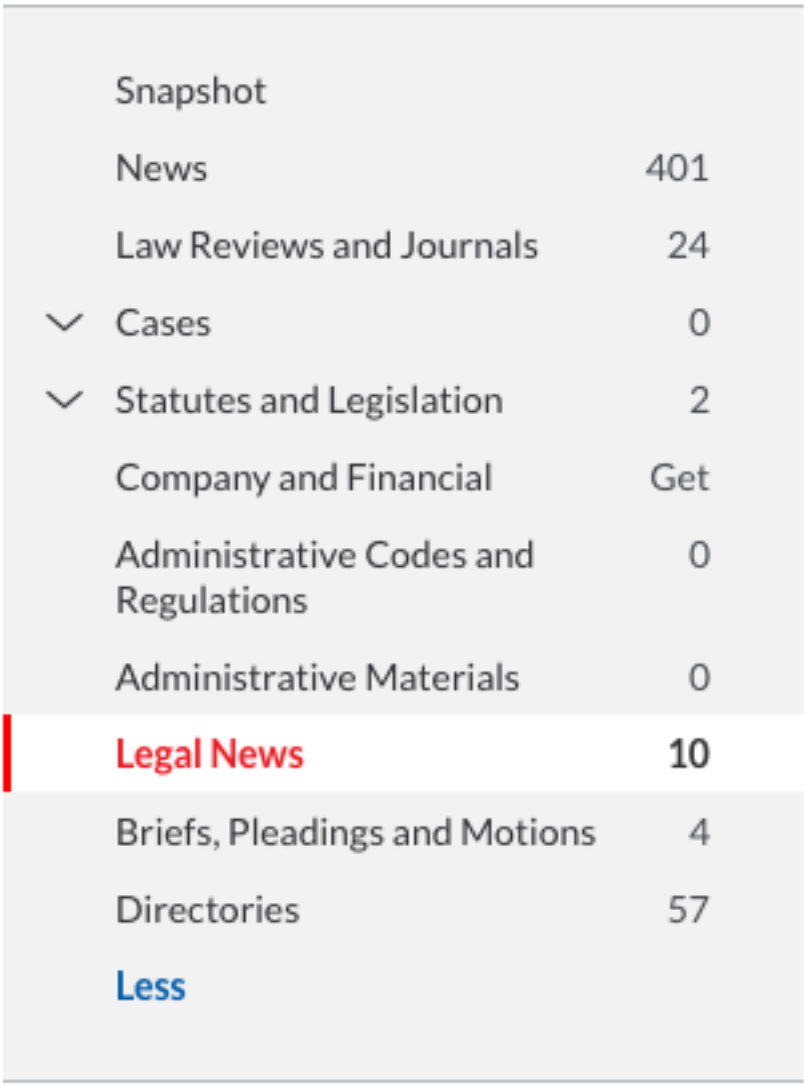


Figure 1: Nexis Uni Results

```
##  ...editions extracted [1.11 secs]

##  ...dates converted [1.12 secs]

##  ...metadata extracted [1.12 secs]

##  ...article texts extracted [1.12 secs]

##  ...superfluous whitespace removed [1.28 secs]

## Elapsed time: 1.28 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")
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 ]
```

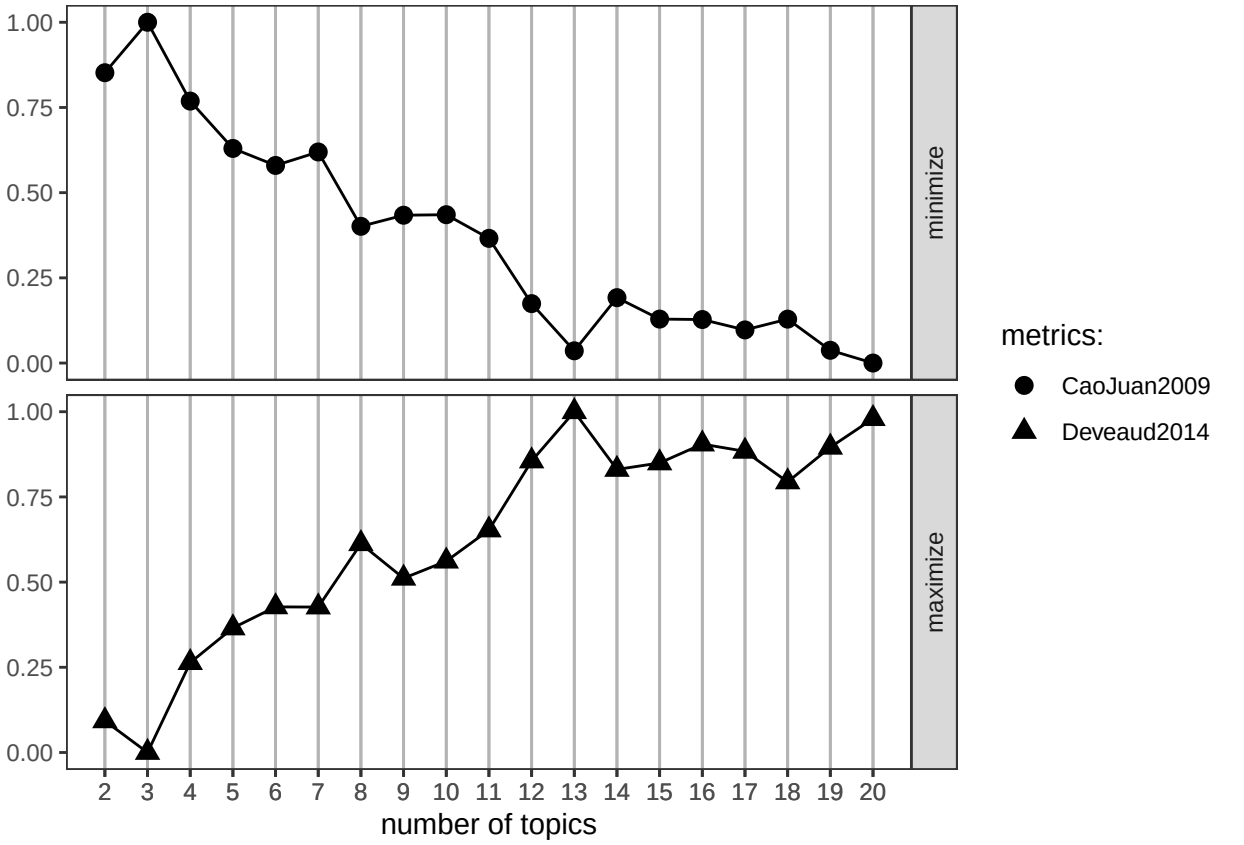
```
#remove rows (docs) with all zeros
sel_idx <- slam::row_sums(dfm) > 0
dfm <- dfm[sel_idx, ]
```

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=4: 50% k=7: 75% k=12: Close to 100%

We ran 3 models based on the number of topics provided by the optimization metrics. We think that k=4, k=7 and k=12 are good values to test for the number of topics according 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=4 instead because of our prior knowledge that climate gentrification does not have that many subtopics.

Topic models for k=4, k=7 and k=12

```
k <- 4

topicModel_k4 <- LDA(dfm, k, method="Gibbs", control=list(iter = 500, verbose = 25))
```

```
## K = 4; 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_4 <- posterior(topicModel_k4)
attributes(tmResult_4)
```

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

```
#nTerms(dfm_comm)
beta_4 <- tmResult_4$terms   # get beta from results
dim(beta_4)                # K distributions over nTerms(DTM) terms# lengthOfVocab
```

```
## [1]     4 13104
```

```
terms(topicModel_k4, 10)
```

```
##       Topic 1     Topic 2 Topic 3     Topic 4
##  [1,] "communiti" "peopl" "climat"    "climat"
##  [2,] "hous"      "work"  "energi"    "flood"
##  [3,] "develop"   "know"  "program"   "citi"
##  [4,] "can"       "go"    "feder"     "miami"
##  [5,] "properti"  "year"  "act"       "chang"
##  [6,] "govern"    "one"   "communiti" "rise"
##  [7,] "use"       "$"     "state"     "will"
##  [8,] "will"      "said"  "nation"    "said"
##  [9,] "resid"     "time"  "fund"      "sea"
## [10,] "land"      "will"  "includ"    "peopl"
```

```
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,] "work"   "know"  "climat"   "communiti"    "miami"   "climat"
##  [2,] "black"  "hous"  "chang"    "hous"         "flood"   "chang"
##  [3,] "art"    "peopl" "will"     "develop"      "citi"    "risk"
##  [4,] "will"   "go"    "properti" "citi"         "said"    "will"
##  [5,] "said"   "die"   "disast"   "resid"        "climat"  "year"
##  [6,] "citi"   "one"   "govern"   "green"        "rise"    "report"
##  [7,] "peopl"  "think" "cost"     "urban"        "home"    "percent"
##  [8,] "miami"  "fair"  "may"      "social"       "sea"     "peopl"
##  [9,] "artist" "now"   "can"      "neighborhood" "hous"    "heat"
## [10,] "cultur" "new"   "state"    "plan"         "florida" "$"
##       Topic 7
##  [1,] "climat"
##  [2,] "energi"
##  [3,] "program"
```

```
##  [4,] "feder"
##  [5,] "act"
##  [6,] "state"
##  [7,] "nation"
##  [8,] "communiti"
##  [9,] "congress"
## [10,] "build"
```

```
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   Topic 7
##  [1,] "miami"   "properti"  "florida" "climat" "de"    "work"    "hous"
##  [2,] "rise"    "law"       "said"    "chang"  "die"   "art"     "afford"
##  [3,] "said"    "right"     "state"   "risk"   "la"    "black"   "resid"
##  [4,] "citi"    "use"       "trump"   "will"   "der"   "artist"  "new"
##  [5,] "sea"     "communiti" "citi"    "impact" "miami" "produc"  "communiti"
##  [6,] "home"    "can"       "new"     "heat"   "en"    "histori" "develop"
##  [7,] "florida" "will"      "via"     "adapt"  "sep"   "$"       "fair"
##  [8,] "flood"   "villag"    "presid"  "year"   "des"   "film"    "home"
##  [9,] "climat"  "also"      "will"    "global" "und"   "will"    "citi"
## [10,] "level"   "peopl"     "polit"   "futur"  "den"   "p.m"     "unit"
##       Topic 8   Topic 9     Topic 10   Topic 11        Topic 12
##  [1,] "percent" "peopl"     "disast"   "communiti"     "energi"
##  [2,] "$"       "know"      "chang"    "green"         "climat"
```

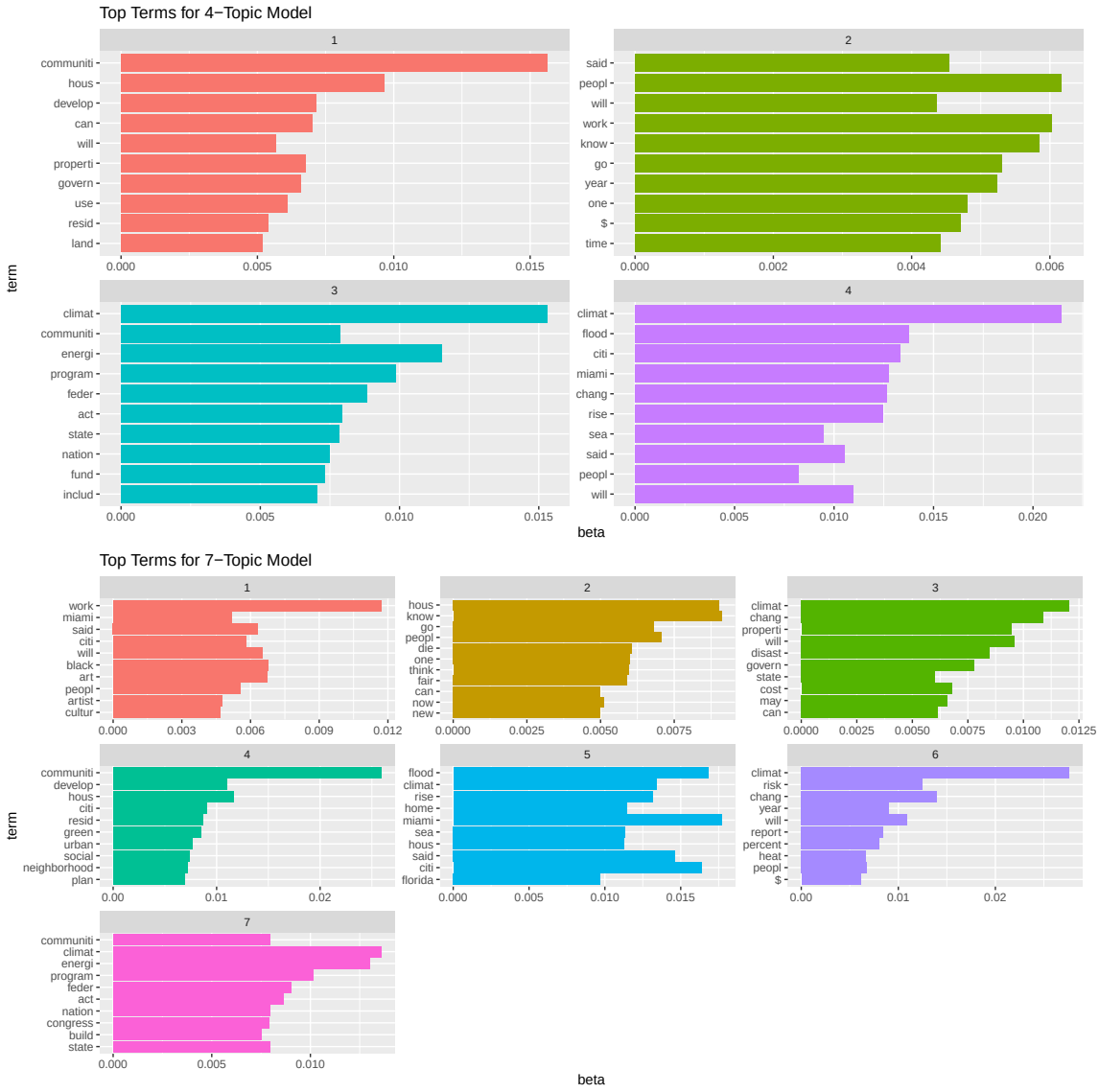
```
##  [3,] "report"  "go"        "flood"    "citi"          "program"
##  [4,] "year"    "think"     "climat"   "develop"       "feder"
##  [5,] "will"    "can"       "cost"     "urban"         "act"
##  [6,] "busi"    "one"       "properti" "environment"   "congress"
##  [7,] "price"   "look"      "will"     "govern"        "build"
##  [8,] "new"     "now"       "govern"   "neighborhood"  "nation"
##  [9,] "market"  "communiti" "land"     "infrastructur" "state"
## [10,] "peopl"   "time"      "state"    "social"        "communiti"
```

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

```
comment_topics_4 <- tidy(topicModel_k4, matrix = "beta")

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


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

top_terms_4 <- comment_topics_4 %>%
  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_4_plot <- top_terms_4 %>%
  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 4-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_4_plot / top_terms_7_plot / top_terms_12_plot
```


Top Terms for 4−Topic Model
1
communiti
hous
develop
can
will
properti
govern
use
resid
land
0.000
0.005
0.010
0.015
2
said
peopl
will
work
know
go
year
one
$
time
0.000
0.002
0.004
0.006
3
climat
communiti
energi
program
feder
act
state
nation
fund
includ
0.000
0.005
0.010
0.015
4
climat
flood
citi
miami
chang
rise
sea
said
peopl
will
0.000
0.005
0.010
0.015
0.020
term
beta
Top Terms for 7−Topic Model
1
work
miami
said
citi
will
black
art
peopl
artist
cultur
0.000
0.003
0.006
0.009
0.012
2
hous
know
go
peopl
die
one
think
fair
can
now
new
0.0000
0.0025
0.0050
0.0075
3
climat
chang
properti
will
disast
govern
state
cost
may
can
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
4
communiti
develop
hous
citi
resid
green
urban
social
neighborhood
plan
0.00
0.01
0.02
5
flood
climat
rise
home
miami
sea
hous
said
citi
florida
0.000
0.005
0.010
0.015
6
climat
risk
chang
year
will
report
percent
heat
peopl
$
0.00
0.01
0.02
7
communiti
climat
energi
program
feder
act
nation
congress
build
state
0.000
0.005
0.010
term
beta
Top Terms for 12−Topic Model
1
climat
miami
rise
sea
said
florida
flood
citi
home
level
0.000
0.005
0.010
0.015
0.020
0.025
2
communiti
properti
law
right
peopl
can
will
use
villag
also
0.000
0.004
0.008
0.012
3
said
florida
citi
new
will
state
trump
via
presid
polit
0.000
0.004
0.008
0.012
4
climat
chang
risk
impact
heat
will
year
adapt
global
temperatur
futur
0.00
0.02
0.04
5
de
die
miami
la
der
en
sep
des
und
den
0.00
0.01
0.02
0.03
6
work
art
$
will
black
artist
produc
histori
film
p.m
0.000
0.005
0.010
7
hous
afford
resid
develop
communiti
citi
home
fair
new
unit
0.00
0.02
0.04
8
percent
$
report
peopl
busi
new
will
price
year
market
compani
0.000
0.005
0.010
0.015
9
know
communiti
go
think
one
peopl
can
look
now
time
0.000
0.005
0.010
0.015
10
climat
chang
disast
flood
properti
cost
govern
land
will
state
0.000
0.003
0.006
0.009
0.012
11
green
develop
communiti
urban
citi
environment
govern
neighborhood
infrastructur
social
0.000
0.005
0.010
0.015
0.020
0.025
12
climat
energi
communiti
program
feder
act
congress
build
state
nation
0.000
0.005
0.010
0.015
term
beta

```
top5termsPerTopic_4 <- terms(topicModel_k4, 5)
topicNames_4 <- apply(top5termsPerTopic_4, 2, paste, collapse=" ")
topicNames_4
```

```
##                               Topic 1                               Topic 2
## "communiti hous develop can properti"            "peopl work know go year"
##                               Topic 3                               Topic 4
##     "climat energi program feder act"       "climat flood citi miami chang"
```

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

```
##                               Topic 1                               Topic 2
##          "work black art will said"           "know hous peopl go die"
##                               Topic 3                               Topic 4
## "climat chang will properti disast" "communiti hous develop citi resid"
##                               Topic 5                               Topic 6
##      "miami flood citi said climat"       "climat chang risk will year"
##                               Topic 7
##   "climat energi program feder act"
```

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

```
##                               Topic 1                               Topic 2
##           "miami rise said citi sea"   "properti law right use communiti"
##                               Topic 3                               Topic 4
##      "florida said state trump citi"      "climat chang risk will impact"
##                               Topic 5                               Topic 6
##                "de die la der miami"       "work art black artist produc"
##                               Topic 7                               Topic 8
##    "hous afford resid new communiti"         "percent $ report year will"
##                               Topic 9                              Topic 10
##            "peopl know go think can"     "disast chang flood climat cost"
##                              Topic 11                              Topic 12
## "communiti green citi develop urban"    "energi climat program feder act"
```

```
library(LDAvis)
library("tsne")
svd_tsne <- function(x) tsne(svd(x)$u)
json <- createJSON(
  phi = tmResult_4$terms,
  theta = tmResult_4$topics,
  doc.length = rowSums(dfm),
  vocab = colnames(dfm),
  term.frequency = colSums(dfm),
  mds.method = svd_tsne,
  plot.opts = list(xlab="", ylab="")
)
```

```
## sigma summary: Min. : 33554432 |1st Qu. : 33554432 |Median : 33554432 |Mean : 33554432 |3rd Qu. : 335

## Epoch: Iteration #100 error is: 8.95315449497482

## Epoch: Iteration #200 error is: 0.0487253639029235

## Epoch: Iteration #300 error is: 0.0487253193137161

## Epoch: Iteration #400 error is: 0.0487253159665143

## Epoch: Iteration #500 error is: 0.0487253110041692

## Epoch: Iteration #600 error is: 0.0487253044043246

## Epoch: Iteration #700 error is: 0.0487252961272695

## Epoch: Iteration #800 error is: 0.0487252861200865

## Epoch: Iteration #900 error is: 0.0487252743173065

## Epoch: Iteration #1000 error is: 0.0487252606326122
```

```
serVis(json)
```

```
## Loading required namespace: servr
```

```
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="")
)
```

```
## sigma summary: Min. : 33554432 |1st Qu. : 33554432 |Median : 33554432 |Mean : 33554432 |3rd Qu. : 335

## Epoch: Iteration #100 error is: 14.6932485329001

## Epoch: Iteration #200 error is: 0.249358315179571

## Epoch: Iteration #300 error is: 0.192993263892333

## Epoch: Iteration #400 error is: 0.162241946795409
```

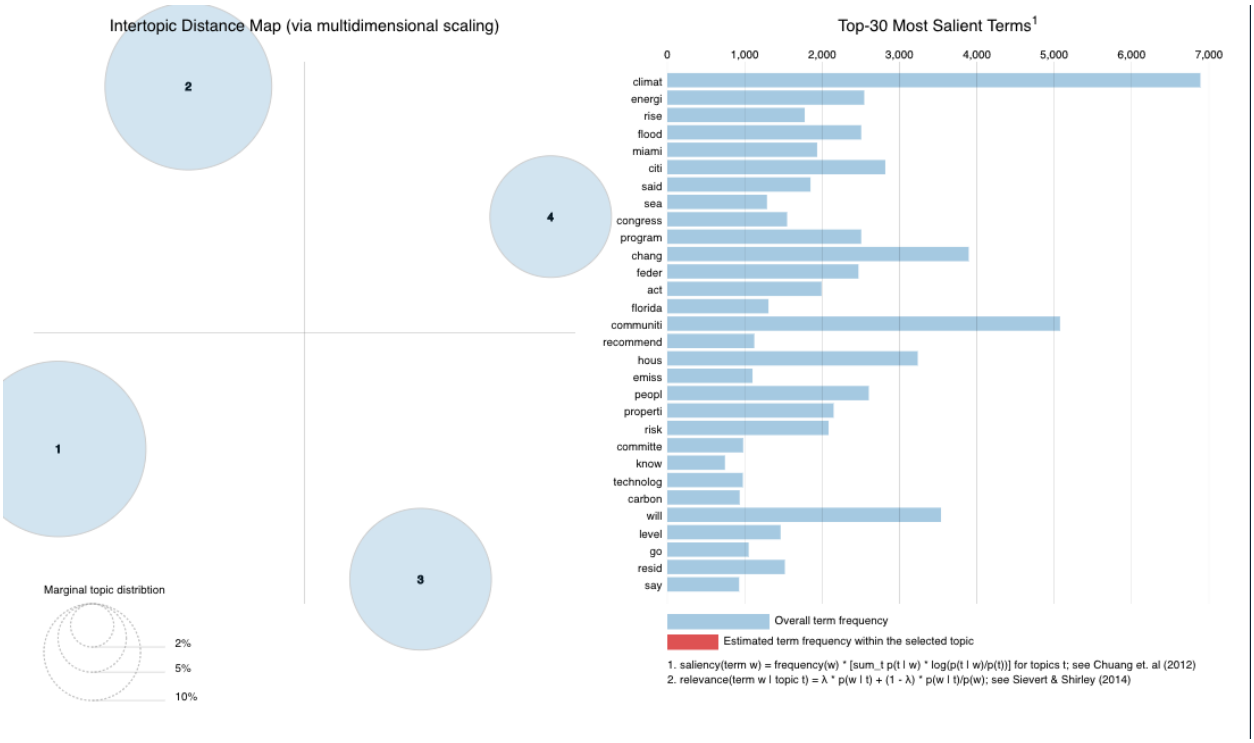


Figure 2: Topic Modelling Intertopic Distance Map for k=4

```
## Epoch: Iteration #500 error is: 0.158652868608222

## Epoch: Iteration #600 error is: 0.158651591120952

## Epoch: Iteration #700 error is: 0.158651589959579

## Epoch: Iteration #800 error is: 0.158651589957977

## Epoch: Iteration #900 error is: 0.158651589956324

## Epoch: Iteration #1000 error is: 0.158651589953862
```

```
serVis(json)
```

```
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="")
)
```

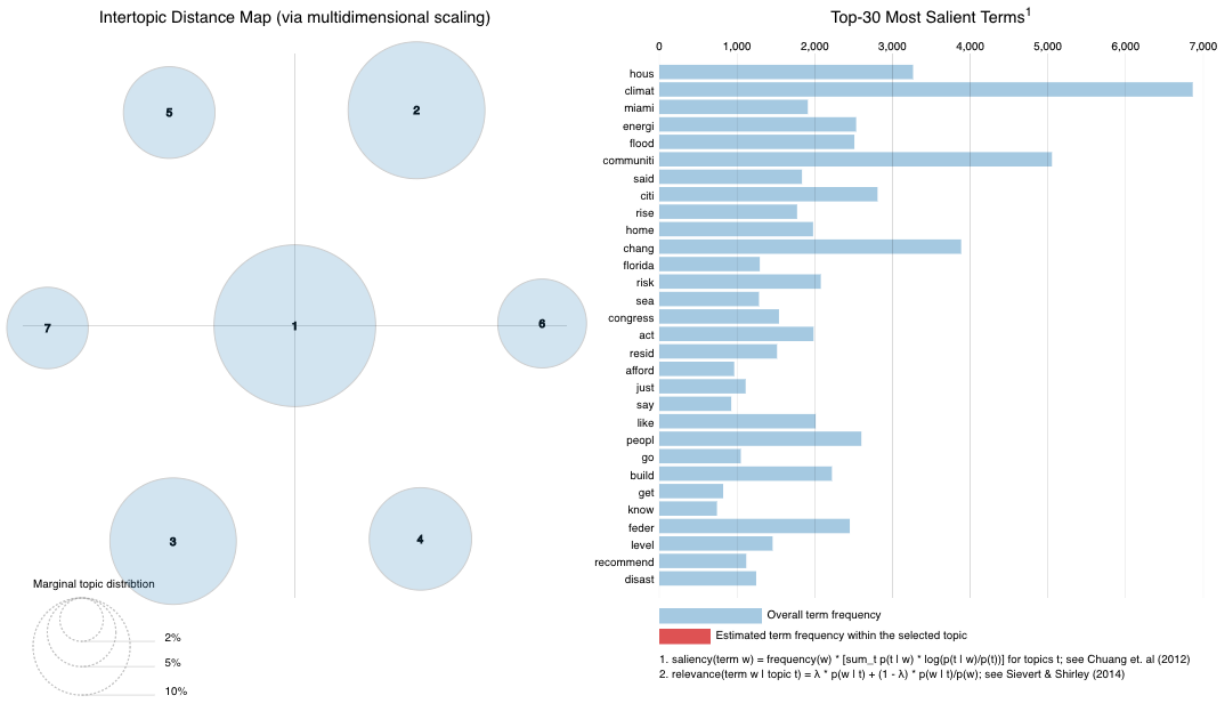
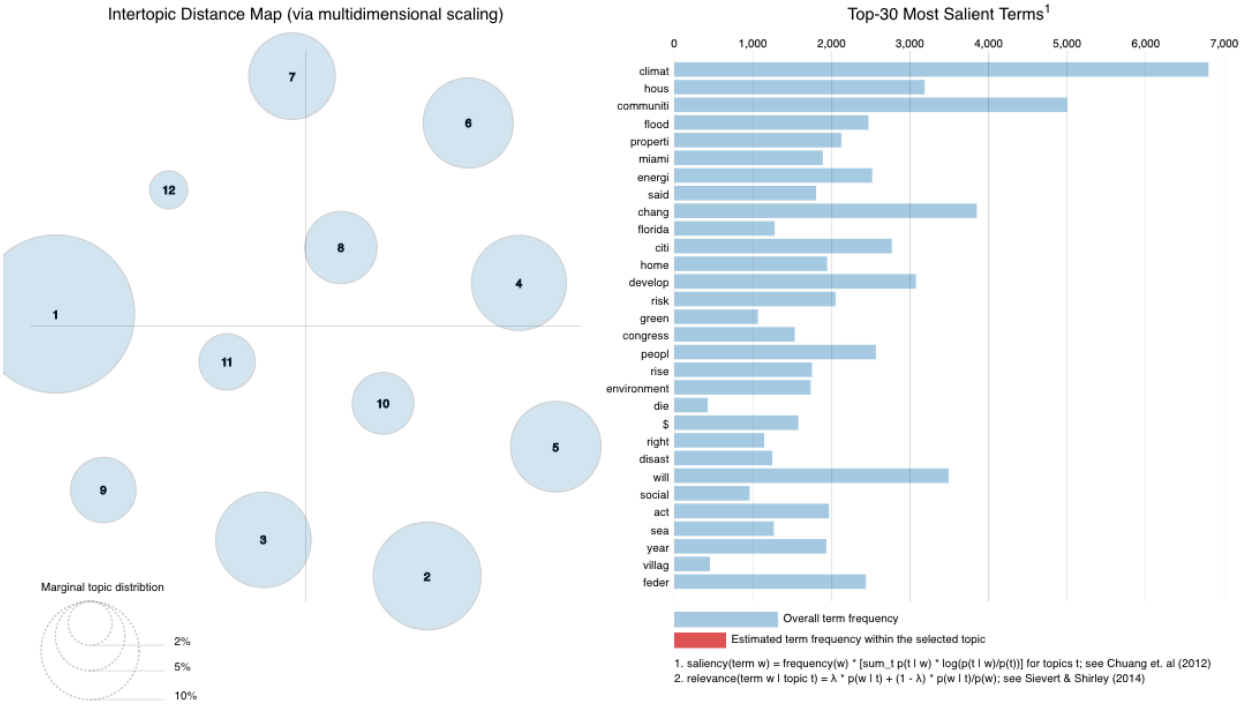


Figure 3: Topic Modelling Intertopic Distance Map for k=7

```
## sigma summary: Min. : 33554432 |1st Qu. : 33554432 |Median : 33554432 |Mean : 33554432 |3rd Qu. : 335

## Epoch: Iteration #100 error is: 12.9090229114709

## Epoch: Iteration #200 error is: 0.445253755699508

## Epoch: Iteration #300 error is: 0.310407135192461

## Epoch: Iteration #400 error is: 0.300021629947652

## Epoch: Iteration #500 error is: 0.290316581788047

## Epoch: Iteration #600 error is: 0.289117340204242

## Epoch: Iteration #700 error is: 0.288943985995375

## Epoch: Iteration #800 error is: 0.288930077272253

## Epoch: Iteration #900 error is: 0.288928511037823

## Epoch: Iteration #1000 error is: 0.288928424802804
```

```
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=4 model has a good spread in the intertopic distance map but it only scored 50% 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 75% on the optimization metric, we believe that the optimal number of topics is 7 as strikes a good balance between detail of the topics and usefulness. However, as mentioned earlier, the newness and specificity of our search term does not make it well suited for topic modelling analysis.