Country Music Project

Matt Shu

2022-05-03

Prereqs

```
# Needed to overcome error found below with Homebrew TBB vs bundled TBB:
# https://github.com/RcppCore/RcppParallel/issues/182
# remotes::install_github("RcppCore/RcppParallel")
library(igraph)
library(tidyverse)
library(stm)
library(RSQLite)
library(RecordLinkage)
library(stringdist)
library(devtools)
library(tm)
devtools::install_github("mikajoh/tidystm", dependencies = TRUE)
library(tidystm)
```

```
conn <- dbConnect(RSQLite::SQLite(), "files/22-04-21-playback-fm-top-country.db")
cleaned_df <- dbGetQuery(conn, 'SELECT * FROM cleaned')
dbDisconnect(conn)
```

```
names(cleaned_df)
```

```
##  [1] "artist_id"          "track_id"           "year"
##  [4] "artist"             "track"              "rank"
##  [7] "link"               "lyrics"             "artist_appearances"
## [10] "mb_id"              "type"               "area.name"
## [13] "gender"             "life_span.begin"    "life_span.ended"
## [16] "song_id"            "cleaned_lyrics"     "lyrics_alnum"
```

```
cleaned_df <- cleaned_df %>%
    mutate(gender = replace(gender, gender == "other", "non-binary")) %>%
    mutate(gender = replace(gender, is.na(gender), "group"))
```

```
print("Cache the Image files")
```

```
## [1] "Cache the Image files"
```

Preprocessing (and STM exploration)

```
# Dataframe containing the text
docs_df <- cleaned_df %>%
   dplyr::select(track_id, lyrics_alnum) %>%
   # first, remove observation with missing values of the meta variables
   filter(!is.na(lyrics_alnum))  %>%
```

```
   # the objects need to be class "data frame"
   as.data.frame()
```

```
# Dataframe containing (sample) documents' metadata of interest
meta_df <- cleaned_df %>%
   dplyr::select(track_id, rank, artist, track, year, gender) %>%
   # the objects need to be class "data frame"
   as.data.frame()
```

```
processed_docs_1 <- textProcessor(documents = docs_df$lyrics_alnum,
                                  metadata = meta_df,
                                  lowercase = TRUE,
                                  removestopwords = TRUE,
                                  removenumbers = TRUE,
                                  removepunctuation = TRUE,
                                  ucp = TRUE,
                                  stem = TRUE,
                                  striphtml = TRUE,
                                  wordLengths = c(3, Inf),
                                  language = "en")
```

```
meta <- processed_docs_1$meta
vocab <- processed_docs_1$vocab
docs <- processed_docs_1$documents
keep <- !is.na(meta$artist) && !is.na(meta$rank)
```

```
## Warning in !is.na(meta$artist) && !is.na(meta$rank): 'length(x) = 5970 > 1' in
## coercion to 'logical(1)'

## Warning in !is.na(meta$artist) && !is.na(meta$rank): 'length(x) = 5970 > 1' in
## coercion to 'logical(1)'
```

```
meta <- meta[keep,]
docs <- docs[keep]
```

```
prepped_data <- prepDocuments(docs,
                              vocab,
                              meta,
                              # the lower threshold value means that only words
                              # that appear more times than the value (in this
                              # example the value = 3) will be retained; this is
                              # another researcher decision
                              lower.thresh = 2)
```

Old code for removing unusual mismatch with no words despite past filters

```
length(docs_df$lyrics_alnum) # original documents
length(prepped_data$meta$track_id) # off from the preceding count
dif <- setdiff(docs_df$track_id, # original vector of documents
               prepped_data$meta$track_id) # list of documents after prepDocuments
tmp <- docs_df
tmp2 <- tmp[!tmp$track_id %in% dif,]
tmp_doc <- tmp2 %>%
  select(track_id, lyrics_alnum)
length(tmp_doc$track_id)
length(prepped_data$meta$track_id)
```

```
# View the track ids that were removed for some reason (often other language)
tmp3 <- tmp[tmp$track_id %in% dif,]
tmp3
```

See Cleaned Sample!

```
head(cleaned_df)
```

```
##   artist_id track_id year    artist                       track rank
## 1         1        0 1944 Red Foley          Smoke On The Water    1
## 2         1      506 1951 Red Foley                 Hobo Boogie   55
## 3         1      587 1953 Red Foley                    Midnight   14
## 4         1      386 1950 Red Foley      Cincinnati Dancing Pig   13
## 5         1      374 1950 Red Foley Chattanoogie Shoe Shine Boy    1
## 6         1      620 1953 Red Foley                   Hot Toddy   47
##                                                             link
## 1          /charts/country/video/1944/red-foley-smoke-on-the-water
## 2                /charts/country/video/1951/red-foley-hobo-boogie
## 3                   /charts/country/video/1953/red-foley-midnight
## 4      /charts/country/video/1950/red-foley-cincinnati-dancing-pig
## 5 /charts/country/video/1950/red-foley-chattanoogie-shoe-shine-boy
## 6                  /charts/country/video/1953/red-foley-hot-toddy
##
## 1
## 2
## 3
## 4
## 5 Chattanoogie Shoe Shine Boy LyricsHave you ever passed the corner of Forth and Grand? Where a litt
## 6
##   artist_appearances                                mb_id   type     area.name
## 1                 33 aff932c2-ec30-4ee9-9125-5f761aae61a4 Person United States
## 2                 33 aff932c2-ec30-4ee9-9125-5f761aae61a4 Person United States
## 3                 33 aff932c2-ec30-4ee9-9125-5f761aae61a4 Person United States
## 4                 33 aff932c2-ec30-4ee9-9125-5f761aae61a4 Person United States
## 5                 33 aff932c2-ec30-4ee9-9125-5f761aae61a4 Person United States
## 6                 33 aff932c2-ec30-4ee9-9125-5f761aae61a4 Person United States
##   gender life_span.begin life_span.ended song_id
## 1   male      1910-06-17            true   14519
## 2   male      1910-06-17            true   11892
## 3   male      1910-06-17            true   13445
## 4   male      1910-06-17            true   10833
## 5   male      1910-06-17            true   10810
## 6   male      1910-06-17            true   11966
##
## 1
## 2
## 3
## 4
## 5 Have you ever passed the corner of Forth and Grand? Where a little ball o' rhythm has a shoe-shine
## 6
##
## 1
## 2
## 3
```

```
## 4
## 5 Have you ever passed the corner of Forth and Grand  Where a little ball o  rhythm has a shoe shine
## 6
```

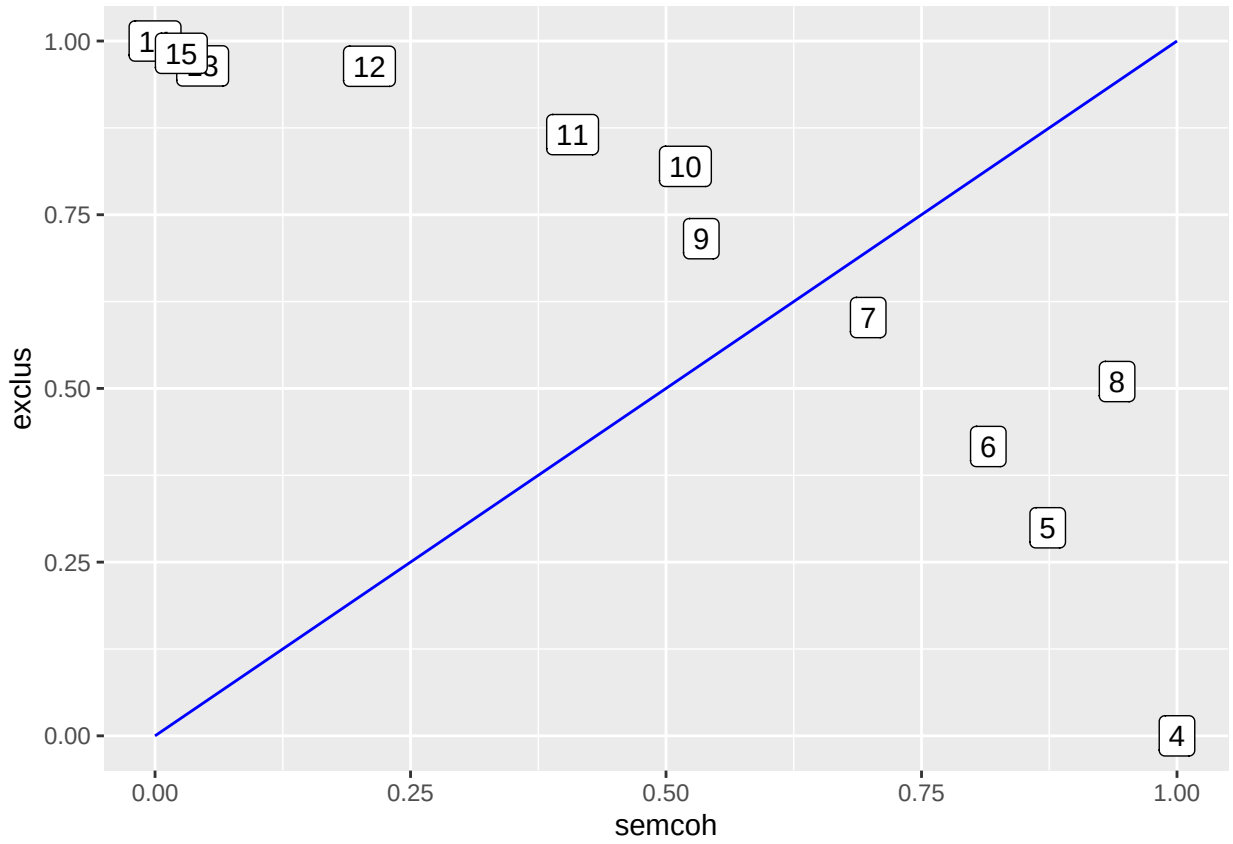
Find K

```
k_seq = seq(4, 15, 1)
```

```
## You can "watch" the algorithm model topics in the console
searched = searchK(prepped_data$documents,
                   prepped_data$vocab,
                   K = k_seq,
                   data = prepped_data$meta,
                   seed = 183654)
saveRDS(searched, file = "files/22-04-29-searchK.RData")
```

Show K

```
searched <- readRDS("files/22-04-29-searchK.RData")
# Get values from `searchK` output
semcoh <- unlist(searched$results$semcoh)
exclus <- unlist(searched$results$exclus)

# Max/min semantic cohesion
max_sc <- max(semcoh)
min_sc<-min(semcoh)

# Max/min exclusivity
max_ex<-max(exclus)
min_ex<-min(exclus)

# Min-max normalization is (value - min)/(max - min)
x_vals <- (semcoh-min_sc)/(max_sc-min_sc)
y_vals <- (exclus-min_ex)/(max_ex-min_ex)
# add semantic cohesion and exclusivity together weighted evenly
search_plot_df <- tibble(id = k_seq,
                   semcoh = x_vals,
                   exclus = y_vals,
                   combine = x_vals*0.5 + y_vals*0.5)
# Plot
ggplot(search_plot_df, mapping = aes(x = semcoh, y = exclus)) +
  xlim(0,1) +
  ylim(0,1) +
  ggplot2::annotate("segment", x = 0, xend = 1, y = 0, yend = 1, color = "blue") +
  geom_label(aes(label=id))
```



Model Work

```
# 6 topics seems to also work nice, with a strong "Country" category
num_topics <- 7 # Chosen after above search and some playing around
out_covariates_7 <- stm(prepped_data$documents,
                        prepped_data$vocab,
                        K = num_topics,
                        prevalence = ~ rank + year * gender,
                        max.em.its = 500,
                        data = prepped_data$meta,
                        seed = 592669)
```

```
head(out_covariates_7$theta) # each row is each document
```

```
##             [,1]        [,2]        [,3]       [,4]        [,5]        [,6]
## [1,] 0.183253946 0.007069806 0.017951578 0.14437433 0.034003629 0.351454015
## [2,] 0.059540942 0.026539450 0.028958227 0.64243815 0.174697316 0.040460910
## [3,] 0.465211708 0.018478813 0.071580114 0.01325611 0.089937091 0.305363030
## [4,] 0.006668523 0.007024786 0.008411345 0.90506787 0.004507195 0.031606214
## [5,] 0.040222095 0.034340099 0.068512467 0.40076420 0.030625939 0.198152810
## [6,] 0.032443538 0.613071173 0.034814173 0.01632440 0.038018646 0.009980522
##             [,7]
## [1,] 0.26189269
## [2,] 0.02736501
## [3,] 0.03617314
## [4,] 0.03671407
## [5,] 0.22738239
```

```
## [6,] 0.25534755
```

```
# To find each artists, link the songs to the artists and then take the average for each artists, for e
head(prepped_data$meta) # same order between dataframes
```

```
##   track_id rank    artist                       track year gender
## 1        0    1 Red Foley          Smoke On The Water 1944   male
## 2      506   55 Red Foley                 Hobo Boogie 1951   male
## 3      587   14 Red Foley                    Midnight 1953   male
## 4      386   13 Red Foley      Cincinnati Dancing Pig 1950   male
## 5      374    1 Red Foley Chattanoogie Shoe Shine Boy 1950   male
## 6      620   47 Red Foley                   Hot Toddy 1953   male
```

```
track_topic_df <- cbind(prepped_data$meta, out_covariates_7$theta)
```

```
terms = labelTopics(out_covariates_7, n = 10)
terms$prob # rows are topics; columns are most probable words (in order)
```

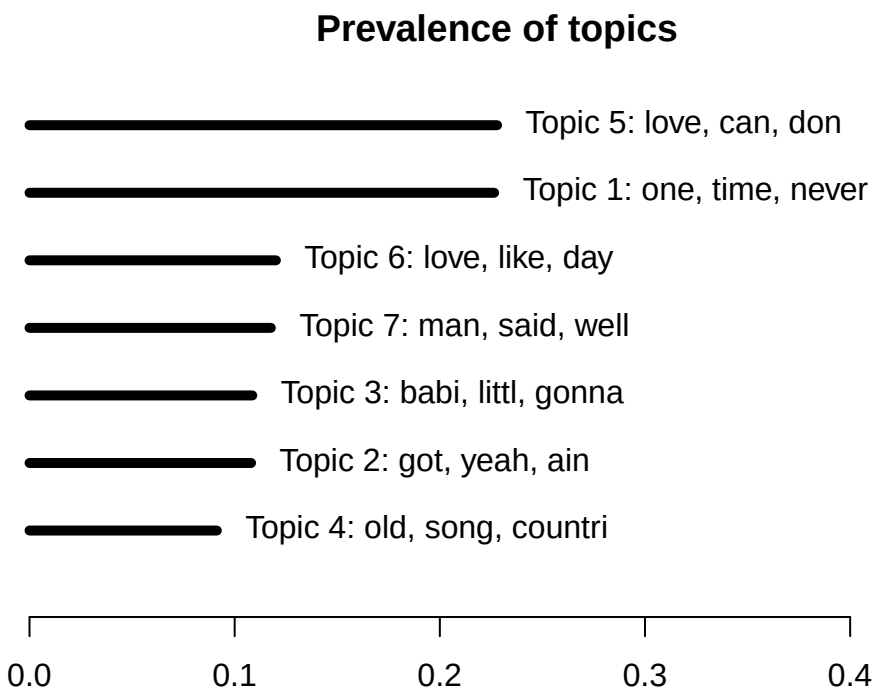
```
##      [,1]   [,2]    [,3]      [,4]    [,5]    [,6]    [,7]   [,8]    [,9]
## [1,] "one"  "time"  "never"   "now"   "heart" "still" "say"  "just"  "gone"
## [2,] "got"  "yeah"  "ain"     "like"  "girl"  "good"  "get"  "wanna" "just"
## [3,] "babi" "littl" "gonna"   "come"  "time"  "night" "get"  "take"  "back"
## [4,] "old"  "song"  "countri" "roll"  "back"  "town"  "road" "ride"  "sing"
## [5,] "love" "can"   "don"     "know"  "just"  "want"  "let"  "make"  "feel"
## [6,] "love" "like"  "day"     "dream" "night" "eye"   "blue" "sweet" "rain"
## [7,] "man"  "said"  "well"    "old"   "daddi" "boy"   "big"  "mama"  "just"
##      [,10]
## [1,] "cri"
## [2,] "can"
## [3,] "right"
## [4,] "like"
## [5,] "need"
## [6,] "light"
## [7,] "got"
```

```
terms$frex # rows are topics; columns are most FREX words (in order)
```

```
##      [,1]       [,2]     [,3]         [,4]     [,5]        [,6]     [,7]
## [1,] "fool"     "goodby" "cri"        "lone"   "heartach" "memori" "tear"
## [2,] "ooh"      "huh"    "boo"        "yeah"   "whoa"     "nothin" "ain"
## [3,] "bye"      "babi"   "honey"      "bit"    "gonna"    "shake"  "danc"
## [4,] "countri" "boogi"  "hillbilli" "cowboy" "crank"    "cha"    "tonk"
## [5,] "want"     "hold"   "need"       "lose"   "fall"     "love"   "believ"
## [6,] "angel"    "rain"   "heaven"     "sail"   "sea"      "storm"  "rainbow"
## [7,] "mom"      "dad"    "wife"       "hero"   "father"   "twenti" "america"
##      [,8]      [,9]       [,10]
## [1,] "miss"    "lie"      "still"
## [2,] "nobodi" "lovin"    "woah"
## [3,] "step"    "littl"    "batter"
## [4,] "jone"    "tennesse" "doo"
## [5,] "easi"    "give"     "don"
## [6,] "sunshin" "wing"     "sky"
## [7,] "daddi"   "sir"      "famili"
```

```
# Parameters modified from: https://milesdwilliams15.github.io/Better-Graphics-for-the-stm-Package-in-R
par(bty="n",lwd=5)
plot(out_covariates_7,
```

```
    type = "summary",
    main = "Prevalence of topics")
```




```
docs_examples_covar <- findThoughts(out_covariates_7,
                                    texts = tmp_doc$track_id,
                                    n = 10,
                                    topics = c(1:num_topics))

for(topic_num in c(1:num_topics)) {
  print(paste("Topic ", topic_num))
  for(track in docs_examples_covar$docs[[topic_num]]) {
    print(cleaned_df$track[cleaned_df$track_id == track])
  }
  print("")
}
```

```
## [1] "Topic  1"
## [1] "Rag Mop"
## [1] "Something Old, Something New"
## [1] "I Forgot To Remember To Forget"
## [1] "All Alone in This World without You"
## [1] "Fool Fool Fool"
## [1] "Happy Journey"
## [1] "Careless Darlin'"
## [1] "Sweetheart You Done Me Wrong"
## [1] "You're The One"
## [1] "Things Aren't Funny Anymore"
## [1] ""
## [1] "Topic  2"
## [1] "Desperate Man"
```

```
## [1] "Gimmie That Girl"
## [1] "My Bucket's Got a Hole in it"
## [1] "Just the Way"
## [1] "Just The Way"
## [1] "Uh-Huh--Mm"
## [1] "Uh-Huh-mm"
## [1] "She Ain't Your Ordinary Girl"
## [1] "Glad You Exist"
## [1] "Drinkin' Beer. Talkin' God. Amen."
## [1] ""
## [1] "Topic  3"
## [1] "Swing"
## [1] "Trademark"
## [1] "Little Bit of Life"
## [1] "Little Bit Of Life"
## [1] "Baby Let's Play House"
## [1] "Penny Arcade"
## [1] "Whole Lotta Shakin' Goin' On"
## [1] "Shine, Shave, Shower (It's Saturday)"
## [1] "Be-Bop-A-Lula"
## [1] "Last Minute Late Night"
## [1] ""
## [1] "Topic  4"
## [1] "Teenage Boogie"
## [1] "Redneck Yacht Club"
## [1] "Cincinnati Dancing Pig"
## [1] "Ragtime Cowboy Joe"
## [1] "Long Live"
## [1] "Mule Train"
## [1] "She Cranks My Tractor"
## [1] "The Rhumba Boogie"
## [1] "Smokey Mountain Boogie"
## [1] "Hula Rock"
## [1] ""
## [1] "Topic  5"
## [1] "Love Can't Wait"
## [1] "Don't Underestimate My Love For You"
## [1] "Don't Underestimate My Love for You"
## [1] "I Want To Know You Before We Make Love"
## [1] "Count on Me"
## [1] "A Lover's Question"
## [1] "Mr. Lovemaker"
## [1] "It Matters to Me"
## [1] "It Matters To Me"
## [1] "Fall into Me"
## [1] ""
## [1] "Topic  6"
## [1] "Ring Of Fire"
## [1] "Your Name Is Beautiful"
## [1] "Sweet Summer Lovin'"
## [1] "Mockin' Bird Hill"
## [1] "The Red Strokes"
## [1] "A Fallen Star"
## [1] "Would You Lay With Me (In A Field Of Stone)"
```

```
## [1] "Kentucky Waltz"
## [1] "Beautiful Brown Eyes"
## [1] "My Special Angel"
## [1] ""
## [1] "Topic  7"
## [1] "What's Your Mama's Name"
## [1] "Life Of A Poor Boy"
## [1] "(Margie's At) The Lincoln Park Inn"
## [1] "No Charge"
## [1] "History Repeats Itself"
## [1] "Poor, Poor Pitiful Me"
## [1] "Deck Of Cards"
## [1] "Po' Folks"
## [1] "Shiftwork"
## [1] "None Of My Business"
## [1] ""
```

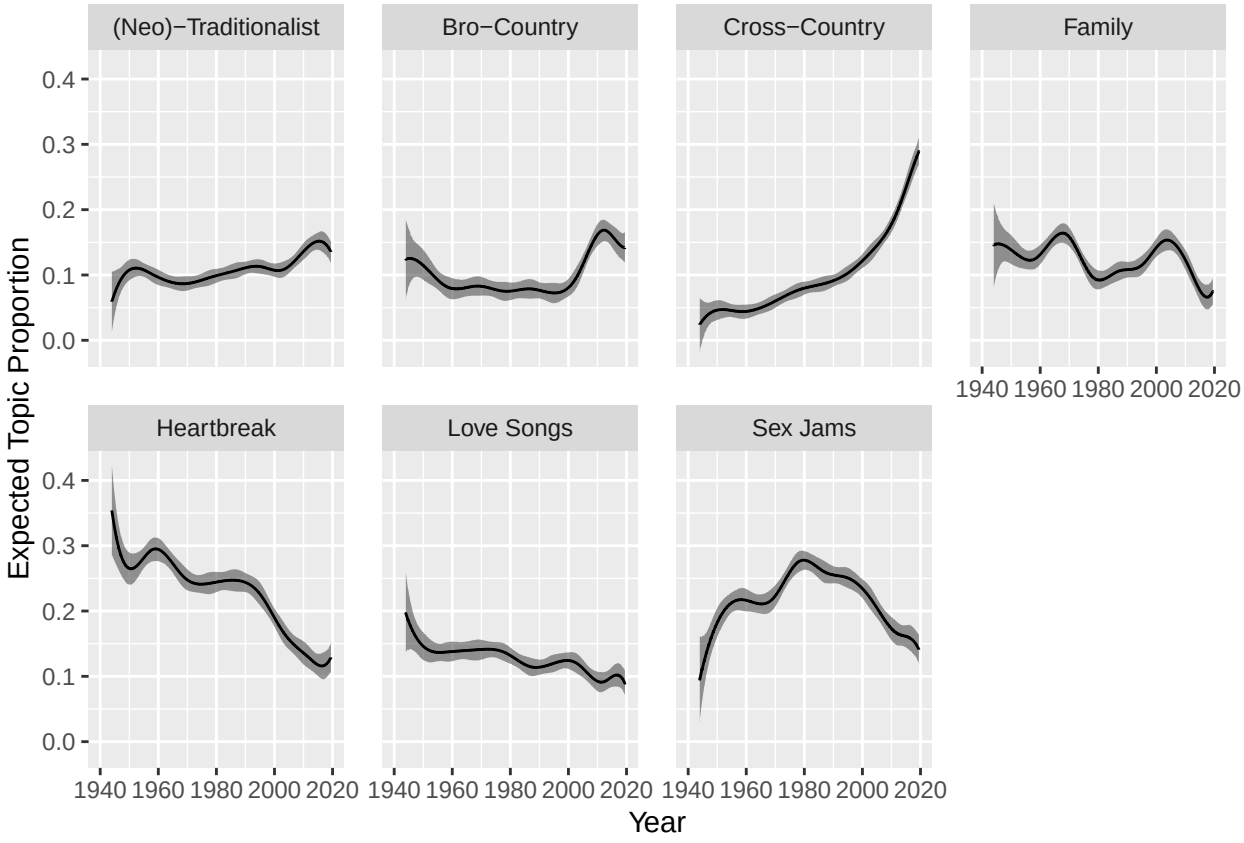
```
# Topic 1: Heartbreak Songs
# Topic 2: Cross-Country (Country Rock/Pop)
# Topic 3: Traditionalist Country (Pardi, Hank Williams)
# Topic 4: Bro-Country
# Topic 5: Sex Jams
# Topic 6: Love songs
# Topic 7: Family
topic_labels <- c("Heartbreak", "Cross-Country", "(Neo)-Traditionalist", "Bro-Country", "Sex Jams", "Lo
```

```
eff1 <- estimateEffect(formula = c(1:num_topics) ~ s(year),
                       # the line above matches the model specification we used
                       stmobj = out_covariates_7,
                       meta = prepped_data$meta,
                       uncertainty = "Global")

# plot.estimateEffect(eff1,
#      covariate = "year",
#      topics = c(1:num_topics),
#      model = out_covariates_7,
#      method = "continuous",
#      xlab = "Year",
#      ylim=c(0, .4),
#      xlim=c(1940, 2020),
#      main = "Effect of Year on Topic Proportion")
```

```
effect <- lapply(c(0, 1), function(i) {
  extract.estimateEffect(eff1,
     covariate = "year",
     topics = c(1:num_topics),
     model = out_covariates_7,
     method = "continuous")
})
effect <- do.call("rbind", effect)
effect <- effect %>% mutate(label = recode(topic, "1"=topic_labels[1], "2" = topic_labels[2], "3" = topi
## And, for example, plot it with ggplot2 and facet by topic instead.
library(ggplot2)

ggplot(effect, aes(x = covariate.value, y = estimate,
```

```
                    ymin = ci.lower, ymax = ci.upper)) +
  facet_wrap(~ label, nrow = 2) +
  geom_ribbon(alpha = .5) +
  geom_line() +
  labs(x = "Year",
       y = "Expected Topic Proportion") +
  scale_x_continuous(breaks=c(1940, 1960, 1980, 2000, 2020),
   labels=waiver(), lim=c(1940,2020)) +
  theme(panel.spacing = unit(1, "lines"))
```

```
## Warning: Removed 4 row(s) containing missing values (geom_path).
```




```
eff <- estimateEffect(formula = c(1:num_topics) ~ year,
                      # the line above matches the model specification we used
                      stmobj = out_covariates_7,
                      meta = prepped_data$meta,
                      uncertainty = "Global")

# Second, plot the results
plot(eff,
     covariate = "year",
     topics = c(1:num_topics),
     model = out_covariates_7,
     method = "continuous",
     xlab = "Year",
     main = "Effect of Year on Topic Proportion")
```

```
library(huge)
```

```
## Registered S3 methods overwritten by 'huge':
##   method    from
##   plot.sim  lava
##   print.sim lava
```

```
topic_corr <- topicCorr(out_covariates_7, method = "huge")
```

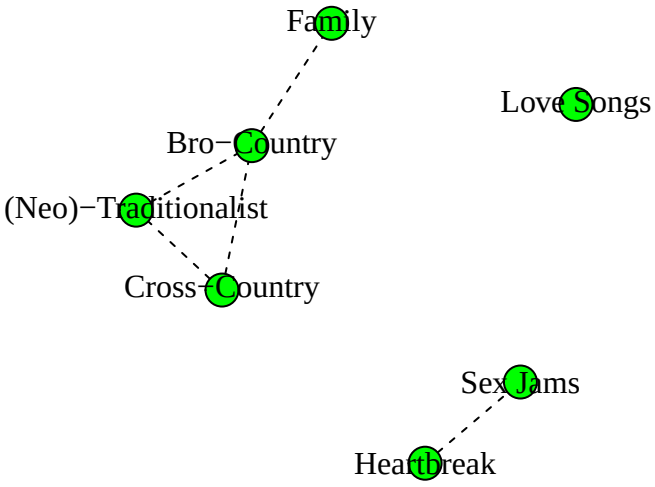
```
## Conducting the nonparanormal (npn) transformation via shrunkun ECDF....done.
## Conducting Meinshausen & Buhlmann graph estimation (mb)....done
## Conducting rotation information criterion (ric) selection....done
## Computing the optimal graph....done
```

```
topic_corr
```

```
## $posadj
## 7 x 7 sparse Matrix of class "dgCMatrix"
##
## [1,] . . . . 1 . .
## [2,] . . 1 1 . . .
## [3,] . 1 . 1 . . .
## [4,] . 1 1 . . . 1
## [5,] 1 . . . . . .
## [6,] . . . . . . .
## [7,] . . . 1 . . .
##
## $poscor
## 7 x 7 sparse Matrix of class "dgCMatrix"
##
## [1,] .          .          .           .           0.03444665 . .
## [2,] .          .          0.138984635 0.052890434 .          . .
## [3,] .          0.13898463 .           0.002650995 .          . .
## [4,] .          0.05289043 0.002650995 .           .          . 0.07017762
## [5,] 0.03444665 .          .           .           .          . .
## [6,] .          .          .           .           .          . .
## [7,] .          .          .           0.070177618 .          . .
##
## $cor
## 7 x 7 Matrix of class "dgeMatrix"
##             [,1]        [,2]         [,3]         [,4]        [,5]       [,6]
## [1,]  0.00000000 -0.34030035 -0.282685092 -0.383999053  0.03444665 -0.1442173
## [2,] -0.34030035  0.00000000  0.138984635  0.052890434  0.00000000 -0.2355104
## [3,] -0.28268509  0.13898463  0.000000000  0.002650995  0.00000000 -0.2104287
## [4,] -0.38399905  0.05289043  0.002650995  0.000000000 -0.41828640 -0.1191443
## [5,]  0.03444665  0.00000000  0.000000000 -0.418286396  0.00000000  0.0000000
## [6,] -0.14421725 -0.23551039 -0.210428750 -0.119144260  0.00000000  0.0000000
## [7,]  0.00000000  0.00000000 -0.132888013  0.070177618 -0.35784909 -0.1596963
##             [,7]
## [1,]  0.00000000
## [2,]  0.00000000
## [3,] -0.13288801
## [4,]  0.07017762
## [5,] -0.35784909
## [6,] -0.15969634
```

```
## [7,]  0.00000000
##
## attr(,"class")
## [1] "topicCorr"
```

```
set.seed(5)
plot(topic_corr,
  vlabels = topic_labels, vertex.label.cex = 1, layout =  layout.auto)
```



Topics 3, 2, 4, 7 are all related. This is an interesting finding! This suggests that traditionalist country especially seems related to both country rock/pop songs Topic 2?: Country Rock/Pop Topic 3: Traditionalist Country Topic 4: Bro-Country Topic 7: Family

More on Topic Models

Questions/Interests

- How would I see where individual artists fell in terms of topics?
- In general, seeing prevalence of certain
- Would it be, taking the top x documents for different topics and counting from there? ### More to Do?
- Plot covariate interaction!
  - Particularly interested in tracking gender * year interactions!