Homework 4

Aaron Coates

5/24/2019

```r
setwd("~/Documents/GitHub/MMSS_311_2")

library(tidytext)
library(tm)
```

```
## Loading required package: NLP
```

```r
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(stringr)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following object is masked from 'package:NLP':
##
##     annotate
```

```r
library(proxy)
```

```
##
## Attaching package: 'proxy'
```

```
## The following objects are masked from 'package:stats':
##
##     as.dist, dist
```

```
## The following object is masked from 'package:base':
##
##     as.matrix
```

```r
library(fields)
```

```
## Loading required package: spam
```

```
## Loading required package: dotCall64
```

```
## Loading required package: grid
```

```
## Spam version 2.2-2 (2019-03-07) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.
```

```
##
## Attaching package: 'spam'
```

```
## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve
```

```
## Loading required package: maps
```

```
## See https://github.com/NCAR/Fields for
##  an extensive vignette, other supplements and source code
```

```r
library(mixtools)
```

```
## mixtools package, version 1.1.0, Released 2017-03-10
## This package is based upon work supported by the National Science Foundation under Grant No. SES-0518
```

```
##
## Attaching package: 'mixtools'
```

```
## The following object is masked from 'package:grid':
##
##     depth
```

```r
library(topicmodels)
library(stm)
```

```
## stm v1.3.3 (2018-1-26) successfully loaded. See ?stm for help.
##  Papers, resources, and other materials at structuraltopicmodel.com
```

1.1

```r
set.seed(732)
Ari <- read_csv("/Users/aaroncoates/Downloads/tx_deathrow_full.csv")
```

```
## Parsed with column specification:
## cols(
##   Execution = col_double(),
##   `Date of Birth` = col_date(format = ""),
##   `Date of Offence` = col_date(format = ""),
##   `Highest Education Level` = col_double(),
##   `Last Name` = col_character(),
##   `First Name` = col_character(),
##   `TDCJ
## Number` = col_double(),
##   `Age at Execution` = col_double(),
##   `Date Received` = col_date(format = ""),
##   `Execution Date` = col_date(format = ""),
##   Race = col_character(),
##   County = col_character(),
##   `Eye Color` = col_character(),
##   Weight = col_double(),
##   Height = col_character(),
##   `Native County` = col_character(),
##   `Native State` = col_character(),
##   `Last Statement` = col_character()
## )
```

```
#Removing those who gave no statement
Ari <- Ari[is.na(Ari$'Last Statement')!=1, ]

Taylor <- Ari %>%
  unnest_tokens(word, 'Last Statement') %>%
  anti_join(stop_words) %>%
  group_by(Execution) %>%
  count(word) %>%
  cast_dtm(Execution, word, n) %>%
  as.matrix()
```

```
## Joining, by = "word"
```

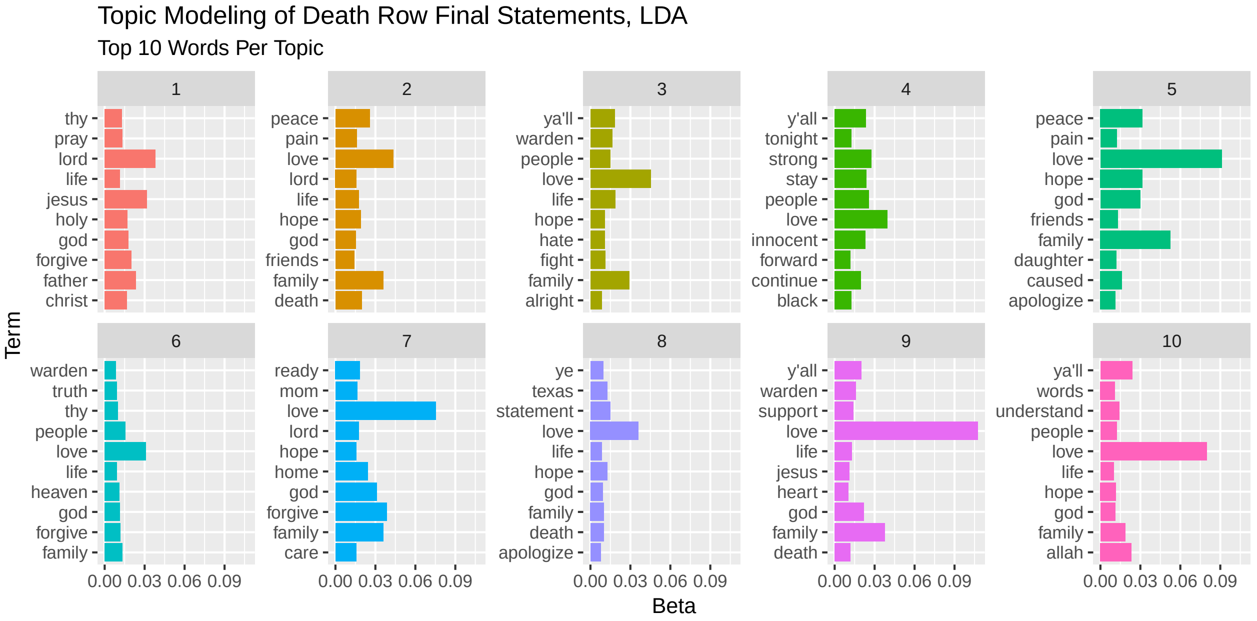
1.2

```
Britney <- LDA(Taylor, k=10)
```

1.3

```
Kelly <- Britney %>%
  tidy() %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup()

CarlyRae <- ggplot(Kelly, aes(term, beta, fill= as.factor(topic))) +
  facet_wrap(~ topic, scales= 'free_y', nrow = 2) + coord_flip() +
  geom_col(show.legend = FALSE) + xlab('Term') + ylab('Beta') +
  labs(title="Topic Modeling of Death Row Final Statements, LDA", subtitle="Top 10 Words Per Topic")
```



2.1

```
Beyonce <- stm::readCorpus(Taylor, type = 'dtm')
```

2.2

```r
#Removing rows that were deleted during pre-processing from the original dataframe
Normani <- as.data.frame(Taylor)
Normani$ExecutionNumber <- as.numeric(rownames(Normani))
Camila <- semi_join(Ari, Normani, by = c('Execution' = 'ExecutionNumber'))

Adele <- stm(documents = Beyonce$documents, vocab = Beyonce$vocab,
              K = 10, prevalence = ~ Race, data = Camila)
```

2.3

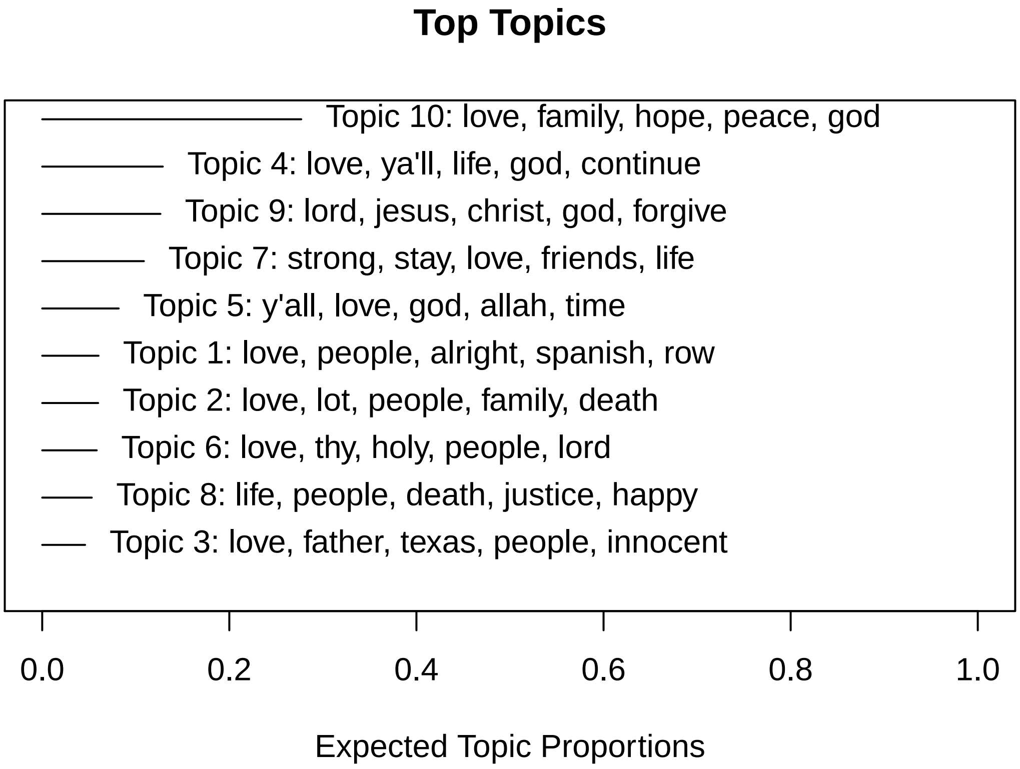
```r
summary(Adele)
```

```
## A topic model with 10 topics, 441 documents and a 3098 word dictionary.

## Topic 1 Top Words:
##       Highest Prob: love, people, alright, spanish, row, tonight, y'all
##       FREX: spanish, alright, stuff, uncle, adams, yall, ya
##       Lift: ernest, otis, grudges, 7, threat, hire, cruz
##       Score: adams, donna, spanish, house, uncle, ya, stuff
## Topic 2 Top Words:
##       Highest Prob: love, lot, people, family, death, life, heart
##       FREX: mexico, lot, guess, eardmann, seek, hatred, shown
##       Lift: caring, cadena, clayton, mitchell, party, dell, lasted
##       Score: eardmann, mexico, clayton, mitchell, party, seek, guess
## Topic 3 Top Words:
##       Highest Prob: love, father, texas, people, innocent, ye, world
##       FREX: ye, texas, found, rejoice, gene, pinkerton, woe
##       Lift: rejoice, showered, chamber, entered, gene, pinkerton, recited
##       Score: ye, woe, process, gene, pinkerton, beatitudes, weep
## Topic 4 Top Words:
##       Highest Prob: love, ya'll, life, god, continue, innocent, care
##       FREX: ya'll, fight, learn, continue, kids, sisters, supporters
##       Lift: kindness, judgment, deep, doug, joey, 39, errors
##       Score: ya'll, fight, prophet, learn, continue, bury, allah
## Topic 5 Top Words:
##       Highest Prob: y'all, love, god, allah, time, life, father
##       FREX: boswell, y'all, smile, allah, officers, asdadu, accident
##       Lift: mumbled, returns, population, tdcj, official, earlier, absolutely
##       Score: y'all, allah, boswell, smile, asdadu, officer, returns
## Topic 6 Top Words:
##       Highest Prob: love, thy, holy, people, lord, heaven, forgive
##       FREX: thy, lynching, marching, holy, america, black, green
##       Lift: avenged, executions, lambert, march, revolution, tapes, dow
##       Score: thy, lynching, black, holy, thou, america, marching
## Topic 7 Top Words:
##       Highest Prob: strong, stay, love, friends, life, statement, truth
##       FREX: stay, strong, mother, dna, positive, truth, supporting
##       Lift: altered, records, hour, linda, pam, cool, mouthed
##       Score: strong, stay, mother, positive, dna, altered, records
## Topic 8 Top Words:
##       Highest Prob: life, people, death, justice, happy, ready, call
##       FREX: ride, polunsky, call, glad, answers, phone, dungeon
##       Lift: carl, powell, america's, anybody's, blooded, blow, campo
##       Score: dungeon, joanna, blooded, equal, hollered, answers, polunsky
## Topic 9 Top Words:
```

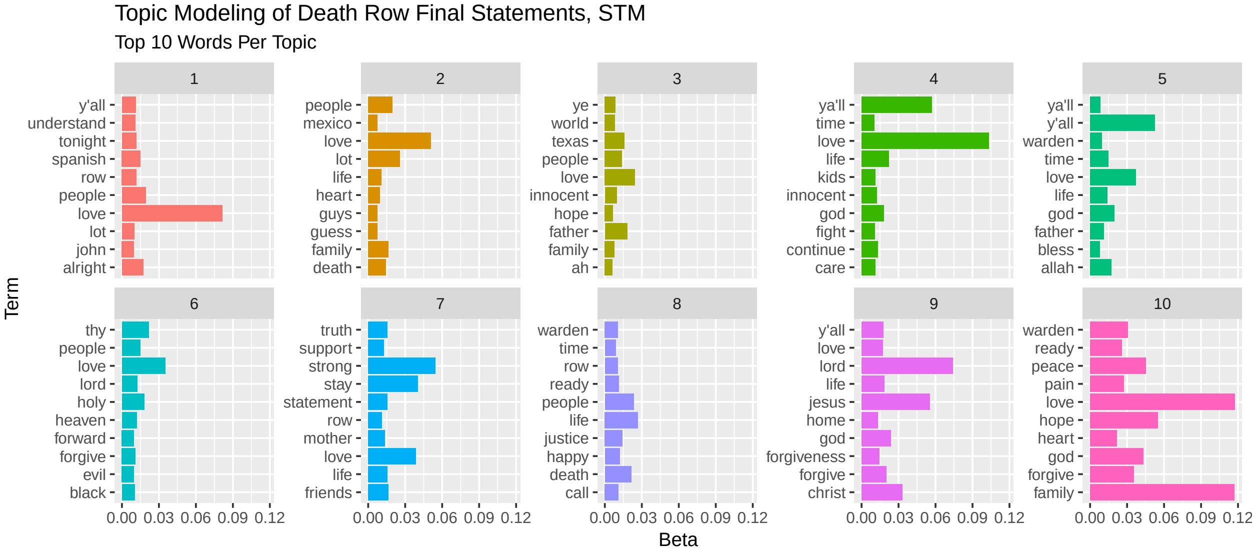
```
##       Highest Prob: lord, jesus, christ, god, forgive, life, y'all
##       FREX: christ, jesus, lord, sins, mama, praise, rest
##       Lift: saving, disappointment, closer, counts, remains, allah's, texans
##       Score: christ, lord, jesus, y'all, sins, gomez, veronica
## Topic 10 Top Words:
##       Highest Prob: love, family, hope, peace, god, forgive, warden
##       FREX: peace, hope, family, pain, apologize, victim's, ready
##       Lift: closeness, bruce, causing, kami, ahh, background, educated
##       Score: peace, hope, pain, ready, apologize, forgive, victim's
```

```
plot.STM(Adele, type = "summary", n=5, xlim=c(0,1))
```



```
Miley <- tidy(Adele) %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup()

Demi <- ggplot(Miley, aes(term, beta, fill= as.factor(topic))) +
  facet_wrap(~ topic, scales= 'free_y', nrow = 2) + coord_flip() +
  geom_col(show.legend = FALSE) + xlab('Term') + ylab('Beta') +
  labs(title="Topic Modeling of Death Row Final Statements, STM", subtitle="Top 10 Words Per Topic")
```



2.4

The topics we find when conditioning on race are somewhat different from those we find when using traditional LDA. For instance, using traditional LDA, the 10 topics are not easily differentiated. For instance, the themes of “love”, “god”, and “apologize” are prevalent in most, if not all, of the topics. The only distinct topic seems to be topic 1, where we see a focus on police, crime, and the law.

When using STM and conditioning on race, the topics are mostly similar, as we see a dominance of topics relating to “love”, “god”, and “apologize”. However, the results are slightly different because topics now feature certain words that may be distinct among certain regions and ethnicities. For instance, we see “Mexico”, “Allah”, and “black” appear in the top 10 lists for certain topics. This helps to illuminate certain topics and words that members of distinct races may have mentioned in their last statements.