Homework 4

Tina Zhang

2019/5/20

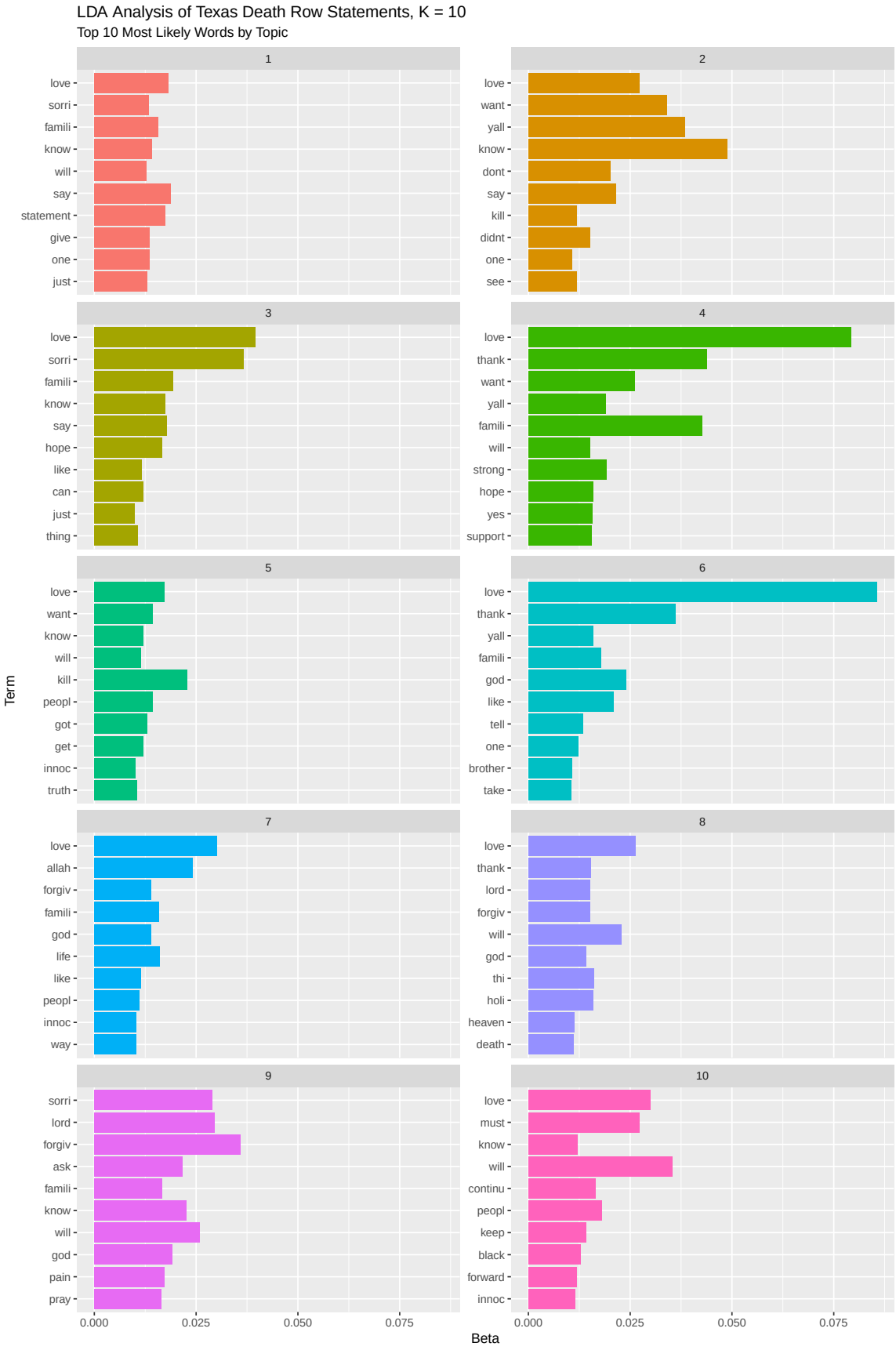
```
knitr::opts_chunk$set(echo = TRUE)
packages <- c("readr","proxy","dplyr","tidytext","ggplot2","SnowballC","topicmodels",
              "stm","tidyr","mixtools","tm")
load.packages <- function(x) {
  if (!require(x, character.only = TRUE)) {
    install.packages(x, dependencies = TRUE)
    library(x, character.only = TRUE)
  }
}
lapply(packages, load.packages)
setwd("C:/Users/tzwhi/Desktop/Northwestern/311-2")
```

Question 1

```
#1
inmates <- read.csv("tx_deathrow_full.csv", stringsAsFactors = FALSE, encoding = "UTF-8") %>%
  mutate(doc_id = row_number(), text = Last.Statement) %>%
  select(doc_id, text, everything())
corp <- VCorpus(DataframeSource(inmates)) %>%
  tm_map(removePunctuation) %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(removeWords, stopwords("english")) %>%
  tm_map(stemDocument)
dtm <- DocumentTermMatrix(corp)
empty_rows <- which(rowSums(as.matrix(dtm))==0)
dtm <- corp[-empty_rows] %>% DocumentTermMatrix()

#2
lda <- LDA(dtm, k=10, control = list(seed=25))

#3
tidy(lda) %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  mutate(term = reorder(term, beta)) %>%
  ggplot(aes(term, beta, fill=factor(topic))) + geom_col(show.legend = FALSE) +
  facet_wrap(~topic, scales = "free_y", nrow = 5) + coord_flip() +
  xlab("Term") + ylab("Beta") +
  labs(title = "LDA Analysis of Texas Death Row Statements, K = 10",
                       subtitle = "Top 10 Most Likely Words by Topic")
```


LDA Analysis of Texas Death Row Statements, K = 10
Top 10 Most Likely Words by Topic
1
love
sorri
famili
know
will
say
statement
give
one
just
2
love
want
yall
know
dont
say
kill
didnt
one
see
3
love
sorri
famili
know
say
hope
like
can
just
thing
4
love
thank
want
yall
famili
will
strong
hope
yes
support
5
love
want
know
will
kill
peopl
got
get
innoc
truth
6
love
thank
yall
famili
god
like
tell
one
brother
take
7
love
allah
forgiv
famili
god
life
like
peopl
innoc
way
8
love
thank
lord
forgiv
will
god
thi
holi
heaven
death
9
sorri
lord
forgiv
ask
famili
know
will
god
pain
pray
10
love
must
know
will
continu
peopl
keep
black
forward
innoc
0.000
0.025
0.050
0.075
Term
Beta

Question 2

```
#1
out <- stm::readCorpus(dtm, type = "slam")

#2
mod.out <- stm(documents = out$documents, vocab = out$vocab, K = 10,
               prevalence = ~Race, data = inmates[-empty_rows, ])
```

```
#3
summary(mod.out)
```

```
## A topic model with 10 topics, 442 documents and a 2722 word dictionary.

## Topic 1 Top Words:
##       Highest Prob: like, famili, one, good, said, will, statement
##       FREX: like, goodby, regret, statement, said, true, last
##       Lift: approach, argu, asid, bicker, brazzil, butterfli, cemeteri
##       Score: regret, goodby, like, doug, true, buri, famili
## Topic 2 Top Words:
##       Highest Prob: lord, forgiv, god, jesus, ask, life, know
##       FREX: christ, lord, jesus, pray, home, pleas, ask
##       Lift: 1991, dell, devil, heal, jenni, kyle, roman
##       Score: jesus, christ, lord, forgiv, sin, home, pleas
## Topic 3 Top Words:
##       Highest Prob: love, yall, know, want, tell, take, dont
##       FREX: yall, care, stay, strong, tell, kid, didnt
##       Lift: 2003, abus, amenia, amigo, ashle, aunti, awe
##       Score: yall, strong, care, stay, love, didnt, kid
## Topic 4 Top Words:
##       Highest Prob: famili, sorri, hope, will, can, say, pain
##       FREX: sorri, apolog, pain, hope, caus, victim, famili
##       Lift: €em, afraid, amanda, ambit, amount, audrey, barbado
##       Score: sorri, bye, famili, hope, pain, apolog, forgiv
## Topic 5 Top Words:
##       Highest Prob: thank, love, want, support, friend, life, say
##       FREX: thank, support, happi, spanish, receiv, spiritu, spirit
##       Lift: coldblood, dee, equal, greater, holler, kin, leo
##       Score: thank, mexican, coldblood, equal, holler, remuner, spiritu
## Topic 6 Top Words:
##       Highest Prob: allah, will, thi, forgiv, holi, god, love
##       FREX: thi, lead, trespass, allah, holi, thou, amen
##       Lift: anointest, arnott, bread, bump, chantal, damien, green
##       Score: allah, holi, thi, thou, trespass, unto, lead
## Topic 7 Top Words:
##       Highest Prob: peac, love, find, heart, death, will, hope
##       FREX: peac, chang, closur, find, texa, heart, row
##       Lift: aunt, commend, herebi, intend, join, pastor, posit
##       Score: peac, find, dungeon, texa, aunt, joanna, herebi
## Topic 8 Top Words:
##       Highest Prob: love, will, now, know, bless, come, shall
##       FREX: final, prophet, your, veronica, unintellig, shall, alon
```

```
##       Lift: beatitud, behold, corinthian, propheci, veronica, weep, 1231b
##       Score: final, veronica, prophet, shall, chapter, woe, gomez
## Topic 9 Top Words:
##       Highest Prob: innoc, that, got, one, peopl, get, just
##       FREX: nobodi, got, innoc, ahead, chanc, that, evid
##       Lift: boswel, complet, dealer, hawthorn, hunt, job, joke
##       Score: boswel, evid, nobodi, price, job, alter, got
## Topic 10 Top Words:
##       Highest Prob: will, must, peopl, love, kill, know, keep
##       FREX: black, lynch, march, must, forward, america, liber
##       Lift: 100, liber, suit, tight, 180, 27th, 300
##       Score: black, lynch, march, must, america, forward, liber
```

4. The topics found when conditioning on venue appear more differentiated than those found using standard LDA. For example, topics 8 and 5 appear tied to thankfulness and love, while topics 2 and 6 appear tied to religion and forgiveness. In contrast, there's a lot of overlap in the top words for each topic in LDA, and it's harder to identify what each topic is about. For example, the word "love" appears in 9 of the 10 topics, while words like "god", "forgiv", and "sorri", and "thank" also appear in many of the topics.