Untitled

TA: Leslie Huang

Course: Text as Data

Date: 4/18/2017

Recitation 11: Unsupervised Learning IIb

```
rm(list=ls())

setwd("/Users/Lingyi/TAD/lab/Text-as-Data-Lab-Spr2018/W11_04_12_18/")

set.seed(1234)

# Possibly new packages to install
#install.packages("lattice")
#install.packages("bursts")
#install.packages("tidytext")
```

```
libraries <- c("Matrix", "ldatuning", "topicmodels", "readtext", "dplyr", "stm", "quanteda", "lda", "bu
lapply(libraries, require, character.only = T)
```

```
## Loading required package: Matrix
## Loading required package: ldatuning
## Loading required package: topicmodels
## Loading required package: readtext
## Loading required package: 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
## Loading required package: stm
## stm v1.3.3 (2018-1-26) successfully loaded. See ?stm for help.
##  Papers, resources, and other materials at structuraltopicmodel.com
## Loading required package: quanteda
## quanteda version 1.0.0
## Using 3 of 4 threads for parallel computing
```

```
##
## Attaching package: 'quanteda'

## The following object is masked from 'package:utils':
##
##     View

## Loading required package: lda

## Loading required package: bursts

## Loading required package: tidytext

## Warning: package 'tidytext' was built under R version 3.4.4

## Loading required package: ggplot2

## Loading required package: lattice

##
## Attaching package: 'lattice'

## The following object is masked from 'package:stm':
##
##     cloud

## Loading required package: quanteda.corpora

## [[1]]
## [1] TRUE
##
## [[2]]
## [1] TRUE
##
## [[3]]
## [1] TRUE
##
## [[4]]
## [1] TRUE
##
## [[5]]
## [1] TRUE
##
## [[6]]
## [1] TRUE
##
## [[7]]
## [1] TRUE
##
## [[8]]
## [1] TRUE
##
## [[9]]
## [1] TRUE
##
## [[10]]
## [1] TRUE
##
## [[11]]
```

```
## [1] TRUE
##
## [[12]]
## [1] TRUE
##
## [[13]]
## [1] TRUE
```

1 Correlated Topic Models (CTM)

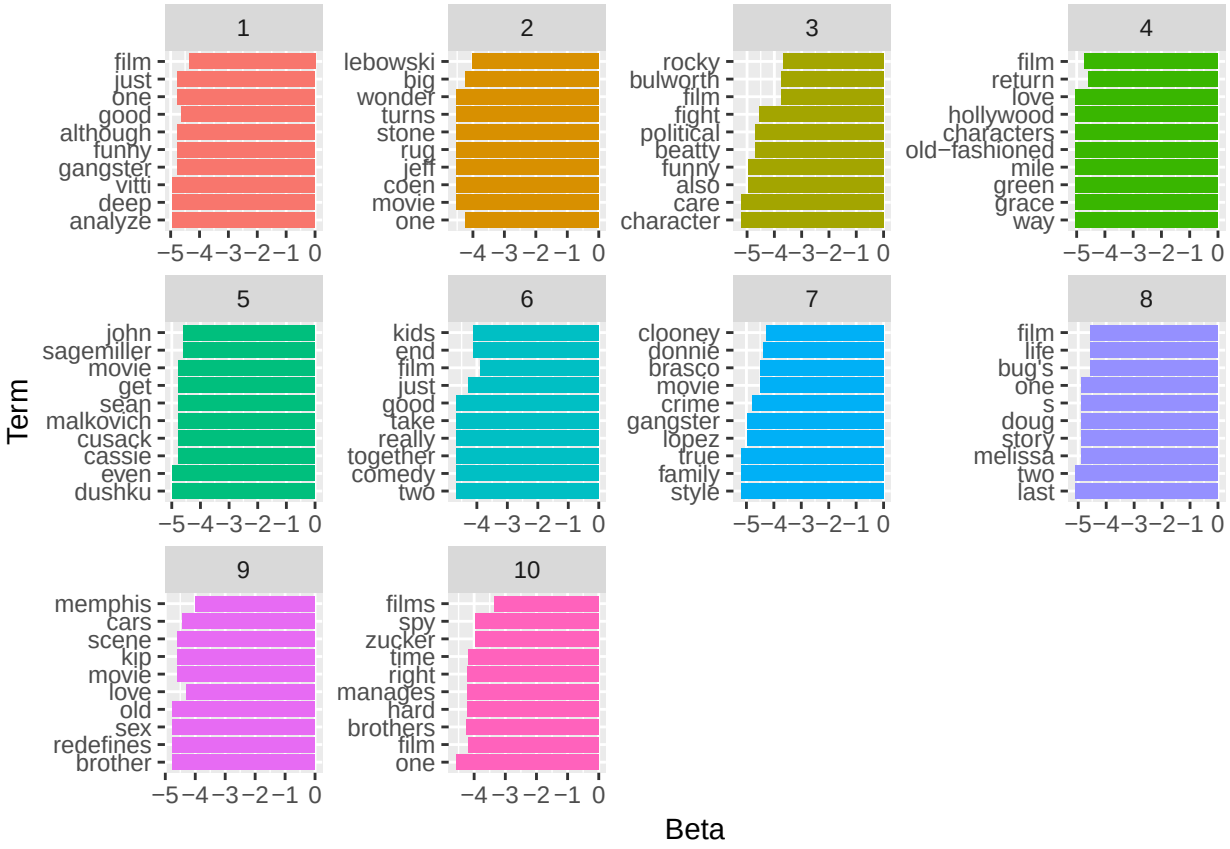
```
# What is a CTM?

# Loading the data and creating a DFM
data("data_corpus_movies", package = "quanteda.corpora")

movies_corp <- corpus_sample(data_corpus_movies, size = 20)

movies_dfm <- dfm(movies_corp, remove = stopwords("english"), remove_punct = T)

# FYI, this function takes a long time to run
as_ctm <- CTM(movies_dfm, k = 10, control = list(seed = 1234))

# Summarization
betas <- t(as_ctm@beta)

words <- as_ctm@terms
```

```
# Gets the top terms
get_terms(as_ctm,10)
```

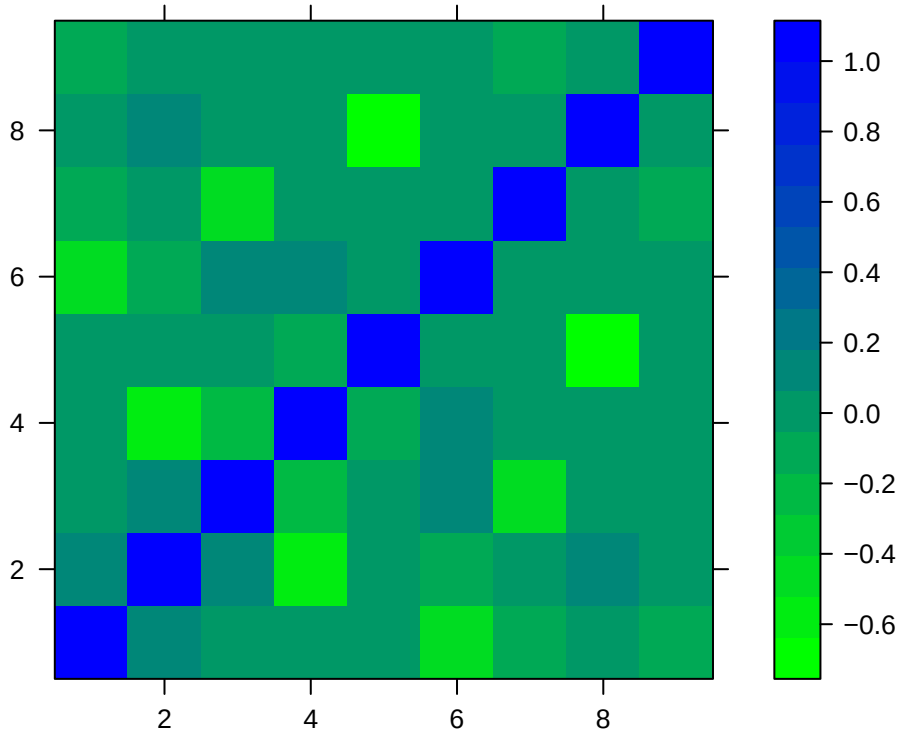
```
##       Topic 1    Topic 2    Topic 3     Topic 4         Topic 5
##  [1,] "film"     "lebowski" "rocky"     "return"        "john"
##  [2,] "good"     "big"      "bulworth"  "film"          "sagemiller"
##  [3,] "just"     "one"      "film"      "hollywood"     "get"
##  [4,] "although" "wonder"   "fight"     "characters"    "malkovich"
##  [5,] "gangster" "turns"    "political" "love"          "cusack"
##  [6,] "one"      "rug"      "beatty"    "old-fashioned" "cassie"
##  [7,] "funny"    "coen"     "also"      "grace"         "sean"
##  [8,] "analyze"  "jeff"     "funny"     "green"         "movie"
##  [9,] "vitti"    "stone"    "care"      "mile"          "even"
## [10,] "deep"     "movie"    "character" "way"           "dushku"
##       Topic 6    Topic 7    Topic 8   Topic 9     Topic 10
##  [1,] "film"     "clooney"  "life"    "memphis"   "films"
##  [2,] "kids"     "donnie"   "bug's"   "love"      "spy"
##  [3,] "end"      "brasco"   "film"    "cars"      "zucker"
##  [4,] "just"     "movie"    "s"       "scene"     "film"
##  [5,] "good"     "crime"    "doug"    "kip"       "time"
##  [6,] "take"     "lopez"    "story"   "movie"     "right"
##  [7,] "two"      "gangster" "melissa" "old"       "manages"
##  [8,] "really"   "true"     "one"     "sex"       "hard"
##  [9,] "together" "family"   "last"    "brother"   "brothers"
## [10,] "comedy"   "style"    "two"     "redefines" "one"
```

```
# Since tidytext doesn't work with CTM, I manually extract the top 10 terms for each topic with this fu
df <- lapply(1:ncol(betas), function(x){data.frame(Topic = x, Term = words, Beta = betas[,x]) %>% arrang

df_ctm <- bind_rows(df)
```

```
df_ctm %>%
  mutate(Term = reorder(Term, Beta)) %>%
  ggplot(aes(Term, Beta, fill = factor(Topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ Topic, scales = "free") +
  coord_flip()
```




```
# Sigma is the Variance Covariance Matrix of the all Topics
topic_var_matrix<-as_ctm@Sigma
topic_cor_matrix <- cov2cor(topic_var_matrix)

#Visualizing correlation matrix between topics
rgb.palette <- colorRampPalette(c("green", "blue"), space = "rgb")
levelplot(topic_cor_matrix, xlab="", ylab="", col.regions=rgb.palette(120))
```



2 Structural Topic Models (STM)

```
# What is an STM?

# Loading data: Political blogs from the 2008 election on a conservative-liberal dimension

data(poliblog5k)

head(poliblog5k.meta)
```

```
##            rating day blog
## 6787 Conservative 182   ha
## 8150 Conservative 299   ha
## 4961      Liberal 345   db
## 5767 Conservative  90   ha
## 2803 Conservative 321   at
## 2359 Conservative 271   at
##                                                 text
## 6787 How happy do you think Team Barry is that, thanks
## 8150  Tough stuff, but conservative passions this year
## 4961 Epic Ideological Failby digbyBefore you listen to
## 5767    Excellent work as usual from a guy who s never
## 2803   The headline at the Los Angeles Times blog says
## 2359 To those of us of a certain age, Paul Newman will
```

```
head(poliblog5k.voc)
```

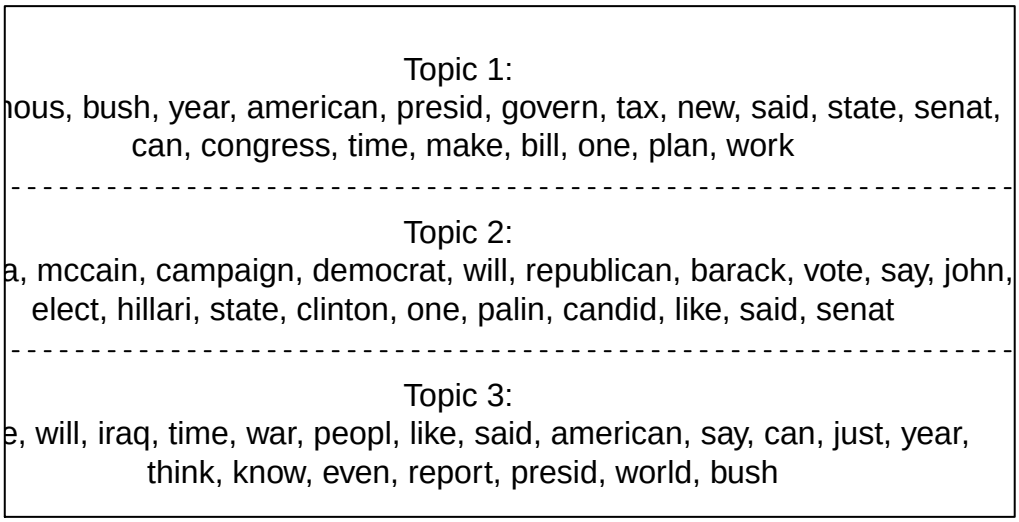
```
## [1] "abandon" "abc"     "abil"    "abl"     "abort"   "abroad"
```

```r
# Fits an STM model with 3 topics
blog_stm <- stm(poliblog5k.docs, poliblog5k.voc, 3,
            prevalence=~rating + s(day), data=poliblog5k.meta)
```

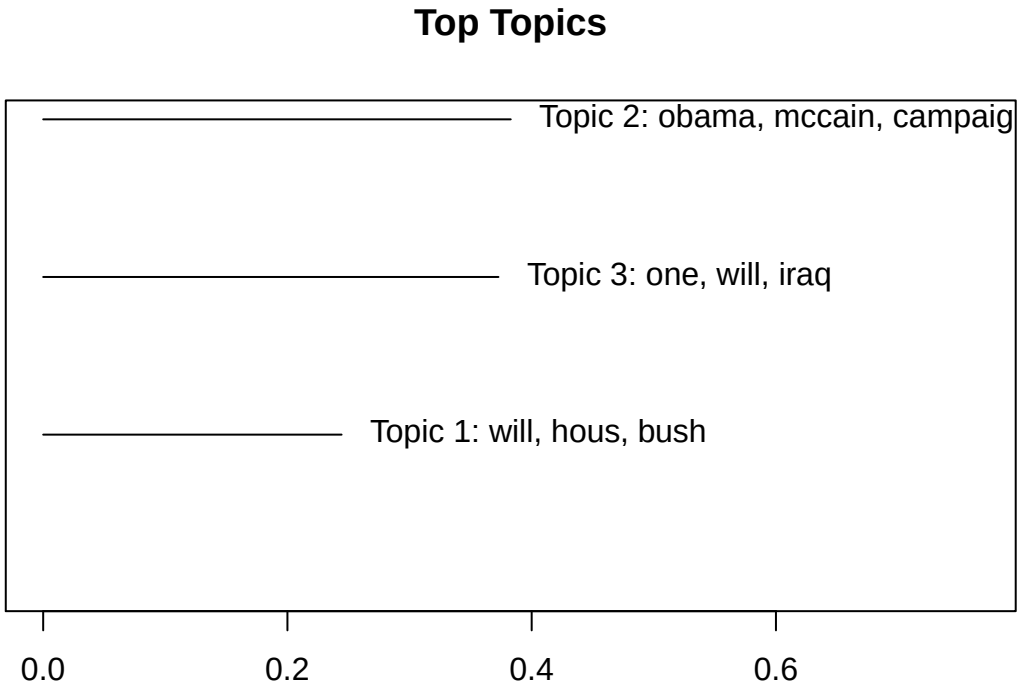
```
## Beginning Spectral Initialization
##   Calculating the gram matrix...
##   Finding anchor words...
##      ...
##   Recovering initialization...
##      ......................
## Initialization complete.
## ..................................................................................................
## Completed E-Step (1 seconds).
## Completed M-Step.
## Completing Iteration 1 (approx. per word bound = -7.197)
## ..................................................................................................
## Completed E-Step (1 seconds).
## Completed M-Step.
## Completing Iteration 2 (approx. per word bound = -7.125, relative change = 9.980e-03)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 3 (approx. per word bound = -7.109, relative change = 2.154e-03)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 4 (approx. per word bound = -7.104, relative change = 7.021e-04)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 5 (approx. per word bound = -7.102, relative change = 3.116e-04)
## Topic 1: will, bush, hous, presid, year
##  Topic 2: obama, mccain, campaign, will, democrat
##  Topic 3: one, will, iraq, time, like
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 6 (approx. per word bound = -7.101, relative change = 1.590e-04)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 7 (approx. per word bound = -7.100, relative change = 8.800e-05)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 8 (approx. per word bound = -7.100, relative change = 5.155e-05)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 9 (approx. per word bound = -7.100, relative change = 3.140e-05)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 10 (approx. per word bound = -7.100, relative change = 1.964e-05)
```

```
## Topic 1: will, bush, hous, year, presid
##  Topic 2: obama, mccain, campaign, democrat, will
##  Topic 3: one, will, iraq, time, war
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Completing Iteration 11 (approx. per word bound = -7.100, relative change = 1.256e-05)
## ..................................................................................................
## Completed E-Step (0 seconds).
## Completed M-Step.
## Model Converged
```

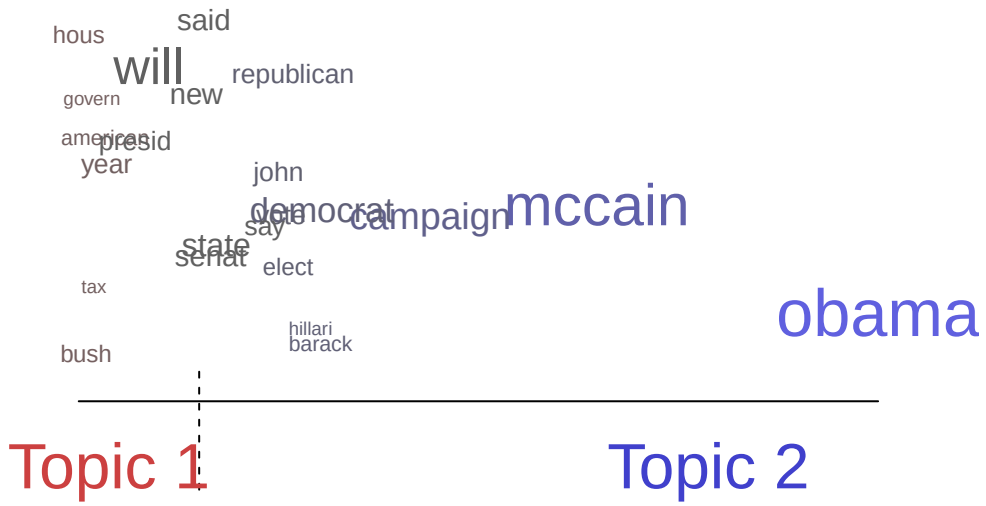
```r
# A plot that summarizes the topics by what words occur most commonly in them
plot(blog_stm,type="labels")
```



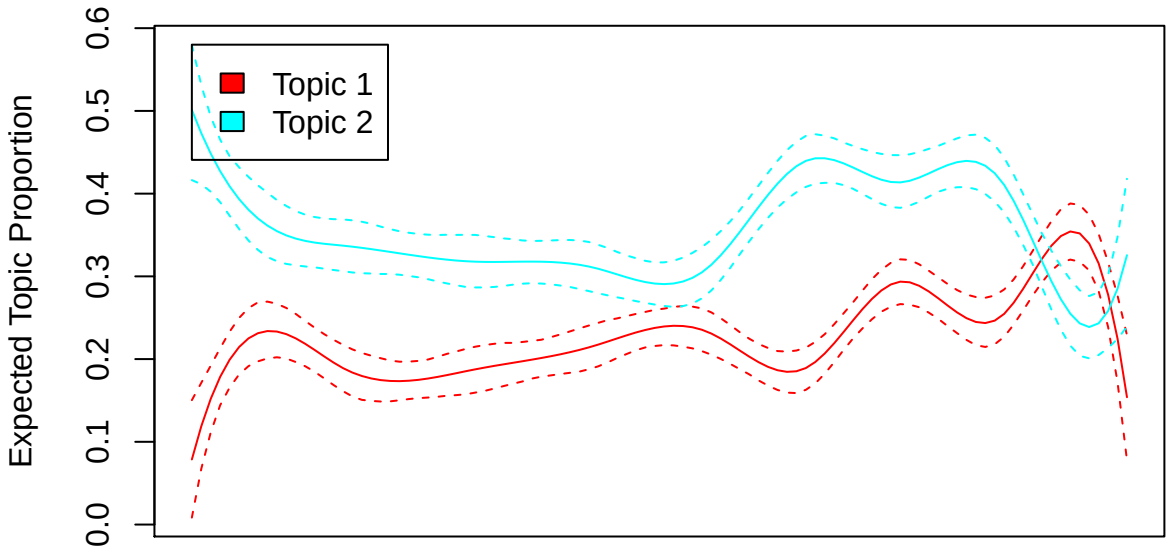

```r
# A summary plot of the topics that ranks them by their average proportion in the corpus
plot(blog_stm, type="summary")
```




```
# A visualization of what words are shared and distinctive to two topics
plot(blog_stm, type="perspectives", topics=c(1,2))
```




```
# Estimates a regression with topics as the dependent variable and metadata as the independent variable
prep <- estimateEffect(1:3 ~ rating + s(day) , blog_stm, meta=poliblog5k.meta)

# Plots the distribution of topics over time
plot(prep, "day", blog_stm, topics = c(1,2),
     method = "continuous",xaxt = "n", xlab = "Date")
```



Date

```
# Plots the Difference in coverage of the topics according to liberal or conservative ideology
plot(prep, "rating", model=blog_stm,
     method="difference",cov.value1="Conservative",cov.value2="Liberal")
```