Text Analysis Project

Lee, Woo Chan

11/10/2021

LDA (Topic modeling)

```r
# Load full twitter dataset grouped by tickers and quarterly (date)
twt_q <- read_csv("/Users/lee14257/Development/CMU/Text Analysis/Project/CBE2/twt_ticker_quarter.csv")
```

```
## Rows: 50 Columns: 3

## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (3): ticker_symbol, post_date, body

##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Create new column doc_id, which represents the ticker symbol + date
twt_q$doc_id <- paste0(twt_q$ticker_symbol, "_", twt_q$post_date)
twt_q <- twt_q %>%
  rename('text' = 'body')
```

```r
# Subset the MSFT and TSLA tickers
twt_msft <- twt_q %>% subset(ticker_symbol == "MSFT")
twt_tsla <- twt_q %>% subset(ticker_symbol == "TSLA")
```

LDA for TSLA

```r
# Create token object for TSLA
twt_tsla_tkn <- twt_tsla %>%
  corpus() %>%
  tokens(what="fastestword", remove_punct = TRUE, remove_symbols = TRUE,
         remove_numbers=TRUE, remove_url=TRUE, remove_separators=TRUE,
         split_hyphens=TRUE
          ) %>%
  tokens_remove(c('\\$[a-z0-9]+', '\\#[a-z0-9]+', '[0-9]+\\%', '\\@[a-z0-9]+'),
                valuetype='regex') %>%
```

```r
  tokens_remove(c(stopwords("english"), "tsla", "tesla",
                "tesla's", "btindle:", "200:1", "10:45", "w/code", "4x,",
                "5x,", "leech-boy"))
```

```r
# Create dfm for TSLA
# Define min_termfreq and max_termfreq to restrict dfm
twt_tsla_dfm <- twt_tsla_tkn %>% dfm() %>%
              dfm_trim(min_termfreq = 30, max_termfreq = 85)
```

```r
# LDA for TSLA
set.seed(2023)
tsla_lda <- textmodel_lda(twt_tsla_dfm, k = 6)

# Overview of top 30 words for each topic
tsla_30 <- as.data.frame(terms(tsla_lda, 30))
```

```r
# Print top words for each topics
print(tsla_30)
```

```
##                topic1        topic2               topic3          topic4
## 1                 itb        unroll            nowfunded      stockguy22
## 2                swks        dallas                com'g           2006.
## 3          priceclick          pdt,                  bks positivestocks:
## 4           sizeclick   retaliation    cryptocurrencies           plan'
## 5      deteriorating.      pmsource:              platts         'master
## 6            skyworks        (thanks             yearend       linkfest:
## 7               ipath        boeing,             bigauto tesla/solarcity
## 8                 xle        barrie     immaterialscale          trend:
## 9                 tbt    chartwatch         downsideagm        lowfloat
## 10           ultrapro     mortgaged                  dd:     supernovapt
## 11               jnug      sequence              moresee           lol:d
## 12                xlk      "leaked"              2018=>          merger.
## 13                jnk            xi                   6'        stochrsi:
## 14        supertrades         buggy                dirt.          10-day
## 15        sentiquant:         harms             3x/-3x     callputratio
## 16                slv artkocapital:             2018hiv      jones2000:
## 17               f/v.      closures          congressman          plan':
## 18                chk           nl:          usualmodel3        highread
## 19                gld        faking            services:            6-7.
## 20               7-10     repayment              keybanc       jimmybob:
## 21           lol.....      blocked.           effective,           site!
## 22                70d     narrator:          anymorewill         \\u2026
## 23            high...       belgium scale&profitability             hod,
## 24               ftse           12/              dummest         merger,
## 25        changeclick            /w                  jpn         e*trade
## 26           members!      webinar,            downturns          norman
## 27                dia           /es               burn'g         brodeur
```

```
## 28              tvix       bonuses    hopeless!funding         hedges:
## 29 freeport-mcmoran,        wheels.           solarthat         mclaren
## 30         bosocial:   recognizing             keycorp       classical
##        topic5        topic6
## 1      folks!         laws.
## 2       wins.       threats
## 3     jealous           ol'
## 4       mars?         bets.
## 5    snapshot        quikfo
## 6         2x,   statements,
## 7        310c    elsewhere.
## 8     today!!    directors.
## 9  presenting       elon...
## 10   in-depth      finance,
## 11      churn         (disc:
## 12       guru embarrassment
## 13     modify         duped
## 14        3pm          309.
## 15    pennant      dealbook:
## 16  breakout,          puke
## 17   rounding      retained
## 18  bloodbath    converting
## 19   falling.     humanity.
## 20      heres          tall
## 21   breached          reps
## 22    coolest        coffin
## 23         6.     anecdotal
## 24       up..   misleading.
## 25        lag          doj,
## 26       ripe        kindly
## 27      mazda     frequency
## 28  porsche's  productivity
## 29       cagr        amazes
## 30       iihs        duties
```

```r
# Store significant / relevant words in tsla_30 in vector form
topic_composition_tsla <- data.frame(topic_num = NA, words = NA)
topic_composition_tsla[1,] <- c("Topic 1", "[ deteriorating, priceclick,
                                itb, supertrades, xle ]")
topic_composition_tsla[2,] <- c("Topic 4", "[ positivestocks, tesla/solarcity,
                                merger, plan, hedges]")
topic_composition_tsla[3,] <- c("Topic 5", "[ mars, breached, bloodbath,
                                falling, breakout ]")
topic_composition_tsla[4,] <- c("Topic 6", "[ laws, threats, embarassment,
                                elon, statements ]")
```

Table 1: Topic Composition for TSLA

Topics	Key Tokens
Topic 1	[deteriorating, priceclick, itb, supertrades, xle]
Topic 4	[positivestocks, tesla/solarcity, merger, plan, hedges]
Topic 5	[mars, breached, bloodbath, falling, breakout]
Topic 6	[laws, threats, embarassment, elon, statements]

```r
# Assign each doc_id to the topics
data.frame(doc_id = twt_tsla$doc_id, Topic = topics(tsla_lda))
```

```
##         doc_id  Topic
## 1  TSLA_2015_1 topic1
## 2  TSLA_2015_2 topic1
## 3  TSLA_2016_1 topic1
## 4  TSLA_2016_2 topic4
## 5  TSLA_2017_1 topic4
## 6  TSLA_2017_2 topic5
## 7  TSLA_2018_1 topic5
## 8  TSLA_2018_2 topic6
## 9  TSLA_2019_1 topic6
## 10 TSLA_2019_2 topic6
```

Plot topic in the time series graph for TSLA

```r
# Read in docuscope-tagged dfm, and filter TSLA
tsla_docuscope <- read_csv("/Users/lee14257/Development/CMU/Text Analysis/Project/CBE2/twt_docuscope_no
  filter(ticker == "TSLA")
```

```
## Rows: 300 Columns: 41
```

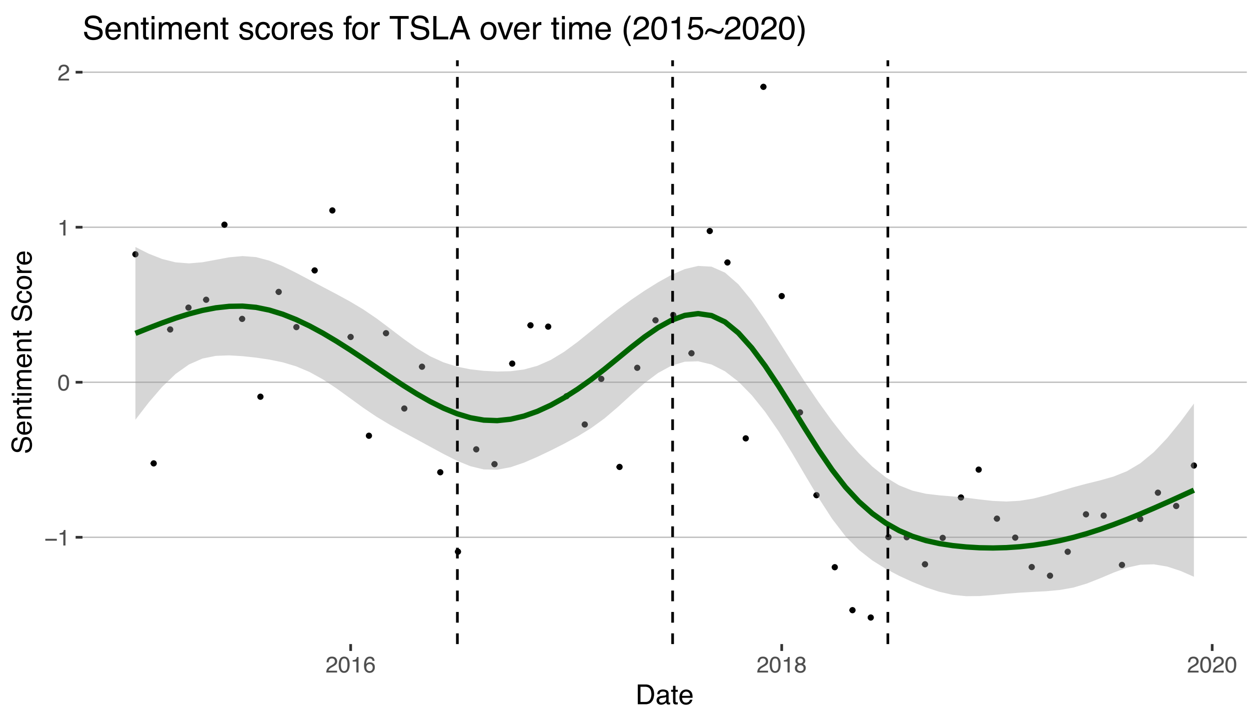
```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): ticker, doc_id
## dbl (39): year, academicterms, academicwritingmoves, character, citation, ci...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Transform dfm to feed to ggplot
tsla_sentiment <- tsla_docuscope %>%
  mutate(
    ticker_symbol = str_extract(doc_id, "^[A-Z]+"),
    date = as.Date(paste0(word(doc_id, 2, sep = "_"), '-01'), format='%Y-%m-%d')
    ) %>%
```

```
  dplyr::select(ticker_symbol, date, sentiment_score) %>%
  filter(date >= "2015-01-01")
```

```
# Graphing the time series plot for TSLA
ggplot(tsla_sentiment, aes(x=date, y=sentiment_score)) +
    geom_point(size = .5) +
    geom_smooth(method = "gam", formula = y ~ s(x, bs = "cs"), size=1,
                level=0.95, se=T, colour="darkgreen") +
    labs(x="Date", y = "Sentiment Score",
         title="Sentiment scores for TSLA over time (2015~2020)")+
    theme(panel.grid.minor.x=element_blank(),
          panel.grid.major.x=element_blank()) +
    theme(panel.grid.minor.y=element_blank(),
          panel.grid.major.y=element_line(colour = "gray",size=0.25)) +
    theme(rect = element_blank()) +
    theme(legend.title=element_blank()) +
    geom_vline(xintercept = c(ymd("2016/06/30"),
                              ymd("2017/06/30"),
                              ymd("2018/06/30")), linetype = 2)
```



LDA for MSFT

```
# Create token for MSFT
twt_msft_tkn <- twt_msft %>%
  corpus() %>%
  tokens(what="fastestword", remove_punct = TRUE, remove_symbols = TRUE,
         remove_numbers=TRUE, remove_url=TRUE, remove_separators=TRUE,
```

```r
          split_hyphens=TRUE
           ) %>%
  tokens_remove(c('\\$[a-z0-9]+', '\\#[a-z0-9]+', '[0-9]+\\%', '\\@[a-z0-9]+'),
                valuetype='regex') %>%
  tokens_remove(c(stopwords("english"), "microsoft", "microsoft's", "msft"))
```

```r
# Create dfm for msft
twt_msft_dfm <- twt_msft_tkn %>% dfm() %>%
              dfm_trim(min_termfreq = 25, max_termfreq = 95)
```

```r
# LDA model
set.seed(222)
msft_lda <- textmodel_lda(twt_msft_dfm, k = 6)

# Overview of top 30 words for each topic
msft_30 <- as.data.frame(terms(msft_lda, 30))
```

```r
# Print top words for each topics
print(msft_30)
```

```
##          topic1       topic2     topic3          topic4          topic5
## 1           opt   (otc:hiph)     racist            lite      revitalize
## 2    tweaktown:      mktloss       cnet            ban.           cboe:
## 3      strangle    nowfunded apologizes       crackdown       bosocial:
## 4          llc;  alzheimer's    appeals            tata              ($
## 5      dominion     lobbying     cable. vulnerabilities          hello,
## 6        parent     pattern.       wand        'project measureschart:
## 7         also,  discovering     grants    installation       saturday,
## 8     partnered     quarters   undersea           remix        analyze:
## 9      outsells     reuters:      like,          geneva             11,
## 10         (min   'dreamers'    foxconn          africa         roundup
## 11     settling markfidelman     turner         laptop,      declining.
## 12      charles    aol&yahoo  linkedin:      chromebook       document,
## 13     however,   premarket:       sues            pro,           2015:
## 14         pete    immigrant   swiftkey             11.            8:00
## 15     waverton        flow:      fable        hexadite         fading.
## 16    pressured        1962,     ruling     marketplace deteriorating.
## 17          ema    monocular       wwdc          harman            acnv
## 18       locked       trials       slew              x,     researcher.
## 19       times.        hyped       slim        tuesday.          cierre
## 20         jedi        1975,    answers          tackle       nicohof1:
## 21    beginner's        read:   closely:       installing        measures
## 22 continuation     leaders:   youtube,   collaborates          month!
## 23        lotto       model,       360.  administration      billions):
## 24    valuation,       4x-40x     idiots      kubernetes          banked
## 25        eagle    alternate    italian            east              mt
## 26     patterns         why:    revamps           newly       strategy,
```

```
## 27       retest competitive!    hours*:         commits          cheer
## 28       (nyse:          tech?    slashes            pro:        access:
## 29       names.        2018hiv        ie,          campus     shrinking.
## 30      tariffs satyanadella       b...            peek     am_alerts:
##                                               topic6
## 1                                             intune
## 2                                             slack.
## 3                                              genee
## 4                                         battlefield
## 5                                             covers
## 6                                           high-end
## 7                                             crispr
## 8                                        10/27/2016.
## 9                                    transformation.
## 10                                       invitation.
## 11                                      accidentally
## 12                                              boot
## 13                                           floater
## 14                                             lands
## 15                                            backup
## 16                                              p.t.
## 17                                           builds.
## 18                                              hub.
## 19                                     configuration
## 20                                           studio,
## 21 ...http://mobileinteractive.com/stockstation/
## 22                                              1.4m
## 23                                        regulators
## 24                                            broker
## 25                                          partner.
## 26                                          lowfloat
## 27                                              1.5m
## 28                                           toolkit
## 29                                          finzine:
## 30                                               ad.
```

```r
# Store significant / relevant words in msft_30 in vector form
topic_composition <- data.frame(topic_num = NA, words = NA)
topic_composition[1,] <- c("Topic 1", "[ opt, parent, partnered, outsells,
                           valuation ]")
topic_composition[2,] <- c("Topic 2", "[ alzheimers, nowfunded, reuters,
                           hyped, pattern ]")
topic_composition[3,] <- c("Topic 3", "[ racist, apologizes, appeals,
                           sues, grants ]")
topic_composition[4,] <- c("Topic 4", "[ ban, crackdown, vulnerabilities,
                           africa, chromebook ]")
topic_composition[5,] <- c("Topic 5", "[ revitalize, analyze, declining,
                           saturday, roundup ]")
topic_composition[6,] <- c("Topic 6", "[ slack, battlefield, covers,
                           transformation, crispr ]")
```

Table 2: Cluster Composition

Topics	Key Tokens
Topic 1	[opt, parent, partnered, outsells, valuation]
Topic 2	[alzheimers, nowfunded, reuters, hyped, pattern]
Topic 3	[racist, apologizes, appeals, sues, grants]
Topic 4	[ban, crackdown, vulnerabilities, africa, chromebook]
Topic 5	[revitalize, analyze, declining, saturday, roundup]
Topic 6	[slack, battlefield, covers, transformation, crispr]

```
# Assign each doc_id to the topics
data.frame(doc_id = twt_msft$doc_id, Topic = topics(msft_lda))
```

```
##         doc_id  Topic
## 1  MSFT_2015_1 topic5
## 2  MSFT_2015_2 topic5
## 3  MSFT_2016_1 topic5
## 4  MSFT_2016_2 topic3
## 5  MSFT_2017_1 topic6
## 6  MSFT_2017_2 topic4
## 7  MSFT_2018_1 topic2
## 8  MSFT_2018_2 topic1
## 9  MSFT_2019_1 topic1
## 10 MSFT_2019_2 topic1
```

```
# Load docuscope-tagged dfm for MSFT
msft_docuscope <- read_csv("/Users/lee14257/Development/CMU/Text Analysis/Project/CBE2/twt_docuscope_no
  filter(ticker == "MSFT")
```

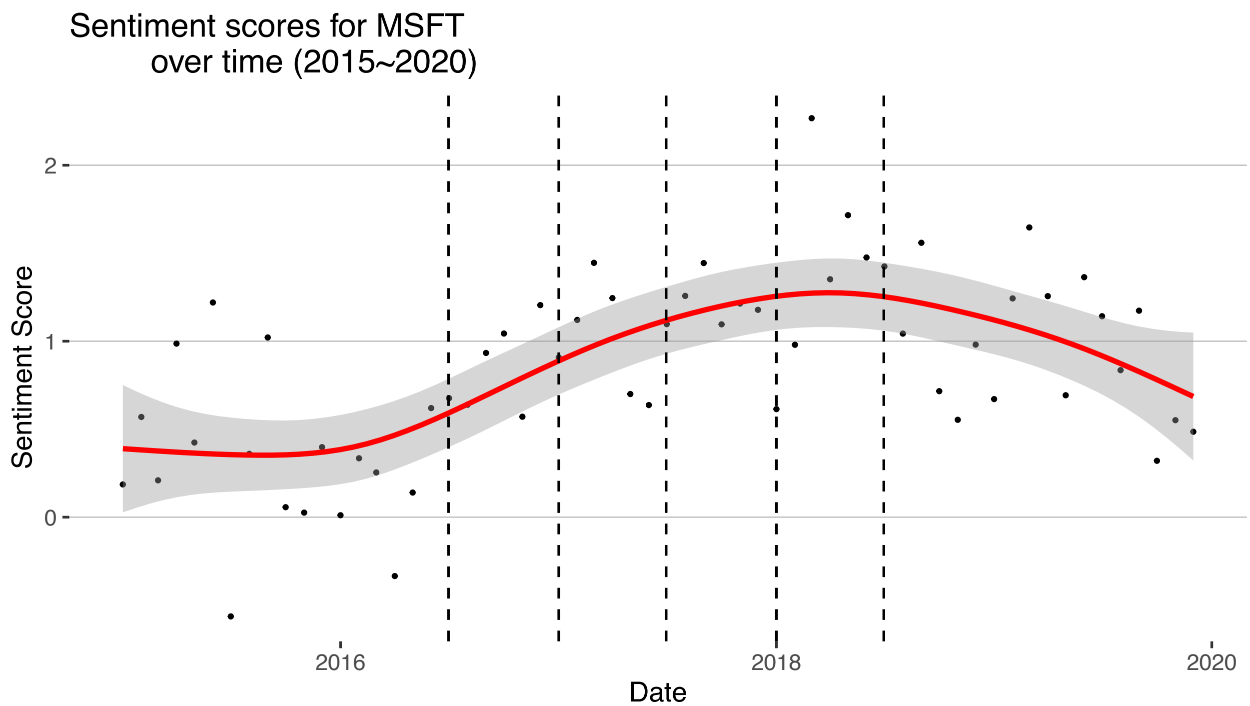
```
## Rows: 300 Columns: 41
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): ticker, doc_id
## dbl (39): year, academicterms, academicwritingmoves, character, citation, ci...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# Transform dfm to feed to ggplot
msft_sentiment <- msft_docuscope %>%
  mutate(
    ticker_symbol = str_extract(doc_id, "^[A-Z]+"),
    date = as.Date(paste0(word(doc_id, 2, sep = "_"), '-01'), format='%Y-%m-%d')
    ) %>%
  dplyr::select(ticker_symbol, date, sentiment_score) %>%
  filter(date >= "2015-01-01")
```

```r
# Graphing the time series plot for MSFT
ggplot(msft_sentiment, aes(x=date, y=sentiment_score)) +
    geom_point(size = .5) +
    geom_smooth(method = "gam", formula = y ~ s(x, bs = "cs"), size=1,
                level=0.95, se=T, colour="red") +
    labs(x="Date", y = "Sentiment Score", title="Sentiment scores for MSFT
         over time (2015~2020)")+
    theme(panel.grid.minor.x=element_blank(),
          panel.grid.major.x=element_blank()) +
    theme(panel.grid.minor.y=element_blank(),
          panel.grid.major.y=element_line(colour = "gray",size=0.25)) +
    theme(rect = element_blank()) +
    theme(legend.title=element_blank()) +
    geom_vline(xintercept = c(ymd("2016/06/30"),
                              ymd("2017/01/01"), ymd("2017/06/30"),
                              ymd("2018/01/01"), ymd("2018/06/30")),
               linetype = 2)
```



Multidimension Analysis (TSLA vs MSFT)

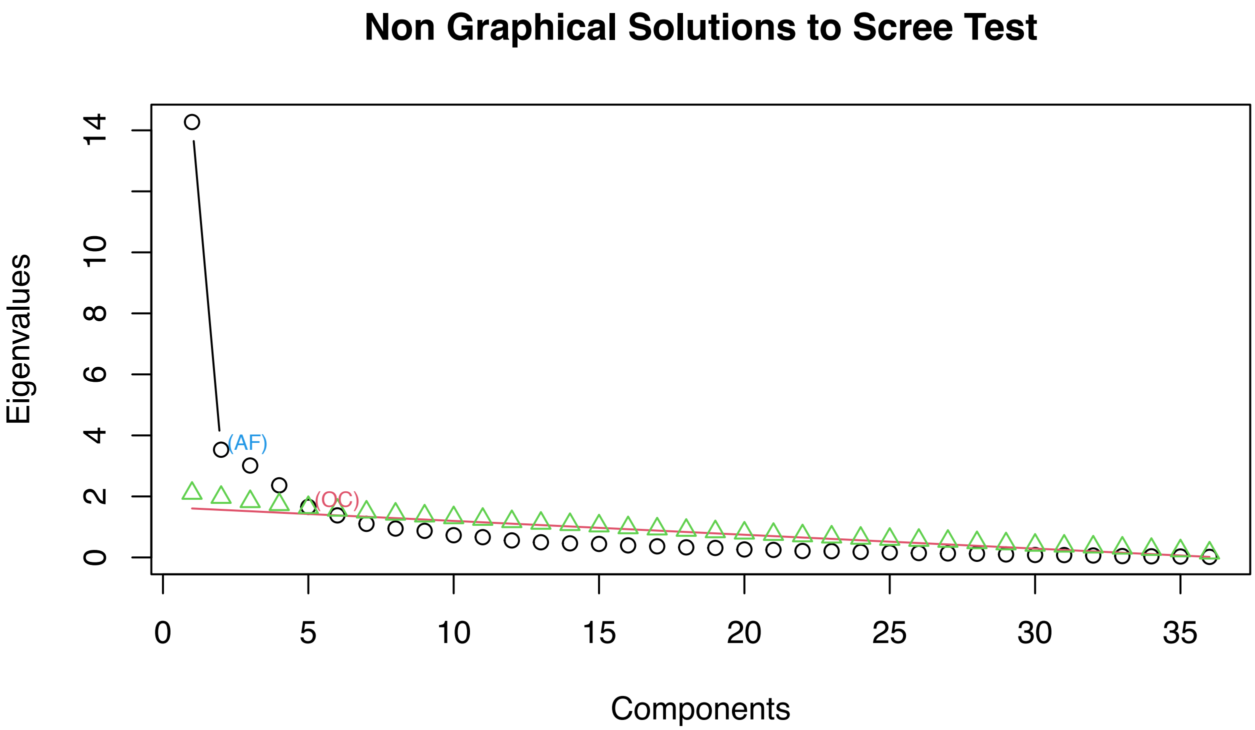
```r
# Create docuscope-tagged, normalized dfm appropriate for MDA
twt_year <- read_csv("/Users/lee14257/Development/CMU/Text Analysis/Project/CBE2/twt_docuscope_normaliz
  filter(ticker == 'TSLA' | ticker == 'MSFT') %>%
  mutate(
    ticker = as.factor(paste0(ticker, "_", year))
  ) %>% dplyr::select(-year, -sentiment_score, -citationhedged) %>%
  column_to_rownames("doc_id")
```

```
## Rows: 300 Columns: 41
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (2): ticker, doc_id
## dbl (39): year, academicterms, academicwritingmoves, character, citation, ci...
```

```
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Scree plot to select optimum number of factors
screeplot_mda(twt_year)
```




```r
# Calculate factor loadings
twt_mda <- mda_loadings(twt_year, n_factors = 5)
```

```r
# Table for factor loadings in factor1, factor2 and factor3
knitr::kable(attr(twt_mda, 'loadings'), caption =
               "Foctor loadings for midterm corpus", booktabs = T,
             linesep = "", digits = 2)
```

```r
# Compare significance of the three factors
f1_lm <- lm(Factor1 ~ group, data = twt_mda)
names(f1_lm$coefficients) <- names(coef(f1_lm)) %>% str_remove("group")
f2_lm <- lm(Factor2 ~ group, data = twt_mda)
```

Table 3: Foctor loadings for midterm corpus

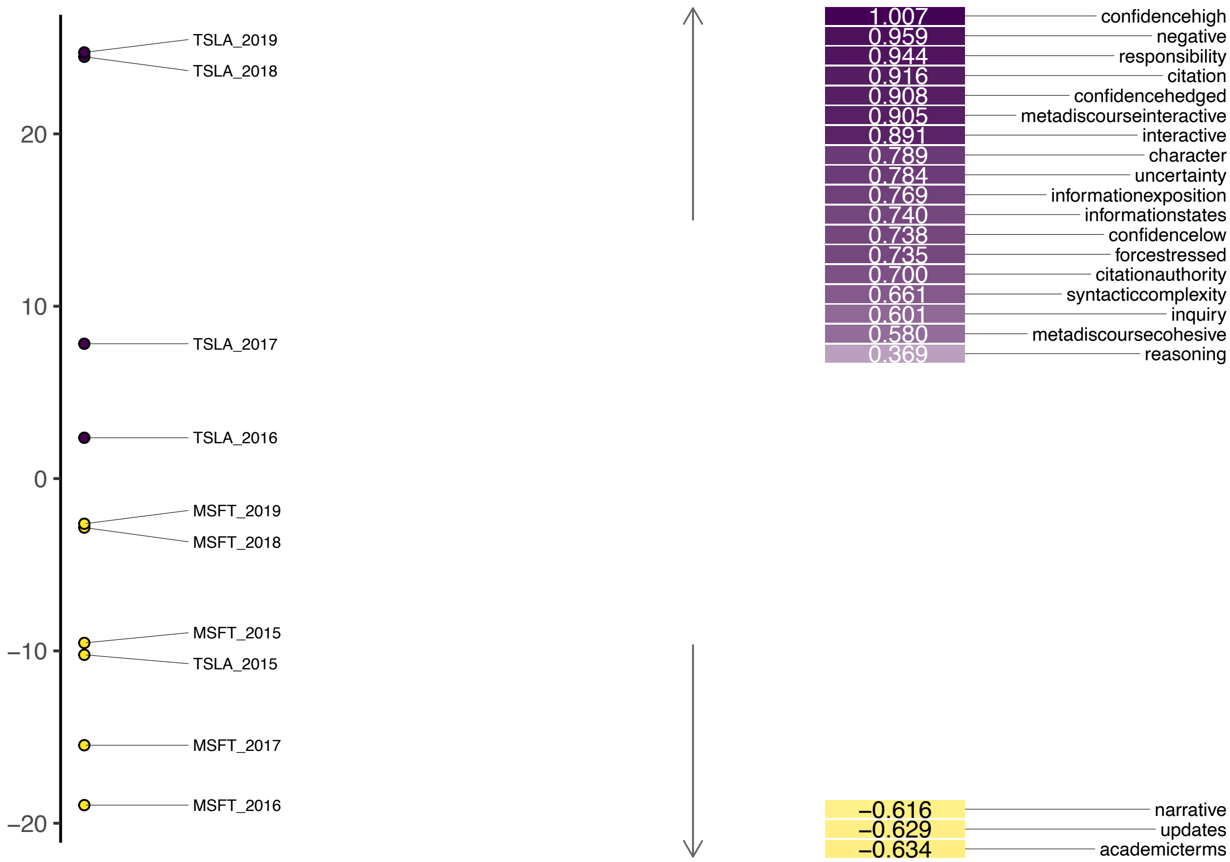
	Factor1	Factor2	Factor3	Factor4	Factor5
academicterms	-0.63	0.28	-0.08	-0.26	-0.33
academicwritingmoves	0.08	0.14	-0.01	0.02	-0.43
character	0.79	0.17	0.27	0.02	0.64
citation	0.92	0.15	-0.23	-0.14	0.07
citationauthority	0.70	-0.11	-0.02	-0.13	-0.03
confidencehedged	0.91	0.10	-0.05	-0.01	0.01
confidencehigh	1.01	0.10	-0.07	-0.04	0.09
confidencelow	0.74	0.20	-0.09	0.07	-0.04
contingent	0.10	-0.79	0.56	-0.13	0.00
description	0.23	0.03	-0.28	-0.06	-0.05
facilitate	-0.21	0.33	0.35	-0.03	-0.21
firstperson	0.23	-0.13	-0.10	0.00	-0.07
forcestressed	0.73	0.02	-0.05	0.39	0.02
future	0.31	-0.38	-0.10	0.30	-0.07
informationchange	-0.26	0.26	0.24	0.91	0.01
informationchangenegative	0.15	0.15	0.38	0.07	0.52
informationchangepositive	-0.33	0.40	0.32	0.24	-0.10
informationexposition	0.77	-0.09	0.40	0.25	0.05
informationplace	0.14	0.50	-0.01	-0.26	0.19
informationreportverbs	0.05	0.00	-0.63	-0.12	-0.21
informationstates	0.74	0.10	-0.19	-0.03	0.05
informationtopics	-0.34	-0.17	0.90	0.07	0.09
inquiry	0.60	0.08	-0.18	0.00	-0.07
interactive	0.89	0.14	-0.35	-0.07	0.17
metadiscoursecohesive	0.58	-0.25	-0.24	-0.10	0.11
metadiscourseinteractive	0.91	0.08	0.03	-0.03	0.06
narrative	-0.62	-1.06	-0.06	0.01	-0.03
negative	0.96	0.12	-0.26	-0.08	0.03
positive	0.24	-0.48	0.44	-0.24	-0.10
publicterms	-0.15	0.60	0.03	0.27	-0.26
reasoning	0.37	-0.11	-0.21	0.51	-0.04
responsibility	0.94	0.38	-0.06	-0.04	-0.05
strategic	-0.03	-0.04	-0.04	0.57	0.03
syntacticcomplexity	0.66	-0.10	0.18	0.37	0.04
uncertainty	0.78	-0.20	0.14	-0.04	0.01
updates	-0.63	-0.01	-0.47	0.18	0.00

```r
names(f2_lm$coefficients) <- names(coef(f2_lm)) %>% str_remove("group")
f3_lm <- lm(Factor3 ~ group, data = twt_mda)
names(f3_lm$coefficients) <- names(coef(f3_lm)) %>% str_remove("group")
f4_lm <- lm(Factor4 ~ group, data = twt_mda)
names(f4_lm$coefficients) <- names(coef(f4_lm)) %>% str_remove("group")
f5_lm <- lm(Factor5 ~ group, data = twt_mda)
names(f5_lm$coefficients) <- names(coef(f5_lm)) %>% str_remove("group")
```

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
(Intercept)	-9.54 ***	2.69 **	-3.20 ***	5.49 ***	-0.95
MSFT_2016	-9.41 ***	4.19 ***	-1.56	-10.01 ***	-1.59 *
MSFT_2017	-5.94 *	1.68	2.94 **	-8.58 ***	-0.69
MSFT_2018	6.69 **	-4.60 ***	10.14 ***	-5.00 ***	1.43
MSFT_2019	6.91 **	-6.33 ***	8.98 ***	-5.01 ***	1.42
TSLA_2015	-0.69	-5.34 ***	-0.82	-7.48 ***	1.47
TSLA_2016	11.90 ***	-5.24 ***	-0.59	-6.24 ***	1.58 *
TSLA_2017	17.36 ***	-8.31 ***	3.95 ***	-4.48 ***	1.30
TSLA_2018	34.01 ***	-2.72 *	4.46 ***	-3.52 ***	2.19 **
TSLA_2019	34.26 ***	-0.58	4.59 ***	-3.85 ***	2.53 **
DF	110.00	110.00	110.00	110.00	110.00
R2	0.86	0.69	0.72	0.60	0.32
F statistic	78.09	27.24	31.76	18.13	5.86

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

```r
# Heatmap for factor 1 (chosen)
mda.biber::heatmap_mda(twt_mda, n_factor = 1)
```


20
10
0
−10
−20
TSLA_2019
TSLA_2018
TSLA_2017
TSLA_2016
MSFT_2019
MSFT_2018
MSFT_2015
TSLA_2015
MSFT_2017
MSFT_2016
1.007 confidencehigh
0.959 negative
0.944 responsibility
0.916 citation
0.908 confidencehedged
0.905 metadiscourseinteractive
0.891 interactive
0.789 character
0.784 uncertainty
0.769 informationexposition
0.740 informationstates
0.738 confidencelow
0.735 forcestressed
0.700 citationauthority
0.661 syntacticcomplexity
0.601 inquiry
0.580 metadiscoursecohesive
0.369 reasoning
−0.616 narrative
−0.629 updates
−0.634 academicterms