TAD Week 9

Tommy Klein

Working Directory

```r
setwd('/Users/tklein/Desktop/Desktop_tpk/JHU_Classes/text_as_data/week9')
```

Library

```r
library(tidyverse)
library(tidytext)
source('../functions/helper_functions.R')
library(stats)
library(cluster)
library(factoextra)
```

```r
reddit_data <- read_csv('../getting_reddit_data/updated_posts_with_text.csv')
```

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   X1 = col_double(),
##   title = col_character(),
##   score = col_double(),
##   id = col_character(),
##   subreddit = col_character(),
##   url = col_character(),
##   num_comments = col_double(),
##   body = col_character(),
##   created = col_double(),
##   cluster = col_double()
## )
```

```r
reddit_data %>% glimpse()
```

```
## Rows: 5,304
## Columns: 10
## $ X1           <dbl> 16271, 16264, 16262, 16261, 16255, 16246, 16245, 16232, 1~
## $ title        <chr> "Making Bitcoin Secure to Quantum attacks.", "Is storing ~
## $ score        <dbl> 33, 96, 6, 1, 0, 3, 6, 3, 1, 87, 3, 1, 3, 0, 1, 5, 6, 8, ~
## $ id           <chr> "rgbudo", "rgfddy", "rgi8tc", "rgijhy", "rgkgk5", "rgrz7e~
## $ subreddit    <chr> "BitcoinBeginners", "BitcoinBeginners", "BitcoinBeginners~
## $ url          <chr> "https://www.reddit.com/r/BitcoinBeginners/comments/rgbud~
## $ num_comments <dbl> 78, 451, 86, 68, 67, 75, 105, 61, 64, 203, 68, 93, 71, 49~
## $ body         <chr> "I read this article [https://www2.deloitte.com/nl/nl/pag~
## $ created      <dbl> 1639501125, 1639510643, 1639518590, 1639519408, 163952465~
## $ cluster      <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
```

```r
reddit_data <- reddit_data[1:1000,]



reddit_corpus <- csv_to_corpus(
  '../getting_reddit_data/updated_posts_with_text.csv',
  text_col = 'body'
  )

reddit_corpus <- reddit_corpus[1:1000]

reddit_corpus %>% head()
```

```
## Corpus consisting of 6 documents and 9 docvars.
## 1 :
## "I read this article [https://www2.deloitte.com/nl/nl/pages/i..."
##
## 2 :
## "Im planning on just keeping my coins on Kraken until I have ..."
##
## 3 :
## "Listened to a video the other day that if I use an exchange ..."
##
## 4 :
## "Guys please check your wallet, I.e, trust wallet, what’s goi..."
##
## 5 :
## "What is the easiest way to pay with bitcoin?  Not looking to..."
##
## 6 :
## "Finally going to move from an app based wallet to cold stora..."
```

Training Data

```r
crypto_dfm <- corp_to_dfm(reddit_corpus, stem = T)
```

```
## Warning: 'stem' is deprecated; use dfm_wordstem() instead
```

```r
training_dfm <- dfm_trim(
  crypto_dfm,
  min_termfreq = 20,
  max_docfreq = .8,
  docfreq_type = 'quantile',
  termfreq_type = 'count'
  )


training_matrix <- as.matrix(crypto_dfm)

training_matrix[is.nan(training_matrix)] = 0
```

```
reddit_kmeans = kmeans(
  x = training_matrix, # All operations are done on our DFM
  centers = 5
)



str(reddit_kmeans)
```

```
## List of 9
##  $ cluster     : Named int [1:1000] 1 1 1 1 1 1 1 1 1 1 ...
##   ..- attr(*, "names")= chr [1:1000] "1" "2" "3" "4" ...
##  $ centers     : num [1:5, 1:4818] 0.0628 0 0 0.1389 0 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:5] "1" "2" "3" "4" ...
##   .. ..$ : chr [1:4818] "read" "articl" "say" "use" ...
##  $ totss       : num 87488
##  $ withinss    : num [1:5] 50976 3676 0 16251 0
##  $ tot.withinss: num 70903
##  $ betweenss   : num 16585
##  $ size        : int [1:5] 923 3 1 72 1
##  $ iter        : int 4
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
reddit_data$cluster <- reddit_kmeans$cluster



reddit_data %>%
  group_by(cluster) %>%
  mutate(rank = row_number()) %>%
  filter(rank < 3) %>%
  select(cluster, body) %>%
  arrange(cluster)
```

```
## # A tibble: 8 x 2
## # Groups:   cluster [5]
##   cluster body
##     <int> <chr>
## 1       1 "I read this article [https://www2.deloitte.com/nl/nl/pages/innovatie~
## 2       1 "Im planning on just keeping my coins on Kraken until I have over $50~
## 3       2 "&#x200B;\n\nThe number one question I get asked is how to find those~
## 4       2 "I wanted to make a shortlist of things beginners can do to avoid sca~
## 5       3 "I'm currently trying out the Ledgers experimental feature where I ca~
## 6       4 "**WARNING**: [Blockchain.com](https://blockchain.com/)'s Customer Su~
## 7       4 "I recently made my first purchase of BTC via my Coinbase account. I ~
## 8       5 "1/27/22, 4:54 PM 1/2 16:35, Jan 27 You: Withdrawal help 16:35, Jan 2~
```

```
reddit_data %>%
  pull(cluster) %>%
  table()
```

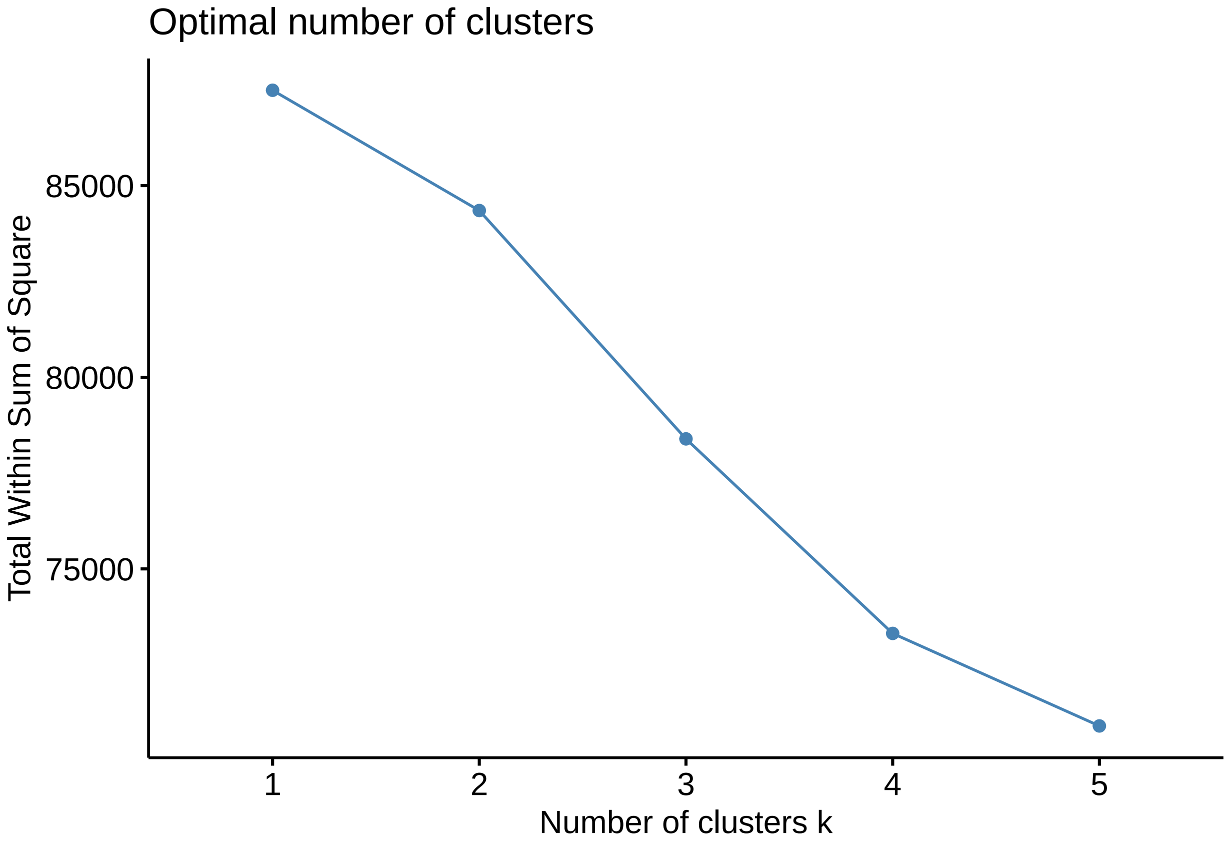
```
## .
##   1   2   3   4   5
## 923   3   1  72   1
```

```
reddit_data %>%
  group_by(cluster) %>%
  summarize(count = n())
```

```
## # A tibble: 5 x 2
##   cluster count
##     <int> <int>
## 1       1   923
## 2       2     3
## 3       3     1
## 4       4    72
## 5       5     1
```

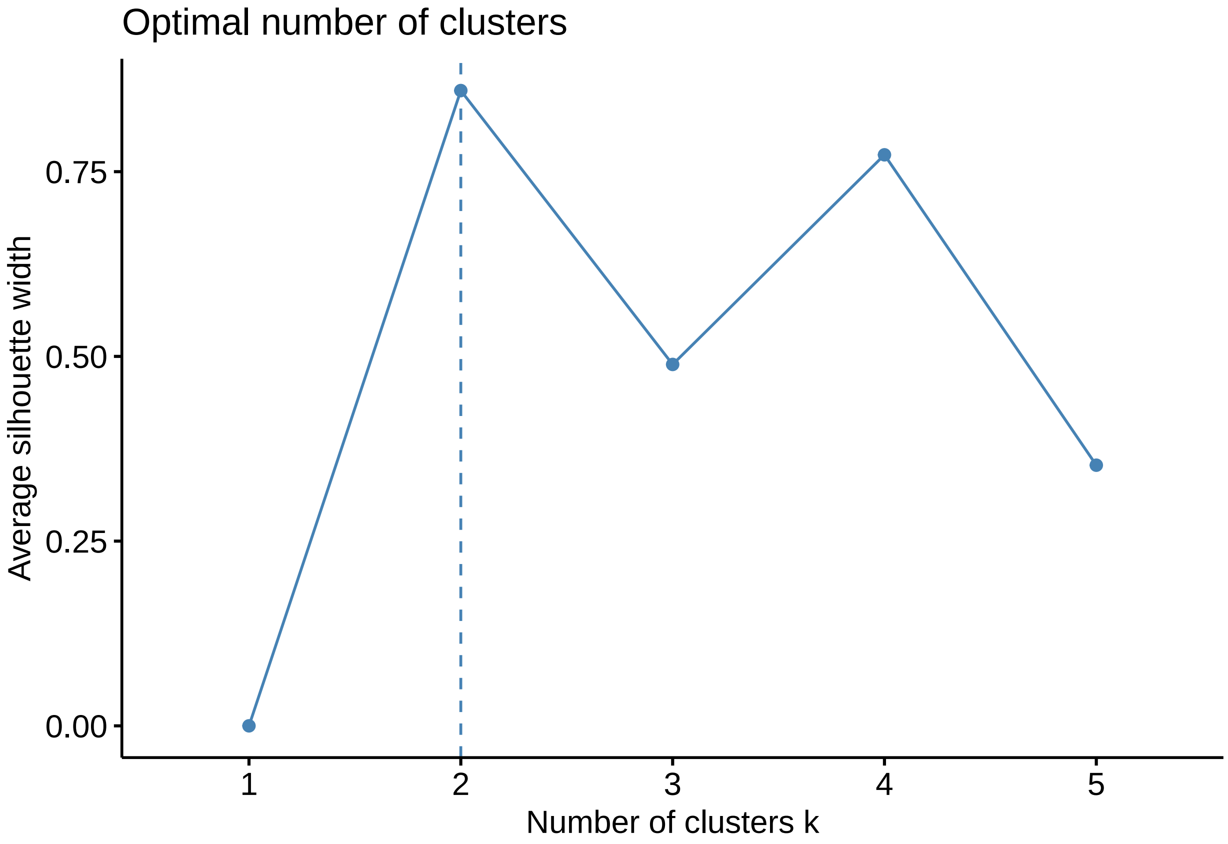
Looks like there is one big group, one smaller group, and then three irrellevant groups with only a few documents.

```
elbow = fviz_nbclust(training_matrix,
                    kmeans,
                    method='wss',
                    k.max = 5,
                    verbose=TRUE)

elbow
```



Looks like 3 or 4 might be ideal. Lets see what the silhouette method says.

```
silhouette = fviz_nbclust(training_matrix,
                    kmeans,
                    method='silhouette',
                    k.max = 5,
                    verbose=TRUE)

silhouette
```



The silhouette says 2, so we'll go with that.

```
reddit_kmeans = kmeans(
  x = training_matrix, # All operations are done on our DFM
  centers = 2
)



str(reddit_kmeans)
```

```
## List of 9
##  $ cluster     : Named int [1:1000] 2 2 2 2 2 2 2 2 2 2 ...
##   ..- attr(*, "names")= chr [1:1000] "1" "2" "3" "4" ...
##  $ centers     : num [1:2, 1:4818] 0 0.0681 0 0.019 1 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:2] "1" "2"
##   .. ..$ : chr [1:4818] "read" "articl" "say" "use" ...
##  $ totss       : num 87488
##  $ withinss    : num [1:2] 0 81264
##  $ tot.withinss: num 81264
##  $ betweenss   : num 6224
##  $ size        : int [1:2] 1 999
##  $ iter        : int 1
##  $ ifault      : int 0
##  - attr(*, "class")= chr "kmeans"
```

```
reddit_data$cluster <- reddit_kmeans$cluster



reddit_data %>%
  group_by(cluster) %>%
  mutate(rank = row_number()) %>%
  filter(rank < 3) %>%
  select(cluster, body) %>%
  arrange(cluster)
```

```
## # A tibble: 3 x 2
## # Groups:   cluster [2]
##   cluster body
##     <int> <chr>
## 1       1 "1/27/22, 4:54 PM 1/2 16:35, Jan 27 You: Withdrawal help 16:35, Jan 2~
## 2       2 "I read this article [https://www2.deloitte.com/nl/nl/pages/innovatie~
## 3       2 "Im planning on just keeping my coins on Kraken until I have over $50~
```

```
reddit_data %>%
  pull(cluster) %>%
  table()
```

```
## .
##   1   2
##   1 999
```

```
reddit_data %>%
  group_by(cluster) %>%
  summarize(count = n())
```

```
## # A tibble: 2 x 2
##   cluster count
##     <int> <int>
## 1       1     1
## 2       2   999
```

Well that just made one cluster, which is not very helpful.

I'm not really sure why k-means isn't returning anything helpful. All of the rows in my data have text. I'm not sure what feature engineering I could do here to get a better result.