TAD Week 10 Assignment

Tommy Klein

4/12/2022

Working Directory

```
setwd('/Users/tklein/Desktop/Desktop_tpk/JHU_Classes/text_as_data/week11')
```

Library

```
library(ndjson)
library(SentimentAnalysis)
```

```
##
## Attaching package: 'SentimentAnalysis'
```

```
## The following object is masked from 'package:base':
##
##     write
```

```
library(RedditExtractoR)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()  masks stats::filter()
## x purrr::flatten() masks ndjson::flatten()
## x dplyr::lag()     masks stats::lag()
```

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

```
## stm v1.3.6 successfully loaded. See ?stm for help.
##  Papers, resources, and other materials at structuraltopicmodel.com
```

```
library(tidytext)
source('../functions/helper_functions.R')
```

```
## Package version: 3.2.0
## Unicode version: 13.0
## ICU version: 69.1
```

```
## Parallel computing: 4 of 4 threads used.
```

```
## See https://quanteda.io for tutorials and examples.
```

```
library(e1071)
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.1.2
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'lattice'
```

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

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(cluster)
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
library(quanteda)
library(dbscan)
```

```
## Warning: package 'dbscan' was built under R version 4.1.2
```

Reading in data

I collected thousands of reddit posts from multiple different sub-reddits: r/Bitcoin, r/Ethereum, r/CryptoCurrency, r/BitcoinBegginers, and r/Coinbase. I'm going to use a naive-bayes model to see if I can predict which sub-reddit a post was posted in based on the text used in the post.

```
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()
## )
```

```
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, ~
```

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

DBscan

```
reddit_data$subreddit = factor(reddit_data$subreddit)

reddit_dfm <- corp_to_dfm(reddit_corpus)
```

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

```
reddit_dfm_trimmed <- reddit_dfm %>% quanteda::dfm_trim(
  max_termfreg = .8, termfreq_type = 'prop',
  min_docfreq = 20, docfreq_type = 'count'
  )

reddit_dfm_matrix <- as.matrix(reddit_dfm_trimmed)

reddit_dfm_matrix[is.nan(reddit_dfm_matrix)] = 0


dbscan_reddit = dbscan::dbscan(x=reddit_dfm_matrix,
                               eps=1, # Gues at initial eps value
                               minPts = 5) # minimum number of data points per cluster
dbscan_reddit
```

```
## DBSCAN clustering for 5304 objects.
## Parameters: eps = 1, minPts = 5
## The clustering contains 6 cluster(s) and 5158 noise points.
##
##    0    1    2    3    4    5    6
## 5158   73   29    8    7   23    6
##
## Available fields: cluster, eps, minPts
```

DBscan creates 6 clusters, which are all relatively tiny compared to the number of noise points - the largest cluster has only 73 documents, while there are 5,158 noise points.

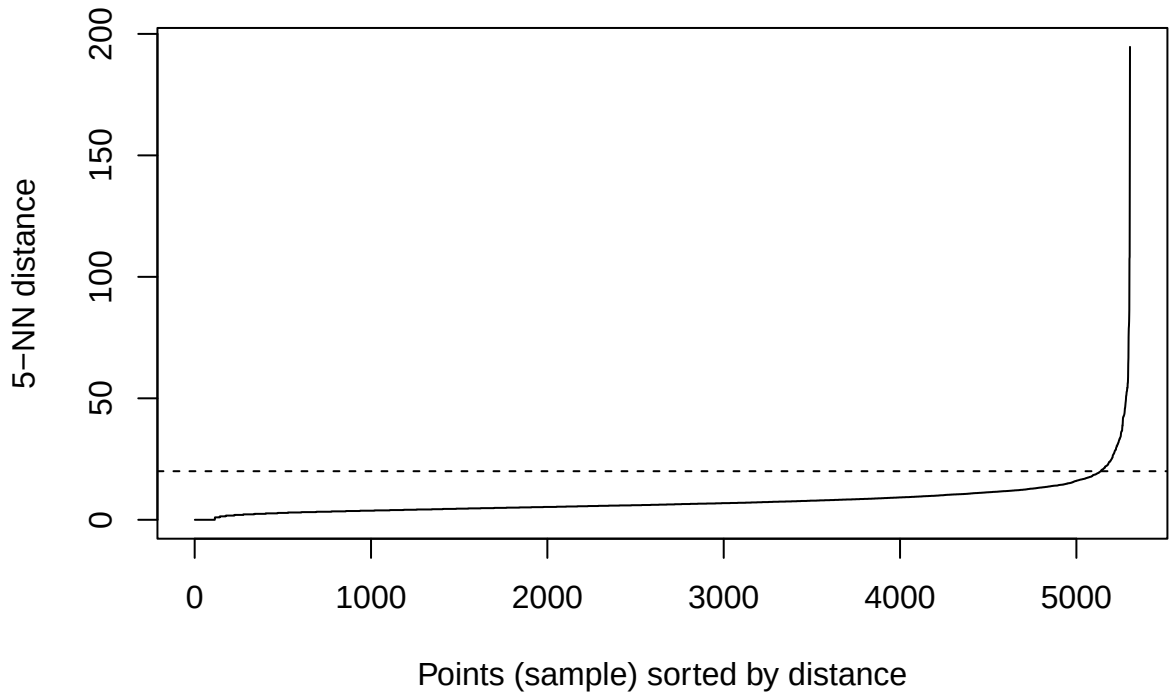
```
reddit_data$dbscan_cluster <- dbscan_reddit$cluster


reddit_data %>%
  filter(dbscan_cluster !=0) %>%
  group_by(dbscan_cluster) %>%
  mutate(rank = rank(dbscan_cluster, ties.method = 'first')) %>%
  filter(rank < 5) %>%
  select(dbscan_cluster, body) %>%
  arrange(dbscan_cluster)
```

```
## # A tibble: 24 x 2
## # Groups:   dbscan_cluster [6]
##    dbscan_cluster body
##             <int> <chr>
##  1              1 "[https://www.youtube.com/watch?v=BSFBYzwCG90](https://www.yo~
##  2              1 "?"
##  3              1 "tank you"
##  4              1 "As title says :)"
##  5              2 "[deleted]\n\n[View Poll](https://www.reddit.com/poll/slih8u)"
##  6              2 "[deleted]\n\n[View Poll](https://www.reddit.com/poll/slknjo)"
##  7              2 "[removed]\n\n[View Poll](https://www.reddit.com/poll/slkqz5)"
##  8              2 "[removed]\n\n[View Poll](https://www.reddit.com/poll/slxxid)"
##  9              3 "**Welcome to the Weekly Discussion. Please read the disclaim~
## 10              3 "**Welcome to the Weekly Discussion. Please read the disclaim~
## # ... with 14 more rows
```

The clusters do seem to be similar, so that is good. The bad new is that none of them seem to be a "typical" post - e.g. one cluster is just a weekly discussion thread in one of the subreddits.

```
kNNdistplot(reddit_dfm_matrix, k =  5)+abline(h = 20, lty = 2)
```




```
## integer(0)
```

Looks like somewhere around 20.

```
dbscan_reddit = dbscan::dbscan(x=reddit_dfm_matrix,
                               eps=20, # Gues at initial eps value
                               minPts = 5) # minimum number of data points per cluster
dbscan_reddit
```

```
## DBSCAN clustering for 5304 objects.
## Parameters: eps = 20, minPts = 5
## The clustering contains 2 cluster(s) and 156 noise points.
##
##    0    1    2
##  156 5140    8
##
## Available fields: cluster, eps, minPts
```

So this model reduces the noise points, but ends up creating only two clusters, which really isn't better.

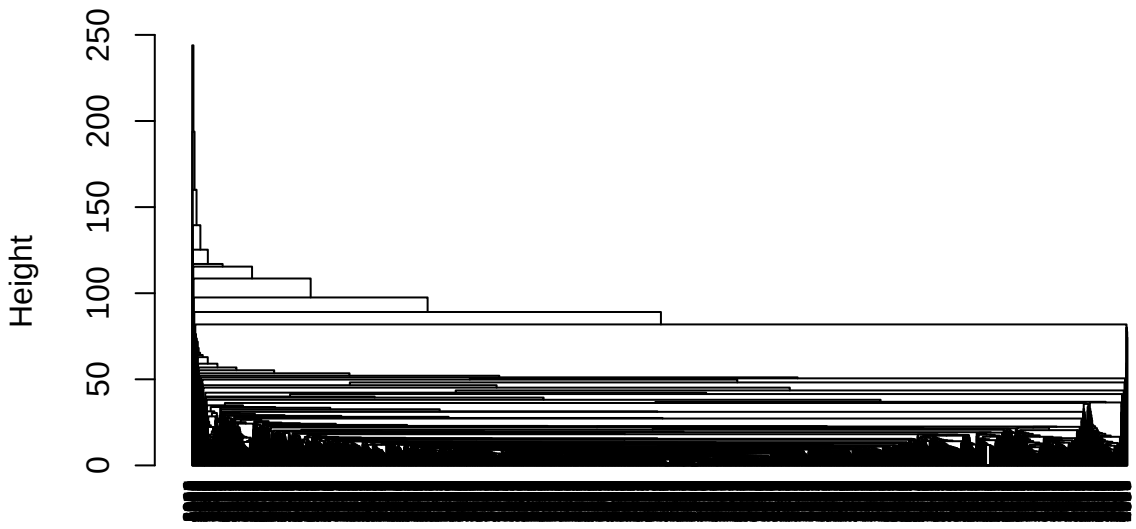
Hierarchical Cluster

```
hierarchical_reddit <- hclust(dist(reddit_dfm_matrix,method='euclidian'),
                              method = "complete" )
```

```
hierarchical_reddit
```

```
##
## Call:
## hclust(d = dist(reddit_dfm_matrix, method = "euclidian"), method = "complete")
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 5304
```

```
plot(hierarchical_reddit, cex = 0.6, hang = -1)
```




```
reddit_data$hier_cluster <- cutree(hierarchical_reddit, h=75)


reddit_data %>%
  group_by(hier_cluster) %>%
  mutate(rank = rank(hier_cluster, ties.method = 'first')) %>%
  filter(rank < 5) %>%
  select(hier_cluster, body) %>%
  arrange(hier_cluster)
```

```
## # A tibble: 28 x 2
## # Groups:   hier_cluster [16]
##    hier_cluster body
##           <int> <chr>
##  1            1 "I read this article [https://www2.deloitte.com/nl/nl/pages/inn~
##  2            1 "Im planning on just keeping my coins on Kraken until I have ov~
##  3            1 "Listened to a video the other day that if I use an exchange (w~
##  4            1 "Guys please check your wallet, I.e, trust wallet, what's going~
##  5            2 "\n\n## Updated User Agreement\n\n   \n\n# Coinbase User Agreem~
##  6            3 "What's up everyone, I'm fairly new to trading in general and t~
##  7            3 "\n\nDaily thread to discuss the Super trend for large cap coin~
##  8            3 "\n\nDaily thread to discuss the Super trend for large cap coin~
##  9            4 "We provide over 100+ FREE crypto articles on our SubStack! :D ~
## 10            5 "\n\n# Are we being fair?\n\n  \n**An open letter proposal to t~
## # ... with 18 more rows
```

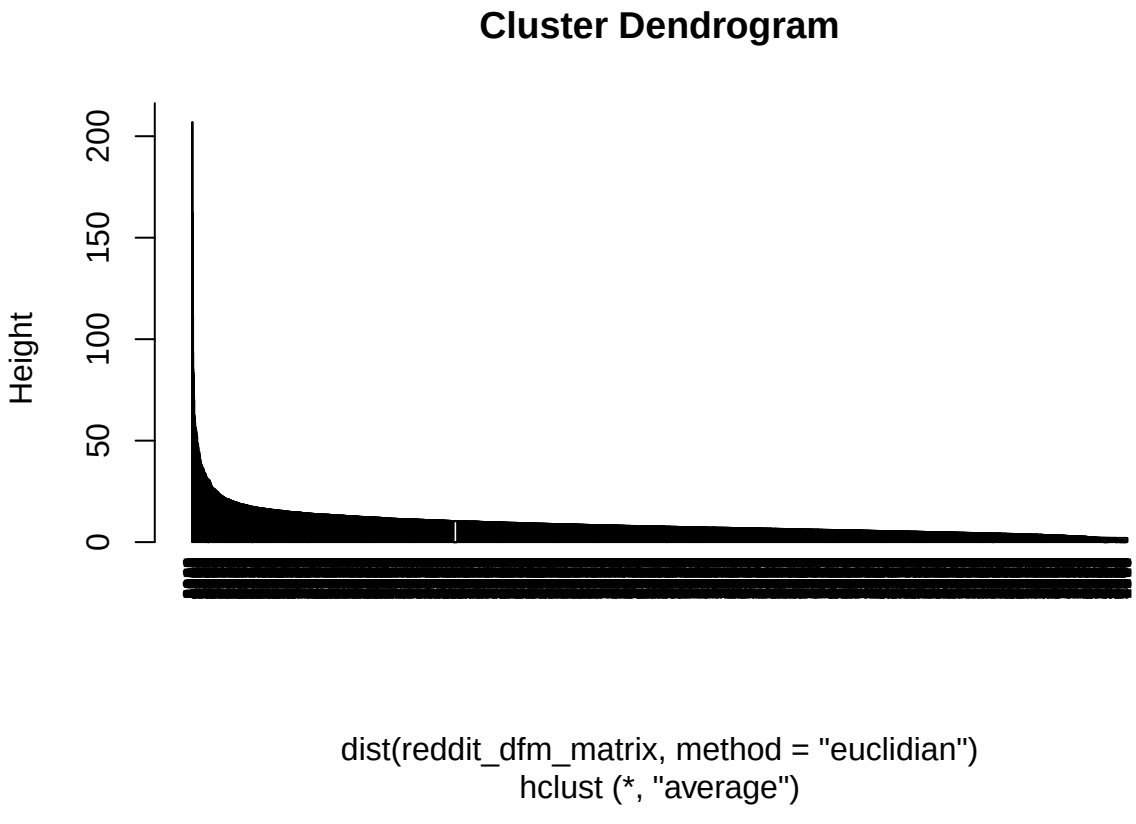
```
reddit_data %>%
  group_by(hier_cluster) %>%
  summarize(count = n())
```

```
## # A tibble: 16 x 2
##    hier_cluster count
##           <int> <int>
##  1            1  5280
##  2            2     1
##  3            3     3
##  4            4     1
##  5            5     3
##  6            6     4
##  7            7     2
##  8            8     1
##  9            9     2
## 10           10     1
## 11           11     1
## 12           12     1
## 13           13     1
## 14           14     1
## 15           15     1
## 16           16     1
```

Like the other clustering algorithms, most of the documents went into one cluster.

```
hierarchical_reddit <- hclust(dist(reddit_dfm_matrix,method='euclidian'),
                              method = "average" )

plot(hierarchical_reddit, cex = 0.6, hang = -1)
```



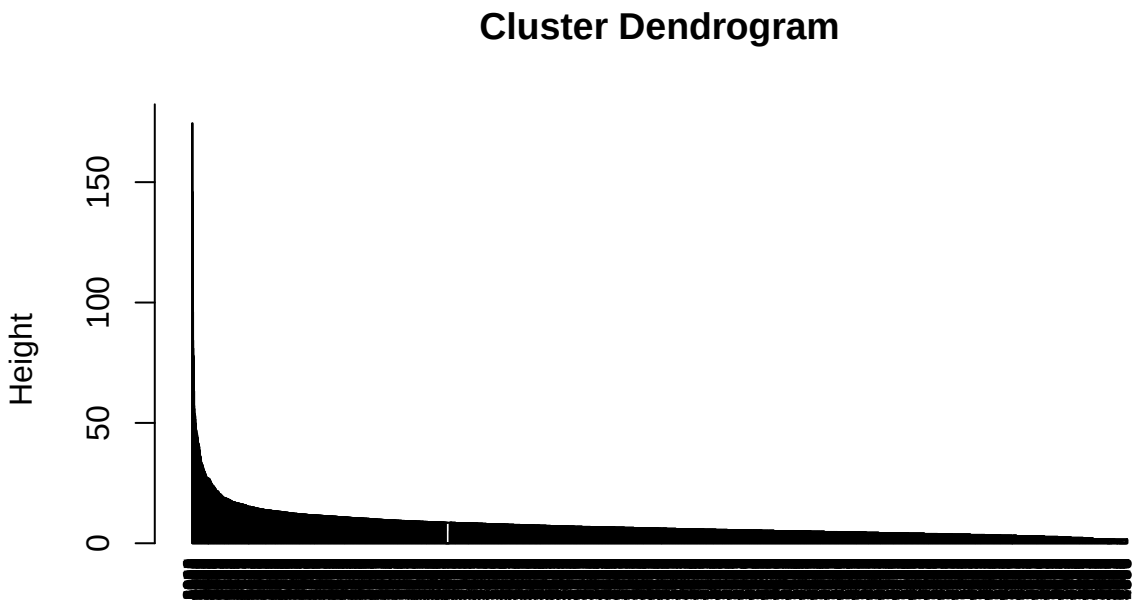

```
reddit_data$hier_cluster <- cutree(hierarchical_reddit, h=75)
```

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

```
## # A tibble: 10 x 2
##    hier_cluster count
##           <int> <int>
##  1            1  5295
##  2            2     1
##  3            3     1
##  4            4     1
##  5            5     1
##  6            6     1
##  7            7     1
##  8            8     1
##  9            9     1
## 10           10     1
```

```
hierarchical_reddit <- hclust(dist(reddit_dfm_matrix,method='euclidian'),
                              method = "single" )

plot(hierarchical_reddit, cex = 0.6, hang = -1)
```




```
reddit_data$hier_cluster <- cutree(hierarchical_reddit, h=75)

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

```
## # A tibble: 9 x 2
##   hier_cluster count
##          <int> <int>
## 1            1  5296
## 2            2     1
## 3            3     1
## 4            4     1
## 5            5     1
## 6            6     1
## 7            7     1
## 8            8     1
## 9            9     1
```

```
hierarchical_reddit <- hclust(dist(reddit_dfm_matrix,method='euclidian'),
                              method = "median" )

plot(hierarchical_reddit, cex = 0.6, hang = -1)
```