TAD Week 8 Assignment

Tommy Klein

3/30/2022

Working Directory

```
setwd('/Users/tklein/Desktop/Desktop_tpk/JHU_Classes/text_as_data/week8')
```

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 3.3.3     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)
```

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/psaw_crypto_posts_with_body.csv')
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
##
## -- 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()
## )
```

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

```
## Rows: 16,274
## Columns: 9
## $ X1           <dbl> 7, 8, 11, 13, 22, 23, 24, 25, 28, 29, 31, 32, 33, 35, 38,~
## $ title        <chr> "Benefits of POW over POS", "#cleanupbitcoin and #changet~
## $ score        <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, ~
## $ id           <chr> "trygjz", "tryfur", "try72a", "trxskd", "trwbio", "trw9oa~
## $ subreddit    <chr> "Bitcoin", "Bitcoin", "Bitcoin", "Bitcoin", "Bitcoin", "B~
## $ url          <chr> "https://www.reddit.com/r/Bitcoin/comments/trygjz/benefit~
## $ num_comments <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 2, 8, 0, 0, 0, ~
## $ body         <chr> "I am new to BTC and want to learn more about proof of wo~
## $ created      <dbl> 1648640412, 1648640348, 1648639444, 1648638013, 164863304~
```

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

Training Data

Now I want to subset my data to use a portion of it for training.

```
set.seed(42)

rand_sample <- sample(x = 16274, size = 5000)
```

```
training_df <- reddit_data[rand_sample, ]

training_corpus <- reddit_corpus[rand_sample,]

training_dfm <- corp_to_dfm(training_corpus)
```

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

```
training_matrix <- as.matrix(training_dfm)

labels = training_df$subreddit %>% as.factor()

levels(labels)
```

```
## [1] "Bitcoin"          "BitcoinBeginners" "CoinBase"         "CryptoCurrency"
## [5] "ethereum"
```

```
training_df %>%
  group_by(subreddit) %>%
  summarize(count = n())
```

```
## # A tibble: 5 x 2
##   subreddit        count
##   <chr>            <int>
## 1 Bitcoin            790
## 2 BitcoinBeginners  1450
## 3 CoinBase          1200
## 4 CryptoCurrency     824
## 5 ethereum           736
```

Training Model

```
nb = e1071::naiveBayes(
  x=training_matrix,
  y=labels,
  method='class'
)
```

Predictions

```
nb_prediction = predict(nb, training_matrix)
```

Results

```
results = data.frame(
  Predictions = nb_prediction,
  Actuals = labels
)

results %>% head(100)
```

```
##          Predictions          Actuals
## 1   BitcoinBeginners         CoinBase
## 2   BitcoinBeginners BitcoinBeginners
## 3   BitcoinBeginners          Bitcoin
## 4   BitcoinBeginners   CryptoCurrency
## 5   BitcoinBeginners         CoinBase
## 6   BitcoinBeginners          Bitcoin
## 7   BitcoinBeginners BitcoinBeginners
## 8   BitcoinBeginners         CoinBase
## 9   BitcoinBeginners         CoinBase
## 10  BitcoinBeginners BitcoinBeginners
## 11  BitcoinBeginners BitcoinBeginners
## 12  BitcoinBeginners BitcoinBeginners
## 13          ethereum BitcoinBeginners
## 14  BitcoinBeginners          Bitcoin
## 15  BitcoinBeginners BitcoinBeginners
## 16  BitcoinBeginners BitcoinBeginners
## 17  BitcoinBeginners BitcoinBeginners
## 18  BitcoinBeginners         CoinBase
## 19  BitcoinBeginners BitcoinBeginners
## 20  BitcoinBeginners         ethereum
## 21  BitcoinBeginners         CoinBase
## 22  BitcoinBeginners         CoinBase
## 23  BitcoinBeginners BitcoinBeginners
## 24  BitcoinBeginners BitcoinBeginners
## 25  BitcoinBeginners   CryptoCurrency
## 26  BitcoinBeginners          Bitcoin
## 27  BitcoinBeginners         CoinBase
## 28  BitcoinBeginners         CoinBase
## 29  BitcoinBeginners BitcoinBeginners
## 30  BitcoinBeginners   CryptoCurrency
## 31  BitcoinBeginners BitcoinBeginners
## 32  BitcoinBeginners         ethereum
## 33  BitcoinBeginners         CoinBase
## 34  BitcoinBeginners BitcoinBeginners
## 35  BitcoinBeginners          Bitcoin
## 36          ethereum   CryptoCurrency
## 37          ethereum          Bitcoin
## 38  BitcoinBeginners BitcoinBeginners
## 39  BitcoinBeginners         CoinBase
## 40  BitcoinBeginners   CryptoCurrency
## 41          ethereum         CoinBase
## 42  BitcoinBeginners BitcoinBeginners
## 43  BitcoinBeginners          Bitcoin
## 44  BitcoinBeginners BitcoinBeginners
## 45  BitcoinBeginners BitcoinBeginners
## 46  BitcoinBeginners BitcoinBeginners
## 47  BitcoinBeginners         CoinBase
## 48  BitcoinBeginners         CoinBase
## 49          CoinBase BitcoinBeginners
## 50  BitcoinBeginners         CoinBase
## 51  BitcoinBeginners   CryptoCurrency
## 52  BitcoinBeginners   CryptoCurrency
## 53  BitcoinBeginners         CoinBase
## 54  BitcoinBeginners          Bitcoin
## 55  BitcoinBeginners         ethereum
## 56  BitcoinBeginners         CoinBase
## 57  BitcoinBeginners         CoinBase
## 58  BitcoinBeginners          Bitcoin
## 59  BitcoinBeginners          Bitcoin
## 60  BitcoinBeginners         CoinBase
## 61  BitcoinBeginners BitcoinBeginners
## 62  BitcoinBeginners          Bitcoin
## 63  BitcoinBeginners         ethereum
## 64  BitcoinBeginners   CryptoCurrency
## 65          ethereum BitcoinBeginners
## 66  BitcoinBeginners          Bitcoin
## 67          ethereum   CryptoCurrency
## 68  BitcoinBeginners BitcoinBeginners
## 69          ethereum BitcoinBeginners
## 70           Bitcoin         ethereum
## 71  BitcoinBeginners         CoinBase
## 72          CoinBase BitcoinBeginners
## 73  BitcoinBeginners BitcoinBeginners
## 74  BitcoinBeginners BitcoinBeginners
## 75  BitcoinBeginners BitcoinBeginners
## 76          ethereum          Bitcoin
## 77  BitcoinBeginners         CoinBase
## 78          ethereum BitcoinBeginners
## 79          ethereum          Bitcoin
## 80  BitcoinBeginners   CryptoCurrency
## 81  BitcoinBeginners         CoinBase
## 82          ethereum   CryptoCurrency
## 83  BitcoinBeginners         CoinBase
## 84  BitcoinBeginners          Bitcoin
## 85  BitcoinBeginners         ethereum
## 86          CoinBase BitcoinBeginners
## 87  BitcoinBeginners   CryptoCurrency
## 88  BitcoinBeginners   CryptoCurrency
## 89          CoinBase         ethereum
## 90  BitcoinBeginners         CoinBase
## 91  BitcoinBeginners          Bitcoin
## 92  BitcoinBeginners BitcoinBeginners
## 93  BitcoinBeginners BitcoinBeginners
## 94  BitcoinBeginners         CoinBase
## 95  BitcoinBeginners BitcoinBeginners
## 96  BitcoinBeginners   CryptoCurrency
## 97  BitcoinBeginners          Bitcoin
## 98          ethereum BitcoinBeginners
## 99  BitcoinBeginners         CoinBase
## 100 BitcoinBeginners         CoinBase
```

```r
results %>%
  mutate(correct = Predictions == Actuals) %>%
  group_by(Predictions, Actuals, correct) %>%
  summarize(count = n()) %>%
  pivot_wider(names_from = correct, values_from = count, values_fill = 0)
```

```
## `summarise()` has grouped output by 'Predictions', 'Actuals'. You can override using the `.groups` a
```

```
## # A tibble: 23 x 4
## # Groups:   Predictions, Actuals [23]
##    Predictions      Actuals          `FALSE` `TRUE`
##    <fct>            <fct>              <int>  <int>
##  1 Bitcoin          BitcoinBeginners       1      0
##  2 Bitcoin          CoinBase               5      0
##  3 Bitcoin          CryptoCurrency         5      0
##  4 Bitcoin          ethereum               6      0
##  5 BitcoinBeginners Bitcoin              694      0
##  6 BitcoinBeginners BitcoinBeginners       0   1250
##  7 BitcoinBeginners CoinBase            1080      0
##  8 BitcoinBeginners CryptoCurrency       674      0
##  9 BitcoinBeginners ethereum             677      0
## 10 CoinBase         Bitcoin               22      0
## # ... with 13 more rows
```

The model did not perform very well! Looks like it predicted the most popular class (BitcoinBeginners) way more than all the other classes. That makes sense from a mathematical perspective, but is not great from a modeling perspective.