RStudio CH2 DGA Text Analysis for Hana

* Manually cleaned data removing all unnecessary punctuation, whitespace, and numbers.
* Saved cleaned text as a .txt file
* Read the data into RStudio
  + library(readtext)
  + DGA\_CH2\_raw <- readtext("/Users/connornorton/Desktop/DGA\_CH\_2\_clean.txt")
* Add text to a Corpus
  + DGA\_CH2\_Corpus <- Corpus(VectorSource(DGA\_CH2\_raw))
* Preprocess the Corpus
  + DGA\_CH2\_Corpus <- tm\_map(DGA\_CH2\_Corpus, tolower)
  + DGA\_CH2\_Corpus <- tm\_map(DGA\_CH2\_Corpus, removePunctuation)
  + DGA\_CH2\_Corpus <- tm\_map(DGA\_CH2\_Corpus, removeNumbers)
  + Removing stopwords
    - library(stopwords)
    - stop\_words <- stopwords("en")
    - stop\_words <- stop\_words[!stop\_words %in% c("can", "cannot", "could", "may", "must", "should", "shall", "would", "are", "be", "able", "to", "need")]
    - DGA\_CH2\_Corpus <- tm\_map(DGA\_CH2\_Corpus, removeWords, stop\_words)
* Document Term Matrix
  + Create a Document Term Matrix from the text in the Corpus
    - DGA\_CH2\_DTM <- DocumentTermMatrix(DGA\_CH2\_Corpus)
  + Find modal verb frequency using *tm* library
    - library(tm)
    - modal\_verb\_freq\_CH2 <- colSums(as.matrix(DGA\_CH2\_DTM[, c("can", "cannot", "could", "may", "must", "should", "shall", "would")]))
      * function could cause error if each word isn’t in text, may have to remove some modal verbs
    - Check for modal verb PHRASE frequency as well
      * library(stringr)
      * Convert data into a string
        + DGA\_CH2\_string <- as.String(DGA\_CH2\_Corpus)
      * freq\_areAbleTo\_CH2 <- str\_count(DGA\_CH2\_string, "are able to")
      * freq\_beAbleTo\_CH2 <- str\_count(DGA\_CH2\_string, "be able to")
    - Combine all modal verb frequencies into one vector
      * total\_modal\_verb\_freq\_CH2 <- append(total\_modal\_verb\_freq\_CH2, freq\_beAbleTo\_CH2)
      * names(total\_modal\_verb\_freq\_CH2)[10] <- "be able to"
    - Generate Table
      * table\_CH2\_modal\_verb\_freq <- data.frame(values = unlist(total\_modal\_verb\_freq\_CH2))

Table

Description automatically generated

* + Find all word frequency
    - library(tm)
    - all\_term\_freq\_CH2 <- colSums(as.matrix(DGA\_CH2\_DTM))
* Create individual word tokens from the Corpus
  + all\_CH2\_tokens <- text\_tokens(DGA\_CH2\_Corpus)
* Check the term stats of the Corpus
  + term\_stats(DGA\_CH2\_Corpus, subset = !term %in% stop\_words)

Table

Description automatically generated

* Generate word clouds
  + library(wordcloud)
  + Of all terms
    - All Black
      * Only the terms that appear < 3 times removed after custom stopwords are removed
        + wordcloud(names(all\_term\_freq\_CH2), all\_term\_freq\_CH2, min.freq = 3)
        + min freq = 3 because otherwise wordcloud would be too large

A close-up of a document

Description automatically generated with medium confidence

* + - * Only the terms that appear < 6 times removed after custom stopwords are removed
        + wordcloud(names(all\_term\_freq\_CH2), all\_term\_freq\_CH2, min.freq = 6)

A close-up of a newspaper

Description automatically generated with medium confidence

* + **Of only modal verbs**
    - Basic (all black)
      * wordcloud(names(total\_modal\_verb\_freq\_CH2), total\_modal\_verb\_freq\_CH2, min.freq = 1)

Text

Description automatically generated Logo

Description automatically generated with low confidence Text

Description automatically generated

* + - Color (RdBu)
      * wordcloud(names(total\_modal\_verb\_freq\_CH2), total\_modal\_verb\_freq\_CH2, min.freq = 1, colors = brewer.pal(6,"RdBu"))

**Logo, company name

Description automatically generated** **Logo, company name

Description automatically generated**

**Logo, company name

Description automatically generated**