Text Mining

Blake Pappas

2023-12-17

Text Mining in R

Load the following packages:

```
library(tm)
library(wordcloud)
```

In this exercise, we use the "superbowl.csv" file. It contains 2000 tweets about the Super Bowl.

Import the dataset as a corpus:

```
superbowl_text = read.csv("superbowl.csv")

superbowl_corpus = Corpus(VectorSource(superbowl_text$Tweet))
```

Examine the 100th tweet in this corpus. What does it say?

```
superbowl_corpus[[100]]$content
```

```
## [1] "Not a bad spot to catch the big game today #SuperBowl #udanationals http://t.co/8hbAWIuWj1"
```

Pre-Processing: Remove Punctuation

```
superbowl_corpus = tm_map(superbowl_corpus, removePunctuation)
```

Pre-Processing: Lower-Casing

```
superbowl_corpus = tm_map(superbowl_corpus, tolower)
```

Pre-Processing: Remove Stopwords

```
superbowl_corpus = tm_map(superbowl_corpus, removeWords,
                     stopwords('english'))
```

Pre-Processing: Word Stemming

```
superbowl_corpus = tm_map(superbowl_corpus, stemDocument)
```

Pre-Processing: Remove Excessive Blank Spaces

```
superbowl_corpus = tm_map(superbowl_corpus, stripWhitespace)
```

Obtain the Term-Document Matrix:

```
dtm = DocumentTermMatrix(superbowl_corpus)

dtm_matrix = as.matrix(dtm)
```

Find the top 5 most popular words in these tweets:

```
word_freq = colSums(as.matrix(dtm))

word_freq_sorted = sort(word_freq, decreasing = TRUE)

word_freq_sorted[1:5]
```

```
## superbowl    bronco   seahawk     super       win
##       828       790       710       556       539
```

Now look at top 10 most popular words.

Do you see anything unusual?

```
word_freq_sorted[1:10]
```

```
##                                     superbowl
##                                           828
##                                        bronco
##                                           790
##                                       seahawk
##                                           710
##                                         super
##                                           556
##                                           win
##                                           539
##                                          bowl
##                                           537
##                                          will
##                                           422
##                                       footbal
##                                           375
##                                          news
##                                           304
## 00000unknownunknownunknownunknownunknown
##                                           262
```

Answer: Yes, I see something unusual. The tenth most popular word isn't a word. It's more of an error message called "00000unknownunknownunknownunknownunknown".

Run the following lines of code to get rid of the unusual word:

```
superbowl_corpus = tm_map(superbowl_corpus, removeWords,
                          "00000unknownunknownunknownunknownunknown")
```

Update the word_freq variable:

```
dtm = DocumentTermMatrix(superbowl_corpus)

dtm_matrix = as.matrix(dtm)

word_freq = colSums(as.matrix(dtm))
```

```r
word_freq_sorted = sort(word_freq, decreasing = TRUE)

word_freq_sorted[1:10]
```

```
##  superbowl     bronco    seahawk      super        win       bowl       will
##        828        790        710        556        539        537        422
##    footbal       news sportscent
##        375        304        257
```

Plot a WordCloud of top 50 most popular words.

Fill in the first two parameters: "words" and "freq".

The last two parameters specify the size of words to display and how many words to show.

```r
wordcloud(words = names(word_freq_sorted),
          freq = word_freq_sorted,
          scale = c(2, 0.25),
          max.words = 50)
```