Word Clouds

Building a Word Cloud

Use text mining to build a word cloud

- Extract text from a speech (or any other text)
- Build a term-document matrix (from text)
- Find frequent words and associations from the matrix.
- A word cloud is used to present frequently occurring words in documents.

Text Mining Packages

Text mining packages in R include:

- tm: provides functions for text mining
- wordcloud: visualizes results

Reading Text

One way to read in text in R

> sbaFile <-
"http://www.historyplace.com/speeches/anthony.htm"

> sba <- readLines(sbaFile)

> str(sba)

chr [1:15] "Friends and fellow citizens: I stand before you tonight under indictment for the alleged crime of having voted at the last pres"| __truncated__ …

Corpus = “Bag of Words”

We first need to build a corpus

- A corpus is a “bag of words.”
- We coerce our text file vector (‘sba’) into a custom "Class" provided by the tm package called a "Corpus".
- The Corpus Class defines the most fundamental object that text miners care about, a corpus containing a collection of documents.

Text Transformations

Four transformations

- Making all of the letters lowercase
- Removing the punctuation
- Removing numbers
- Taking out the "stop" words
  - Words such as *the*, *a*, and *at* appear in so many different parts of the text that they are useless for differentiating between documents.

Text Transformations in R

Four transformations:

```
> words.vec <- VectorSource(sba)
> words.corpus <- Corpus(words.vec)
> words.corpus
<<VCorpus>>
Metadata:  corpus specific: 0, document level (indexed): 0
Content:  documents: 15

> words.corpus <- tm_map(words.corpus,
content_transformer(tolower))
> words.corpus <- tm_map(words.corpus, removePunctuation)
> words.corpus <- tm_map(words.corpus, removeNumbers)
> words.corpus <- tm_map(words.corpus, removeWords,
stopwords("english"))
```

A Term-Document Matrix

- A rectangular data structure with terms (words) as the rows and documents as the columns

- A term may be a single word, for example, *biology*, or it could also be a compound word, such as *data analysis*.

A Term-Document Matrix (continued)

- If a term like *data* appears once in the first document, twice in the second, and not at all in the third document, then the column for the term *data* will contain 1, 2, 0.

- Most term document matrices are quite sparse—the overwhelming number of cells that contain zero—indicating that the term does not appear in a document.

Creating a TermDocumentMatrix

```
> tdm <- TermDocumentMatrix(words.corpus)
> tdm
<<TermDocumentMatrix (terms: 189,
                                    documents: 15)>>

Non-/sparse entries: 225/2610
Sparsity           : 92%
Maximal term length: 20
Weighting          : term frequency (tf)
```

The wordcloud Function

- Expects two vectors as input arguments:
  - The first a list of the terms
  - The second a list of the frequencies of occurrence of the terms
- The list of terms and frequencies must be sorted with the most frequent terms first.

  → We first have to coerce our text data back into a plain data matrix so that we can sort it by frequency.

Creating a Word Cloud In R

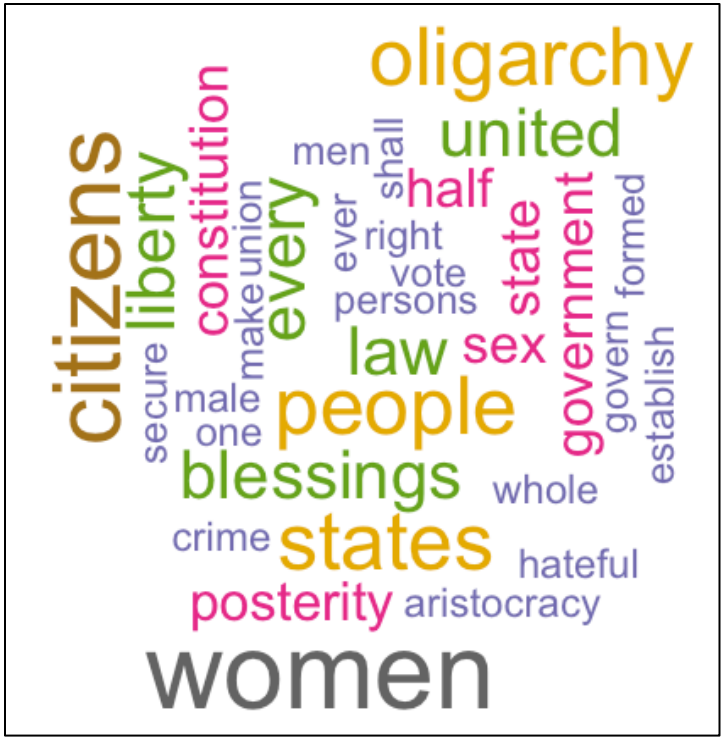
```
> m <- as.matrix(tdm)
> wordCounts <- rowSums(m)
> wordCounts <- sort(wordCounts, decreasing=TRUE)
> head(wordCounts)
   women  citizens oligarchy    people    states blessings
       7         6          5         5         5         4

> cloudFrame<-data.frame( +
  word=names(sortedMatrix),freq=sortedMatrix)
> wordcloud(cloudFrame$word,cloudFrame$freq)
```

Word Cloud Example

Another Word Cloud Example

```
wordcloud(names(wordCounts), wordCounts, min.freq=2, +
max.words=50, rot.per=0.35, colors=brewer.pal(8, "Dark2"))
```



Question:

How useful are word clouds?

When are they appropriate to use?

Sentiment Analysis

Conceptual Methodology

- Load Positive and Negative word Lists
- **Count** positive words and negative words (in entire document or part of a document)
- **Compute the ratio** of positive to negative words

Example R Code

Get the Positive and Negative Word Files

Access to positive and negative words:
https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

```
> pos <- "positive-words.txt"
> neg <- "negative-words.txt"

> #read the files
> p <- scan(pos, character(0),sep = "\n") # separate each word
Read 2040 items

> n <- scan(neg, character(0),sep = "\n") # separate each word
Read 4817 items
```

Clean Up the Word Files

#remove the first 34 lines (header info)

```
> p <- p[-1:-34]
> n <- n[-1:-34]
```

```
> head(p, 10)
[1] "a+"         "abound"     "abounds"    "abundance"  "abundant"   "accessable"
[7] "accessible" "acclaim"    "acclaimed"  "acclamation"
```

```
> head(n,10)
[1] "2-faced"    "2-faces"    "abnormal"   "abolish"    "abominable" "abominably"
[7] "abominate"  "abomination" "abort"     "aborted"
```

More R Code

Get the total number of words:

```
#calculate the total number of words
> totalWords <- sum(wordCounts)

#have a vector that just has all the words
words <- names(wordCounts)

> matched <- match(words, p, nomatch = 0)
> head(matched,10)
[1]    0    0    0    0    0    0    0    0 1083    0
```

More R Code

Explore the matched words:

```
> matched[9]
[1] 1083

> p[1083]
[1] "liberty"

> words[9]
[1] "liberty”
```

Calculate the Positive Word Count

```
> mCounts <- wordCounts[which(matched != 0)]
> length(mCounts)
[1] 12

> mWords <- names(mCounts)
> nPos <- sum(mCounts)
> nPos
[1] 17
```

Calculate the Negative Word Count

```
> matched <- match(words, n, nomatch = 0)
> nCounts <- wordCounts[which(matched != 0)]
> nNeg <- sum(nCounts)
> nWords <- names(nCounts)
> nNeg
[1] 13
> length(nCounts)
[1] 11
```

Calculate the Sentiment!

```
> #calculate the % of words that are positive or negative
> totalWords <- length(words)
> ratioPos <- nPos/totalWords
> ratioPos
[1] 0.08994709

> ratioNeg <- nNeg/totalWords
> ratioNeg
[1] 0.06878307
```

Given this, we can see that Susan B. Anthony's speech was about **9% positive** and a little less than **7% negative**.

Question

- **Does this truly measure sentiment?**
- **Where could it go wrong?**


SYRACUSE UNIVERSITY
SUOS
CULTORES
SCIENTIA
CORONAT