Text Analysis with R, part 1

Michele Claibourn, UVA Library's StatLab

October 16, 2018

Motivations

Text is everywhere!!!

- Government documents
- Social media
- Literature
- Digital archives
- News, letters, speeches, transcripts, articles, online reviews, ...

So let's use it...

Roadmap for today

Today we'll look at some core functions and key concepts in text analysis, including

- a little bit on acquiring data
- reading text data into R (and some data cleaning)
- creating corpora and metadata
- exploring a corpus
- processing text data, tokenizing, n-grams
- processing text data, reducing dimensionality (stopwords, lowercase, stem)
- creating a document-feature matrix (dfm)
- analysis on dfms, frequencies, relative frequencies and weighting, keyness

And implementation using auanteda!

Assumptions of Text as Data

- Texts represent an observable implication of some underlying characteristic of interest
- Texts can be represented through extracting their features
- A document-feature matrix can be analyzed using quantitative methods to produce meaningful and valid estimates of the underlying characteristic of interest

Corpora, Documents, Features

- Corpus: a large and structured set of texts for analysis
- Document: documentary unit of analysis, selected by researcher
- Feature: word or word stem, but also linguistic features (POS) or multi-word expressions (ngrams), selected phrases, human-annotated segments; to be converted into a quantitative matrix

Principles of Quanteda

quanteda has opinions!

- Corpus texts should remain unchanged during subsequent analysis and processing.
- A corpus should be capable of holding additional objects that will be associated with the corpus, such as dictionaries, stopword, and phrase lists.
- Objects should record histories of the operations applied to them.
- A dfm should always be structued as documents (or document groups) in rows and features in columns.
- Encoding of texts should always be UTF-8.

How quanteda compares. . .

Today's Text

1. Comments submitted to Ours to Shape

- scraped with `acquire_comments.R` (in materials)

2. Tweets from @realDonaldTrump from January 20, 2017 to September 30, 2018

- pulled from Brendan Brown's Trump Twitter Archive
- as part of the broader Public Presidency Project

Acquiring text

So many ways of acquiring text

- Databases and archives, e.g., Lexis-Nexis, JSTOR Data for Research, HathiTrust, Project Gutenberg, arXiv, PLOS, ...
- Twitter, reddit, yelp and other social media, generally via API's
- Scrapeable web content (not all sites allow scraping; here's a past workshop)
- Government documents – party manifestos, congressional speeches, central bank annoucements, ...
- Digitize your own (if archives allow scanning)
- Have MTurkers transcribe for you

R Script Supplements | Readability

Flesch-Kincaid Readababability: a measure of text complexity based on syllables and sentence length

$$0.39\left(\frac{words}{sentences}\right) + 11.8\left(\frac{syllables}{words}\right)$$

R Script Supplements | Processing to Reduce Dimensionality

1. Removing punctuation, special characters
2. Removing capitalization
3. Removing stopwords, words with no substantive content
4. Stemming, removes the ends of words to creqte equivalencies, e.g., family, familie, families', familial == famili (an approximation to lemmatization)
5. Trimming, remove words occuring in less than/more than X documents, or less than/more than Y times

R Script Supplements | Keyness

A measure of distinctiveness of words, or ability of words to distinguish documents by some category. Based on rates of word use in some group/category relative to average rate across corpus. A word which is positively keyed occurs more often than would be expected by chance in a group of doucments relative to a reference corpus.

R Script Supplements | Lexical Diversity

Carroll's Corrected Type-Token Ratio: a measure of text "richness" or complexity, based on how many differnet words are used (varied vocabulary) with little repetition

$$\frac{V}{\sqrt{2N}}$$

As the number of words increases, all TTR measures generate smaller values (longer texts generally require more repetition), so better at comparing documents of similar length.