Quantitative Text Analysis with Quanteda

Michele Claibourn, UVA Library's StatLab

April 10, 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 talk about key concepts in text analysis, including

- corpora and documents; operations on corpora (subsetting, reshaping, kwic, readability)
- document feature matrix (dfm); reducing dimensionality through preprocessing (lowercase, stemming, stopwords)
- analysis on dfms; term frequencies, weighting, similarity, lexicons, and more

And implementation using auanteda!

- quanteda implements some classification and scaling approaches, and feeds neatly into multiple topic model libraries (e.g., stm, topicmodel), but these are beyond our scope today.

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 a documents (or document groups) in rows by features in columns.
- Encoding of texts should always be UTF-8.

Trump Tweets

Today's text: Tweets from @realDonaldTrump from January 20, 2017 to April 9, 2018

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

Let's explore these a little (and review some functions in R while we're at it)!

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 | Similarity

With documents presented as vectors (e.g., vectors of word frequencies), measure the degree of similarity of two documents as the correlation between their corresponding vectors - the cosine of the angle between the two vectors.

Distance/Similarity is a reasonably big topic with lots of choices (weighting of vectors, measures of distance); distance measures are frequently used in clustering algorithms which also provide lots of choices. It's a big topic!