Computational Text Analysis

Summer Institute in Computational Social Science

Oxford University, June 2019

Taylor W. Brown

@taywhittenbrown

WHAT IS COMPUTATIONAL TEXT ANALYSIS?

COMPUTATIONAL TEXT ANALYSIS

Having to do with computers.

COMPUTATIONAL TEXT ANALYSIS

Any object that can be "read."

COMPUTATIONAL TEXT ANALYSIS

Systematic examination of the structure or mechanisms of something.

COMPUTATIONAL TEXT ANALYSIS

Systematic, computer-assisted examination of the structure or mechanisms of readable content.

COMPUTATIONAL TEXT ANALYSIS

within social science

"Policy makers or computer scientists may be interested in finding the needle in the haystack, (such as a potential terrorist threat or the right web page to display from a search), but social scientists are more commonly interested in characterizing the haystack."
(Hopkins & King, 2010, p. 230)

1600s Catholic church tracks proportion of nonreligious printed texts

1934 Laswell produces first key-word count

1940s Social scientists use similar methods

1950 Turin applies AI to text

1952 Bereleson publishes first textbook on Content Analysis

1954 First automatic translation of text (Georgetown Experiment)

1966 Stone & Bales use mainframe computer to measure psychometric text properties

1980 Machine learning applied to NLP

1985 Schrodt introduces automated event coding

1986 Pennebaker develops LIWC

1989 Franzosi brings Quantitative Narrative Analysis to social science

1998 First Topic Models Developed

1998 Mohr conducts first Quantitative Analysis of Worldviews

1999 Bearman et al. apply Network Methods to Narratives

2001 Blei et al. develop LDA

2003 MALLET created

2005 Quin et al use analyze political speeches using topic models

2010 King/Hopkins Bring Topic Models into mainstream

2014 Margaret Roberts, et al. develop Structural Topic Models

June 19, 2019

GETTING THE DATA…

Where is the data?

TWITTER

Barbera (2015). *Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data*. Political Analysis.

Munger (2017). *Tweetment Effects on the Tweeted: Experimentally Reducing Racist Harassment.* Political Behavior.

Tan, Lee, & Pang (2014). *The effect of wording on message propagation: Topic- and author-controlled natural experiments on Twitter.* arXiv.org.

REDDIT

Chandrasekhara et al. (2017). *You Can't Stay Here: The Efficacy of Reddit's 2015 Ban Examined Through Hate Speech.* ACM-HCI.

FACEBOOK

Bail, Brown, Mann (2017). *Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation*. ASR.

KICKSTARTER

Mitra & Gilbert (2014). *The Language That Gets People to Give: Phrases That Predict Success on Kickstarter.* CSCW.

AIRBNB

Ma et al. (2017). *Self-Disclosure and Perceived Trustworthiness of Airbnb Host Profiles.* CSCW.

OTHER

King, Pan, & Roberts (2013). *How Censorship in China Allows Government Criticism but Silences Collective Expression.* American Political Science Review.

Where is the data?

OPEN-ENDED SURVEYS

Roberts et al. (2014). *Structural Topic Models for Open-Ended Survey Responses.* American Journal of Political Science.

HISTORICAL ARCHIVES

Bearman & Stovel (2000). *Becoming a Nazi: A model for narrative networks*. Poetics.

Miller (2013). *Rebellion, crime and violence in Qing China, 1722–1911: A topic modeling approach.* Poetics.

ENRON EMAILS

Prabhakaran & Rambow (2017). *Dialog Structure Through the Lens of Gender, Gender Environment, and Power.* Dialogue & Discourse.

POLITICAL DOCUMENTS

Rule, Cointet, Bearman (2015). *Lexical shifts, substantive changes, and continuity in State of the Union discourse, 1790–2014*. PNAS.

Mohr, Wagner-Pacifici, Breiger, & Bogdanov (2013). *Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics.* Poetics.

NEWSPAPERS

DiMaggio, Nag, Blei (2013). *Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding.* Poetics.

Andrews & Caren (2010). *Making the News: Movement Organizations, Media Attention, and the Public Agenda*. ASR.

Where is the data?

Google Ngrams

English-Corpora.org

The Manifesto Project

Spinn3r

InternetArchive

How do you get it?

Open Source / API

Private agreement

Purchased

Scraped

TERMS OF USE.
TERMS OF USE.
TERMS OF USE.

PREPARING THE DATA…

Preprocessing

https://frieze.com/article/time-interrupted-toulouses-contemporary-art-biennial[\n \n Springtime in September \u2013 the defining concept of the Toulouse contemporary art biennial \u2013 seems, right now, like an immensely pleasing prospect. At the blustery conclusion of a summer that has been, depending on your predicament or point of view, either a blissful heatwave haze or an unsettling foretaste of the apocalyptic climate to come, there might be few things more attractive than the impossible dream of turning back time, putting the weather in rewind, hitting refresh on spring. Such, it might seem, is the fanciful aspiration of \u2018Le Printemps de Septembre\u2019. Under the directorship of veteran curator Christian Bernard, this richly varied biennial once again invites us to share in an autumn-replacing printemps: an abundant flowering of art across a city and its environs \u2013 an offer that has, on the face of it, an upbeat, benevolent civic generosity. Art, we might presume, is like spring: with it comes revival, new life, more light. \n But what if \u2013 to darken the mood \u2013 this bonus springtime doesn\u2019t cancel the imminent autumn and instead coincides with it, complicating it?...

Under the directorship of veteran curator Christian Bernard, \t this richly varied biennial once again invites us to share in an autumn-replacing printemps: an abundant flowering of art across a city and its environs \u2013 an offer that has, on the face of it, an upbeat, benevolent civic generosity.

GREP, which stands for "Globally search a Regular Expression and Print."

```
text <- "Under the directorship of veteran curator
Christian Bernard, \t this richly varied biennial
once again invites us to share in an autumn-
replacing <em>printemps</em>: an abundant
flowering of art across a city and its environs
\u2013 an offer that has, on the face of it, an
upbeat, benevolent civic generosity."
```

gsub("\t", "", text)

```
"Under the directorship of veteran curator
Christian Bernard, this richly varied biennial
once again invites us to share in an autumn-
replacing <em>printemps</em>: an abundant
flowering of art across a city and its environs
\u2013 an offer that has, on the face of it, an
upbeat, benevolent civic generosity."
```

```
cleanFun <- function(htmlString) {
  return(gsub("<.*?>", "", htmlString))
}
```

GREP cheatsheet:
http://www.rstudio.com/wp-content/uploads/2016/09/RegExCheatsheet.pdf

Under the directorship of veteran curator Christian Bernard, this richly varied biennial once again invites us to share in an autumn-replacing printemps: an abundant flowering of art across a city and its environs an offer that has, on the face of it, an upbeat, benevolent civic generosity.

Preprocessing

Stop word removal

Removing words that are extremely common but unrelated to the quantity of interest. This can include function words (e.g. “and”, “the”, “then”, “at”, “or”, etc), but can also be corpus specific.

Stemming and lemmatization

Stemming removes the endings of conjugated verbs or plural pronouns, returning only the ‘stem’

running → run
saw (v.) → saw

Lemmatization identifies the base form of the word and groups these words together, returning only the ‘lemma’

saw (n.) → saw
saw (v.) → see

Preprocessing

N-grams

unigram: ‘new’ ‘york’ ‘city’
bigram(s): ‘new_york’ ‘york_city’
trigram(s): ‘new_york_city’

Identifying parts-of-speech

she_prp
sells_vbz
seashells_nns

on_in
the_dt
seashore_nn

Identifying named-entities

a subtask of information extraction that seeks to locate and classify named entity mentions in unstructured text into pre-defined categories such as the person names, organizations, locations, etc

Now we'll move over to R and the tutorial on preprocessing, but to give a preview of our next session…

ANALYZING THE DATA…

Supervised: Given a labeled training sample, supervised learning trains a function in some function family F, with the goal that f(x) predicts the true label y on future data x.

Unsupervised: No labels providing supervision as to how individual instances should be handled.

Semi-supervised: A combination of these two, for example with pseudo-labeling.

Supervised: Given a labeled training sample, supervised learning trains a function in some function family F, with the goal that f(x) predicts the true label y on future data x.

Unsupervised: No labels providing supervision as to how individual instances should be handled.

Semi-supervised: A combination of these two, for example with pseudo-labeling.

When we come back…

Topic Models

Word Embedding

Network Analysis