

ECPR Methods Summer School: Automated Collection of Web and Social Data

Pablo Barberá

School of International Relations
University of Southern California
pablobarbera.com

Networked Democracy Lab
www.netdem.org

Course website:
github.com/pablobarbera/ECPR-SC103

Text as data

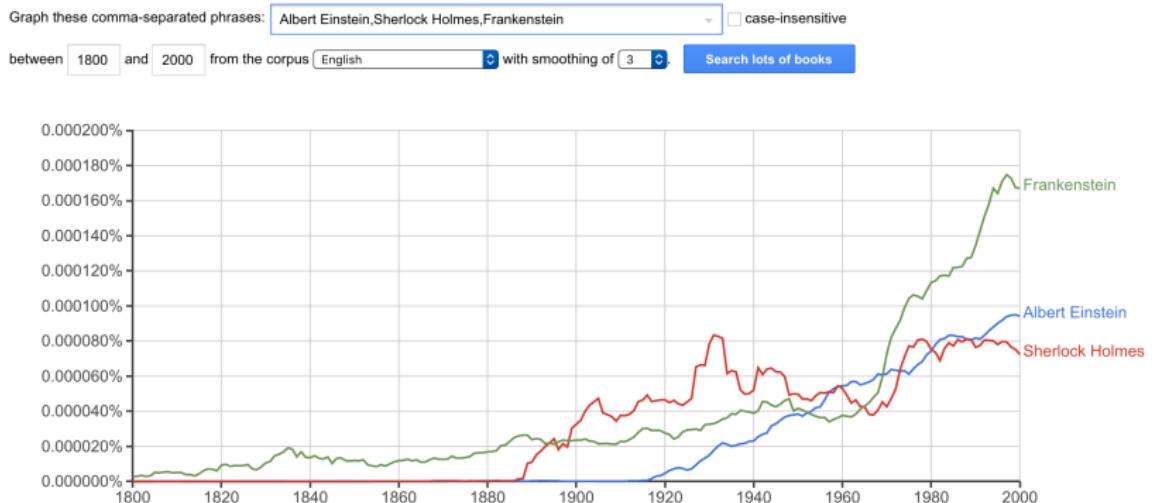


Text as data



Text as data

Google Books Ngram Viewer



Text as data



Overview of text as data methods

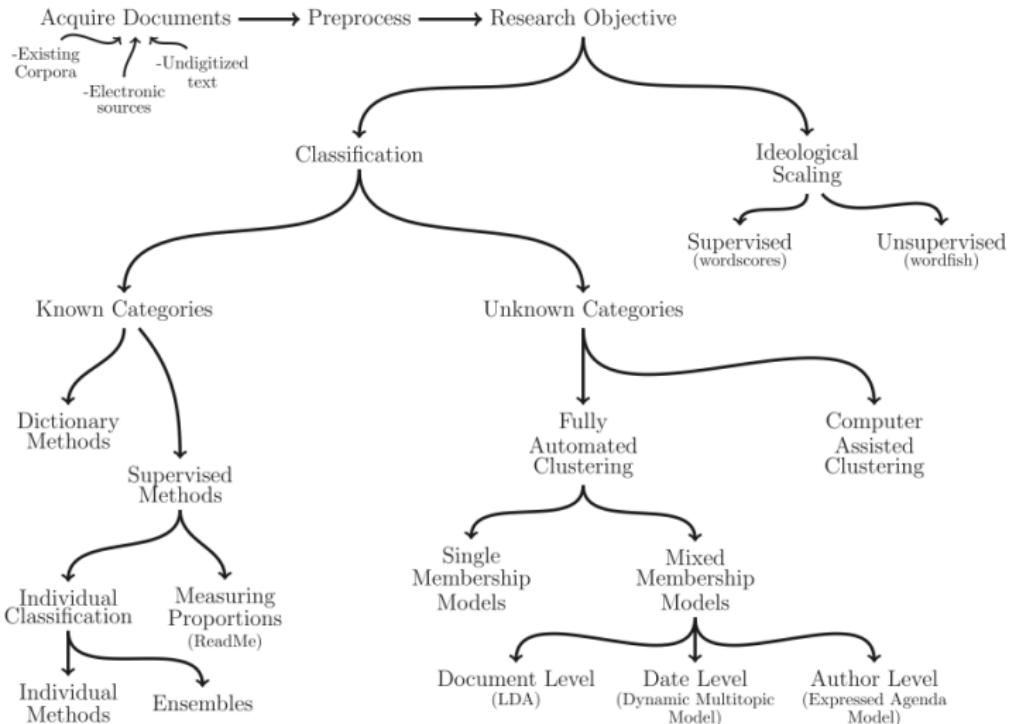


Fig. 1 in Grimmer and Stewart (2013)

From words to numbers

1. Bag-of-words assumption
2. Pre-processing text
 - ▶ Capitalization, cleaning digits/URLs, removing stopwords and sparse words, etc.
 - ▶ Stemming / lemmatization
 - ▶ Part-of-speech tagging
3. Document-term matrix
 - ▶ \mathbf{W} : matrix of N documents by M unique words
 - ▶ W_{im} = number of times m -th words appears in i -th document.
 - ▶ Usually large matrix, but sparse (so it fits in memory)

From words to numbers

From words to numbers

1. Preprocess text: lowercase, remove stopwords and punctuation, stem, tokenize into unigrams and bigrams (bag-of-words assumption)

[@, thank, congratul, you'r, best, #ep2014, @ thank, thank congratul, congratul you'r, you'r best, best, best #ep2014]

[@, you'r, idiot, never, vote, @ you'r, you'r idiot, idiot never, never vote]

2. Document-term matrix:

- \mathbf{W} : matrix of N documents by M unique n-grams
- w_{im} = number of times m -th n-gram appears in i -th document.

	@	thank	congratul	you'r	#ep2014	@ thank	:	M words
Document 1	1	1	1	1	1	1	...	
Document 2	1	0	0	1	0	0	...	
...								
Document n	0	1	1	0	0	0	...	

Dictionary methods

Classifying documents when categories are known using dictionaries:

- ▶ Lists of words that correspond to each category:
 - ▶ Positive or negative, for sentiment
 - ▶ Sad, happy, angry, anxious... for emotions
 - ▶ Insight, causation, discrepancy, tentative... for cognitive processes
 - ▶ Sexism, homophobia, xenophobia, racism... for hate speech
- ▶ many others: see LIWC, VADER, SentiStrength, LexiCoder...
- ▶ Count number of times they appear in each document
- ▶ Normalize by document length (optional)
- ▶ Validate, validate, validate.
 - ▶ Check sensitivity of results to exclusion of specific words
 - ▶ Code a few documents manually and see if dictionary prediction aligns with human coding of document