Text Analysis Techniques

A Brief Overview

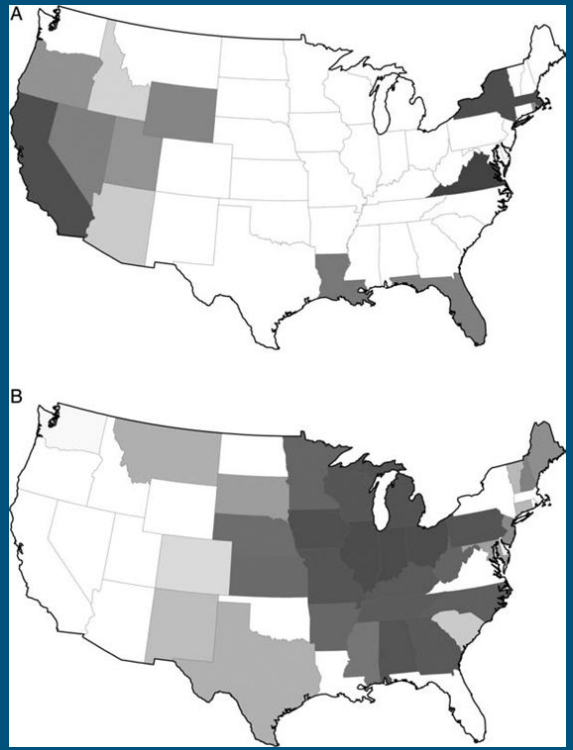
Text Analysis demystified: It's just counting.

Sense and Sensibility		Moby Dick	
to	4063	the	13721
the	3861	of	6536
of	3565	and	6024
...			
secret	19	cold	30
discourse	19	jaws	30
hoped	19	won	30
age	19	harpoons	30

Techniques

- Natural Language Processing (takes into account context)
  - Part of Speech, Named Entities, Concordances, Word Similarities, semantics

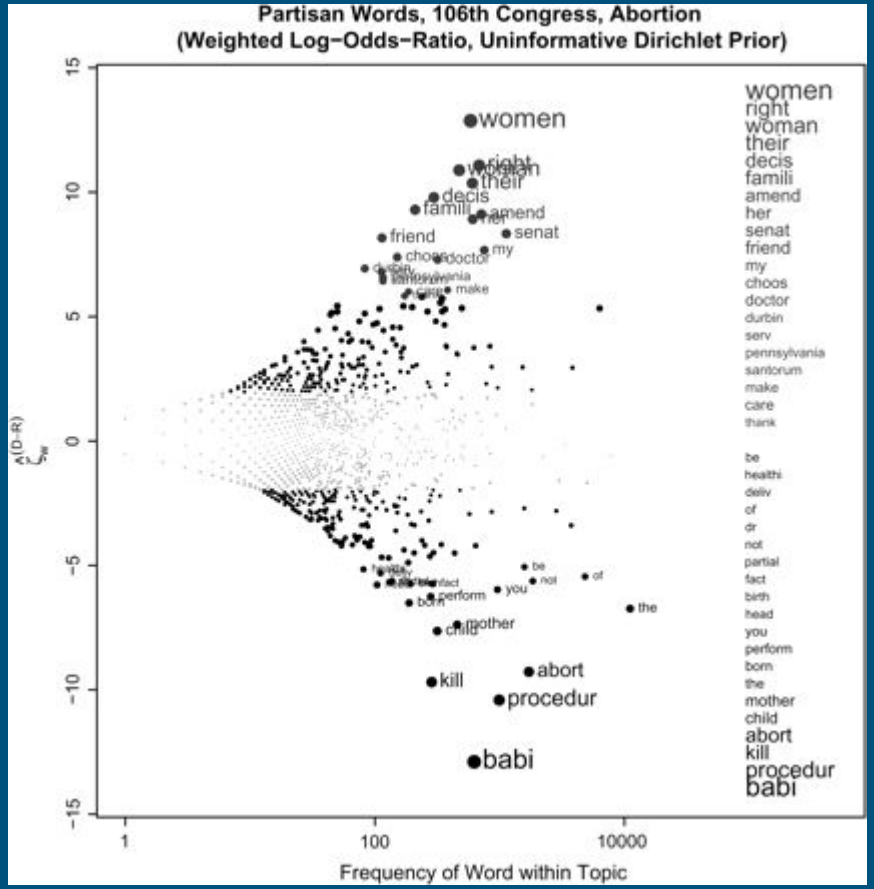
Fig. 7: Dunning log-likelihood values for named-location counts in the full corpus measured against mean state populations, 1850–80. (a) States overrepresented relative to their populations; (b) underrepresented states. Darker shades indicate larger absolute values, hence greater under- or overrepresentation.



Wilkens, Matthew. 2013. “The Geographic Imagination of Civil War-Era American Fiction.” *American Literary History* 25(4): 803-840.

Techniques

- Natural Language Processing (takes into account context)
  - Part of Speech, Named Entities, Concordances, Word Similarities, semantics
- Distinguishing Words
  - Difference of Proportions, TF-IDF, chi-squared, Dunning Log-Likelihood)

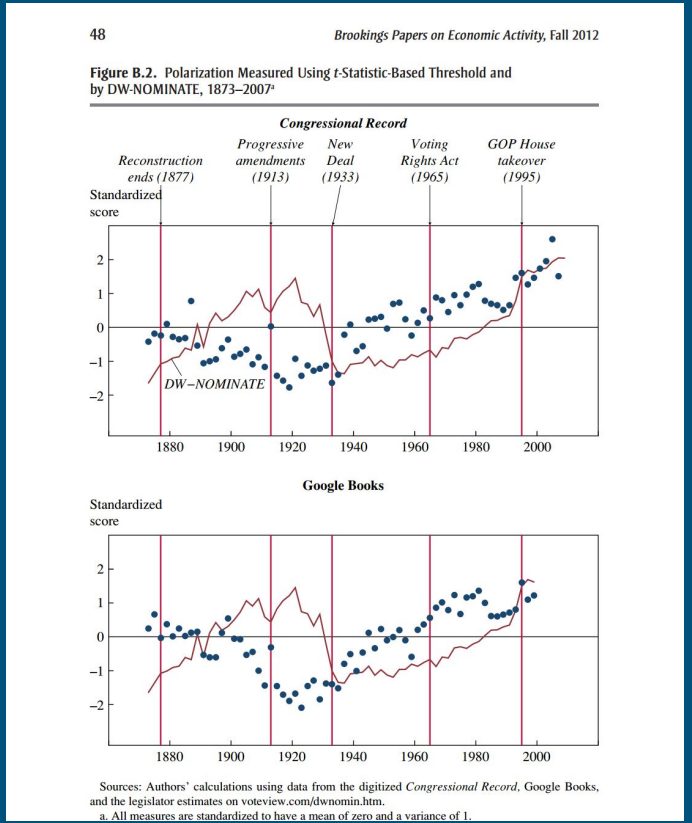


From: Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict
Polit Anal. 2009;16(4):372-403.
doi:10.1093/pan/mpn018

Techniques

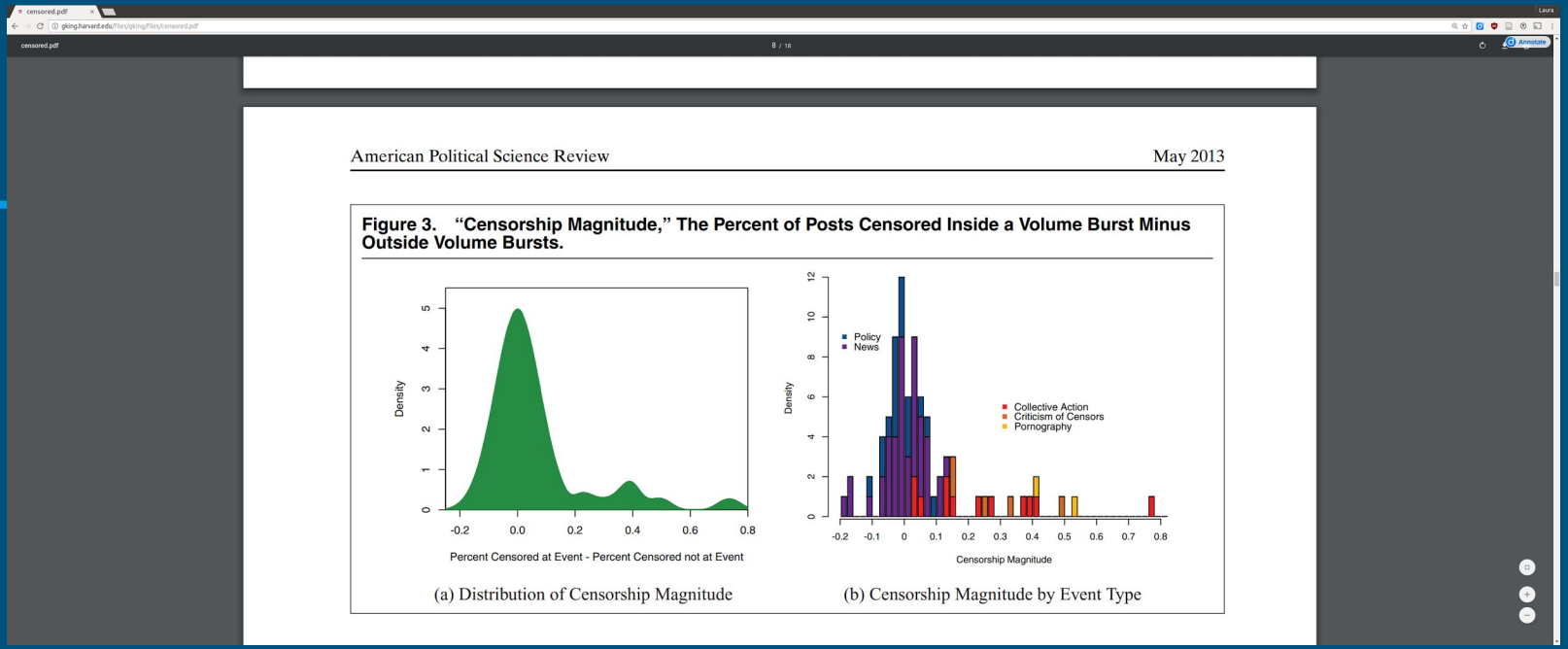
- Natural Language Processing (takes into account context)
  - Part of Speech, Named Entities, Concordances, Word Similarities, semantics
- Distinguishing Words
  - Difference of Proportions, TF-IDF, chi-squared, Dunning Log-Likelihood)
- Dictionary Methods (themes)

Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. 2012. “Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech.” *Brookings Papers on Economic Activity*.



Techniques

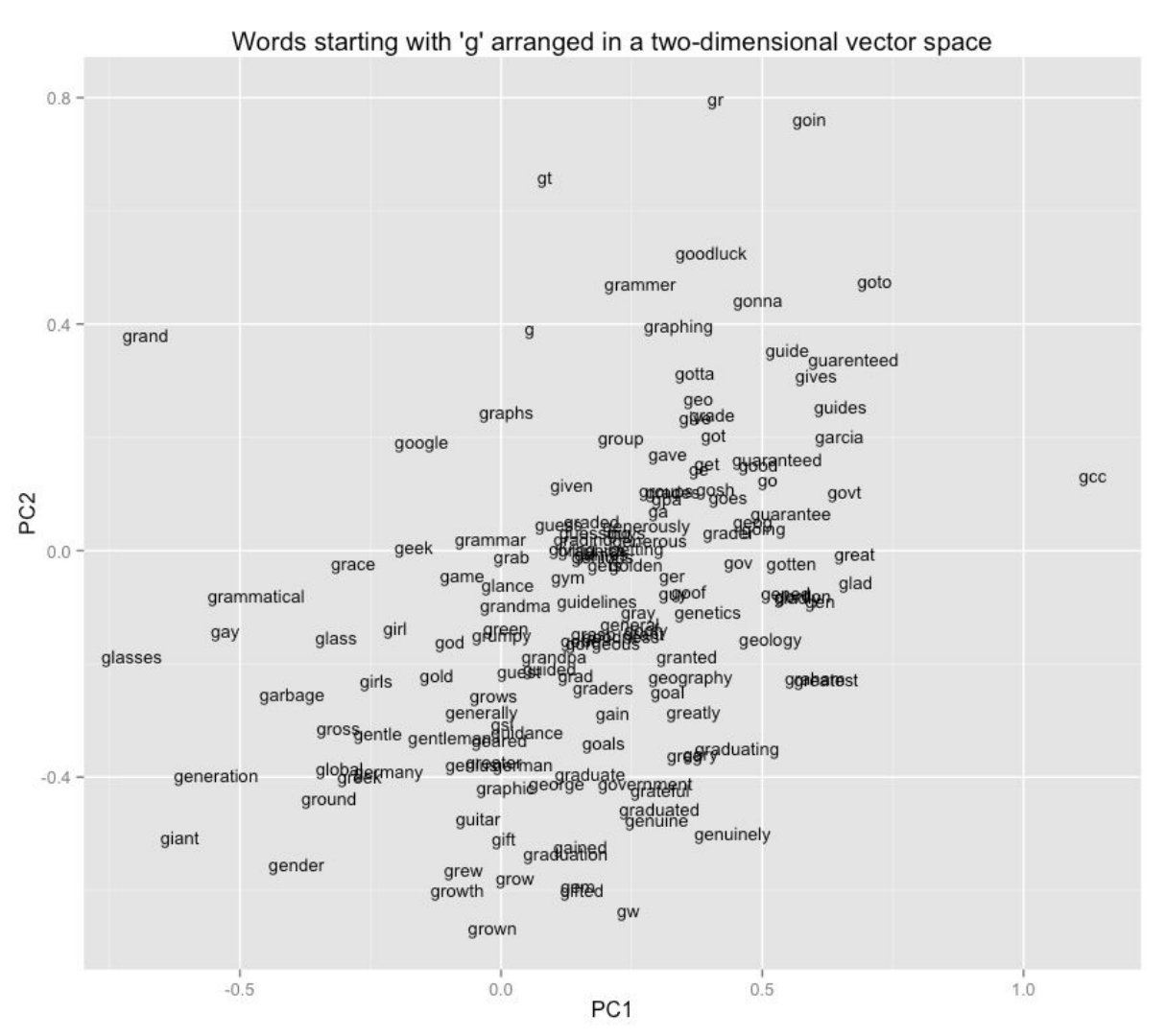
- Natural Language Processing (takes into account context)
  - Part of Speech, Named Entities, Concordances, Word Similarities, semantics
- Distinguishing Words
  - Difference of Proportions, TF-IDF, chi-squared, Dunning Log-Likelihood)
- Dictionary Methods (themes)
- Supervised Machine Learning (classification)



King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* 107(2): 1-18.

Techniques

- Natural Language Processing (takes into account context)
  - Part of Speech, Named Entities, Concordances, Word Similarities, semantics
- Distinguishing Words
  - Difference of Proportions, TF-IDF, chi-squared, Dunning Log-Likelihood)
- Dictionary Methods (themes)
- Supervised Machine Learning (classification)
- Vector Space Models
  - Clustering, document distance



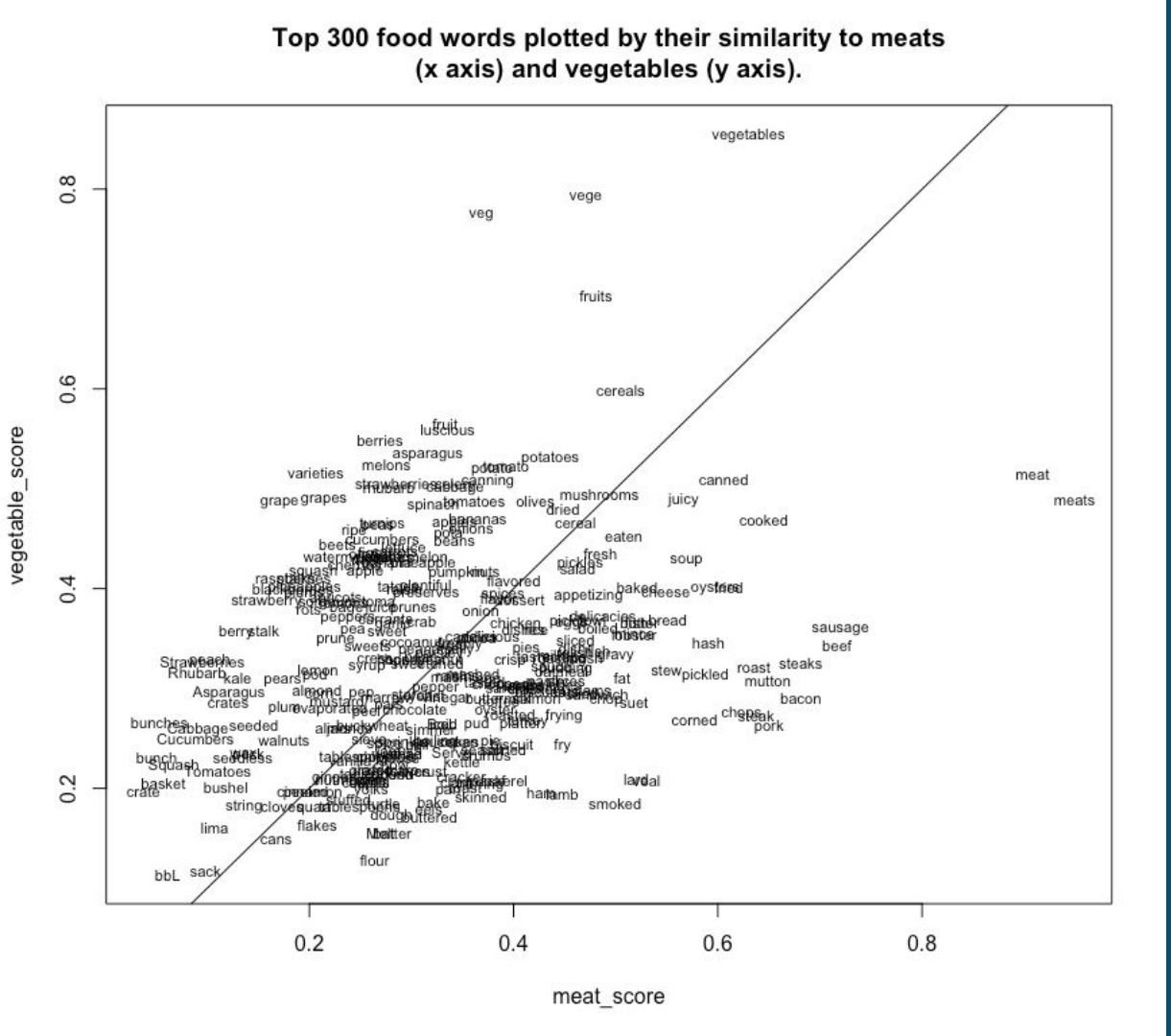
Ben Schmidt, Word Embeddings Models.

Techniques

- Natural Language Processing (takes into account context)
  - Part of Speech, Named Entities, Concordances, Word Similarities, semantics
- Distinguishing Words
  - Difference of Proportions, TF-IDF, chi-squared, Dunning Log-Likelihood)
- Dictionary Methods (themes)
- Supervised Machine Learning (classification)
- Vector Space Models
  - Clustering, document distance
- Topic Modeling

Techniques

- Natural Language Processing (takes into account context)
  - Part of Speech, Named Entities, Concordances, Word Similarities, semantics
- Distinguishing Words
  - Difference of Proportions, TF-IDF, chi-squared, Dunning Log-Likelihood)
- Dictionary Methods (themes)
- Supervised Machine Learning (classification)
- Vector Space Models
  - Clustering, document distance
- Topic Modeling
- Word embeddings
  - Word2Vec, GloVe



Ben Schmidt, Word Embeddings Models.

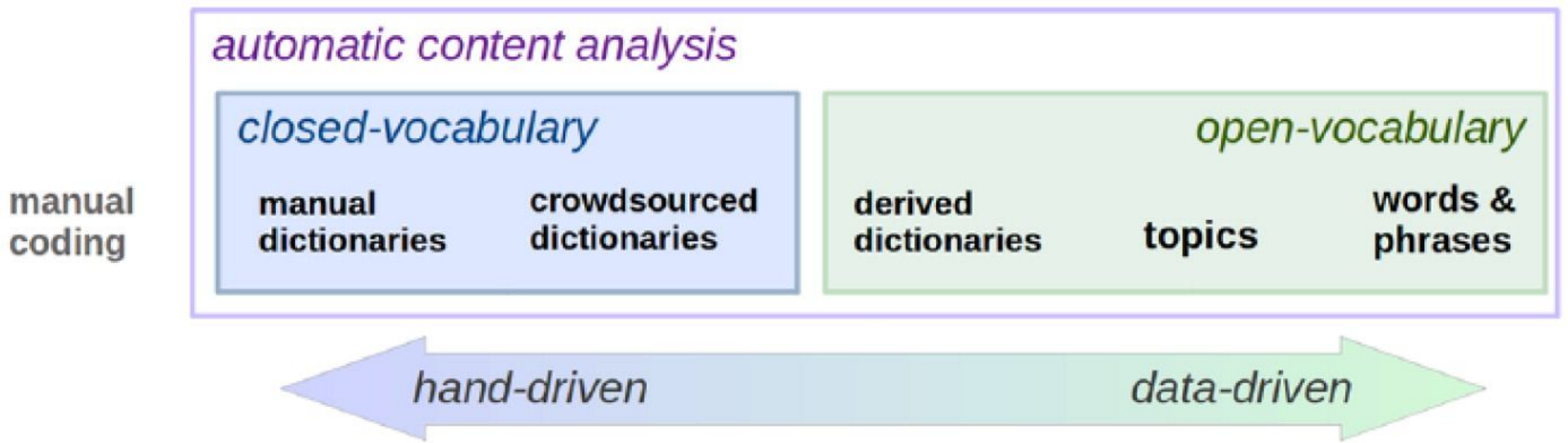


Figure 1 Categorization of Content Analysis Techniques

Published in: Dhavan V. Shah; Joseph N. Cappella; W. Russell Neuman;
Published in: H. Andrew Schwartz; Lyle H. Ungar; The ANNALS of the American Academy of Political and Social Science 659, 78-94.


LET'S BEGIN