Tools for Text

University of Washington/Seattle

14-15 June 2010

Sponsored by NSF, Society for Political Methodology,
& the Center for American Politics and Public Policy

A Very Brief Introduction to Text as Data

Philip A. Schrodt and John D. Wilkerson
schrodt@psu.edu
jwilker@u.washington.edu

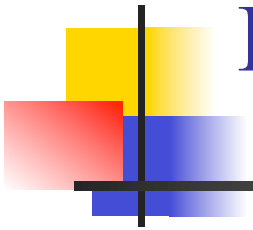
Tools for Text Workshop
University of Washington/Seattle

14-15 June 2010

Key points to be made:

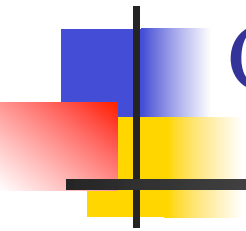
- Text analysis has developed substantially in the past two decades
- There are a variety well-developed technologies from a number of different fields for systematically analyzing text; text has regular statistical characteristics
- The Web has made a tremendous amount of data available in machine readable form, at your desktop, for free. However, there can be considerable variation in how easy it is to acquire and analyze.

In short: this is low-hanging fruit—an under-utilized method that can be applied to numerous interesting problems



But first…

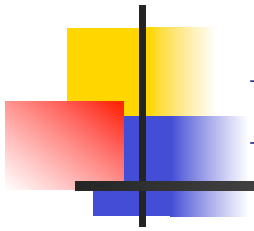
- You won't understand everything here the first time
- But you will understand it the k-th time, where k is a small integer
  - Geek humor
- So even if you don't understand something, you are making progress towards understanding it
- Even if this sometimes seems not to be true…



Contemporary Content Analysis

Levels of content analysis

Analytical Term	Linguistic Term	Methodology
Thematic	Lexical	Analysis of words and phrases. “bag of words”
Syntactic	Syntactic	Use grammatical rules to determine role of words
Network	Semantic	Use relationships between words to disambiguate meanings



Research in other fields

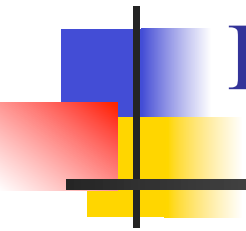
Library science	Automated indexing
Computational Linguistics	Automated translation and natural language processing generally
Psychology	Personality tests
Communications Studies	Content of popular culture —books, movie and television scripts
Education	Automated grading
Business	Automated evaluation of resumés, aptitude tests

Potential text sources relevant to political behavior

- News reports
- Legislation
- Campaign platforms and party manifestos
- Campaign web sites
- Editorials
- Open ended survey questions
- New media: blogs, tweets, social networking sites

Advantages of text as a data source

- Text is one of the primary methods of communicating political information
- The source material is intentional: it was created for some political purpose, either to *persuade, inform,* or *implement*
- Text is unaffected by the act of measurement
- Web-based text can be collected in near-real-time at very little cost
- Machine-assisted coding dramatically decreases any text analysis project, even when it is largely human coded
- A single individual can create an original, customized data set with little or no funding



Human and Automated Coding

Reliability in content analysis

- **Stability**—the ability of a coder to consistently assign the same code to a given text;
- **Reproducibility**—intercoder reliability;
- **Accuracy**—the ability of a group of coders to conform to a standard.

Source: Weber (1990:17)

Advantages of automated coding

- Fast and inexpensive
- Transparent: coding rules are explicit in the dictionaries
- Reproducible: a coding system can be consistently maintained over a period of time without the "coding drift" caused by changing teams of coders.
- Coding dictionaries are also be shared between institutions
- The coding of individual reports is not affected by the biases of individual coders. Dictionaries, however, can be so affected.
- Can create rules for difficult tecchnical and cultural vocabulary that is otherwise difficult to learn

Disadvantages of automated coding

- Automated thematic coding has problems with disambiguation; automated syntactic coding makes errors on complex sentences.
- Requires a properly formatted, machine-readable source of text, therefore older paper and microfilm sources are difficult to code.
- Development of new coding dictionaries is time-consuming—KEDS/PANDA initial dictionary development required 2-labor-years. (Modification of existing dictionaries, however, requires far less effort)

Human and machine coding tradeoffs

- Machine coding uses only information that is explicit in the text; human coders are likely to use implicit knowledge of the situation.
- Machine coding is not affected by boredom and fatigue
- Human coders can more effectively interpret idiomatic and metaphorical text, provided they are familiar with the context
- Human coders can more effectively deal with complex subordinate phrases and other unexpected grammatical constructions

Summary

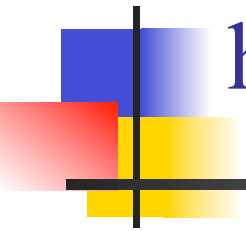
Advantage to human coding

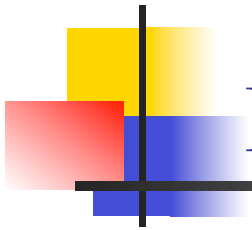
- Small data sets
- Data coded only one time at a single site
- Existing dictionaries cannot be modified
- Complex sentence structure
- Metaphorical, idiomatic, or time- dependent text
- Money available to fund coders and supervisors

Advantage to machine coding

- Large data sets
- Data coded over a period of time or across projects
- Existing dictionaries can be modified
- Simple sentence structures
- Literal, present-tense text
- Money is limited

Basic challenges that make text analysis hard



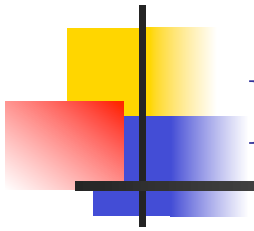


Example: Some text

For everywhere we look, there is work to be done. The state of the economy calls for action, bold and swift, and we will act - not only to create new jobs, but to lay a new foundation for growth. We will build the roads and bridges, the electric grids and digital lines that feed our commerce and bind us together. We will restore science to its rightful place, and wield technology's wonders to raise health care's quality and lower its cost. We will harness the sun and the winds and the soil to fuel our cars and run our factories. And we will transform our schools and colleges and universities to meet the demands of a new age. All this we can do. And all this we will do.

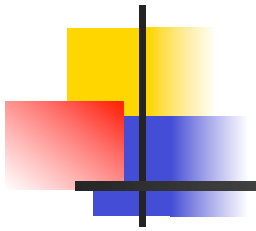
What it actually looks like

- For everywhere we look, there is work to be done. The state of our economy calls for action, bold and swift. And we will act, not only to create new jobs, but to lay a new foundation for growth. We will build the roads and bridges, the electric grids and digital lines that feed our commerce and bind us together. We'll restore science to its rightful place, and wield technology's wonders to raise health care's quality and lower its cost. We will harness the sun and the winds and the soil to fuel our cars and run our factories. And we will transform our schools and colleges and universities to meet the demands of a new age. All this we can do. All this we will do.



Pre-processing

- We see distinct words, sentences, punctuation and make distinctions between entities and actions.
- A computer sees none of this… and more
- It must be instructed to distinguish what we think is important – "tokenizing"



Features

- Next step is to get machine to pay attention to the tokens that are relevant to our goals – and to ignore those that are not.
- Explicitly delineate relevant tokens – keywords

- Remove generally irrelevant stuff
  - Stop words
  - Stemming

Stop words and Stemming

Stop words

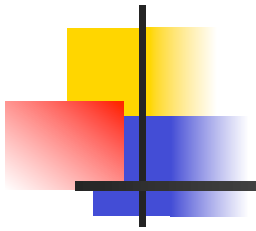
- a about above across after again against all almost alone along…. t take taken than that the their them then ….

Stemming

- Suffix stripping
  - "ing"
- N-grams
  - "post office"
- Lemmatization
  - "meeting"
- It's complicated!

What *information* is in this text?

For everywhere we look, there is work to be done. The state of the economy calls for action, bold and swift, and we will act - not only to create new jobs, but to lay a new foundation for growth. We will build the roads and bridges, the electric grids and digital lines that feed our commerce and bind us together. We will restore science to its rightful place, and wield technology's wonders to raise health care's quality and lower its cost. We will harness the sun and the winds and the soil to fuel our cars and run our factories. And we will transform our schools and colleges and universities to meet the demands of a new age. All this we can do. And all this we will do.



Examples

- Presence (or absence) of specific words
- Word frequencies
- Frequencies of words with similar meanings
- Words in context
- People, places, etc
- Topics, issues, etc
- Attitude, emotion, opinion
- Rhetoric or framing
- Relationships
- Gender, Personality, Intelligence
- ?

What can we *do* with this information?

- Find (search)
- Group (categorize)
- Compare (across or over time)
- Scale
- Predict (attitudes, behavior)
  - And much, much more….

May not be easy…

A bill to prohibit tobacco sales

A bill to prohibit gun sales

A bill to promote safe streets

A bill to promote smoking cessation

Use text similarities and differences to group these bills

Same but not relevant

A bill to prohibit tobacco sales

A bill to prohibit gun sales

A bill to promote safe streets

A bill to promote smoking cessation

Same and probably not relevant

A bill to prohibit tobacco sales

A bill to prohibit gun sales

A bill to promote safe streets

A bill to promote smoking cessation

Probably relevant but not the same

A bill to prohibit tobacco sales

A bill to prohibit gun sales

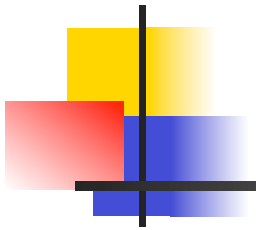
A bill to promote safe streets

A bill to promote smoking cessation

Word frequency in English

% of usage	# of words
40%	50
60%	2,300
85%	8,000
99%	16,000

- Total words in American English: about 600,000
- Total words in technical English (all fields): about 3-million



Functional Words

Very short words such as

- Articles: `a an the`
- Interogatives: `who what when where why how`
- Prepositions: `to from at in above below`
- Auxillary verbs: `have has was were been`
- Markers: `by in at to` (French `de`, German `du`, Arabic `fi`)
- Pronouns: `I you he she him her his hers`

In English, the specificity of a word is *generally* proportional to its length. These short will typically be in the stop word list, though a few longer words (e.g. “though” and “although”) also will be stop words

Marker words have multiple uses: *Random House College Dictionary* lists 29 meanings for “by,” 31 for “in,” 25 for “to,” and 15 for “for.”

It gets harder: Disambiguation (“Bat”)

- Noun
  - wooden (or aluminum) cylinder used in the game of baseball
  - small flying mammal
- Verb
  - act of batting (“at bat”)
  - blinking (“bat an eye”)
- Idiomatic uses
  - “go to bat for”: defending or interceding;
  - “right off the bat”: immediately;
  - “bats in the belfry”: commentary on an individual’s cognitive ability
- Foreign phrases
  - “bat mitzvah”: a girl’s coming-of-age ceremony (Hebrew).

Disambiguation, cont.

- Any of these uses might be encountered in an English-language text. Multiple uses might be found in a single sentence:

  "The umpire didn't bat an eye as Sarah lowered her bat to watch the bat flying around the pitcher."

Disambiguation, cont.

- Words can also change from verbs to nouns without modification. Consider
  - `I plan to drive to the store, then wash the car.`
  - `When John returned from the car wash, he parked his car in the drive.`

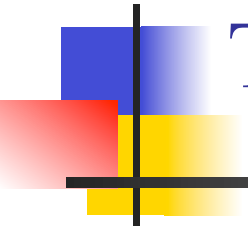
- In summary:

"Verbing weirds language."

Bill Watterson, *Calvin and Hobbes*

Memes, idioms, metaphors and slang

- Political text frequently uses distinct idiomatic phrases
  - "Right to life", "right to choice"
- Memes can have a high frequency for brief periods of time
  - "lipstick on a pig"
  - "top kill", "junk shot", "Deep Horizon"
- Military metaphors are common in political (and sports) rhetoric
  - "Tea Party insurgency", "battleground state"
- OMG! WTF! Like IMHO slang expressions are common, and rapidly changing, in new media (lol…)



Text as a Statistical Object

What does a document look like as a statistical object?

- Mathematically:
  - it is a high-dimensional, sparse feature vector where the elements of the vector are the frequencies of specific words and phrases in the document
- Geometrically:
  - it is a point in a high-dimensional space.

- Upshot:
  - Anything you can do with points, you can do with documents

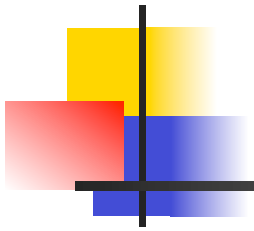
What does a document look like as a statistical object?

	T1	T2	T3	T4	T5	T6	T7	T8
Doc1	2	0	4	3	0	1	0	2
Doc2	0	2	4	0	2	3	0	0
Doc3	4	0	1	3	0	1	0	1
Doc4	0	1	0	2	0	0	1	0
Doc5	0	0	2	0	0	4	0	0
Doc6	1	1	0	2	0	1	1	3
Doc7	2	1	3	4	0	2	0	2

The Term-Document Matrix

Reduction of Dimensionality

- Computational incentives
  - Eliminate information that does not distinguish between documents: stop words
  - Combine words that have the same information: stemming
- Conceptual incentives:
  - Deductive: identify groups of words or phrases that are consistently associated with the concepts you are trying to code
  - Inductive: give a set of related texts, find the common language, which may not be obvious
- Statistical incentives
  - Words that occur everywhere are noise and may make documents seem more similar than they are
  - Words that almost never occur are not useful for machine learning, even if they are very meaningful for a human coder



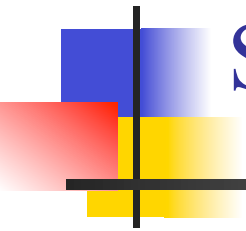
Zipf's Law (a.k.a. rank-size law)

"The frequency of the occurrence of a word in a natural language is inversely proportional to its rank in frequency"

- In mathematics: $f_i \propto 1/r_i$
- In English (or any other natural language): A small number of words account for most of word usage

Zipf's Law collides with statistical analysis

- Information theory:
  - the information contained in an item of data is proportional to $\log(f_i)$
- Statistical Efficiency:
  - the standard error of a parameter estimate is inversely proportional to the square root of the sample size
- Upshot: Any content analysis must balance the high level of information contained in low-frequency words with the requirements of getting a sample of those words sufficiently large for reasonable parameter estimation



Statistical Methods of Text Analysis

Statistical Methods I:
Reduction of dimensionality

Objective: Approximate the high dimensional space with a space of lower dimensionality while preserving as much of the variance as possible in the original space

- Factor analysis: correlation metric
- Principal components: Euclidean metric
- Correspondence analysis: chi-square metric

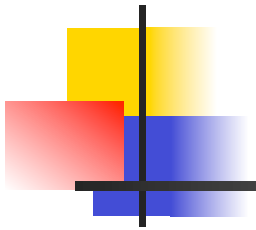
Result: Document can be characterized by a small number of composite indicators

Statistical Methods II:
Cluster analysis

Objective: Determine clusters of documents that are similar to each other based on their feature vectors

- Nearest neighbor methods—K-Means, KNN
- Contextual Clustering
- Decision trees

Result: Documents can be clustered in groups that have credible substantive interpretations



Statistical Methods III: Classification algorithms

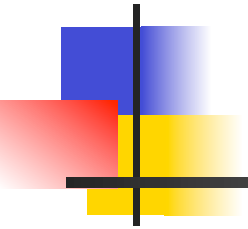
Objective: identify the characteristics of documents that are most useful in differentiating them into categories that have been specified a priori

- Discriminant analysis
- SVM
- Neural networks
- Numerous text-specific methods—naïve Bayes, tf-idf—we will be discussing over the next two days

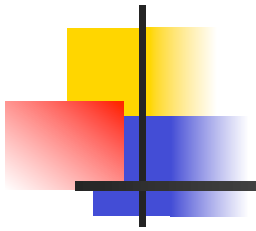
Result: documents can be used to classify cases into a set of categories

Some additional comments

- From the perspective of a political scientist, text analysis often requires unfamiliar methods—not everything can be done with linear models
  - Some, but not all, of these methods are very standard and well understood in other fields, however
- The distinction between statistical and machine-learning methods is very fuzzy
- Usually if a method works for English, it will work for any other language as well
  - Coding Finnish is a point of pride in some projects…



Some additional considerations



Source and consistency of text

How much work will be involved in getting the text into a form you can use?

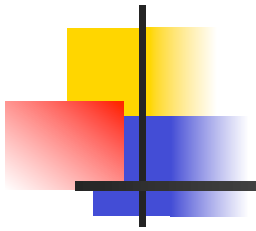
- ASCII/UniCode text, for example news reports
- HTML
  - web formats change frequently ☹
- PDF
- Scanned/OCR text
- Proprietary word processing formats (Word, WordPerfect)
- New media sources such as blogs and tweets

Style of language

- News reports and official documents are usually formal, syntactically-correct English
- Quotations and letters are a mix of formal and informal
- Open-ended responses range from formal to very fragmentary
- New media sources are often very informal and abbreviated
- Variants of English and changes in usage over time (e.g. slang, memes)
- Languages other than English

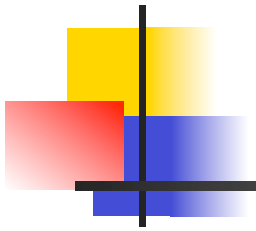
Intellectual property

- Copyright law is generally open to "fair use" in research and education. However, institutional contracts with data providers are more limited
  - Information does not necessarily want to be free
- Human subjects considerations—and therefore IRB review—apply to identifiable data
- A lot of the legal issues, particularly involving content on the web, are still *very* open
  - You probably do not want to be a test case



And of course, costs

- Is the information source already formatted?
  - Spinn3R, Thomas ☺
  - Web pages vary dramatically in ease of downloading
- How much data do you actually need?
  - Text data sets are frequency much, much larger than typical political science data such as surveys and national indicators
  - Will just a sample be sufficient?
  - Are you coding more information than you will actually use?
- How much time will it take to code each document?
  - Who's going to train and supervise the manual coders?
  - How much can be fully automated?
  - How good is good enough?



The next two days

- Very little on *getting* and *preparing* text as data
- Focus is on *introducing* tools for *analyzing* text as data

- Manual Annotation
- Unsupervised learning
- Supervised learning
- Dimensional scaling
- Behavior and prediction

- Using them well requires dedication (but well worth it!)