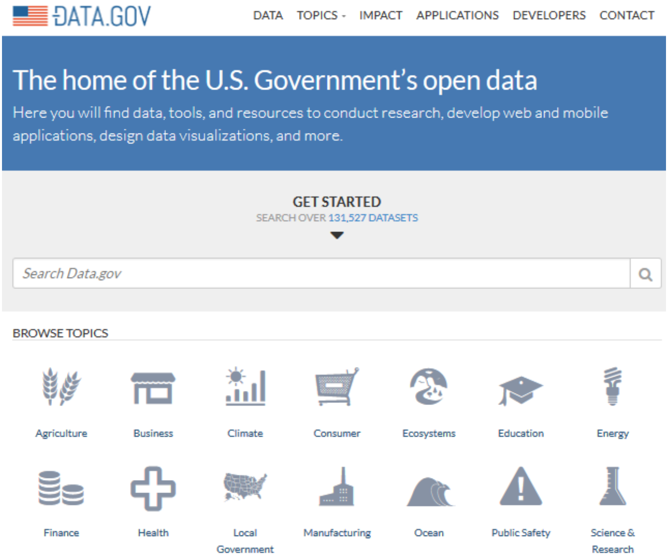
Government Text as Data: Opportunities and Challenges

John Wilkerson, Andreu Casas
University of Washington
jwilker@uw.edu

June 22, 2015

CAP Text as Data Workshop

A World of Possibility



DATA.GOV

DATA TOPICS IMPACT APPLICATIONS DEVELOPERS CONTACT

The home of the U.S. Government's open data

Here you will find data, tools, and resources to conduct research, develop web and mobile applications, design data visualizations, and more.

GET STARTED

SEARCH OVER 131,527 DATASETS

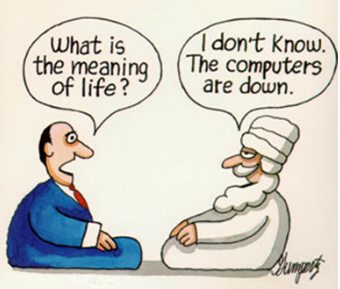
Search Data.gov

BROWSE TOPICS

Agriculture, Business, Climate, Consumer, Ecosystems, Education, Energy, Finance, Health, Local Government, Manufacturing, Ocean, Public Safety, Science & Research

First hour

- Steps in the process
- Takeaways
- Example applications
- Popular analytic methods

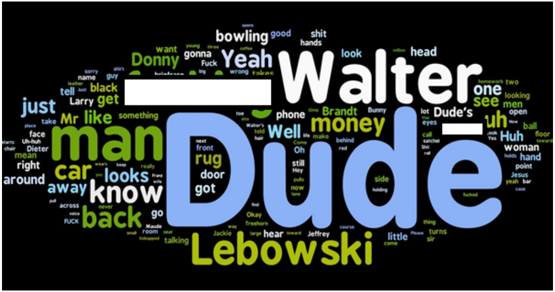

What is the meaning of life?
I don't Know. The computers are down.

Second hour

- Run some scripts
- Other things to discuss?

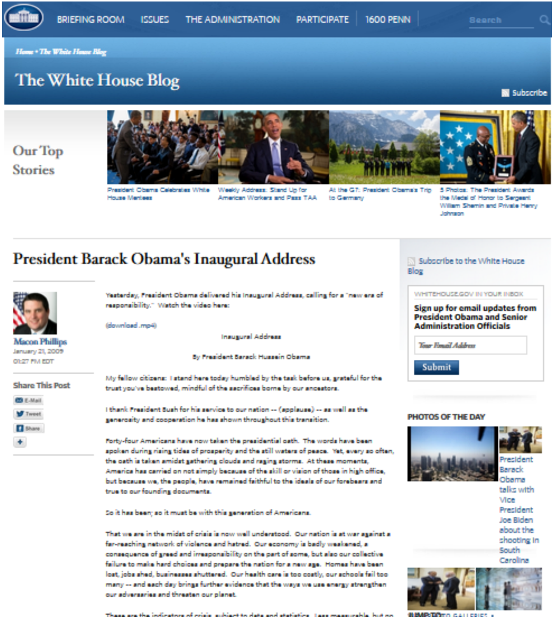
Typical Steps in Working with Text as Data

- Get text
- Specify its important "features"
- Convert features to numeric data
- Analyze the data quantitatively
- Visualize the results?



Takeaways

- Someone already wrote the code - just need to find and adapt it
- Data prep is often the most time consuming part of a project.
- Humans are more perceptive than computers
- Advantage lies in scale - be ambitious or don’t bother
- Validate!



The White House Blog

Our Top Stories

President Obama Celebrates White House Mentees

Weekly Address: Stand Up for American Workers and Pass TAA

At the G7: President Obama's Trip to Germany

5 Photos: The President Awards the Medal of Honor to Sergeant William Shemin and Private Henry Johnson

President Barack Obama's Inaugural Address

Macon Phillips
January 21, 2009
01:27 PM EDT

Yesterday, President Obama delivered his Inaugural Address, calling for a "new era of responsibility." Watch the video here:

(download .mp4)

Inaugural Address

By President Barack Hussein Obama

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you've bestowed, mindful of the sacrifices borne by our ancestors.

I thank President Bush for his service to our nation -- (applause) -- as well as the generosity and cooperation he has shown throughout this transition.

Forty-four Americans have now taken the presidential oath. The words have been spoken during rising tides of prosperity and the still waters of peace. Yet, every so often, the oath is taken amidst gathering clouds and raging storms. At these moments, America has carried on not simply because of the skill or vision of those in high office, but because we, the people, have remained faithful to the ideals of our forebears and true to our founding documents.

So it has been; so it must be with this generation of Americans.

That we are in the midst of crisis is now well understood. Our nation is at war against a far-reaching network of violence and hatred. Our economy is badly weakened, a consequence of greed and irresponsibility on the part of some, but also our collective failure to make hard choices and prepare the nation for a new age. Homes have been lost, jobs shed, businesses shuttered. Our health care is too costly, our schools fail too many -- and each day brings further evidence that the ways we use energy strengthen our adversaries and threaten our planet.

These are the indicators of crisis, subject to data and statistics. Less measurable, but no

Subscribe to the White House Blog

WHITEHOUSE.GOV IN YOUR INBOX

Sign up for email updates from President Obama and Senior Administration Officials

PHOTOS OF THE DAY

President Barack Obama talks with Vice President Joe Biden about the shooting in South Carolina

Read in file(s)

Download to a desktop directory to be accessed in R or Python

What is DownThemAll?

DownThemAll (or just dTa) is a powerful yet easy-to-use Mozilla Firefox extension that adds **new advanced download capabilities** to your browser.

DownThemAll lets you **download all the links or images contained in a webpage** and much more: you can refine your downloads by fully customizable criteria to get only what you really want.

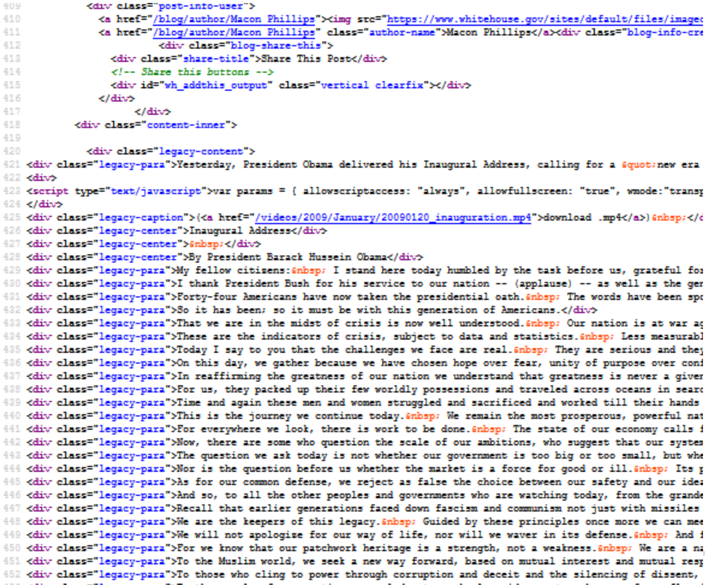
DownThemAll is all you can desire from a download manager: it features an advanced accelerator that **increases speed up to 400%**, it allows you to pause and resume downloads at any time and, last but not least, it's fully integrated into your **favorite browser**!

Install DownThemAll now!

Latest stable version: 2.0.18

What the web page actually looks like

If there’s an API or raw text file, use it!

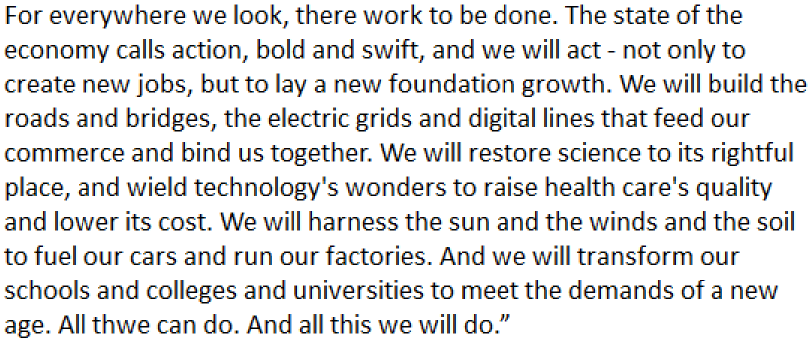


```
            <div class="post-info-user">
              <a href="/blog/author/Macon Phillips"><img src="https://www.whitehouse.gov/sites/default/files/imagec
              <a href="/blog/author/Macon Phillips" class="author-name">Macon Phillips</a><div class="blog-info-cre
                        <div class="blog-share-this">
                <div class="share-title">Share This Post</div>
                <!-- Share this buttons -->
                <div id="wh_addthis_output" class="vertical clearfix"></div>
              </div>
                    </div>
          <div class="content-inner">

            <div class="legacy-content">
<div class="legacy-para">Yesterday, President Obama delivered his Inaugural Address, calling for a &quot;new era
<div>
<script type="text/javascript">var params = { allowscriptaccess: "always", allowfullscreen: "true", wmode:"transp
</div>
<div class="legacy-caption">(<a href="/videos/2009/January/20090120_inauguration.mp4">download .mp4</a>)&nbsp;</c
<div class="legacy-center">Inaugural Address</div>
<div class="legacy-center">&nbsp;</div>
<div class="legacy-center">By President Barack Hussein Obama</div>
<div class="legacy-para">My fellow citizens:&nbsp; I stand here today humbled by the task before us, grateful for
<div class="legacy-para">I thank President Bush for his service to our nation -- (applause) -- as well as the ger
<div class="legacy-para">Forty-four Americans have now taken the presidential oath.&nbsp; The words have been spo
<div class="legacy-para">So it has been; so it must be with this generation of Americans.</div>
<div class="legacy-para">That we are in the midst of crisis is now well understood.&nbsp; Our nation is at war ag
<div class="legacy-para">These are the indicators of crisis, subject to data and statistics.&nbsp; Less measurabl
<div class="legacy-para">Today I say to you that the challenges we face are real.&nbsp; They are serious and they
<div class="legacy-para">On this day, we gather because we have chosen hope over fear, unity of purpose over conf
<div class="legacy-para">In reaffirming the greatness of our nation we understand that greatness is never a giver
<div class="legacy-para">For us, they packed up their few worldly possessions and traveled across oceans in searc
<div class="legacy-para">Time and again these men and women struggled and sacrificed and worked till their hands
<div class="legacy-para">This is the journey we continue today.&nbsp; We remain the most prosperous, powerful nat
<div class="legacy-para">For everywhere we look, there is work to be done.&nbsp; The state of our economy calls f
<div class="legacy-para">Now, there are some who question the scale of our ambitions, who suggest that our system
<div class="legacy-para">The question we ask today is not whether our government is too big or too small, but whe
<div class="legacy-para">Nor is the question before us whether the market is a force for good or ill.&nbsp; Its p
<div class="legacy-para">As for our common defense, we reject as false the choice between our safety and our idea
<div class="legacy-para">And so, to all the other peoples and governments who are watching today, from the grande
<div class="legacy-para">Recall that earlier generations faced down fascism and communism not just with missiles
<div class="legacy-para">We are the keepers of this legacy.&nbsp; Guided by these principles once more we can mee
<div class="legacy-para">We will not apologize for our way of life, nor will we waver in its defense.&nbsp; And f
<div class="legacy-para">For we know that our patchwork heritage is a strength, not a weakness.&nbsp; We are a na
<div class="legacy-para">To the Muslim world, we seek a new way forward, based on mutual interest and mutual resp
<div class="legacy-para">To those who cling to power through corruption and deceit and the silencing of dissent,
```

Strip the formatting language

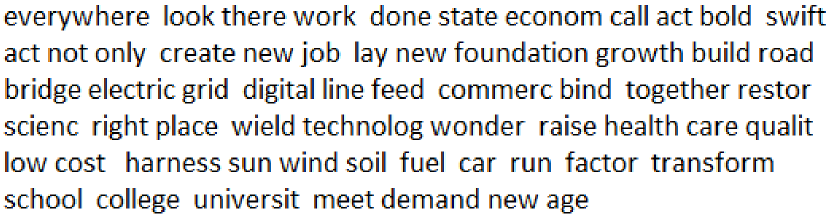
(Might also want to ’parse’ by sentence, paragraph etc)



For everywhere we look, there work to be done. The state of the economy calls action, bold and swift, and we will act - not only to create new jobs, but to lay a new foundation 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 thwe can do. And all this we will do.”

Define relevant features

Remove things that are not relevant, such as punctuation, "stopwords," sparse (rare) words, and stem words



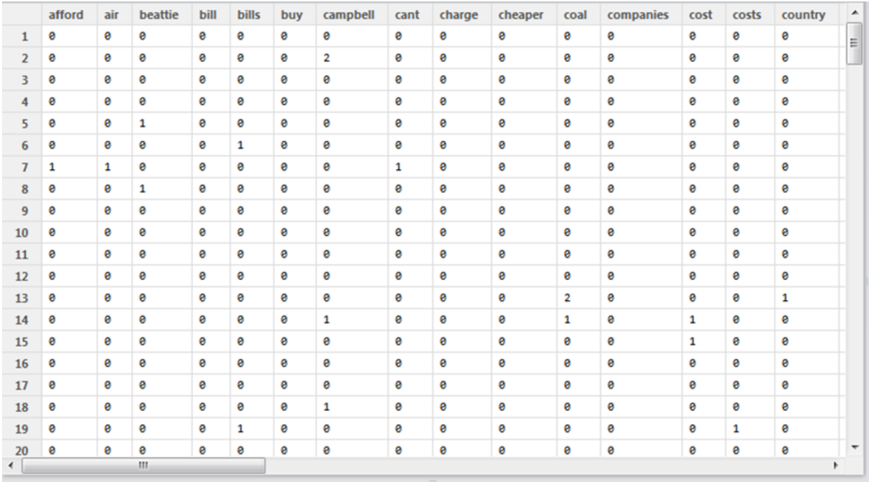
everywhere look there work done state econom call act bold swift act not only create new job lay new foundation growth build road bridge electric grid digital line feed commerc bind together restor scienc right place wield technolog wonder raise health care qualit low cost harness sun wind soil fuel car run factor transform school college universit meet demand new age

Aside: Endless possibilities for features

- individual words (unigrams)
- word combinations (bigrams, n-grams)
- word weighting?
- external information (e.g. hair color; date of birth)

Convert to Document Term Matrix or...

Each row is a document (e.g. sentence or hearing); each column is a feature (e.g. term).



	afford	air	beattie	bill	bills	buy	campbell	cant	charge	cheaper	coal	companies	cost	costs	country
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
7	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0
8	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1
14	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
19	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Aside: Regular expressions are the bomb!

- Sequence of characters that define a search pattern
- Can use to extract or parse by dates, words etc.

([0-9]1,2[0-9]1,2[0-9]4)
e.g. 06/22/2014

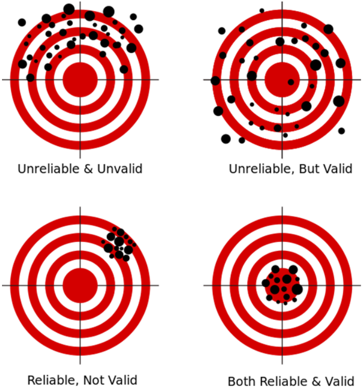
https://docs.python.org/2/library/re.html

Popular Text as Data methods

- keywords
- crowdsourcing
- unsupervised machine learning
- supervised machine learning
- natural language processing

As with any method...

- Validity: Are you measuring what you want to measure?
- Reliability: Are you doing so consistently?



Unreliable & Unvalid

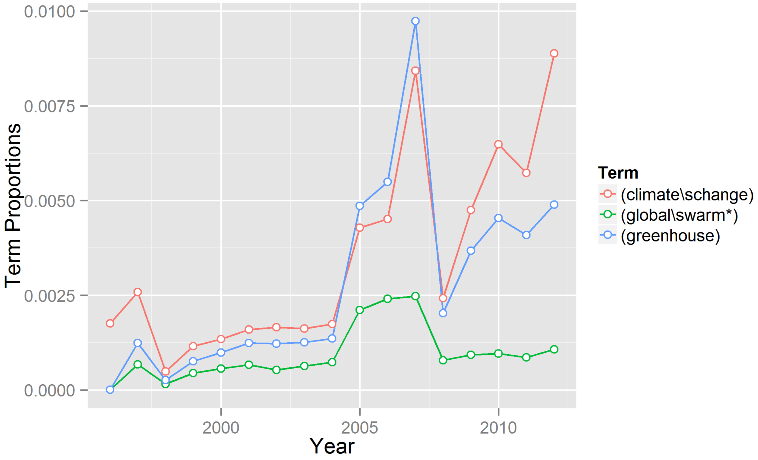
Unreliable, But Valid

Reliable, Not Valid

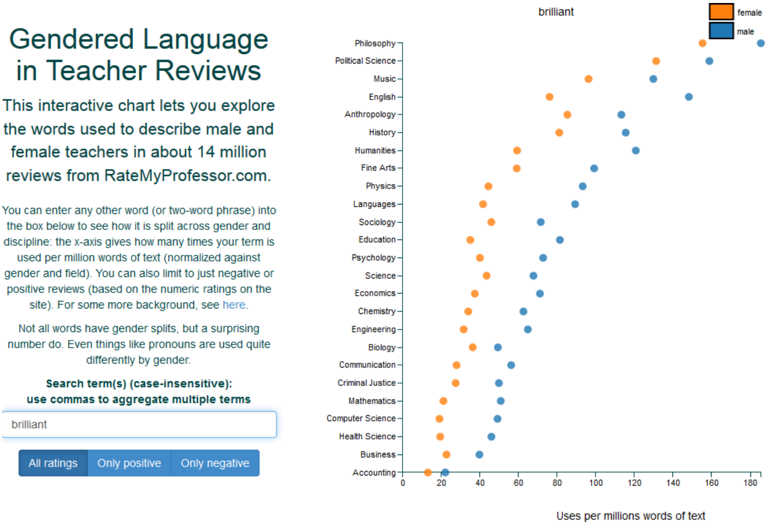
Both Reliable & Valid

Keywords

- Do you really need fancy?
- Methodology is explicit, replicable



Keywords



Gendered Language in Teacher Reviews

This interactive chart lets you explore the words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com.

You can enter any other word (or two-word phrase) into the box below to see how it is split across gender and discipline: the x-axis gives how many times your term is used per million words of text (normalized against gender and field). You can also limit to just negative or positive reviews (based on the numeric ratings on the site). For some more background, see here.

Not all words have gender splits, but a surprising number do. Even things like pronouns are used quite differently by gender.

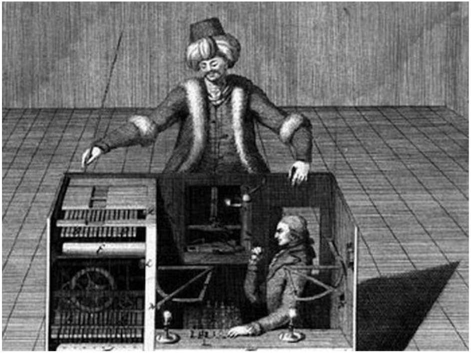
Search term(s) (case-insensitive): use commas to aggregate multiple terms

brilliant

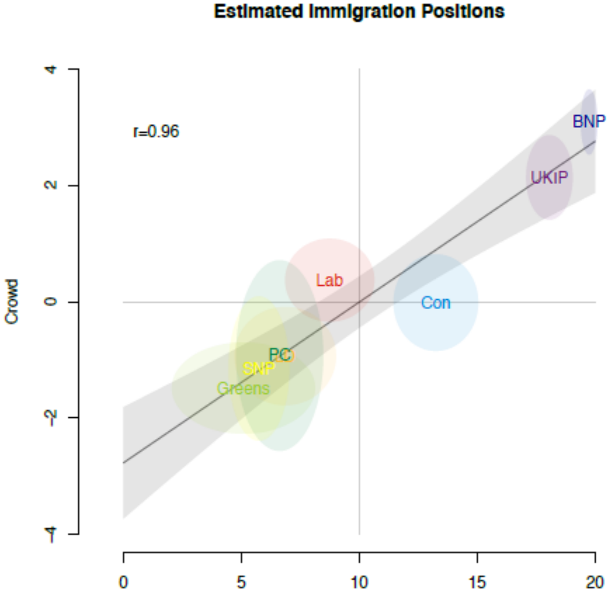
http://benschmidt.org/profGender/

Crowdsourcing

- "Artificial artificial intelligence"
- Inexpensive option for the right task
- Companies will design a project including quality control

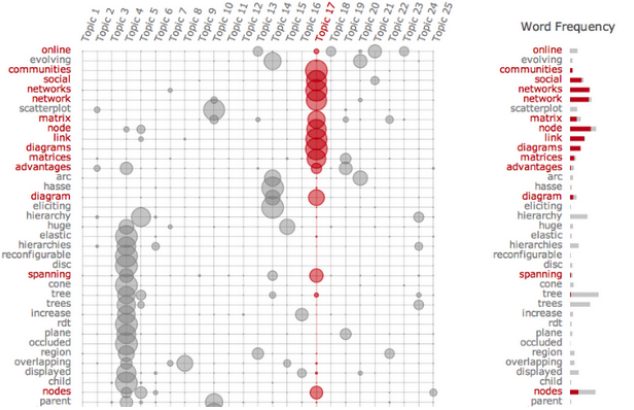


Crowdsourcing (Benoit et al.)



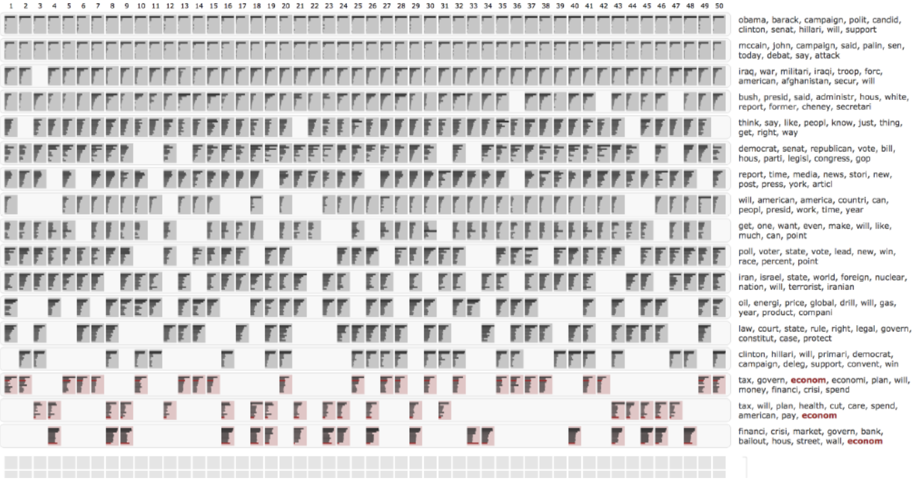
Unsupervised ML (e.g. topic modeling)

- Cluster documents based on shared features
- Easy to implement, sounds cool (Dirichlet)
- Not intended for hypothesis testing
- Validation is a concern



The Termite system. A tabular view (left) displays term-topic distributions for an LDA topic model. A bar chart (right) shows the marginal probability of each term.

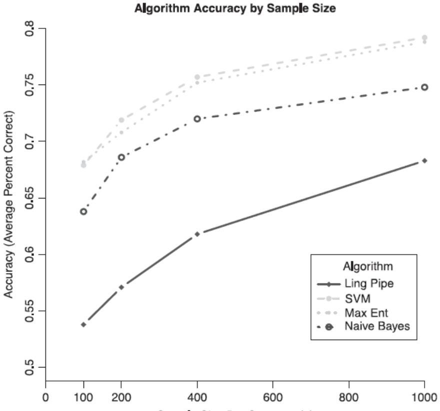
Ensemble validation (Chuang et al.)



Supervised Machine Learning

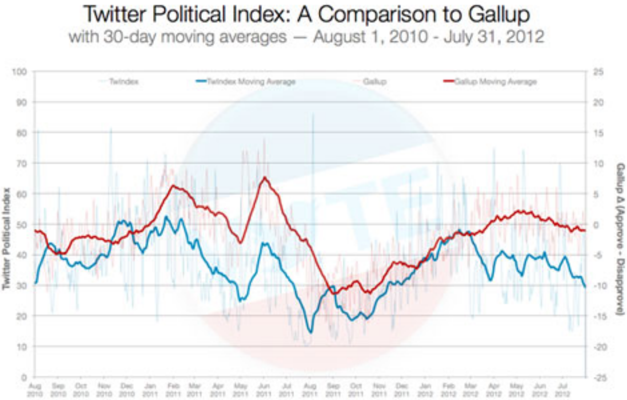
- Train predictive algorithm using human-labeled examples
  High startup costs (usually)
- Validation is central to the method
- RTextTools package is available!

FIGURE 1. Algorithm performance and training sample size.



Beyond the 'bag of words:' Natural Language Processing

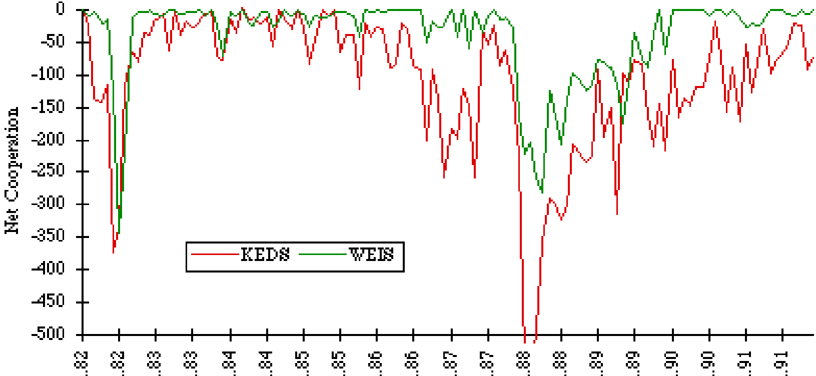
- Incorporate meaning (e.g. sentiment)
- Scale documents based on keyword frequencies
- Train algorithm using labeled examples (movie reviews)



Natural Language Processing (NLP)

- Incorporate sentence structure (P.O.S.)
- Disinguish entities, verbs, adjectives etc.
- Use thesaurus to group similar words

Israel->Palestinian Net Cooperation, 1982-1991

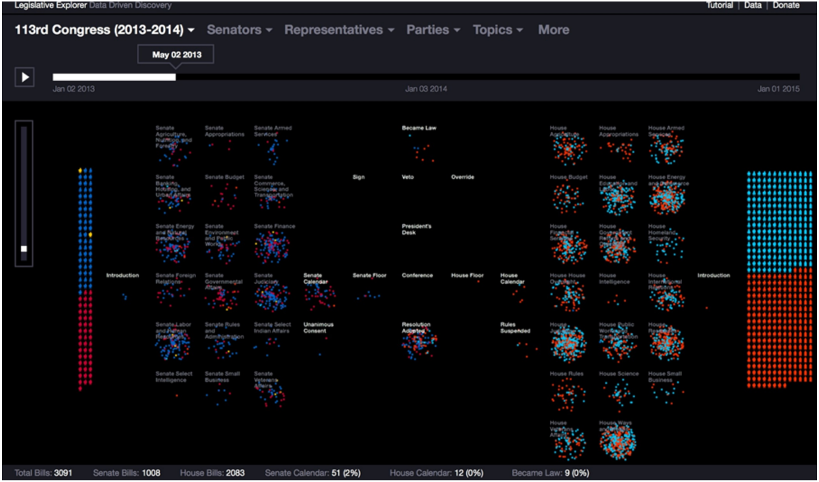


Part II: Programming introduction

- `http://faculty.washington.edu/jwilker/CAP/CAP2015Workshop.doc`
- Getting text
- Pre-processing
- Creating and exporting data matrices
- Analyzing text as data

Structured data: Legislative Explorer

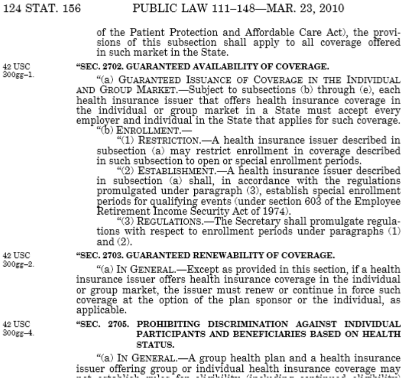
APIs to ’scrape’ Library of Congress website



Total Bills: 3091 Senate Bills: 1008 House Bills: 2083 Senate Calendar: 51 (2%) House Calendar: 12 (0%) Became Law: 9 (0%)

Semi-structured: Bill ”sections” as ideas

- A "single proposition of enactment"
- A conferral of "authority"
- Sections sharing content = same idea



124 STAT. 156 PUBLIC LAW 111–148—MAR. 23, 2010

of the Patient Protection and Affordable Care Act), the provisions of this subsection shall apply to all coverage offered in such market in the State.

42 USC 300gg–1. **“SEC. 2702. GUARANTEED AVAILABILITY OF COVERAGE.**

“(a) GUARANTEED ISSUANCE OF COVERAGE IN THE INDIVIDUAL AND GROUP MARKET.—Subject to subsections (b) through (e), each health insurance issuer that offers health insurance coverage in the individual or group market in a State must accept every employer and individual in the State that applies for such coverage.

“(b) ENROLLMENT.—

“(1) RESTRICTION.—A health insurance issuer described in subsection (a) may restrict enrollment in coverage described in such subsection to open or special enrollment periods.

“(2) ESTABLISHMENT.—A health insurance issuer described in subsection (a) shall, in accordance with the regulations promulgated under paragraph (3), establish special enrollment periods for qualifying events (under section 603 of the Employee Retirement Income Security Act of 1974).

“(3) REGULATIONS.—The Secretary shall promulgate regulations with respect to enrollment periods under paragraphs (1) and (2).

42 USC 300gg–2. **“SEC. 2703. GUARANTEED RENEWABILITY OF COVERAGE.**

“(a) IN GENERAL.—Except as provided in this section, if a health insurance issuer offers health insurance coverage in the individual or group market, the issuer must renew or continue in force such coverage at the option of the plan sponsor or the individual, as applicable.

42 USC 300gg–4. **“SEC. 2705. PROHIBITING DISCRIMINATION AGAINST INDIVIDUAL PARTICIPANTS AND BENEFICIARIES BASED ON HEALTH STATUS.**

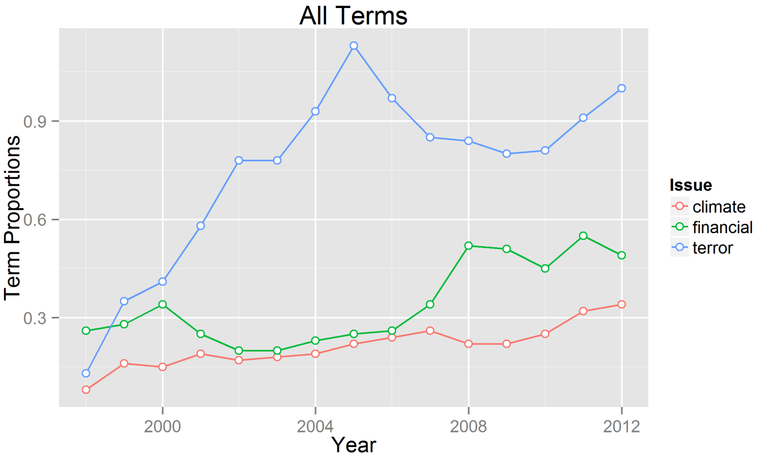
“(a) IN GENERAL.—A group health plan and a health insurance issuer offering group or individual health insurance coverage may not establish rules for eligibility (including continued eligibility)

Shared idea example from ACA

111 S. 1244	111 H.R. 3590
ing mothers a in general section 7 of the fair labor standards act——— 29 usc 207 is amended by adding at the end the following r 1 an employer shall provide—- reasonable break time for an employee to express breast milk for her nursing child for 1 year after the childs birth each time such employee has need to express the milk the employer shall make reasonable efforts to provide a place other than a bathroom that is shielded from view and free from intrusion from coworkers and the public which may be used by an employee to express breast milk– an employer shall not be required to compensate an employee————————— ——— for any work time spent for such purpose 2 for purposes of this subsection the term employer means an employ	ing mothers——— - section 7 of the fair labor standards act of 1938 29 usc 207 is amended by adding at the end the following r 1 an employer shall provide a a reasonable break time for an employee to express breast milk for her nursing child for 1 year after the childs birth each time such employee has need to express the milk and — ————————————b a place other than a bathroom that is shielded from view and free from intrusion from coworkers and the public which may be used by an employee to express breast milk 2 an employer shall not be required to compensate an employee receiving reasonable break time under paragraph 1 for any work time spent for such purpose 3 ——————— –an employer —-that employ

Messy Data: .GOV (Internet Archive)

1 billion webpages of differing formats



General Observations

- Think big! That’s where the leverage is
- Choose projects where 90% (e.g.) is good enough?
- Partner?
- Validate! Validate!
- Avoid Topic Models (personal opinion!)