Can computational systems help investigative journalists to identify patterns of intellectual influence that shape the creation of public policy?

Hyunwoo Park
Gopal Vaswani
Jeremy Whitaker
Michael Manoochehri

Advisor: Eric Kansa

Partners: Rob Ennals, Intel Research
Laurent El Ghaoui, EECS
Saheli Datta
The Importance of Journalism
"Statements by more than a dozen lawmakers were ghostwritten, in whole or in part, by Washington lobbyists...The lobbyists, employed by Genentech and by two Washington law firms, were remarkably successful in getting the statements printed in the Congressional Record under the names of different members of Congress. Genentech, a subsidiary of the Swiss drug giant Roche, estimates that 42 House members picked up some of its talking points — 22 Republicans and 20 Democrats, an unusual bipartisan coup for lobbyists."

Robert Pear
N.Y.T, Nov. 14th, 2009
This is a complex intersection where private interests clash over ideas and seek influence public policy.
Corporate Speech

- PR Documents
- Lobbyists

Expert Knowledge

- Think Tanks
- Academia

Government Records

- Congressional
- Presidential
- Hearings
<table>
<thead>
<tr>
<th>Corporate Speech</th>
<th>Constructed speech designed to influence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert Knowledge</td>
<td>A bridge between academia and policy</td>
</tr>
<tr>
<td>Government Records</td>
<td>Policy and the opinion of policy makers</td>
</tr>
<tr>
<td>Corporate Speech</td>
<td>Campaign Finance Law, penalties</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------------------------------</td>
</tr>
<tr>
<td>Expert Knowledge</td>
<td>Should behave like a 501c3</td>
</tr>
<tr>
<td>Government Records</td>
<td>Accountable to “informed constituency”</td>
</tr>
</tbody>
</table>
Migratory Words
Computational Tools for Investigative Political Research
≠
IDENTITY SIGNALS

Health Care Takeover
Corporate Welfare
Military Industrial Complex
Death Tax
We are, as language animals, to a substantial degree simply permuters of substrings already well established within the history of the whole language (Wilks, n.d.).
“INTERESTINGNESS”

memorable    sticky    emotional

Drill Baby, Drill
Take for example:

“One of the reasons I have long supported the U.S. biotechnology industry is that it is a homegrown success story that has been an engine of job creation in this country. Unfortunately, many of the largest companies that would seek to enter the biosimilar market have made their money by outsourcing their research to foreign countries like India.”

Robert Pear, 2009
How did Robert Pear and the NYT make the link?

Given Insider Text

Intuition-Based Identification

Line by Line Search
Journalism’s Current Methods

using networks of people
playing details against each other
identifying difficult patterns
aggregating different sources
generating trust with sources
applying human intuition as the investigator
understanding what is interesting to readers
leading the reader through a compelling story-line
(lots of tacit knowledge)
Migratory Words
Computational Tools for Investigative Political Research

Illustration: Gregg Kulick  (C): Wired Magazine, 2010
The “Migratory Words” Tool
They said that the health care bill passed by the House of Representatives “keeps in place the longstanding and widely supported federal policy against government funding of elective abortions and plans that include elective abortions” and “ensures that where federal funds are involved, people are not required to pay for other people’s abortions.” The Senate bill does not maintain this commitment.

In the Senate version, “federal funds will help subsidize, and in some cases a federal agency will facilitate and promote, health plans that cover elective abortions,” the bishops said. “All purchasers of such plans will be required to pay for other people’s abortions in a very direct and explicit way, through a separate premium payment designed solely to pay for abortion. There is no provision for individuals to opt out of this abortion payment in federally subsidized plans, so people will be required by law to pay for other people’s abortions.” The public consensus against abortion funding, said the bishops, “is borne out by many opinion surveys, including the new Quinnipiac University survey December 22 showing 72 percent opposed to public funding of abortion in health care reform legislation.”

“This bill also continues to fall short of the House-passed bill in preventing governmental discrimination against health care providers that decline involvement in abortion,” the bishops said. And it also “includes no conscience protection allowing Catholic and other institutions to provide and purchase health coverage consistent with their moral and religious convictions on other procedures.

U.S. Conference of Catholic Bishops
Dec. 22 2009
They said that the health care bill passed by the House of Representatives “keeps in place the longstanding and widely supported federal policy against government funding of elective abortions and plans that include elective abortions” and “ensures that where federal funds are involved, people are not required to pay for other people’s abortions.” The Senate bill does not maintain this commitment.

In the Senate version, “federal funds will help subsidize, and in some cases a federal agency will facilitate and promote, health plans that cover elective abortions,” the bishops said. “All purchasers of such plans will be required to pay for other people’s abortions in a very direct and explicit way, through a separate premium payment designed solely to pay for abortion. There is no provision for individuals to opt out of this abortion payment in federally subsidized plans, so people will be required by law to pay for other people’s abortions.” The public consensus against abortion funding, said the bishops, “is borne out by many opinion surveys, including the new Quinnipiac University survey December 22 showing 72 percent opposed to public funding of abortion in health care reform legislation.”

“This bill also continues to fall short of the House-passed bill in preventing governmental discrimination against health care providers that decline involvement in abortion,” the bishops said. And it also “includes no conscience protection allowing Catholic and other institutions to provide and purchase health coverage consistent with their moral and religious convictions on other procedures.

U.S. Conference of Catholic Bishops
Dec. 22 2009
Methods: Corpus Construction
CORPUS INDEXING

Indexed over 1.2 million documents

Over 1.7 billion terms

19 repositories of publicly available text
Computational Tools for Investigative Political Research

PRNewsWire (HTML)

- All News Releases - Sept 09 - Mar 10 (6mo)
  - All Health News
  - All Financial News
  - All Current Events

LouisDB

- Bills and Resolutions
- Congressional Reports
- Congressional Record
- Congressional Hearings
- Federal Register
- Presidential Documents

Corporate Text via "Hill Archive" HTML

- All articles provided by "Hill" White-Paper database
  - Health Care
  - Health Care Reform
  - Public Health
  - Finance

PRNewsWire via data dump

- All News Releases - 2006-2007 (2yrs)

PR Feeds

- PRNewswire United Business Media

Government Text

- LOUISdb Library of Unified Information Services

Position Papers

- THINK TANKS

Think Tank (HTML and PDF)

- RAND
- CATO
- Fraser
- Center for Global Development
- Brookings
- NCPA
- Civitas
- Center for Foreign Relations
- American Enterprise Institute
- NBER
- Urban Institute
DIFERENT PARTS OF HTML
**Computational Tools for Investigative Political Research**

**Migratory Words**

**Even worse. PDFs.**

---

**Paul Krugman's Nostalgianomics**

_Economic Policies, Social Values, and Income Inequality_

**Brink Lindsey**

---

**Even worse. PDFs.**

---

...
## Different Metadata Schemes

<table>
<thead>
<tr>
<th>corpus</th>
<th>doc_id</th>
<th>trackback</th>
<th>published_time</th>
<th>publisher</th>
<th>author</th>
<th>title</th>
<th>keywords</th>
<th>categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>bills</td>
<td>[filename]</td>
<td>&lt;id&gt;</td>
<td>&lt;published&gt;</td>
<td></td>
<td></td>
<td>&lt;title&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cr</td>
<td>[filename]</td>
<td>&lt;id&gt;</td>
<td>&lt;published&gt;</td>
<td></td>
<td></td>
<td>&lt;title&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>fr</td>
<td>[filename]</td>
<td>&lt;id&gt;</td>
<td>&lt;published&gt;</td>
<td></td>
<td></td>
<td>&lt;title&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>hearings</td>
<td>[filename]</td>
<td>&lt;id&gt;</td>
<td>&lt;published&gt;</td>
<td></td>
<td></td>
<td>&lt;title&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>presidential</td>
<td>[filename]</td>
<td>&lt;id&gt;</td>
<td>&lt;published&gt;</td>
<td></td>
<td></td>
<td>&lt;title&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>reports</td>
<td>[filename]</td>
<td>&lt;id&gt;</td>
<td>&lt;published&gt;</td>
<td></td>
<td></td>
<td>&lt;title&gt;</td>
<td></td>
<td></td>
</tr>
<tr>
<td>aei</td>
<td>[filename]</td>
<td>URL:</td>
<td>DC.date:</td>
<td>publisher:</td>
<td>author:</td>
<td>DC.title:</td>
<td>keywords:</td>
<td>taxonomy:</td>
</tr>
<tr>
<td>cato</td>
<td>[filename]</td>
<td>URL:</td>
<td>date:</td>
<td>author:</td>
<td></td>
<td>title:</td>
<td>keywords:</td>
<td>sector:</td>
</tr>
<tr>
<td>cgdev</td>
<td>[filename]</td>
<td>URL:</td>
<td>date:</td>
<td>copyright:</td>
<td>authors:</td>
<td>title:</td>
<td>keywords:</td>
<td>sector:</td>
</tr>
<tr>
<td>fraser</td>
<td>[filename]</td>
<td>URL:</td>
<td>Date Published:</td>
<td></td>
<td>Author(s):</td>
<td></td>
<td></td>
<td>sector:</td>
</tr>
<tr>
<td>thehill</td>
<td>[filename]</td>
<td>URL:</td>
<td></td>
<td>organization:</td>
<td></td>
<td>title:</td>
<td>keywords:</td>
<td>sector:</td>
</tr>
</tbody>
</table>
IT WAS JUST CUTE.
Methods: Filter on <interesting>?
President Obama opens new area for offshore oil, gas exploration

By Juliet Eilperin and Anne E. Kornblut
Washington Post Staff Writer
Thursday, April 1, 2010

Nationwide, a solid majority of Americans support broader energy exploration: In the latest national survey on the subject, 63 percent of Americans told Pew pollsters in February that they favor allowing offshore oil and gas drilling in U.S. waters.

Obama emphasized Wednesday that the policy fits within his overall framework for clean-energy development. The nation will "need to harness traditional sources of fuel even as we ramp up production of new sources of renewable, homegrown energy," he said at Joint Base Andrews Naval Air Facility Washington, formerly Andrews Air Force Base. "So today we're announcing the expansion of offshore oil and gas exploration -- but in ways that balance the need to harness domestic energy resources and the need to protect America's natural resources."

In what could represent the biggest expansion of offshore energy exploration in half a century, Obama announced that he will open the deep-water drilling off Florida, Georgia, Alabama, Mississippi, Louisiana, Texas, Arkansas, New Mexico, and California.
A majority of Americans support broader energy policies. A national survey on the subject, 63 percent of respondents in February that they favor allowing more drilling in U.S. waters.

Ted Ra sent that the policy fits within his overall energy development. The nation will "need to keep fuel even as we ramp up production of renewable, homegrown energy," he said at Joint Base by Washington, formerly Andrews Air Force Base announcing the expansion of offshore oil and gas that balance the need to harness domestic energy and need to protect America's natural resources."
ramp up
up production
production of
of new
new sources
sources of
of renewable
renewable, , homegrown
homegrown energy
energy, , he
he said
said at

FILTER: REMOVE NOISE

• digits
• n-grams with many common words
• n-grams with really long terms (like URLs)
• etc...
Computational Tools for Investigative Political Research

Migratory Words

Score for "Unlikeness"

ramp up
up production
new sources
homegrown energy

up appears 1,047,124 times
homegrown appears 1,911 times
production appears 502,350 times
energy appears 1,106,644 times
Count

“ramp up”
2,757 documents

“homegrown energy”
67 documents
“Homegrown Energy”
IT’S A LITTLE MORE COMPLICATED

“Does the phrase have more that 50% stopwords?”

“Does one term appear < 30 times in corpus?”

“Does a term have more two hyphens in it?”

“Does the n-gram end with a stopword?”
LIMITATIONS

SEMANTIC:
“biological nuclear national security missile regional stability firearms”

STATISTICAL:
“appropriations subcommittee in january”
Can computational systems help investigative journalists to identify patterns of intellectual influence that shape the creation of public policy?

Type or paste your text into the search box and let Migratory Words find the most interesting language as it has evolved and changed over time.

Read more ...

<table>
<thead>
<tr>
<th>Favorite phrases by users</th>
<th>Council on Foreign Relations</th>
<th>Cato Institute</th>
<th>Fraser Institute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slash-and-burn and hit-and-run negative attack</td>
<td>Renewable-energy technologies</td>
<td>Minority-serving institutions</td>
<td>Full-time-equivalent employees</td>
</tr>
<tr>
<td>Run roughshod over the constitution</td>
<td>National-security information</td>
<td>Notice-and-comment procedures</td>
<td>Administration environmental</td>
</tr>
<tr>
<td>Combating global warming</td>
<td>Nationally-representative sample</td>
<td>Information-technology products</td>
<td>Non-governmental organization</td>
</tr>
<tr>
<td>Moral and religious convictions</td>
<td>Next-generation technologies</td>
<td>Energy-conservation measures</td>
<td>Regulatory responsibilities</td>
</tr>
<tr>
<td>Intelligence-related activities</td>
<td>Information-technology industry</td>
<td>Small-government jurisdictions</td>
<td>Telecommunications networks</td>
</tr>
<tr>
<td>Research-and-development spending</td>
<td></td>
<td>Environmental considerations</td>
<td></td>
</tr>
</tbody>
</table>
Future Potential

Studying the “Congressional Remix”

Social Network Integration, Twitter

Automated Content Gathering

Firefox Plug-in for Contextual Search

Semantic and Sentiment Analysis

Linguistic Analysis of Legal Text

Academic Ideological Mapping
Ongoing Policy Challenges: Fair Use

Not for Profit

For Research Purposes

Public Service, Important for Discussion or Debate

Transformative Use

Caching

Link Back to Source
Ongoing Policy Challenges: Transparency

“Targeted Transparency” - Kung et al is not
“Naked Transparency” - Lessig
MigratoryWords.com

Questions?