Political Text Analysis

Lecture 6

Kohei Watanabe
Housekeeping

About final exam
Final exam

• Write an essays on political science subjects
  • You have to use quantitative text analysis methods covered in this course
  • You can write an essay in group
    • Essay should have an overarching theme
    • Each member to perform different analyses
• Essay should be 2,000-3,000 words per person
  • 4,000-6,000 words for a group of two
• Due on 26th June 2019
  • Send to koheiw@aoni.waseda.jp by email
  • You will have 4 weeks from the last day of the course
How to write a good essay

• Try to be get closer to the ideal in writing your essay
  • Ideally, your research should be driven by theory
    1. Identify important but unanswered questions in literature
    2. Collect original data
    3. Perform analysis based on the theoretical framework
  • In reality, your research would be driven by convenience
    1. Identify frequently-discussed topics in lectures
    2. Select research questions that you can answer with existing data
    3. Perform analysis using tools that require minimal manual input
Presentation

• You should present the work-in-progress in week 8
  • About 10-15 min including questions and answers
• Your presentation should include
  1. Background/theory
  2. Research question
  3. Description of data
  4. Analysis methods
  5. (expected) results
Unsupervised models
Wordfish

- Wordfish locates documents and words on a single dimension
- When
  - $y_{ij}$ is frequency of word $i$ in document $j$
  - $\beta_i$ is ideological polarity of word $i$
  - $\theta_j$ ideology of the author of document $j$
  - $\Psi_j$ and $\alpha_i$ heterogeneity of word $i$ and the author of document $j$ (fixed effect)
  - $y_{ij} = \text{Poisson}(\lambda_{ij})$
  - $\lambda_{ij} = \exp(\beta_i \theta_j + \Psi_j + \alpha_i)$
- Expectation maximization (EM) algorithm is used to estimate the four latent variables on the right-hand side
 Expectation Maximization (EM) 

• EM algorithm finds most likely values for unknown parameters iteratively
• When parameters $\alpha, \beta_i$ are unknown but constrained by function $F$

$$
\alpha = F(\beta_1 x_1 + \cdots + \beta_n x_n)
$$

1. Set random initial values to $\alpha, \beta_i$
2. Repeat expectation and maximization
   • Expectation: predict $\alpha$ using the latest $\beta_i$
   • Maximization: estimate most likely value for all $\beta_i$ given $\alpha$
3. Stop when likelihood stops improving (convergence)
Correspondence analysis

• Locate documents or features on multi-dimensional space
• Perform SVD of a residual matrix $R$
  • When
    • $X$ is the DFM
    • $n$ is the total number of words in $X$
    • $P(d_i)$ and $P(f_j)$ are probability of documents and features
      \[ R = \left( \frac{X}{n} - E \right) \]
      \[ E = P(d_i) \times P(f_j) \]
  • $E$ is expected normalized frequency when documents and features are independent
    • $R$ is a residual matrix that records deviation from independence
Singular Value Decomposition (SVD)

\[ X \approx \hat{X} = DST' \]

- \( X \) is the document-term matrix, \( m \times n \)
- \( \hat{X} \) is the approximation, \( m \times k \)
- \( D \) is the diagonal matrix of singular values, \( k \times k \)
- \( S \) is the left singular matrix, \( k \times k \)
- \( T' \) is the right singular matrix, \( k \times n \)
Row
5
10
15
20

Column
Dimensions: 20 x 100
Dimensions: 20 x 100
Topic models

• Classify documents into topics automatically
  • We only need to specify the number of topics $k$
    • Documents can have single or multiple topics
  • Topic models exploit cooccurrences of words to detect topics
• There are many topic models
  • Latent Semantic Analysis (LSA/LSI)
  • Probabilistic Latent Semantic Analysis (pLSA)
  • Latent Dirichlet Allocation (LDA)
  • Structural Topic Models (STM)
Topic words

- Five topics from a 50-topic model found in the Yale Law Journal from 1980-2003 by LDA

<table>
<thead>
<tr>
<th>Topic 1</th>
<th>Topic 2</th>
<th>Topic 3</th>
<th>Topic 4</th>
<th>Topic 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>contractual</td>
<td>employment</td>
<td>female</td>
<td>markets</td>
<td>criminal</td>
</tr>
<tr>
<td>expectation</td>
<td>industrial</td>
<td>men</td>
<td>earnings</td>
<td>discretion</td>
</tr>
<tr>
<td>gain</td>
<td>local</td>
<td>women</td>
<td>investors</td>
<td>justice</td>
</tr>
<tr>
<td>promises</td>
<td>jobs</td>
<td>see</td>
<td>sec</td>
<td>civil</td>
</tr>
<tr>
<td>expectations</td>
<td>employees</td>
<td>sexual</td>
<td>research</td>
<td>process</td>
</tr>
<tr>
<td>breach</td>
<td>relations</td>
<td>note</td>
<td>structure</td>
<td>federal</td>
</tr>
<tr>
<td>enforcing</td>
<td>unfair</td>
<td>employer</td>
<td>managers</td>
<td>see</td>
</tr>
<tr>
<td>supra</td>
<td>agreement</td>
<td>discrimination</td>
<td>firm</td>
<td>officer</td>
</tr>
<tr>
<td>note</td>
<td>economic</td>
<td>harassment</td>
<td>risk</td>
<td>parole</td>
</tr>
<tr>
<td>perform</td>
<td>case</td>
<td>gender</td>
<td>large</td>
<td>inmates</td>
</tr>
</tbody>
</table>
Latent Dirichlet Allocation (LDA)

- LDA models data generating process
  1. There are words related to topics: $\beta_k \sim \text{Dir}(\eta)$
  2. Author decided topics of a document: $\theta_d \sim \text{Dir}(\alpha)$
  3. Author choose topic-related words to write the document

- Assumes Dirichlet distribution of topics over words and documents
  - Multinomial continuous distribution
## Difference between topic models

<table>
<thead>
<tr>
<th>Model</th>
<th>Generative</th>
<th>Membership</th>
<th>Algorithm</th>
<th>Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSA</td>
<td>No</td>
<td>Single</td>
<td>SVD</td>
<td>Identify synonyms (word embedding)</td>
</tr>
<tr>
<td>pLSA</td>
<td>Yes</td>
<td>Single</td>
<td>EM</td>
<td>Document classification</td>
</tr>
<tr>
<td>LDA</td>
<td>No</td>
<td>Multiple</td>
<td>EM/Gibbs sampling</td>
<td>Document classification</td>
</tr>
<tr>
<td>STM</td>
<td>Yes</td>
<td>Multiple</td>
<td>Generalize linear model (EM)</td>
<td>Study relationship between topics and covariates (author, time, etc.)</td>
</tr>
</tbody>
</table>
Slapin & Proksch 2008
A Scaling Model for Estimating Time-Series Party Positions from Texts
Problems

• Wordscore depends on the accuracy of reference text
  • Contradictory rating of reference texts by analysts create to concentration to scores on the center

• Wordscore is not suitable for time-series analysis
  • Political texts are changeable over time (between elections)
Solution

• Don’t use reference text at all
  • Developed an unsupervised document scaling model
entry into force protects safe they/she the

fascism professional ban male violence emancipation pornography

income taxation non-wage labor costs education vouchers
FIGURE 1  Estimated Party Positions in Germany, 1990–2005

(A) Left–Right

(B) Economic Policy

(C) Societal Policy

(D) Foreign Policy

Political Text Analysis 21
## Validation

### Table 2: Cross-Validation: Correlations between German Party Position Estimates

<table>
<thead>
<tr>
<th></th>
<th>Poisson Scaling Model</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Left-Right</td>
</tr>
<tr>
<td><strong>Hand-coding manifestos</strong></td>
<td></td>
</tr>
<tr>
<td>CMP: Left-Right (n = 15, 1990–1998)</td>
<td></td>
</tr>
<tr>
<td>CMP: Markeco (n = 15, 1990–1998)</td>
<td></td>
</tr>
<tr>
<td>CMP: Welfare (n = 15, 1990–1998)</td>
<td></td>
</tr>
<tr>
<td>CMP: Intpeace (n = 15, 1990–1998)</td>
<td></td>
</tr>
<tr>
<td><strong>Expert Survey</strong></td>
<td></td>
</tr>
<tr>
<td>Benoit/Laver 2006: Left-Right (n = 5, 2002)</td>
<td></td>
</tr>
<tr>
<td>Benoit/Laver 2006: Taxes-Spending (n = 5, 2002)</td>
<td></td>
</tr>
<tr>
<td><strong>Wordscores</strong></td>
<td></td>
</tr>
<tr>
<td>Laver et al. 2003: Economic (n = 10, 1990–1994)</td>
<td></td>
</tr>
<tr>
<td>Laver et al. 2003: Social (n = 10, 1990–1994)</td>
<td></td>
</tr>
<tr>
<td>Proksch/Slapin 2006: Economic (n = 5, 2005)</td>
<td></td>
</tr>
<tr>
<td>Proksch/Slapin 2006: Social (n = 5, 2005)</td>
<td></td>
</tr>
</tbody>
</table>
Issues

• The model extracts only the most prominent dimension
• Direction of scale is random
  • You have to flip the scores manually
• It needs multiple documents from the same year
  • Otherwise, extract temporal changes
• It does not work well with large and noisy corpora
Schonhardt-Bailey 2005

Measuring Ideas More Effectively: An Analysis of Bush and Kerry’s National Security Speeches
Research question

• RQ: What were the main themes on national security in the 2004 presidential election
  • Only 3 years after the September 11 attacks
  • US troops were fighting in Afghanistan and Iraq
  • American voters elected George W Bush over John Kerry
Data

- 20 speech transcripts made by George W Bush over John Kerry in 2002-2004
  - A speech by George W Bush is from 2002
- 73,715 words in total
  - On average 3,685 words per speech
Analysis

• The author uses a software program called Alceste
• Make composite documents from sub-units of speeches
  • Segment speeches into sentences (or quasi-sentences)
  • Group them into 7 groups based on similarity
  • Give labels to composite documents manually
• Perform hieratical clustering and correspondence analysis
Figure 1
Tree Graph of the Classes for Bush and Kerry on National and Homeland Security
Figure 2
Correspondence Analysis of Classes for Bush and Kerry on National and Homeland Security
Conclusions

• Kerry focused on US specific security issues, while Bush was more about global.
  • In US specific issues, Burh made more emotional appeals (Gratitude vs Fear) emphasizing terror threat facing the United Sates (War on Terror)
  • Kerry expressed gratitude to the veterans (Fellow Veterans)
• Bush’s emotional speeches made him more popular in the 2004 presidential election
Issues

• Difficult for us to understand analytic procedures of commercial package

• Hierarchical clustering and correspondence analysis produced very similar results
  • Both of them are clustering algorithms
Boussalis & Coan 2016
Text-mining the signals of climate change doubt
Research question

• RQ: How conservative think tanks deny climate change?
  • There are still many legislators in the US Congress who deny climate change
  • Conservative think tanks are obstructing laws related to climate change by misinformation
  • There has been no systematic analysis of conservative think tanks
Data

• Documents published by conservative think tanks in North American
  • Downloaded web pages and PDF files from 19 well-funded conservative think tanks
  • Corpus contains 16,000 documents published in 1998-2013
Table 1
Climate sceptic organizations. The table displays the total count of words (thousands), the number, and type of documents from 19 well-known conservative think-tanks over the period January 1998–August 2013. Documents have been classified as follows: (A) op-eds, articles and blogs; (B) policy/science reports and analyses; (C) speech/interview transcripts; (D) press releases/open letters; (E) scientific reviews.

<table>
<thead>
<tr>
<th>Organization name</th>
<th>Total words thous.</th>
<th>Total docs.</th>
<th>Document type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>American Enterprise Institute (AEI)</td>
<td>1872.53</td>
<td>745</td>
<td>596</td>
</tr>
<tr>
<td>Cato Institute</td>
<td>772.68</td>
<td>768</td>
<td>712</td>
</tr>
<tr>
<td>Center for the Study of Carbon Dioxide and Global Change (CO2Science)</td>
<td>2387.27</td>
<td>4592</td>
<td>713</td>
</tr>
<tr>
<td>Competitive Enterprise Institute (CEI)</td>
<td>1743.02</td>
<td>1461</td>
<td>941</td>
</tr>
<tr>
<td>Committee for a Constructive Tomorrow (CFACT)</td>
<td>738.52</td>
<td>894</td>
<td>882</td>
</tr>
<tr>
<td>Citizens for a Sound Economy (CSE)</td>
<td>88.2</td>
<td>111</td>
<td>105</td>
</tr>
<tr>
<td>Fraser Institute</td>
<td>78.39</td>
<td>81</td>
<td>62</td>
</tr>
<tr>
<td>Foundation for Research on Economics and the Environment (Free-Eco)</td>
<td>76.64</td>
<td>105</td>
<td>105</td>
</tr>
<tr>
<td>Heartland Institute</td>
<td>9900.54</td>
<td>2930</td>
<td>1383</td>
</tr>
<tr>
<td>Heritage Foundation</td>
<td>1825.78</td>
<td>1652</td>
<td>1198</td>
</tr>
<tr>
<td>Hoover Institution</td>
<td>51.06</td>
<td>37</td>
<td>3</td>
</tr>
<tr>
<td>Hudson Institute</td>
<td>124.61</td>
<td>83</td>
<td>81</td>
</tr>
<tr>
<td>Manhattan Institute</td>
<td>315.59</td>
<td>199</td>
<td>183</td>
</tr>
<tr>
<td>George C. Marshall Institute</td>
<td>209.75</td>
<td>101</td>
<td>69</td>
</tr>
<tr>
<td>National Center for Policy Analysis (NCPA)</td>
<td>469.78</td>
<td>451</td>
<td>376</td>
</tr>
<tr>
<td>National Center for Public Policy Research (NCPPR)</td>
<td>393.54</td>
<td>639</td>
<td>378</td>
</tr>
<tr>
<td>Pacific Research Institute</td>
<td>384.68</td>
<td>435</td>
<td>402</td>
</tr>
<tr>
<td>Reason Foundation</td>
<td>397.12</td>
<td>192</td>
<td>179</td>
</tr>
<tr>
<td>Science and Public Policy Institute (SPPI)</td>
<td>3064.88</td>
<td>552</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>24,894.58</td>
<td>16,028</td>
<td>8368</td>
</tr>
</tbody>
</table>
Analysis

- Classify documents using LDA
  - Set the number of topics to be 57 using a hyperparameter optimization technique
  - Only 47 topics were substantively interpretable
- Visualized topics on a two-dimensional space
  - Used multi-dimensional scaling (MDS) for visualization
  - Used Jensen-Shannon divergence to compute similarity between topics
Table 2
A full list of the estimated topics. The table provides each topic’s unique ID, descriptive label (in bold), and top 5 stemmed keywords based on the FREQ score (Roberts et al., 2014). Further, we code whether each topic is related to science (S) or politics and policy (P).

<table>
<thead>
<tr>
<th>Id</th>
<th>S/P</th>
<th>Topic Name</th>
<th>Id</th>
<th>S/P</th>
<th>Topic name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>S</td>
<td>Climate sensitivity to CO2 warm dege cool dioxid warmer</td>
<td>25</td>
<td>P</td>
<td>Economic impact of climate policy baselin discount sector eia mit</td>
</tr>
<tr>
<td>2</td>
<td>P</td>
<td>Fossil fuel production shale barrel oil drill pipelin</td>
<td>26</td>
<td>S</td>
<td>Monckton monckton graph ppmv brenchley humankind</td>
</tr>
<tr>
<td>3</td>
<td>S</td>
<td>Sea level rise antarct greenland glacier melt antarctica</td>
<td>27</td>
<td>S</td>
<td>IPCC integrity chapter ipcc tsd wg summari</td>
</tr>
<tr>
<td>4</td>
<td>S</td>
<td>No scientific consensus consensu denier oresk agw scientif</td>
<td>28</td>
<td>S</td>
<td>Storms cyclon storm hurricane tc frequenc</td>
</tr>
<tr>
<td>5</td>
<td>S</td>
<td>Long-term climate trends holocen millenni quaternari mediev palaeo</td>
<td>29</td>
<td>P</td>
<td>Emissions reduction carbon scheme credit trade dioxid</td>
</tr>
<tr>
<td>6</td>
<td>P</td>
<td>Public opinion Gallup abc pew cnn cb</td>
<td>30</td>
<td>S</td>
<td>Plant impacts seedl leaf mycorrhiz cultivar elev</td>
</tr>
<tr>
<td>7</td>
<td>P</td>
<td>US politics republican sen mccain democrat vote</td>
<td>31</td>
<td>P</td>
<td>Int’l trade &amp; develop india china chines wto asia</td>
</tr>
<tr>
<td>8</td>
<td>P</td>
<td>Renewable energy rp turbin renew wind megawatt</td>
<td>32</td>
<td>P</td>
<td>Tax &amp; spend tax dividend incom fiscal medicaid</td>
</tr>
<tr>
<td>9</td>
<td>P</td>
<td>Govt. intervention approach intervent principle geoingen outcom</td>
<td>33</td>
<td>P</td>
<td>Conservation timber eagl fisheri perc graze</td>
</tr>
<tr>
<td>10</td>
<td>P</td>
<td>Environmentalism lomborg holdren ehrlich evangl simon</td>
<td>34</td>
<td>S</td>
<td>Forest impacts npp ndvi shrub peatland finzi</td>
</tr>
<tr>
<td>11</td>
<td>S</td>
<td>Climate models simul gcm model cmip coupl</td>
<td>35</td>
<td>P</td>
<td>Cap &amp; trade markey waxman lieberman warner cap</td>
</tr>
<tr>
<td>12</td>
<td>S</td>
<td>Solar forcing &amp;cloud models cosmic cloud radiat ray aerosol</td>
<td>36</td>
<td>P</td>
<td>Public transportation rail ridership travel passeng vmt</td>
</tr>
<tr>
<td>13</td>
<td>S</td>
<td>Temperature data station giss ushen fig thermomet</td>
<td>37</td>
<td>P</td>
<td>Climate adaptation goklani adapt stern mitig resili</td>
</tr>
<tr>
<td>14</td>
<td>S</td>
<td>Scientific misconduct cru mcintyr mann hockey email</td>
<td>38</td>
<td>P</td>
<td>EPA cia endager naaq anpr</td>
</tr>
<tr>
<td>15</td>
<td>P</td>
<td>Govt. agencies fy sec gao omb provis</td>
<td>39</td>
<td>P</td>
<td>Law court judici lawsuit constitut suprem</td>
</tr>
<tr>
<td>16</td>
<td>S</td>
<td>Alarmism gore morano romm inconveni depot</td>
<td>40</td>
<td>S</td>
<td>State climate reports viru cessat nile wigley inch</td>
</tr>
<tr>
<td>17</td>
<td>P</td>
<td>Int’l relations militari nato missil afghanistan iran</td>
<td>41</td>
<td>P</td>
<td>State climate policy ghg jersey greenhous wefa rggi</td>
</tr>
<tr>
<td>18</td>
<td>P</td>
<td>Agri. Industry corn ethanol biofuel farmer sugar</td>
<td>42</td>
<td>S</td>
<td>Acidification calcif reef bleach coral phytoplankton</td>
</tr>
<tr>
<td>19</td>
<td>S</td>
<td>Human health ddt precautionari malaria diseas cancer</td>
<td>43</td>
<td>P</td>
<td>Disaster costs insur pension mortgag florida premium</td>
</tr>
<tr>
<td>20</td>
<td>P</td>
<td>Corporations &amp; env. borelli sharehold greenpeac donor philanthropi</td>
<td>44</td>
<td>P</td>
<td>Int’l climate agreements kyoto protocol treati ratifi ratif</td>
</tr>
<tr>
<td>21</td>
<td>P</td>
<td>Urban develop. california ab metropolitan schwarzenegg californian</td>
<td>45</td>
<td>S</td>
<td>Pollution mercuri ozon toxic asthma particul</td>
</tr>
<tr>
<td>22</td>
<td>P</td>
<td>Reuse &amp; recycle bag mtbe bulb cfl reus</td>
<td>46</td>
<td>S</td>
<td>Endangered species butterfli stirl extinct bear polar</td>
</tr>
<tr>
<td>23</td>
<td>P</td>
<td>Nuclear power hydrogen reactor nuclear technolog cell</td>
<td>47</td>
<td>P</td>
<td>Auto. fuel standards cafe nhtsa mpg vehicul car</td>
</tr>
<tr>
<td>24</td>
<td>P</td>
<td>Green jobs job stimulu taxpay subsidy green</td>
<td>48</td>
<td>P</td>
<td>Political Text Analysis</td>
</tr>
</tbody>
</table>
(a) Absolute topic prevalence

Sum of Topic Proportions

Date


Science
Policy
Conclusions

- Conservative think tanks continue to deny climate change
- They respond to external events strongly
  - Al Gore’s Inconvenient Truth (2006)
  - Nobel Peace Prize award of Gore and the IPCC (2007)
  - COP 15 Conference (2009)
- They change their strategy from attacking climate science to debating utility of climate policy in recent years
Issues

• LDA should detect more substantive topics
  • 57 topics is too many for analysis
    • Still 6 in 57 were junk topics
  • Authors need to manually group topics into 5 themes
    • LDA’s topic is not always meaningful to political scientists

• Difficult to find clear trend in the analysis results
  • Topic proportion has too much fluctuation
Mueller & Rauh 2018

Reading Between the Lines: Prediction of Political Violence Using Newspaper Text
Research question

• RQ: How to predict political violence?
  • Prediction of political violence is difficult because they are rare and concentrated to certain countries
Data

• 700,000 newspaper articles that mention country names
  • Published by English-speaking newspapers 1975-2015
    • New York Times
    • The Washington Post
    • The Economist
Analysis

• Use LDA to reduce the high dimensionality of textual data
  • Apply stemming and removed very high and low features
    • Features were reduced from 5 to 0.9 million
  • Fit LDA to further reduce the dimension to 15 topics
    • Topic proportions are aggregated by country-year

• Predict next years’ political violence by a regression model
  • Aggregated topic proportions are used as independent variable
FIGURE 3. Word Clouds of Topics

(a) Conflict 1

(b) Conflict 2

(c) Justice

(d) Economics

Notes: These are the top 50 words of 4 out of 15 topics computed using LDA with $\alpha = 3.33$ and $\beta = 0.01$ for the entire sample until 2013. The size of a term represents its probability within a given topic. The position conveys no information. A list of the 15 topics is exhibited in Appendix Table I.1.
FIGURE 4. ROC Curves for Onset (Topics Model)

a) Civil War

b) Armed Conflict

Notes: Predictions result from a panel estimated as in Equation (2). The topic model contains 15 topics as $\theta_k$, derived using LDA with $\alpha = 3.33$ and $\beta = 0.01$, and are aggregated at the country-year level. The within model is the overall model net of country fixed effects as presented in Equation (3).
Issues

• Authors could form composite documents for country-year
  • Directly estimating topic proportion by LDA seems more correct
• Countries mentioned in news articles are not always where violent events occur
  • Location of places can be better identified using geographical classifier
References


