Notes on Text Analysis

Michael Peress

SUNY-Stony Brook

September 15, 2020

Overview of Text Analysis

The 3 Most Common Goals of Text Analysis in Political Science:

- Topics
  - What topics are being discussed on cable news? (e.g. the Syrian conflict, the election, unemployment)
  - What topics are members of congress discussing? (Quinn et al., 2010)
  - What is the topic of a bill that is being voted on? (comparative agendas project)

Overview of Text Analysis

The 3 Most Common Goals of Text Analysis in Political Science:

- Topics
  - What topics are being discussed on cable news? (e.g. the Syrian conflict, the election, unemployment)
  - What topics are members of congress discussing? (Quinn et al., 2010)
  - What is the topic of a bill that is being voted on? (comparative agendas project)
- Sentiment
  - Soroka (2006): Measure sentiment in economic articles in the London Times
  - Kayser and Peress (2015): Measure sentiment in economic articles in 32 newspapers
  - Hopkins and King (2010): Measure sentiment in blog posts

Overview of Text Analysis

The 3 Most Common Goals of Text Analysis in Political Science:

- Ideology
  - Ideology of political parties from manifesto data (Laver, Benoit and Garry, 2003; Slapin and Proksch, 2010)
  - Ideology of newspapers (Groseclose and Jeffrey, 2005; Gentzkow and Shapiro, 2010)
  - Ideology of TV news shows from transcripts (Martin and Yurukoglu, 2015)
  - Ideology of legislators from speeches (Monroe, Colaresi and Quinn, 2008; Diermeier et al., 2011)

Overview of Text Analysis

Approaches to Text Analysis:

- Human Coding (need to code everything)
  - Soroka (2006): 6000 newspaper articles about the economy
  - Comparative Agendas Project (http://www.comparativeagendas.net/): many bills, newspaper articles, executive orders, etc.

Overview of Text Analysis

Approaches to Text Analysis:

- Human Coding (need to code everything)
  - Soroka (2006): 6000 newspaper articles about the economy
  - Comparative Agendas Project (http://www.comparativeagendas.net/): many bills, newspaper articles, executive orders, etc.
- Dictionary Coding (may need to code dictionary)
  - DeBoef and Kellstedt (2004): 5000 newspaper articles about the economy
  - Soroka, Stecula and Wlezien (2014): 30000 newspaper articles about the economy
  - Kayser and Peress (2015): 2 million newspaper articles about the economy

Overview of Text Analysis

Approaches to Text Analysis:

- Supervised Learning (need to code "training set")
  - Hopkins and King (2010) – measure the sentiment of blog posts

Overview of Text Analysis

Approaches to Text Analysis:

- "Fake" Supervised Learning (model is trained in a different context where training set is already "coded")
  - Party ID of members of congress plus congressional speech from the congressional record is used to train ideology model – applied to measure the ideology of newspapers (Gentzkow and Shapiro, 2010)
  - Ideology of members of congress plus congressional speech from the congressional record is used to train ideology model – used to measure ideology of cable news channels from transcripts (Diermeier et al., 2011; Martin and Yurukoglu, 2015)
  - Reference texts are used to code ideology in party manifestos (Laver, Benoit and Garry, 2003)
  - Texts of known Jihadi's are used to code whether additional texts express Jihadi ideology (Nielsen, 2012)

Overview of Text Analysis

Approaches to Text Analysis:

- Unsupervised Learning (no need to code training set, but interpretation of topics requires human intervention)
  - Clustering
    - LDA (Latent Dirichlet Allocation): Blei, Ng and Jordan (2003)
    - Quinn et al. (2010): measure topics in congressional speech
    - Grimmer (2010): measure topics in congressional press releases
    - Roberts et al. (2014): code topics of open ended survey responses

Overview of Text Analysis

Approaches to Text Analysis:

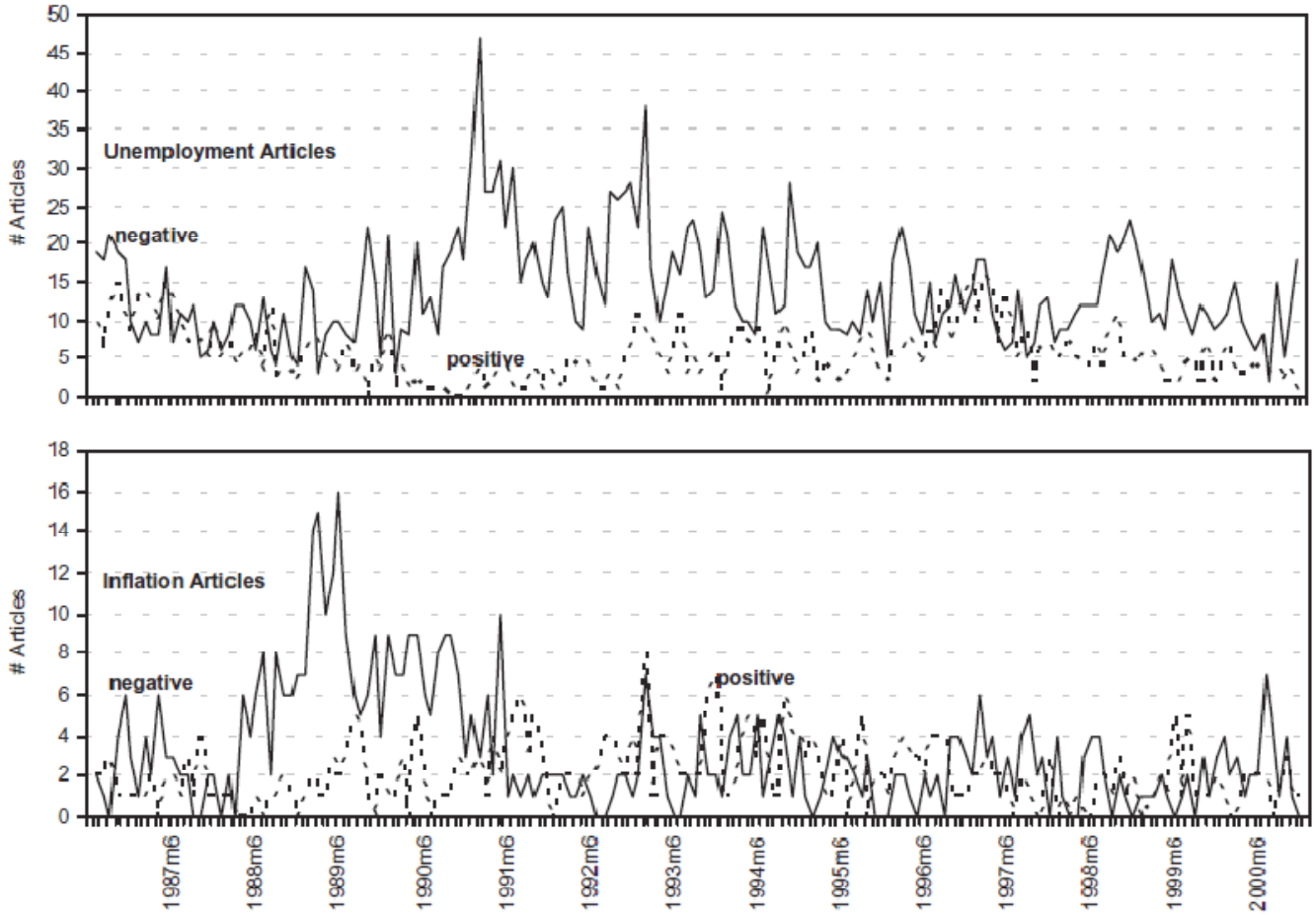
- Unsupervised Learning (no need to code training set, but interpretation of topics requires human intervention)
  - Clustering
    - LDA (Latent Dirichlet Allocation): Blei, Ng and Jordan (2003)
    - Quinn et al. (2010): measure topics in congressional speech
    - Grimmer (2010): measure topics in congressional press releases
    - Roberts et al. (2014): code topics of open ended survey responses
  - Scaling
    - Monroe and Maeda (2004): measure ideology in text
    - Slapin and Proksch (2010): measure positions of German parties from manifesto data over time

Overview of Text Analysis

Task	**Human Coding**	**Dictionary Coding**	**Supervised Learning**		**Unsupervised Learning**	
			"Fake"	Coded Training Data	Clustering	Scaling
Topics	Comparative Agendas Project	Kayser and Peress (2015)	Nielsen (2012)	Hopkins and King (2010)	Grimmer (2010); Quinn et al. (2010); Roberts et al. (2014)	n/a
Sentiment	Soroka (2006)	DeBoef and Kellstedt (2004); Kayser and Peress (2015)	n/a	Hopkins and King (2010)	n/a	n/a
Ideology	n/a	n/a	Laver, Benoit and Garry (2003); Groseclose and Jeffrey (2005); Monroe, Colaresi and Quinn (2008); Gentzkow and Shapiro (2010); Diermeier et al. (2011); Martin and Yurukoglu (2015)	n/a	n/a	Monroe and Maeda (2004); Slapin and Proksch (2010)

Overview of Text Analysis

Soroka (2006):



Overview of Text Analysis

Soroka (2006):

TABLE 1 Economic Indicators and Media Coverage: Unemployment and Inflation

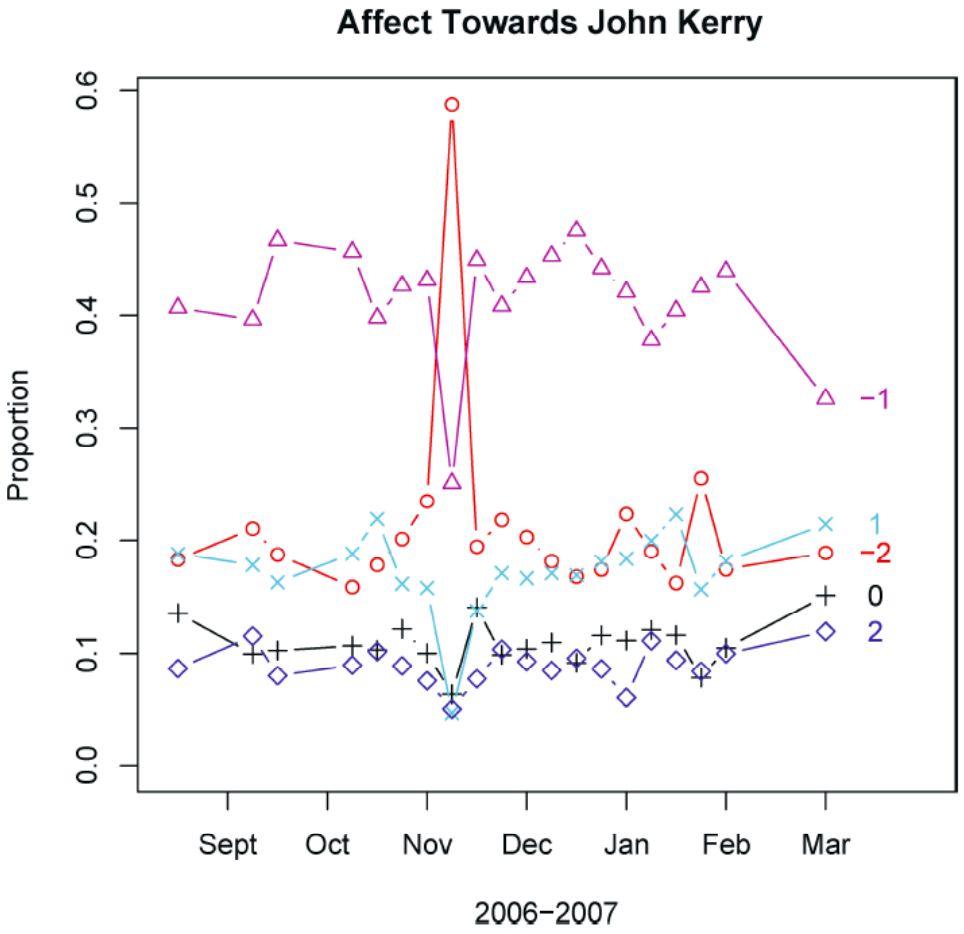
	Dependent Variable											
	Unemployment Articles						Inflation Articles					
	Negative		Positive		Net		Negative		Positive		Net	
Column	Model 1 (1)	Model 2 (2)	Model 1 (3)	Model 2 (4)	Model 1 (5)	Model 2 (6)	Model 1 (7)	Model 2 (8)	Model 1 (9)	Model 2 (10)	Model 1 (11)	Model 2 (12)
Chg in Economy$_t$[a]	16.256*	—	−2.809	—	−22.17*	—	1.367*	—	−1.027*	—	−1.145	—
	(4.234)		(2.215)		(5.317)		(.460)		(.324)		(.612)	
Chg Economy (Worse)$_t$[a]	—	28.396*	—	−2.268	—	−34.10*	—	3.869*	—	−.788	—	−1.450
		(7.875)		(3.629)		(9.139)		(.869)		(.611)		(1.198)
Chg Economy (Better)$_t$[a]	—	8.139	—	−3.270	—	−14.884*	—	−.476	—	−1.208*	—	−.920
		(6.124)		(3.303)		(6.977)		(.708)		(.509)		(.979)
Σ Dependent $_{t-1,4}$	.545*	.502*	.711*	.711*	.541*	.502*	.813*	.774*	.449*	.449*	.774*	.770*
	(.091)	(.093)	(.089)	(.089)	(.092)	(.095)	(.068)	(.066)	(.120)	(.121)	(.078)	(.080)
Constant	6.958*	6.657*	1.349*	1.300*	−4.866*	−4.350*	.576*	.139	1.120*	1.065*	−.276	−.211
	(1.436)	(1.435)	.487	(.554)	(1.058)	1.101	(.282)	(.303)	(.274)	(.299)	(.245)	(.328)
N	170	170	170	170	170	170	163	163	163	163	163	163
Rsq/Adj Rsq	.412/.394	.423/.402	.431/.413	.431/.410	.512/.498	.520/.502	.525/.507	.557/.537	.173/.142		.416/.394	.417/.390
Durbin's *h*	1.217	.212	.205	.396	3.257*	.847	1.237	1.075	.474		4.952*	6.012*

Note: Cells contain OLS regression coefficients with standard errors in parentheses, and standardized coefficients in italics.
[a]*Chg in Economy* is monthly changes in the unemployment rate for unemployment models, and monthly changes in the rate of inflation for inflation models.
*$p > .05$.

Overview of Text Analysis

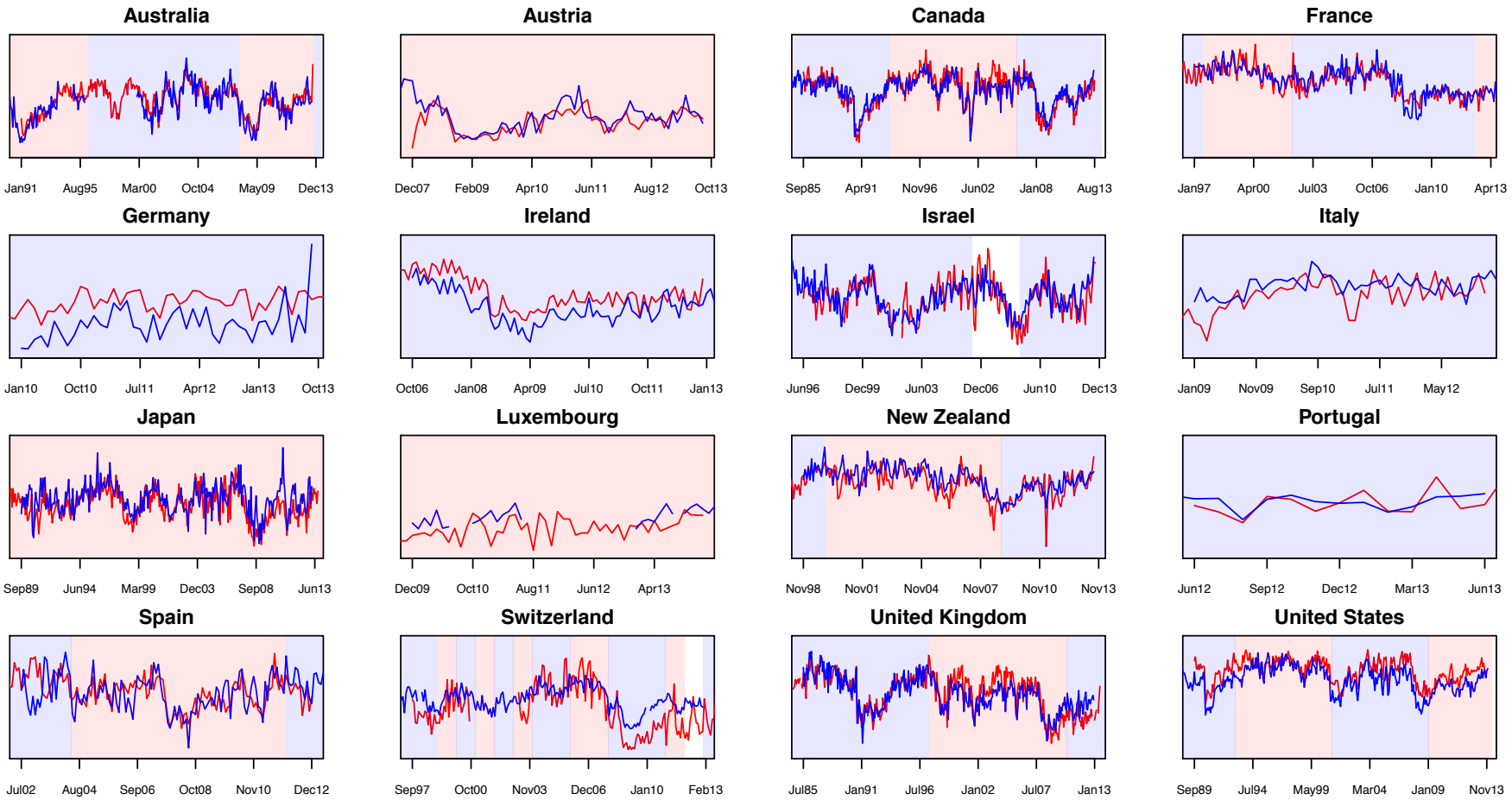
Hopkins and King (2010):

FIGURE 1 Blogosphere Responses to Kerry's Botched Joke



Overview of Text Analysis

Kayser and Peress (2015):



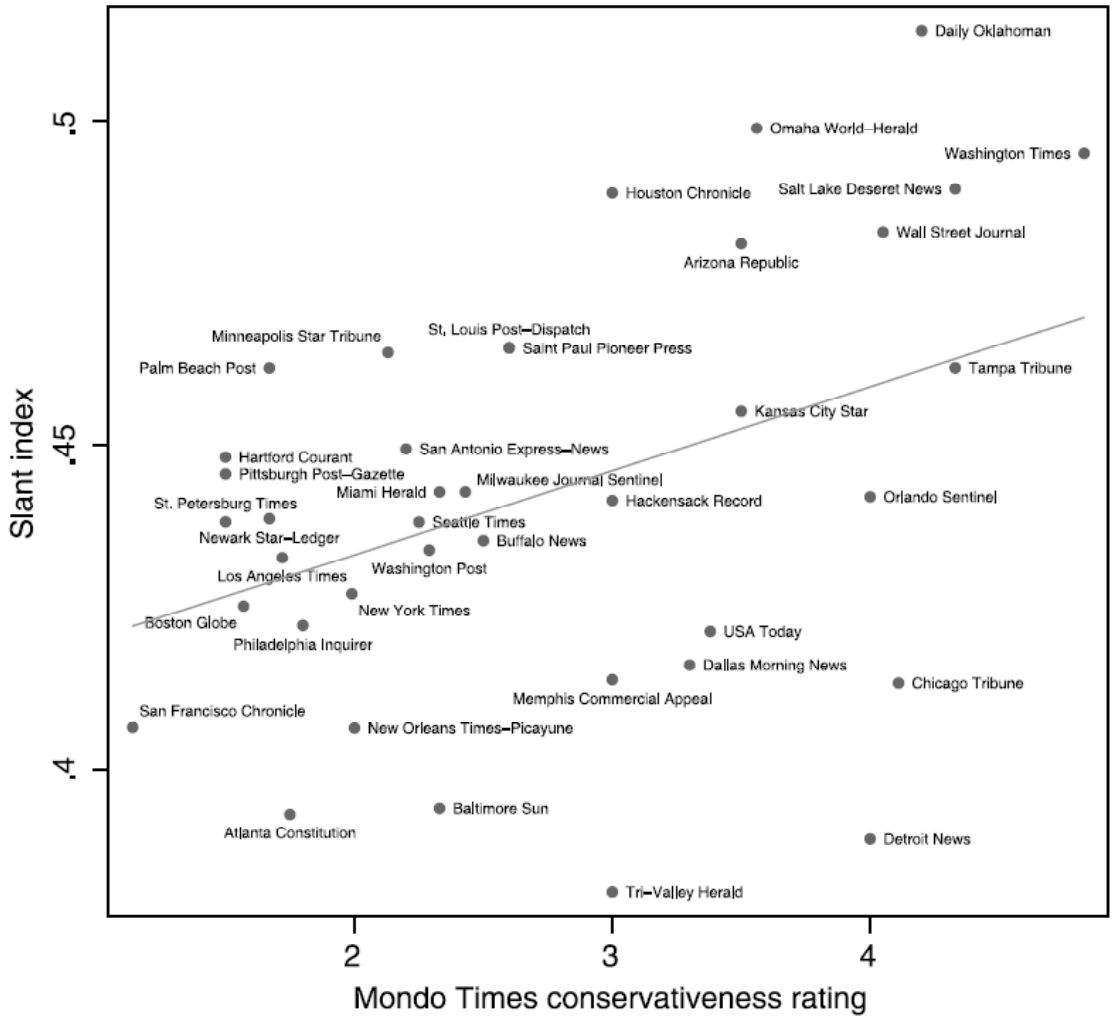
Overview of Text Analysis

Kayser and Peress (2015):

Dependent Variable:	Growth Sent.	Unem. Sent.	Inf. Sent.	Growth Sent.	Unem. Sent.	Inf. Sent.
Independent Variables:						
Ideological Match	-0.013	-0.008	0.011	-0.006	0.000	-0.001
	(0.019)	(0.010)	(0.010)	(0.019)	(0.008)	(0.008)
Growth (yearly)				0.023***	0.012***	-0.007***
				(0.003)	(0.001)	(0.002)
Change in Unem. (yearly)					-0.019***	
					(0.004)	
Change in Inf. (yearly)						-0.011***
						(0.002)
Growth * Ideo. Match				0.000	-0.001	0.004+
				(0.004)	(0.002)	(0.002)
Change in Unem. * Ideo. Match					0.009	
					(0.005)	
Change in Inf. * Ideo. Match						0.005+
						(0.002)
Number of Months	6664	6658	6608	6650	6644	6595
Number of Newspapers	32	32	32	32	32	32
Number of Countries	16	16	16	16	16	16
R-Squared	0.467	0.250	0.249	0.601	0.355	0.285

Overview of Text Analysis

Gentzkow and Shapiro (2010):



Overview of Text Analysis

Gentzkow and Shapiro (2010):

TABLE III

DETERMINANTS OF NEWSPAPER SLANT[a]

	OLS	2SLS	OLS	RE
Share Republican in newspaper's market	0.1460 (0.0148)	0.1605 (0.0612)	0.1603 (0.0191)	0.1717 (0.0157)
Ownership group fixed effects?			X	
State fixed effects?				X
Standard deviation (SD) of ownership effect				0.0062 (0.0037)
Likelihood ratio test that SD of owner effect is zero (p value)				0.1601
Number of observations	429	421	429	429
R^2	0.1859	—	0.4445	—

[a]The dependent variable is slant index ($\hat{y}_n$). Standard errors are given in parentheses. An excluded instrument in the 2SLS model is share attending church monthly or more in the newspaper's market during 1972–1998, which is available for 421 of our 429 observations. The first-stage has coefficient 0.2309 and standard error 0.0450. The RE model was estimated via maximum likelihood. See Section 7.2 for details.

Obtaining and Preparing Text

- Databases: Lexus Nexus, Factiva
- By hand: Scanning and OCR
- "Web Scraping": automatically downloading web pages
  - Data can be in more complicated forms (PDF files, web pages automatically generated using javascript, etc.)

Obtaining and Preparing Text

KEYSTONE PIPELINE

Mr. THUNE. Mr. President, in a moment some of my colleagues will come to the floor and ask to enter into a colloquy and discuss an issue that is important to many of us, especially to those of us who represent States in the West and Midwest.

The issue I wish to speak about has to do with something that over the past 5 years the Obama administration has been particularly active in pursuing.

Mr. REID. Mr. President, will my friend allow me to ask a question through the Chair?

Mr. THUNE. Yes.

Mr. REID. I was in my office when I heard the statement by the Republican leader about Keystone. I direct this question to the Senator from South Dakota, who is a fine Senator and understands energy issues.

We agreed to have a vote on Keystone. My friend, the Republican leader, keeps misdirecting the matter. We can have a vote on Keystone. That was part of the deal we made. We had a bipartisan bill, Portman- Shaheen. They worked on that bill for months, since last fall. They put in amendments that people wanted.

Jeanne Shaheen came here yesterday and said: Let's have a vote on Keystone, but just as long as we can have a vote on energy efficiency. She even suggested we could have a vote using the McConnell rule–a 60- vote threshold–on both of them.

This is so transparent that my friend the Republican leader is doing the bidding again of the Koch brothers, who own the first or second largest tar

[[<a href="/congressional-record/2014/senate-section/page/S3951" >Page S3951]]

sands holding which exists in the world.

I say to my friend from South Dakota: Why can't we just have a vote on both of those–energy efficiency and on Keystone?

Obtaining and Preparing Text

KEYSTONE PIPELINE

Mr. THUNE. Mr. President, in a moment some of my colleagues will come to the floor and ask to enter into a colloquy and discuss an issue that is important to many of us, especially to those of us who represent States in the West and Midwest.

The issue I wish to speak about has to do with something that over the past 5 years the Obama administration has been particularly active in pursuing.

Mr. REID. Mr. President, will my friend allow me to ask a question through the Chair?

Mr. THUNE. Yes.

Mr. REID. I was in my office when I heard the statement by the Republican leader about Keystone. I direct this question to the Senator from South Dakota, who is a fine Senator and understands energy issues.

We agreed to have a vote on Keystone. My friend, the Republican leader, keeps misdirecting the matter. We can have a vote on Keystone. That was part of the deal we made. We had a bipartisan bill, Portman- Shaheen. They worked on that bill for months, since last fall. They put in amendments that people wanted.

Jeanne Shaheen came here yesterday and said: Let's have a vote on Keystone, but just as long as we can have a vote on energy efficiency. She even suggested we could have a vote using the McConnell rule–a 60- vote threshold–on both of them.

This is so transparent that my friend the Republican leader is doing the bidding again of the Koch brothers, who own the first or second largest tar

[[<a href="/congressional-record/2014/senate-section/page/S3951" >Page S3951]]

sands holding which exists in the world.

I say to my friend from South Dakota: Why can't we just have a vote on both of those–energy efficiency and on Keystone?

Obtaining and Preparing Text

- Thune:
  - Mr. President, in a moment some of my colleagues will come to the floor and ask to enter into a colloquy and discuss an issue that is important to many of us, especially to those of us who represent States in the West and Midwest. The issue I wish to speak about has to do with something that over the past 5 years the Obama administration has been particularly active in pursuing.
  - Yes.
- Reid:
  - Mr. President, will my friend allow me to ask a question through the Chair?
  - I was in my office when I heard the statement by the Republican leader about Keystone. I direct this question to the Senator from South Dakota, who is a fine Senator and understands energy issues. We agreed to have a vote on Keystone. My friend, the Republican leader, keeps misdirecting the matter. We can have a vote on Keystone. That was part of the deal we made. We had a bipartisan bill, Portman-Shaheen. They worked on that bill for months, since last fall. They put in amendments that people wanted. Jeanne Shaheen came here yesterday and said: Let's have a vote on Keystone, but just as long as we can have a vote on energy efficiency. She even suggested we could have a vote using the McConnell rule–a 60-vote threshold–on both of them. This is so transparent that my friend the Republican leader is doing the bidding again of the Koch brothers, who own the first or second largest tar sands holding which exists in the world. I say to my friend from South Dakota: Why can't we just have a vote on both of those–energy efficiency and on Keystone?

Obtaining and Preparing Text

- *Regular expressions* are used to match patterns in text
- Consider the example of breaking up the text above by speaker
- Regular expression pattern,
  regex1 <- "\\n [A-Zc]{3,}\\. "
- Looks for the pattern: an endline character, followed by two spaces, followed by a name consisting of three or more uppercase letters or a lower case c (i.e. McCARTY), followed by a period and a space
- To get the names, *str_match_all(text1,regex1)*
- To get the text, *str_split(text1,regex1)*

Obtaining and Preparing Text

Regular Expressions:

- .: any character ("\.at" matches cat and hat)
- [,]: group possible matches ("0-9" could be used to match any number)
- *: Zero or more of the preceding
- +: 1 or more of the preceding ("[A-Z][a-z]+" could be used to match a capitalized word)
- ?: 0 or 1 of the preceding
- ^: match negation ("[^A-Za-z0-9]" could be used to match anything that is not a letter or number)
- |: choice operator ("cat|dog") matches "cat" or "dog")
- ( ): can be used to group expression
- For the most part, the syntax of regular expressions is the same in different languages
- The syntax for applying regular expressions differs by language

Obtaining and Preparing Text

Regular Expressions:

- The *stringr* package in r is one way to apply regular expressions
- *str_match_all* will find all matches of a regular expression within a string
- *str_replace_all* will replaces all matches of a regular expression with a string, within a string
- *str_split* will split a string by a regular expression

Obtaining and Preparing Text

- Let $d = 1, 2, ..., D$ index documents
- Let $i = 1, 2, ..., I$ index features
- The *document feature matrix* X is the matrix with elements X_{di} denoting the value of feature i in document d
- A document could be a newspaper article, a paragraph from a newspaper article, a Senate press release, a blog post, etc.
- Typically, it will be a count of the number of times term i appears in document d
- It could also be the proportion of times term i appears in document d
- Typically, for most d and i, $X_{di} = 0$, i.e. the document feature matrix is a *sparse matrix*

Obtaining and Preparing Text

Dense Format

$$X = \begin{bmatrix} 3 & 0 & 1 & 0 \\ 0 & 0 & 0 & 4 \\ 0 & 1 & 3 & 7 \\ 1 & 0 & 0 & 2 \\ 5 & 0 & 0 & 0 \end{bmatrix}$$

RCV (row-column-value) Format

$$r = \begin{bmatrix} 1 & 1 & 2 & 3 & 3 & 3 & 4 & 4 & 5 & 5 \end{bmatrix}$$

$$c = \begin{bmatrix} 1 & 3 & 4 & 2 & 3 & 4 & 1 & 4 & 1 & 2 \end{bmatrix}$$

$$v = \begin{bmatrix} 3 & 1 & 4 & 1 & 3 & 7 & 1 & 2 & 5 & 9 \end{bmatrix}$$

Obtaining and Preparing Text

- If the document feature matrix (typically) counts the number of terms, what is a *term*?
- Consider the document, "Unemployment did not rise this month."
  - A single word: {unemployment,did,not,rise,this,month}
  - A single word with some modifications: {unemployment,did,NOT_rise,this,month}
  - An *n-gram*: {unemployment_did,did_not,not_rise,rise_this,this_month}
- Text is converted to the same case so that "Unemployment" and "unemployment" are the same terms
- Punctuation is typically ignored (though case be considered as a "modification"
- Text is *tokenized*
- Text may be *stemmed* so that "employment" and "employ" are the same term

Obtaining and Preparing Text

- *Document frequency* refers to the percent of documents in which a term appears
- *Term frequency* refers to the percent of times a give term appears
- Sometimes drop terms with high or low document frequency or term frequency

Blei, D. M., A. Y. Ng and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

DeBoef, Suzanna and Paul M Kellstedt. 2004. "The Political (and Economic) Origins of Consumer Confidence." *American Journal of Political Science* 48:633–649.

Diermeier, Daniel, Jean-Francois Godbout, Bei Yu and Stefan Kaufmann. 2011. "Language and Ideology in Congress." *British Journal of Political Science* 42:31–55.

Gentzkow, Matthew and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence From U.S. Daily Newspapers." *Econometrica* 78:35–71.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18:1–35.

Groseclose, Tim and Jeffrey. 2005. "A Measure of Media Bias." *Quarterly Journal of Economics* 120:1191–1237.

Hopkins, Daniel and Gary King. 2010. "Extracting Systematic Social Science Meaning from Text." *American Journal of Political Science* 54:229–247.

Kayser, Mark Andreas and Michael Peress. 2015. "The Media, the Economy and the Vote." Working Paper.

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97:311–331.

Martin, Gergory J. and Ali Yurukoglu. 2015. "Bias in Cable News: Persuasion and Polarization." Working Paper.

Monroe, Burt and Ko Maeda. 2004. "Talks Cheap: Text-based Estimation of Rhetorical Ideal Points." Working Paper.

Monroe, Burt, Michael Colaresi and Kevin Quinn. 2008. "Fightin Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* 16:372–403.

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54:209–228.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson and David G. Rand. 2014. "Structural Topic

Models for Open-Ended Survey Responses." *American Journal of Political Science* 58:1064–1082.

Slapin, Jonathan B. and Sven-Oliver Proksch. 2010. "A Poisson Scaling Model for Estimating Time-Series Party Position from Texts." *American Journal of Political Science* 52:705–722.

Soroka, Stuart. 2006. "Good News and Bad News: Asymmetric Responses to Economic Information." *Journal of Politics* 68:372–385.

Soroka, Stuart N, Dominik A Stecula and Christopher Wlezien. 2014. "It's (Change in) the (Future) Economy, Stupid: Economic Indicators, the Media, and Public Opinion." *American Journal of Political Science* .