12. Special Topics

DS-GA 1015, Text as Data
Arthur Spirling

April 27, 2021

Where Are We?

Where Are We?

The Lost Realm of
Here was of old the Witch-realm of Angmar
Carn Dûm
Angmar
Ettenmoors
Fornost
Weather Hills
Weathertop (Amon Sûl)
RHUDAUR
Last Bridge
Bree
East-West Road
South Downs
THE SHIRE
Hobbiton
Sarn Ford
CARDOLAN
Mitheithel
Nîn-in-Eilph (Swanfleet)
Tharbad
Glanduin
MINHIRIATH
Baranduin
Harlond
Harlindon
Forlindon
Forlond
Gulf of Lhûn
Grey Havens
Mithlond
Blue Mountains
Tower Hills
Eryn Vorn
Lond Daer
ENEDWAITH
Gwathló (Greyflood)
Dunland
Isengard
North-South Road
Fords of Isen
(Isen)
Adorn
Angren
FANGORN
WESTFOLD
Helm's Deep
ERED NIMRAIS
Pinnath Gelin
Anfalas (Langstrand)
Andrast (Ras Morthil)
The Realm Of
MIDDLE EARTH
Miles
50
100
150
200

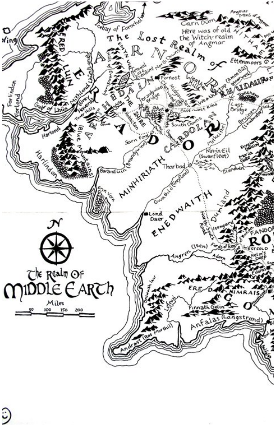
Where Are We?

We've covered the main ideas of text analysis: representing text,

Where Are We?

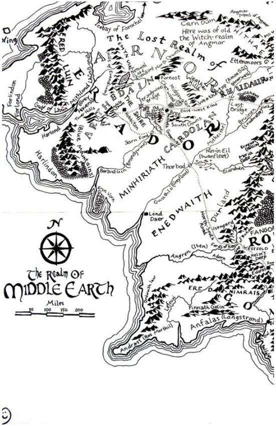
We've covered the main ideas of text analysis: representing text, supervised

Where Are We?



We've covered the main ideas of text analysis: representing text, supervised and unsupervised learning.

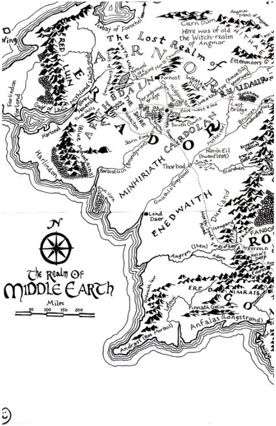
Where Are We?



We've covered the main ideas of text analysis: representing text, supervised and unsupervised learning.

Now look at some 'special topics' on

Where Are We?



We've covered the main ideas of text analysis: representing text, supervised and unsupervised learning.

Now look at some ‘special topics’ on debate,

Where Are We?

We've covered the main ideas of text analysis: representing text, supervised and unsupervised learning.

Now look at some 'special topics' on debate, community behavior,

Where Are We?

We've covered the main ideas of text analysis: representing text, supervised and unsupervised learning.

Now look at some 'special topics' on debate, community behavior, bursts in streams,

Where Are We?

We've covered the main ideas of text analysis: representing text, supervised and unsupervised learning.

Now look at some 'special topics' on debate, community behavior, bursts in streams, memes and spreading of stories/information.

Modeling Debate

Modeling Debate

Modeling Debate

Most politics take place via *debate*

Modeling Debate

Most politics take place via *debate*

i.e. some kind of 'back and forth' between actors with particular roles and views.

Modeling Debate

Most politics take place via *debate*

i.e. some kind of 'back and forth' between actors with particular roles and views.

This is especially common in institutions like legislatures and courts

Modeling Debate

Most politics take place via *debate*

i.e. some kind of 'back and forth' between actors with particular roles and views.

This is especially common in institutions like legislatures and courts but we also see it in 'new media' like twitter.

Modeling Debate

Most politics take place via *debate*

i.e. some kind of 'back and forth' between actors with particular roles and views.

This is especially common in institutions like legislatures and courts but we also see it in 'new media' like twitter.

e.g. Westminster/Parliamentary systems:

Modeling Debate

Most politics take place via *debate*

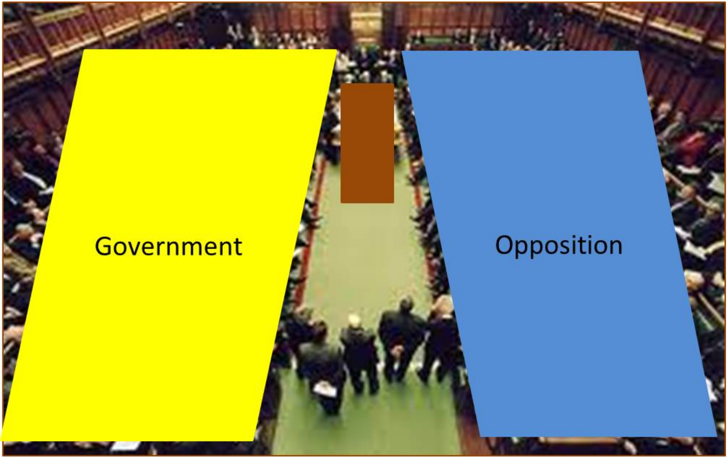
i.e. some kind of 'back and forth' between actors with particular roles and views.

This is especially common in institutions like legislatures and courts but we also see it in 'new media' like twitter.

e.g. Westminster/Parliamentary systems: government-vs-opposition dynamic in which most debate takes place between one party vs the other(s).

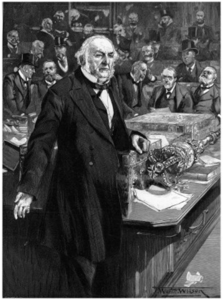
Modern Arrangement

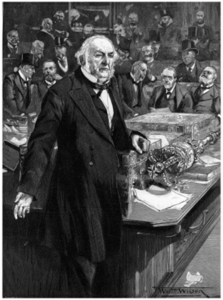
Modern Arrangement



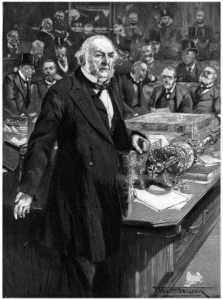
British Political Development

Theory is that between 1880–1902,



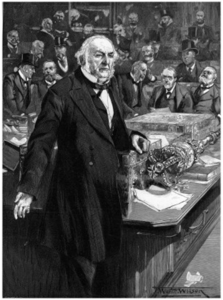


Theory is that between 1880–1902, ministers became more ‘responsive’ to (non-cabinet) MPs.



Theory is that between 1880–1902, ministers became more 'responsive' to (non-cabinet) MPs.

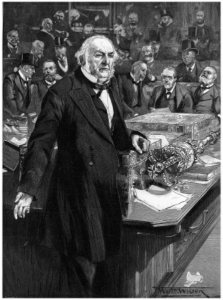
Increase in 'responsiveness' disproportionately accrued to opposition,



Theory is that between 1880–1902, ministers became more ‘responsive’ to (non-cabinet) MPs.

Increase in ‘responsiveness’ disproportionately accrued to opposition, relative to government backbenchers,

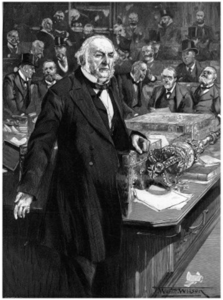
British Political Development



Theory is that between 1880–1902, ministers became more ‘responsive’ to (non-cabinet) MPs.

Increase in ‘responsiveness’ disproportionately accrued to opposition, relative to government backbenchers,

Increase is ‘once-and-for-all’, not response to special circumstances



Theory is that between 1880–1902, ministers became more 'responsive' to (non-cabinet) MPs.

Increase in 'responsiveness' disproportionately accrued to opposition, relative to government backbenchers,

Increase is 'once-and-for-all', not response to special circumstances

→ *need to measure responsiveness*

Three Types of Actors

Three Types of Actors

G_B

Three Types of Actors

G_B government backbencher

Three Types of Actors

G_B government backbencher

G_M

Three Types of Actors

G_B government backbencher

G_M government minister

Three Types of Actors

G_B government backbencher

G_M government minister

O

Three Types of Actors

G_B government backbencher

G_M government minister

O opposition member

Three Types of Actors

G_B government backbencher

G_M government minister

O opposition member

Three Types of Actors

G_B government backbencher

G_M government minister

O opposition member

Three Types of Actors

G_B government backbencher

G_M government minister

O opposition member

What we have

What we have

Every speech made in parliament,
1803–today

What we have

Every speech made in parliament, 1803–today

Focus on period between 1832 (1st RA) and 1915 (4th RA)

What we have

Every speech made in parliament, 1803–today

Focus on period between 1832 (1st RA) and 1915 (4th RA)

1M+ digitized speech records:

What we have

Every speech made in parliament, 1803–today

Focus on period between 1832 (1st RA) and 1915 (4th RA)

1M+ digitized speech records: organized within ‘debates’

What we have

Every speech made in parliament, 1803–today

Focus on period between 1832 (1st RA) and 1915 (4th RA)

1M+ digitized speech records: organized within 'debates'

Problem: party not recorded (needed for Gov/Oppn assignment),

What we have

Every speech made in parliament, 1803–today

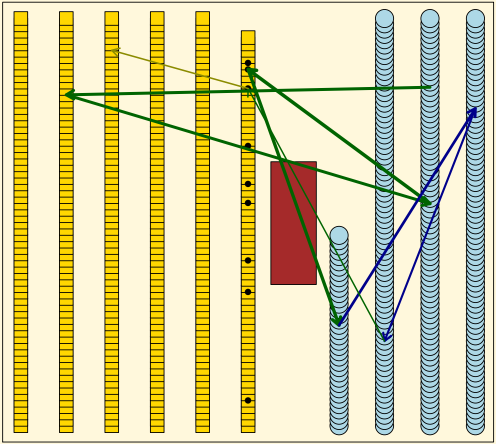
Focus on period between 1832 (1st RA) and 1915 (4th RA)

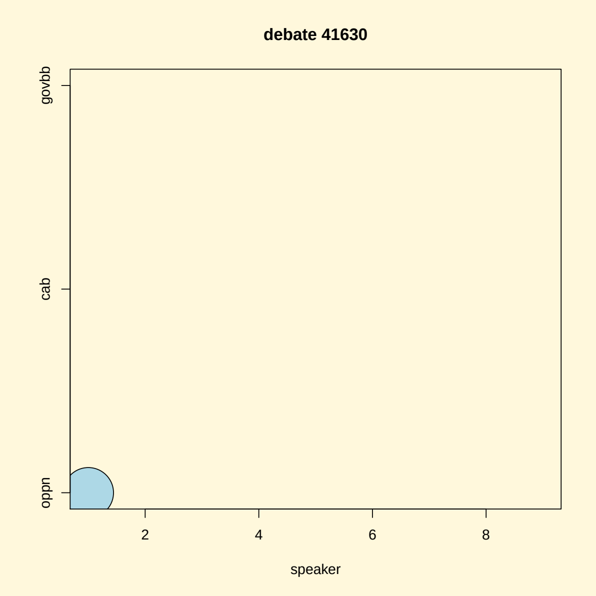
1M+ digitized speech records: organized within ‘debates’

Problem: party not recorded (needed for Gov/Oppn assignment), lots of missingness/errors in terms of MP ids


```
- <p id="S3V0094P0-00140">
  <member>SIR GEORGE GREY</member>
- <membercontribution>
  , in reply, stated that he had not as yet received any reply from the coroner of the
   district, to whom, as well as to the magistrates, he had written; neither had he
   received any communication from the magistrates tending to confirm the charges
   made against the owners of the colliery. He had, in consequence of the statement
   which had been made by the hon. Member for Finsbury respecting the accident,
   addressed a communication to the magistrates and coroner of the district, offering
   any assistance which could be given by the Home Office to forward the inquiry;
   and he had directed the magistrates to inquire rigidly into the means adopted for
   saving the lives of the persons who had been left in the pit, and to investigate the
   substance of the charges made against the proprietors of the colliery. He had just
   received a letter, dated the 6th of July, from the magistrates, in which they stated,
   that in consequence of the letter from the Home Office, they had directed their
   clerk to call a meeting of the magistrates, and that they had heard the statements
   of several parties upon the subjects alluded to in the communication. The result of
   the inquiry was, that they had come to an unanimous opinion as to cause of the
   accident. As that question, however, was still under the consideration of the
   coroner's inquest, he (Sir G. Grey) did not think it would be right for him to state
   the nature of their opinion until the verdict of the coroner's jury should have been
   ascertained. As to the question of the subsequent conduct of the owners of the
   colliery in preventing persons from descending into the pit to rescue those who
   might
  <image src="S3V0094P0I0035" />
  <col>49</col>
  have been left alive in it, the magistrates were convinced that no man left in the pit
   after the explosion could have been alive, and that every exertion that could have
   been made was made to get them out. That letter was signed by five magistrates.
   As he had before stated, he had received no letter from the coroner, whose
   investigation was still proceeding; but he would observe, that the gentleman who
   had been alluded to by the hon. Member for Fins-bury had had every opportunity
   during the inquest of examining and cross-examining any witnesses he chose.
   </membercontribution>
   </p>
- <p id="S3V0094P0-00141">
  <member>MR. DUNCOMBE</member>
  <membercontribution>expressed his astonishment at the hon. Member for Berwick
   denying the grounds for the statement which he had made. He had informed
   Gentlemen who was his authority. The man himself had been in London, and might
   have been examined in the lobby of the House by the hon. Member, had he chosen
   to satisfy himself upon the subject. And now he (Mr. Duncombe) was prepared to
   support the statement he had made. If the masters could have contradicted those
   statements, they had had opportunities of going before the coroner, whose inquiry
   had been adjourned from Thursday last to that very day. But he would state what
   one of the owners, Mr. Robert Lankester, had himself stated. Mr. Robert Lankester
   said the men were bricked up and could not escape.</membercontribution>
   </p>
   </section>
- <section>
```

debate 41630




debate 41630
govbb
cab
oppn
2
4
6
8
speaker

Debate

Debate

Use speech information—

Debate

Use speech information—'to and fro'

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

L_{15}

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

L_{15} $\rightarrow$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \rightarrow L_4$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \rightarrow L_4 \rightarrow$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \rightarrow L_{4} \rightarrow C_{6}$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \rightarrow L_4 \rightarrow C_6 \rightarrow$

Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature

$$L_{15} \to L_4 \to C_6 \to L_2 \to C_5 \to L_2 \to L_{87} \to I_{15} \to C_5 \to C_{13}$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \to L_4 \to C_6 \to L_2 \to C_5 \to L_2 \to L_{87} \to I_{15} \to C_5 \to C_{13}$

O_{15}

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \rightarrow L_4 \rightarrow C_6 \rightarrow L_2 \rightarrow C_5 \rightarrow L_2 \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_5 \rightarrow C_{13}$

$O_{15} \rightarrow$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$L_{15} \to L_4 \to C_6 \to L_2 \to C_5 \to L_2 \to L_{87} \to I_{15} \to C_5 \to C_{13}$$

$$O_{15} \to O_4$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \rightarrow L_{4} \rightarrow C_{6} \rightarrow L_{2} \rightarrow C_{5} \rightarrow L_{2} \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_{5} \rightarrow C_{13}$

$O_{15} \rightarrow O_{4} \rightarrow$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$L_{15} \rightarrow L_4 \rightarrow C_6 \rightarrow L_2 \rightarrow C_5 \rightarrow L_2 \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_5 \rightarrow C_{13}$

$O_{15} \rightarrow O_4 \rightarrow G_{B6}$

Debate

Use speech information—‘to and fro’— to measure how responsive front bench is to legislature

$$L_{15} \rightarrow L_4 \rightarrow C_6 \rightarrow L_2 \rightarrow C_5 \rightarrow L_2 \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_5 \rightarrow C_{13}$$

$$O_{15} \rightarrow O_4 \rightarrow G_{B\,6} \rightarrow O_2 \rightarrow G_{M\,5} \rightarrow O_2 \rightarrow O_{87} \rightarrow O_{31} \rightarrow G_{M\,5} \rightarrow G_{B\,13}$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$L_{15} \rightarrow L_4 \rightarrow C_6 \rightarrow L_2 \rightarrow C_5 \rightarrow L_2 \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_5 \rightarrow C_{13}$$

$$O_{15} \rightarrow O_4 \rightarrow G_{B6} \rightarrow O_2 \rightarrow G_{M5} \rightarrow O_2 \rightarrow O_{87} \rightarrow O_{31} \rightarrow G_{M5} \rightarrow G_{B13}$$

$$O$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$L_{15} \to L_4 \to C_6 \to L_2 \to C_5 \to L_2 \to L_{87} \to I_{15} \to C_5 \to C_{13}$$

$$O_{15} \to O_4 \to G_{B6} \to O_2 \to G_{M5} \to O_2 \to O_{87} \to O_{31} \to G_{M5} \to G_{B13}$$

$$O \to$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$L_{15} \rightarrow L_4 \rightarrow C_6 \rightarrow L_2 \rightarrow C_5 \rightarrow L_2 \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_5 \rightarrow C_{13}$$

$$O_{15} \rightarrow O_4 \rightarrow G_{B6} \rightarrow O_2 \rightarrow G_{M5} \rightarrow O_2 \rightarrow O_{87} \rightarrow O_{31} \rightarrow G_{M5} \rightarrow G_{B13}$$

$$O \rightarrow O$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$L_{15} \rightarrow L_{4} \rightarrow C_{6} \rightarrow L_{2} \rightarrow C_{5} \rightarrow L_{2} \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_{5} \rightarrow C_{13}$$

$$O_{15} \rightarrow O_{4} \rightarrow G_{B6} \rightarrow O_{2} \rightarrow G_{M5} \rightarrow O_{2} \rightarrow O_{87} \rightarrow O_{31} \rightarrow G_{M5} \rightarrow G_{B13}$$

$$O \rightarrow O \rightarrow$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$L_{15} \rightarrow L_4 \rightarrow C_6 \rightarrow L_2 \rightarrow C_5 \rightarrow L_2 \rightarrow L_{87} \rightarrow I_{15} \rightarrow C_5 \rightarrow C_{13}$$

$$O_{15} \rightarrow O_4 \rightarrow G_{B6} \rightarrow O_2 \rightarrow G_{M5} \rightarrow O_2 \rightarrow O_{87} \rightarrow O_{31} \rightarrow G_{M5} \rightarrow G_{B13}$$

$$O \rightarrow O \rightarrow G_B$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$L_{15} \to L_4 \to C_6 \to L_2 \to C_5 \to L_2 \to L_{87} \to I_{15} \to C_5 \to C_{13}$$

$$O_{15} \to O_4 \to G_{B6} \to O_2 \to G_{M5} \to O_2 \to O_{87} \to O_{31} \to G_{M5} \to G_{B13}$$

$$\boxed{O \to O \to G_B \to O \to G_M \to O \to O \to O \to G_M \to G_B}$$

Debate

Use speech information—'to and fro'— to measure how responsive front bench is to legislature

$$\boxed{O \rightarrow O \rightarrow G_B \rightarrow O \rightarrow G_M \rightarrow O \rightarrow O \rightarrow O \rightarrow G_M \rightarrow G_B}$$

If. . .

If. . .

For given debate. . .

If. . .

For given debate. . .

Three states

If. . .

For given debate. . .

Three states: ministers, government backbenchers, opposition

If. . .

For given debate. . .

- Three states: ministers, government backbenchers, opposition
- Sequence of speaker (identities) has Markov property: identity at $t+1$ depends only on speaker at t.

If. . .

For given debate. . .

Three states: ministers, government backbenchers, opposition

Sequence of speaker (identities) has Markov property: identity at $t + 1$ depends only on speaker at t. That is,

If. . .

For given debate. . .

- Three states: ministers, government backbenchers, opposition
- Sequence of speaker (identities) has Markov property: identity at $t+1$ depends only on speaker at t. That is, conditional independence holds.

If. . .

For given debate. . .

- Three states: ministers, government backbenchers, opposition
- Sequence of speaker (identities) has Markov property: identity at $t+1$ depends only on speaker at t. That is, conditional independence holds.

$\rightarrow$ speaker responds to man immediately proceeding him.

If. . .

For given debate. . .

Three states: ministers, government backbenchers, opposition

Sequence of speaker (identities) has Markov property: identity at $t+1$ depends only on speaker at t. That is, conditional independence holds.

$\rightarrow$ speaker responds to man immediately proceeding him. Surely true of interrogatives.

If...

For given debate...

Three states: ministers, government backbenchers, opposition

Sequence of speaker (identities) has Markov property: identity at $t + 1$ depends only on speaker at t. That is, conditional independence holds.

$\rightarrow$ speaker responds to man immediately proceeding him. Surely true of interrogatives.

Resulting Markov chain is discrete:

If. . .

For given debate. . .

Three states: ministers, government backbenchers, opposition

Sequence of speaker (identities) has Markov property: identity at $t+1$ depends only on speaker at t. That is, conditional independence holds.

→ speaker responds to man immediately proceeding him. Surely true of interrogatives.

Resulting Markov chain is discrete: i.e. only relevant information is identity of speaker (not 'how long' we have been in particular state)

If. . .

For given debate. . .

Three states: ministers, government backbenchers, opposition

Sequence of speaker (identities) has Markov property: identity at $t+1$ depends only on speaker at t. That is, conditional independence holds.

$\rightarrow$ speaker responds to man immediately proceeding him. Surely true of interrogatives.

Resulting Markov chain is discrete: i.e. only relevant information is identity of speaker (not 'how long' we have been in particular state)

Resulting Markov chain is (time) homogenous:

If. . .

For given debate. . .

- Three states: ministers, government backbenchers, opposition
- Sequence of speaker (identities) has Markov property: identity at $t+1$ depends only on speaker at t. That is, conditional independence holds.

$\rightarrow$ speaker responds to man immediately proceeding him. Surely true of interrogatives.

- Resulting Markov chain is discrete: i.e. only relevant information is identity of speaker (not 'how long' we have been in particular state)
- Resulting Markov chain is (time) homogenous: i.e. probability of moving from state i to j does not depend on t.

Then. . .

Then. . .

We can characterize chain with the set of transition probabilities

Then. . .

We can characterize chain with the set of transition probabilities

In particular:

Then. . .

We can characterize chain with the set of transition probabilities

In particular:

$$\begin{array}{cc} & \begin{array}{ccc} G_M & G_B & O \end{array} \\ \begin{array}{c} G_M \\ G_B \\ O \end{array} & \begin{pmatrix} m_{MM} & m_{MB} & m_{MO} \\ m_{BM} & m_{BB} & m_{BO} \\ m_{OM} & m_{OB} & m_{OO} \end{pmatrix} \end{array}$$

Then...

We can characterize chain with the set of transition probabilities

In particular:

$$\begin{array}{cc} & \begin{array}{ccc} G_M & G_B & O \end{array} \\ \begin{array}{c} G_M \\ G_B \\ O \end{array} & \begin{pmatrix} m_{MM} & m_{MB} & m_{MO} \\ m_{BM} & m_{BB} & m_{BO} \\ m_{OM} & m_{OB} & m_{OO} \end{pmatrix} \end{array}$$

where m_{ij} is probability of a move from speaker of identity i to speaker of identity j

Task

Task

Must estimate m_{ij}

Task

Must estimate m_{ij}

MLE is well known, and available,

Task

Must estimate m_{ij}

MLE is well known, and available, but we want to include debate and ministry random effects.

Task

Must estimate m_{ij}

MLE is well known, and available, but we want to include debate and ministry random effects.

So use Markov Chain Monte Carlo approach to estimate Generalized Linear Mixed Model.

$\rightarrow$ set up as multinomial logit with random effects

Task

Must estimate m_{ij}

MLE is well known, and available, but we want to include debate and ministry random effects.

So use Markov Chain Monte Carlo approach to estimate Generalized Linear Mixed Model.

$\rightarrow$ set up as multinomial logit with random effects

$$\exp(\boldsymbol{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e})$$

Task

Must estimate m_{ij}

MLE is well known, and available, but we want to include debate and ministry random effects.

So use Markov Chain Monte Carlo approach to estimate Generalized Linear Mixed Model.

$\rightarrow$ set up as multinomial logit with random effects

$$\exp(\boldsymbol{X}\boldsymbol{\beta} + \mathbf{Zu} + \mathbf{e})$$

$\rightarrow$ Predicted probabilities are then (estimates of) transition probabilities

Rearranging Data

Rearranging Data

S_1	S_2	S_3	S_4	S_5	S_6	S_7

Rearranging Data

	S_1		S_2		S_3		S_4		S_5		S_6		S_7
d_2	O	$\rightarrow$	O	$\rightarrow$	G_B	$\rightarrow$	O	$\rightarrow$	G_M	$\rightarrow$	G_M	$\rightarrow$	G_B

Rearranging Data

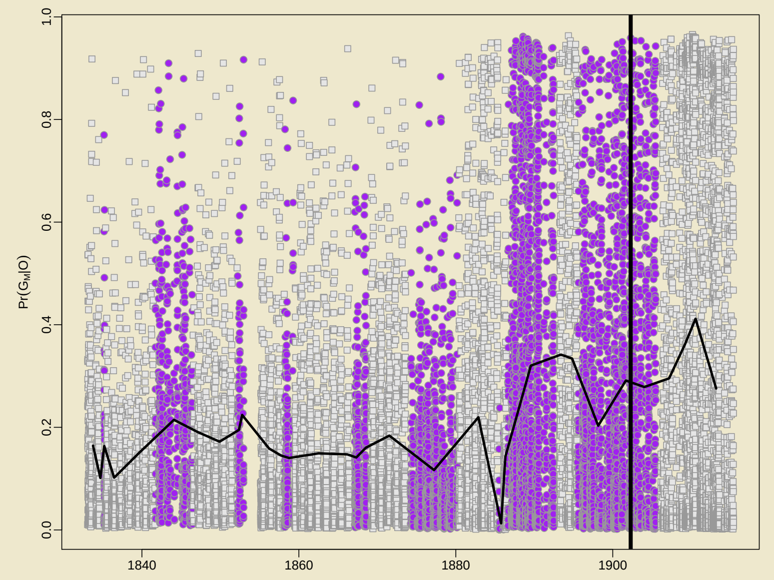
	S_1		S_2		S_3		S_4		S_5		S_6		S_7
d_2	O	$\rightarrow$	O	$\rightarrow$	G_B	$\rightarrow$	O	$\rightarrow$	G_M	$\rightarrow$	G_M	$\rightarrow$	G_B

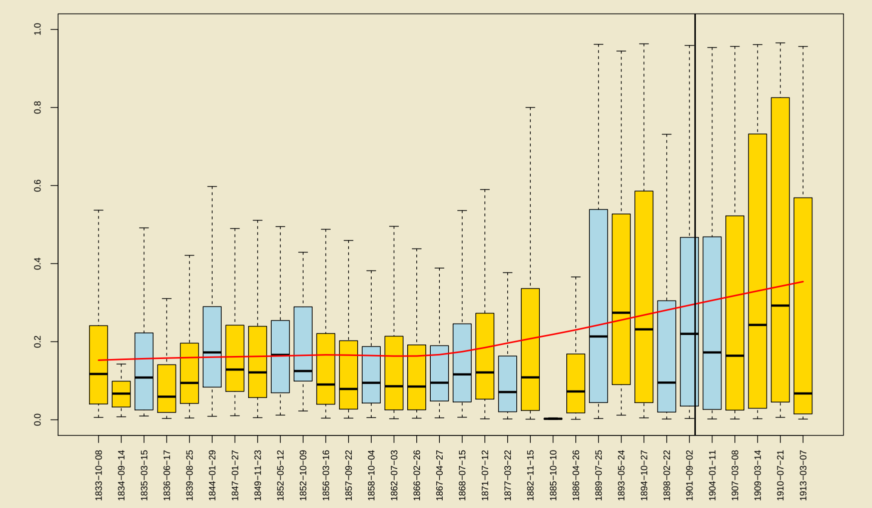
x	x_1	x_2	x_3	x_4	x_5	x_6
	O	O	G_B	O	G_M	G_M

	S_1	S_2	S_3	S_4	S_5	S_6	S_7
d_2	$O \rightarrow$	$O \rightarrow$	$G_B \rightarrow$	$O \rightarrow$	$G_M \rightarrow$	$G_M \rightarrow$	G_B
x	x_1: O	x_2: O	x_3: G_B	x_4: O	x_5: G_M	x_6: G_M	
y		y_1: O	y_2: G_B	y_3: O	y_4: G_M	y_5: G_M	y_6: G_B

Results

Results





0.0
0.2
0.4
0.6
0.8
1.0
1833−10−08
1834−09−14
1835−03−15
1836−06−17
1839−08−25
1844−01−29
1847−01−27
1849−11−23
1852−05−12
1852−10−09
1856−03−16
1857−09−22
1858−10−04
1862−07−03
1866−02−26
1867−04−27
1868−07−15
1871−07−12
1877−03−22
1882−11−15
1885−10−10
1886−04−26
1889−07−25
1893−05−24
1894−10−27
1898−02−22
1901−09−02
1904−01−11
1907−03−08
1909−03−14
1910−07−21
1913−03−07

Danescu-Niculescu-Mizil et al (2013) "No Country for Old Members" study linguistic innovation and adoption in `RateBeer` and `BeerAdvocate`

Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al (2013) "No Country for Old Members" study linguistic innovation and adoption in `RateBeer` and `BeerAdvocate`

Look at users as they 'age' (from first post to last post):

Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al (2013) "No Country for Old Members" study linguistic innovation and adoption in `RateBeer` and `BeerAdvocate`

Look at users as they 'age' (from first post to last post):

Well, early on, users adopt norms of community—until 'linguistic adolescence'—after which they cease to respond to community changes.

Linguistic Change in Online Communities

Danescu-Niculescu-Mizil et al (2013) "No Country for Old Members" study linguistic innovation and adoption in `RateBeer` and `BeerAdvocate`

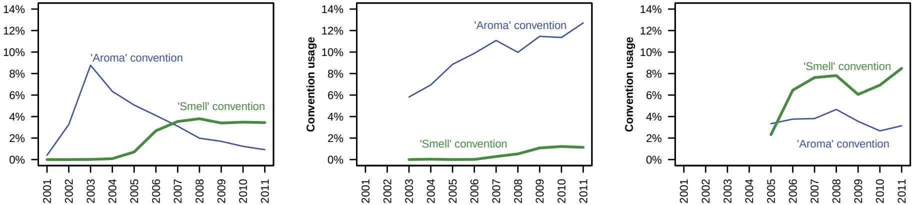
Look at users as they 'age' (from first post to last post):

Well, early on, users adopt norms of community—until 'linguistic adolescence'—after which they cease to respond to community changes.

and results allow them to predict how long user will stay in community from early posts!

Aroma v Smell: overall, 2003 joiners, 2007 joiners

Aroma v Smell: overall, 2003 joiners, 2007 joiners



Intuition

Intuition

'Snapshot language model':

Intuition

'Snapshot language model': use cross-entropy to compare language of a user's (held out) posts to see how different it was to rest of community at the time

Intuition

'Snapshot language model': use cross-entropy to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual bigrams it contained.

Intuition

'Snapshot language model': use cross-entropy to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual bigrams it contained.

When cross entropy is high,

Intuition

'Snapshot language model': use cross-entropy to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual bigrams it contained.

When cross entropy is high, user is using different terms from community.

Intuition

'Snapshot language model': use cross-entropy to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual bigrams it contained.

When cross entropy is high, user is using different terms from community.

Specifically, find that members stop using first person pronouns and start using beer specific vocab.

Intuition

'Snapshot language model': use cross-entropy to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual bigrams it contained.

When cross entropy is high, user is using different terms from community.

Specifically, find that members stop using first person pronouns and start using beer specific vocab.

and all users die 'old' in sense that they stop adopting linguistic innovation proportionally (up to one third of lifespan) to how long they are in the community.

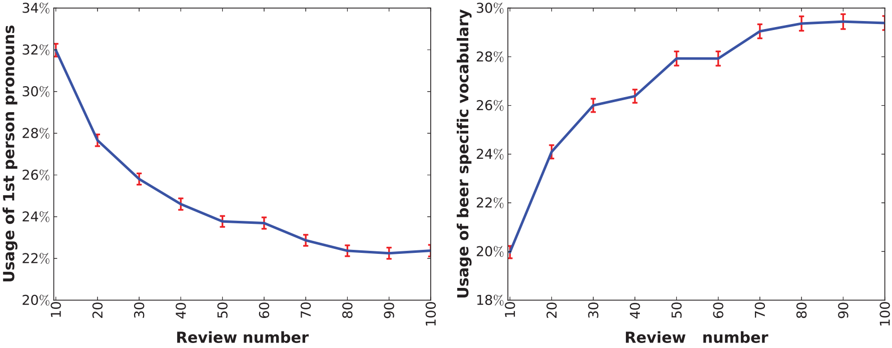
Intuition

'Snapshot language model': use cross-entropy to compare language of a user's (held out) posts to see how different it was to rest of community at the time—in terms of unusual bigrams it contained.

When cross entropy is high, user is using different terms from community.

Specifically, find that members stop using first person pronouns and start using beer specific vocab.

and all users die 'old' in sense that they stop adopting linguistic innovation proportionally (up to one third of lifespan) to how long they are in the community. Implies that 'adult language stability assumption' is *relative* rather than absolute online.



(a) First person sing. pronouns (b) Beer specific vocabulary

Burstiness

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter,

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Surges must be long,

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Surges must be long, and/or intense,

Burstiness

Kleinberg (2002) "Bursty and Hierarchical Structure in Streams" develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As 'gaps' between use of same word become shorter, infer that term is 'bursty'.

Surges must be long, and/or intense, depending on specification of model.

Burstiness

Kleinberg (2002) “Bursty and Hierarchical Structure in Streams” develops methods based on Markov models to detect interesting structure in streams of documents (such as email) arriving over time.

Methods allow one to see which words and phrases have dramatic shifts in usage over time—oftentimes this corresponds to substance but sometimes style.

Idea is to model arrival times of words. As ‘gaps’ between use of same word become shorter, infer that term is ‘bursty’.

Surges must be long, and/or intense, depending on specification of model.

Applying to SOTU, 1790–2002

Applying to SOTU, 1790–2002

Applying to SOTU, 1790–2002

word	burst
gentlemen	1790–1800
british	1809–1814
slaves	1859–1863
japanese	1942–1945
health	1992–1994
help	1998–

Burstiness in Parliament

Burstiness in Parliament

Use burstiness to model MPs as agenda-setters

Burstiness in Parliament

Use burstiness to model MPs as agenda-setters

Burstiness in Parliament

Use burstiness to model MPs as agenda-setters

```
To ask the Secretary
of State for the Home
Department whether, having
regard to the fact that
the women suffragettes now
imprisoned in Holloway Gaol
are political rather than
criminal offenders....
```

(Keir Hardie, Oct 31, 1906)

Burstiness in Parliament

Use burstiness to model MPs as agenda-setters

Look specifically at changes to composition of opposition in terms of concentration of agenda-power.

```
To ask the Secretary
of State for the Home
Department whether, having
regard to the fact that
the women suffragettes now
imprisoned in Holloway Gaol
are political rather than
criminal offenders....
```

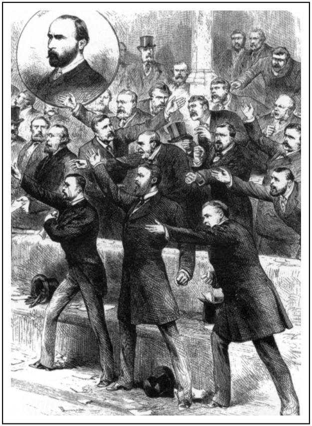
(Keir Hardie, Oct 31, 1906)

Burstiness in Parliament

Use burstiness to model MPs as agenda-setters

Look specifically at changes to composition of opposition in terms of concentration of agenda-power.

→ timing of emergence of small(er) group of opposition agenda-setters is *prima facie* evidence of shadow cabinet (which is otherwise impossible to document)

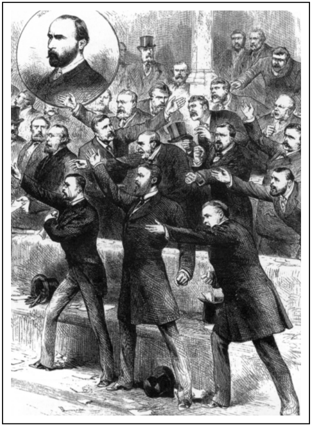


Burstiness in Parliament

Use burstiness to model MPs as agenda-setters

Look specifically at changes to composition of opposition in terms of concentration of agenda-power.

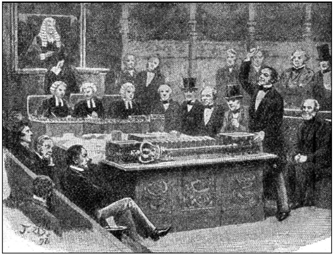
→ timing of emergence of small(er) group of opposition agenda-setters is *prima facie* evidence of shadow cabinet (which is otherwise impossible to document)



Burstiness in Parliament

Use burstiness to model MPs as agenda-setters

Look specifically at changes to composition of opposition in terms of concentration of agenda-power.



→ timing of emergence of small(er) group of opposition agenda-setters is *prima facie* evidence of shadow cabinet (which is otherwise impossible to document)

Details

Details

denote the gap 'times' as x.

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

Details

denote the gap ‘times’ as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters,

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large,

Details

denote the gap ‘times’ as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down),

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down), we have evidence that a burst is occurring (ending).

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T}s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down), we have evidence that a burst is occurring (ending).

In principle,

Details

denote the gap 'times' as x. Suppose that these gaps are exponentially distributed:

$$f(x) = \alpha_i e^{-\alpha_i x}$$

where α_i is the rate

wrt model parameters, $\alpha_i = \frac{n}{T} s^i$, where n is the number of occurrences and T is total time.

i.e. when α_i is large, (mean) wait is short (soon see word again).

If estimated α changes up (down), we have evidence that a burst is occurring (ending).

In principle, s could be estimated, but typically set to 2.

Details II

Details II

Here gaps are number of speeches between uses of a term.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches,

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word,

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

i.e. if $s = 2$, arrival rate has to increase (wrt to $\frac{1}{\alpha_0}$): $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

i.e. if $s = 2$, arrival rate has to increase (wrt to $\frac{1}{\alpha_0}$): $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

$\rightarrow$ This is a geometric process: unless wait time halves (somewhere), we never leave level $i = 0$.

Details II

Here gaps are number of speeches between uses of a term. Obviously discrete, but approximates something continuous (time).

The base rate of word usage is $f_0(x) = \alpha_0 e^{-\alpha_0 x}$ where $\alpha_0 = \frac{n}{T}$.

e.g. there are 100 speeches, 5 of them mention word, $\alpha_0 = \frac{1}{20}$.

For a fixed value of s, going from base level to i=1 to i=2 requires decreasing mean wait time by factor of $\frac{1}{s^i}$ each time.

i.e. if $s = 2$, arrival rate has to increase (wrt to $\frac{1}{\alpha_0}$): $1 \rightarrow \frac{1}{2} \rightarrow \frac{1}{4} \rightarrow \frac{1}{8}$

$\rightarrow$ This is a geometric process: unless wait time halves (somewhere), we never leave level $i = 0$.

$\Leftrightarrow$ this is an infinite state hidden Markov model.

More Intuition

More Intuition

Consider a word like `the`

More Intuition

Consider a word like `the`: used in (almost) every speech,

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small.

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`

More Intuition

Consider a word like the: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like zither: very rarely used,

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long.

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

More Intuition

Consider a word like the: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like zither: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

Consider a word like boundary

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

btw in keeping with original presentation, we treat base level as zero.

More Intuition

- Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).
- Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).
- Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

btw in keeping with original presentation, we treat base level as zero.

NB no need to stem or stop words:

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

btw in keeping with original presentation, we treat base level as zero.

NB no need to stem or stop words: stops should all be level 0 anyway.

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

btw in keeping with original presentation, we treat base level as zero.

NB no need to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

More Intuition

- Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).
- Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).
- Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

btw in keeping with original presentation, we treat base level as zero.

NB no need to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

so this is not tf-idf (or similar):

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

btw in keeping with original presentation, we treat base level as zero.

NB no need to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

so this is not tf-idf (or similar): not based on which words typify documents,

More Intuition

Consider a word like `the`: used in (almost) every speech, so gaps are always small. Would be almost impossible to make term bursty (given we have binary use metric).

Consider a word like `zither`: very rarely used, so gaps are always long. Would be almost impossible to make term bursty (under any reasonable choice of s).

Consider a word like `boundary`: intense bursts of usage in some years when redistribution of seats is discussed.

btw in keeping with original presentation, we treat base level as zero.

NB no need to stem or stop words: stops should all be level 0 anyway. We do make everything lowercase.

so this is not tf-idf (or similar): not based on which words typify documents, but rather on relative intensity of use *within word* over time.

Validation

Validation

Ultimately, this is measurement problem,

Validation

Ultimately, this is measurement problem, so validation matters:

Validation

Ultimately, this is measurement problem, so validation matters:

1. at a given time,

Validation

Ultimately, this is measurement problem, so validation matters:

1 at a given time, when we know certain issues were at stake,

Validation

Ultimately, this is measurement problem, so validation matters:

1. at a given time, when we know certain issues were at stake, they should be bursty

Validation

Ultimately, this is measurement problem, so validation matters:

1 at a given time, when we know certain issues were at stake, they should be bursty

2 given words should be bursty at 'right' times

Validation

Ultimately, this is measurement problem, so validation matters:

1. at a given time, when we know certain issues were at stake, they should be bursty

2. given words should be bursty at 'right' times (when we rescale all terms between 0 and 1)

Validation

Ultimately, this is measurement problem, so validation matters:

1 at a given time, when we know certain issues were at stake, they should be bursty

2 given words should be bursty at 'right' times (when we rescale all terms between 0 and 1)

3 given MPs should be bursty at 'right' times

Validation

Ultimately, this is measurement problem, so validation matters:

1 at a given time, when we know certain issues were at stake, they should be bursty

2 given words should be bursty at 'right' times (when we rescale all terms between 0 and 1)

3 given MPs should be bursty at 'right' times (when we rescale all terms between 0 and 1)

Validation

Ultimately, this is measurement problem, so validation matters:

1. at a given time, when we know certain issues were at stake, they should be bursty
2. given words should be bursty at 'right' times (when we rescale all terms between 0 and 1)
3. given MPs should be bursty at 'right' times (when we rescale all terms between 0 and 1)

Validating I: right words are bursty, given time

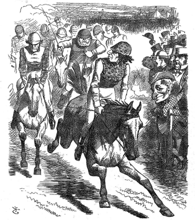
Validating I: right words are bursty, given time

session	1846 (1841, 6)	1866 (1865, 1)	1885 (1885, 1)

Validating I: right words are bursty, given time

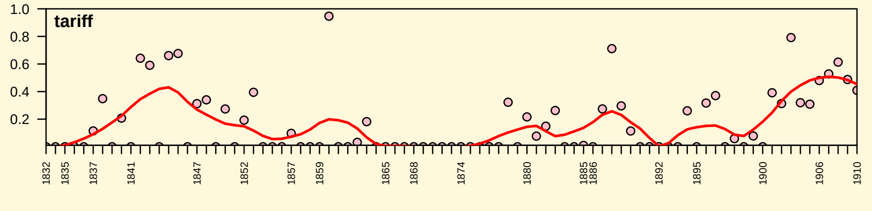
session	1846 (1841, 6)	1866 (1865, 1)	1885 (1885, 1)
		 THE DERBY, 1867. DIZZY WINS WITH "REFORM BILL." MR. PUNCH. "DON'T BE TOO SURE; WAIT TILL HE'S WEIGHED!"	 THE RIVALS.

Validating I: right words are bursty, given time

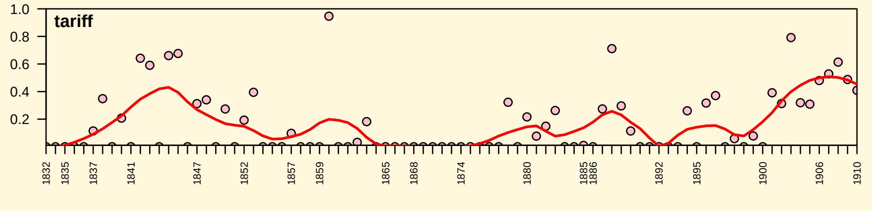
session	1846 (1841, 6)	1866 (1865, 1)	1885 (1885, 1)
		 THE DERBY, 1867. DIZZY WINS WITH "REFORM BILL." MR. PUNCH. "DON'T BE TOO SURE; WAIT TILL HE'S WEIGHED."	 THE RIVALS.
terms (rank)	agriculturists (1) wheat (3) grain (5) farmer (6) prices (7)	suffrage (4) franchise (5) 1832 (7) redistribution (10) seats (11)	irishmen (2) 1782 (3) kingharmon (6) parnell (15) tenant (18)

Validating II: given words are bursty, at right time

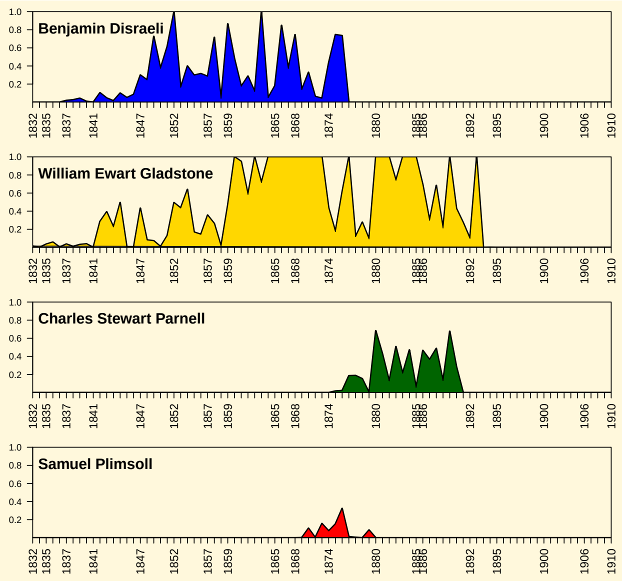
Validating II: given words are bursty, at right time



Validating II: given words are bursty, at right time



Validating II: right MPs are bursty, at right time



"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web:

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

“Meme-tracking and the Dynamics of the News Cycle”, Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are ‘genetic signatures’: mutate over time,

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track!

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track! Key components:

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track! Key components:

1. imitation:

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track! Key components:

1. imitation: news sources copy each others' decisions on what to cover,

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track! Key components:

1. imitation: news sources copy each others' decisions on what to cover, subject to
2. recency,

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track! Key components:

1. imitation: news sources copy each others' decisions on what to cover, subject to
2. recency, in sense that they want 'fresh' content.

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track! Key components:

1. imitation: news sources copy each others' decisions on what to cover, subject to
2. recency, in sense that they want 'fresh' content.

How should we think about what counts as a 'meme'?

"Meme-tracking and the Dynamics of the News Cycle", Leskovec, Backstrom, and Kleinberg (2008)

Want to understand how short, signature phrases—memes—diffuse and propagate over web: from news, to blogs etc.

Memes are 'genetic signatures': mutate over time, which means very hard to track! Key components:

1. imitation: news sources copy each others' decisions on what to cover, subject to
2. recency, in sense that they want 'fresh' content.

How should we think about what counts as a 'meme'? How does attention peak and decay?

Data and Examples

Data and Examples

Data and Examples

"You can put lipstick on a pig,"

Data and Examples

"You can put `lipstick on a pig`," he said as the crowd cheered. "It's still a pig."

Data and Examples

"You can put `lipstick on a pig`," he said as the crowd cheered. "It's still a pig."

Gathered 1 million news articles in run up to 2008 election,

"You can put `lipstick on a pig`," he said as the crowd cheered. "It's still a pig."

Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

Data and Examples

"You can put `lipstick on a pig`," he said as the crowd cheered. "It's still a pig."

Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

From that,

Data and Examples

"You can put `lipstick on a pig`," he said as the crowd cheered. "It's still a pig."

Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

From that, gathered 112 million quoted strings (phrases),

Data and Examples

"You can put `lipstick on a pig`," he said as the crowd cheered. "It's still a pig."

Gathered 1 million news articles in run up to 2008 election, plus 1.6 million blog posts

From that, gathered 112 million quoted strings (phrases), which they have to parse down to phrase clusters to observe over time.

Concept

Concept

A phrase cluster is collection of phrases/quotes that are close textual variants of each other.

Concept

A phrase cluster is collection of phrases/quotes that are close textual variants of each other.

Impose lower bound on how short phrases can be,

Concept

A phrase cluster is collection of phrases/quotes that are close textual variants of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

Concept

A phrase cluster is collection of phrases/quotes that are close textual variants of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

Build a graph of a quote in which each node is one of the phrases actually reported,

Concept

A phrase cluster is collection of phrases/quotes that are close textual variants of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

Build a graph of a quote in which each node is one of the phrases actually reported, which contain some contiguous subsequence of the words in the original quote

Concept

A phrase cluster is collection of phrases/quotes that are close textual variants of each other.

Impose lower bound on how short phrases can be, and how commonly they must appear

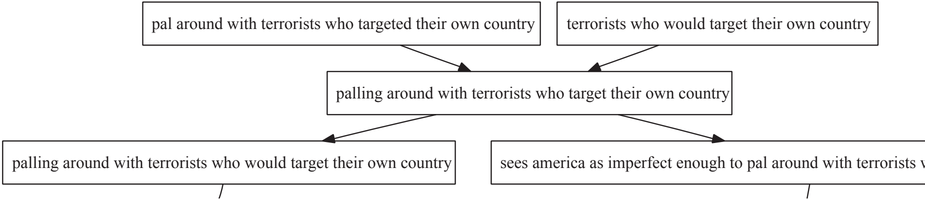
Build a graph of a quote in which each node is one of the phrases actually reported, which contain some contiguous subsequence of the words in the original quote

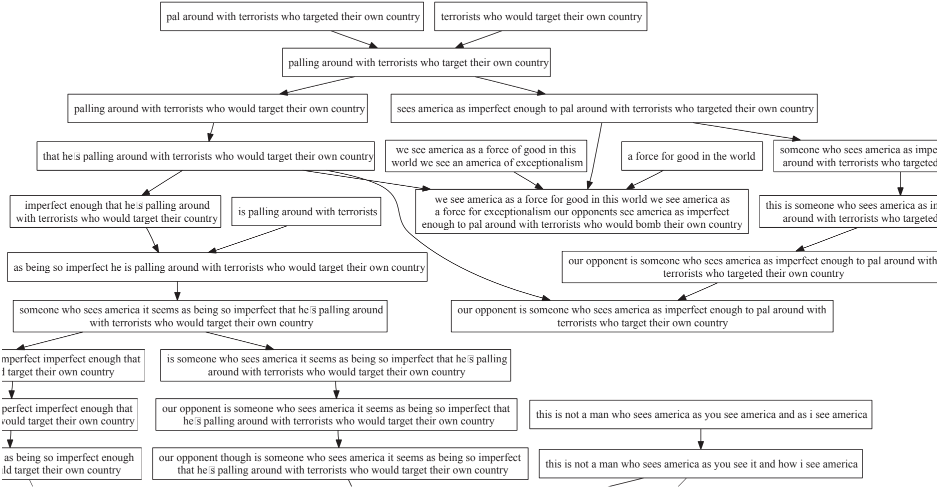
then partition the graph back to a 'meme' of closely related (defined technically) phrases that can be followed through the news media as a thread

Example

Example

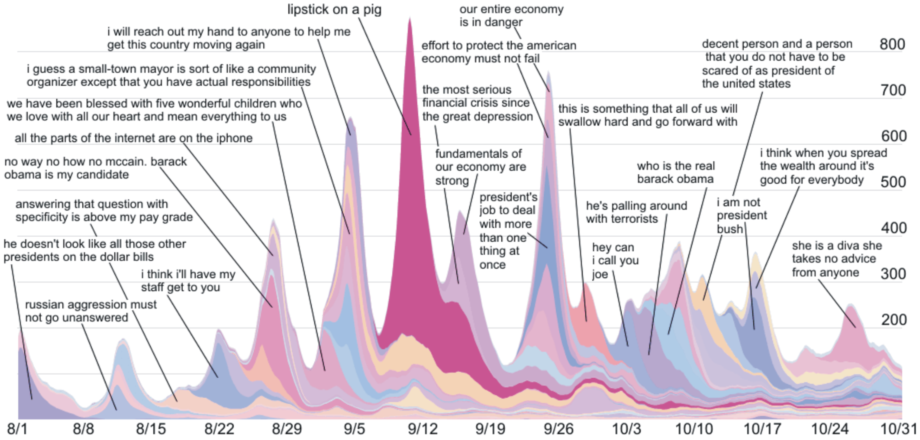
"Our opponent is someone who sees America, it seems, as being so imperfect, imperfect enough that he's palling around with terrorists who would target their own country."


pal around with terrorists who targeted their own country
terrorists who would target their own country
palling around with terrorists who target their own country
palling around with terrorists who would target their own country
sees america as imperfect enough to pal around with terrorists


pal around with terrorists who targeted their own country
terrorists who would target their own country
palling around with terrorists who target their own country
palling around with terrorists who would target their own country
sees america as imperfect enough to pal around with terrorists who targeted their own country
that he□s palling around with terrorists who would target their own country
we see america as a force of good in this world we see an america of exceptionalism
a force for good in the world
someone who sees america as impe around with terrorists who targeted
imperfect enough that he□s palling around with terrorists who would target their country
is palling around with terrorists
we see america as a force for good in this world we see america as a force for exceptionalism our opponents see america as imperfect enough to pal around with terrorists who would bomb their own country
this is someone who sees america as in around with terrorists who targeted
as being so imperfect he is palling around with terrorists who would target their own country
our opponent is someone who sees america as imperfect enough to pal around with terrorists who targeted their own country
someone who sees america it seems as being so imperfect that he□s palling around with terrorists who would target their own country
our opponent is someone who sees america as imperfect enough to pal around with terrorists who target their own country
mperfect imperfect enough that l target their own country
is someone who sees america it seems as being so imperfect that he□s palling around with terrorists who would target their own country
perfect imperfect enough that vould target their own country
our opponent is someone who sees america it seems as being so imperfect that he□s palling around with terrorists who would target their own country
this is not a man who sees america as you see america and as i see america
as being so imperfect enough ld target their own country
our opponent though is someone who sees america it seems as being so imperfect that he□s palling around with terrorists who would target their own country
this is not a man who sees america as you see it and how i see america

Top 50 threads in 2008/9

Top 50 threads in 2008/9



Other Findings

Other Findings

News cycle is a product of imitation and recency.

Other Findings

News cycle is a product of imitation and recency.

Peak and decay is symmetric: not fast build up, and slow decay.

Other Findings

News cycle is a product of imitation and recency.

Peak and decay is symmetric: not fast build up, and slow decay.

Peak blog intensity is around 2.5 hours after news peak ('heart beat' pattern of decline in news, then picked up by blogs).

Other Findings

News cycle is a product of imitation and recency.

Peak and decay is symmetric: not fast build up, and slow decay.

Peak blog intensity is around 2.5 hours after news peak (‘heart beat’ pattern of decline in news, then picked up by blogs).

Generally movement is news $\rightarrow$ blogs,

Other Findings

News cycle is a product of imitation and recency.

Peak and decay is symmetric: not fast build up, and slow decay.

Peak blog intensity is around 2.5 hours after news peak ('heart beat' pattern of decline in news, then picked up by blogs).

Generally movement is news $\rightarrow$ blogs, but some phrases move the other way.

Quotus (Niculae et al, 2015)

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles,

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles, 2274 Obama public speeches.

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is fixed and prominent,

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is fixed and prominent, so question is how outlets choose to quote (may) reflect bias.

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is fixed and prominent, so question is how outlets choose to quote (may) reflect bias.

Main model matrix/graph has 275 media outlets,

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is fixed and prominent, so question is how outlets choose to quote (may) reflect bias.

Main model matrix/graph has 275 media outlets, and their selection of the $267,000$ quotes they could use (binary).

Quotus (Niculae et al, 2015)

There is *prima facie* evidence that outlets of different hues embrace different quotes from Obama's speeches.

Niculae et al set out to examine whether there is a (ideological) bias in reporting.

Data is six billion news articles, 2274 Obama public speeches. Idea is that what Obama says is fixed and prominent, so question is how outlets choose to quote (may) reflect bias.

Main model matrix/graph has 275 media outlets, and their selection of the $267,000$ quotes they could use (binary).

Prima Facie evidence of bias

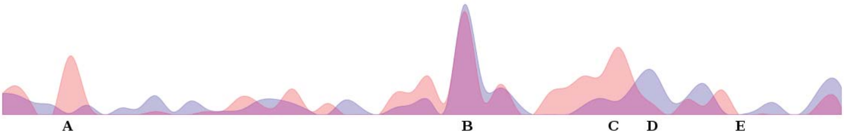


Figure 1: Volume of quotations for each word from a fragment of the 2010 State of the Union Address split by political leaning: conservative outlets shown in red and liberal outlets shown in blue. Quotes from the marked positions are reproduced in Table 1 and shown in the QUOTUS visualization in Figure 2.

Position	Quote from the 2010 State of the Union Address
A	And in the last year, hundreds of al Qaeda's fighters and affiliates, including many senior leaders, have been captured or killed—far more than in 2008.
B	I will work with Congress and our military to finally repeal the law that denies gay Americans the right to serve the country they love because of who they are. It's the right thing to do.
C	Each time lobbyists game the system or politicians tear each other down instead of lifting this country up, we lose faith. The more that TV pundits reduce serious debates to silly arguments, big issues into sound bites, our citizens turn away.
D	Democracy in a nation of 300 million people can be noisy and messy and complicated. And when you try to do big things and make big changes, it stirs passions and controversy. That's just how it is.
E	But I wake up every day knowing that they are nothing compared to the setbacks that families all across this country have faced this year.

Findings

Findings

Authors (basically) reduce the main matrix down to two dimensions,

Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

1. first dimension is independent-mainstream while

Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

1. first dimension is independent-mainstream while
2. second dimension is liberal-conservative.

Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

1. first dimension is independent-mainstream while
2. second dimension is liberal-conservative.

Find that more conservative outlets tend to favor quotes that display negative sentiment (depressing!),

Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

1. first dimension is independent-mainstream while
2. second dimension is liberal-conservative.

Find that more conservative outlets tend to favor quotes that display negative sentiment (depressing!), more negation (controversial topics),

Findings

Authors (basically) reduce the main matrix down to two dimensions, and discover:

1. first dimension is independent-mainstream while
2. second dimension is liberal-conservative.

Find that more conservative outlets tend to favor quotes that display negative sentiment (depressing!), more negation (controversial topics), more conservative topics of interest (e.g. troops rather than health care)

Quote Results

Quote Results

	First dimension of bias
High	The principle that people of all faiths are welcome in this country, and will not be treated differently by their government, is essential to who we are.
	The United States is not, and will never be, at war with Islam. In fact, our partnership with the Muslim world is critical.
	At a time when our discourse has become so sharply polarized [...] it's important for us to pause for a moment and make sure that we are talking with each other in a way that heals, not a way that wounds.
Low	Tonight, we are turning the east room into a bona fide country music hall.
	You guys get two presidents for one, which is a pretty good deal.
	Now, nothing wrong with an art history degree—I love art history.

	Second dimension of bias
High	Those of you who are watching certain news channels, on which I'm not very popular, and you see folks waving tea bags around...
	If we don't work even harder than we did in 2008, then we're going to have a government that tells the American people, "you're on your own."
	By the way, if you've got health insurance, you're not getting hit by a tax.
Middle	Congress passed a temporary fix. A band-aid. But these cuts are scheduled to keep falling across other parts of the government that provide vital services for the American people.
	Keep in mind, nobody is asking them to raise income tax rates. All we're asking is for them to consider closing tax loopholes and deductions.
	The truth is, you could figure out on the back of an envelope how to get this done. The question is one of political will.
Low	By the end of the next year, all U.S. troops will be out of Iraq.
	We come together here in Copenhagen because climate change poses a grave and growing danger to our people.
	Wow, we must come together to end this war successfully.