12. Special Topics (flipped)

DS-GA 1015, Text as Data
Arthur Spirling

April 27, 2021

Housekeeping

Housekeeping

1 Homework coming in on May 4, 2021, at 11pm.

Housekeeping

1. Homework coming in on May 4, 2021, at 11pm.
2. Final papers come in on May 12 . No extensions, no exceptions!

Housekeeping

1. Homework coming in on May 4, 2021, at 11pm.

2. Final papers come in on May 12 . No extensions, no exceptions! Details will follow for uploading etc.

Housekeeping

1. Homework coming in on May 4, 2021, at 11pm.
2. Final papers come in on May 12 . No extensions, no exceptions! Details will follow for uploading etc.
3. Lecture next week is flipped only.

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

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

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

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

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}

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

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

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

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

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

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

$O \rightarrow$

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

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

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

$$O \to O \to G_B$$

Debate

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

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

$$O_{15} \to O_{4} \to G_{B\,6} \to O_{2} \to G_{M\,5} \to O_{2} \to O_{87} \to O_{31} \to G_{M\,5} \to G_{B\,13}$$

$$\boxed{O \to O \to G_{B} \to O \to G_{M} \to O \to O \to O \to G_{M} \to G_{B}}$$

Debate

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

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

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.

→ 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.

$\rightarrow$ 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)

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

Exercise

Exercise

1. Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{array}{cc} G & B \end{array} \\ \begin{array}{c} G \\ B \end{array} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

Exercise

1. Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{array}{cc} G & B \end{array} \\ \begin{array}{c} G \\ B \end{array} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

If yesterday was a 'good' day, what is the probability that today is a 'bad' day?

Exercise

1. Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{array}{cc} G & B \end{array} \\ \begin{array}{c} G \\ B \end{array} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix} \end{array}.$$

   If yesterday was a 'good' day, what is the probability that today is a 'bad' day? How does you prediction for today change if you learn that the day before yesterday was a 'good' day?

Exercise

1. Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{cc} & \begin{array}{cc} G & B \end{array} \\ \begin{array}{c} G \\ B \end{array} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix}. \end{array}$$

   If yesterday was a 'good' day, what is the probability that today is a 'bad' day? How does you prediction for today change if you learn that the day before yesterday was a 'good' day?

2. If today is a 'good' day, what is the probability that the weather two days from now will be good?

Exercise

1. Suppose that the weather {good, bad} evolves according to a Markov chain. In particular:

$$\begin{array}{c c} & \begin{array}{cc} G & B \end{array} \\ \begin{array}{c} G \\ B \end{array} & \begin{pmatrix} 0.8 & 0.2 \\ 0.6 & 0.4 \end{pmatrix}. \end{array}$$

   If yesterday was a 'good' day, what is the probability that today is a 'bad' day? How does you prediction for today change if you learn that the day before yesterday was a 'good' day?

2. If today is a 'good' day, what is the probability that the weather two days from now will be good?

Language Communities

Language Communities

'Snapshot language model':

Language Communities

'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

Language Communities

'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.

Language Communities

'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,

Language Communities

'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.

Language Communities

'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.

Language Communities

'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.

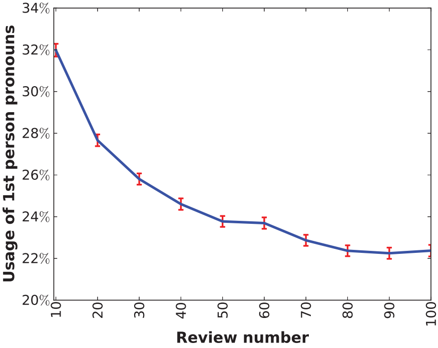
Language Communities

'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.

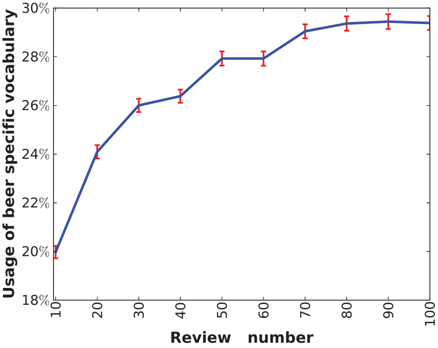
When cross entropy is high, user is using different terms from community.

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

Exercise

Exercise

According to psychologists like Pennebaker, the use of first person singular pronouns ('I', 'me', 'my') is more common in some groups than others.

Exercise

According to psychologists like Pennebaker, the use of first person singular pronouns ('I', 'me', 'my') is more common in some groups than others. Who do you think uses them more (and why?) in blogs, essays etc:

Exercise

According to psychologists like Pennebaker, the use of first person singular pronouns ('I', 'me', 'my') is more common in some groups than others. Who do you think uses them more (and why?) in blogs, essays etc:

1. men or women?

2. people who are happy or people who are depressed?

3. extraverts or introverts?

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.

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.

Exercise

Exercise

1. Do we need to remove stop words when using calculating burstiness of given tokens? Why or why not?

Exercise

1. Do we need to remove stop words when using calculating burstiness of given tokens? Why or why not?

2. Should we stem the words in the texts?

Exercise

1. Do we need to remove stop words when using calculating burstiness of given tokens? Why or why not?

2. Should we stem the words in the texts?

3. How do models of the burstiness of words differ from 'topic models'? Which would you use to study changing subjects of debate over time? Which would you use to study conceptual change?

BTW, Twitter trends

BTW, Twitter trends

How are Trends determined?

Trends are determined by an algorithm and, by default, are tailored for you based on who you follow, your interests, and your location. This algorithm identifies topics that are popular now, rather than topics that have been popular for a while or on a daily basis, to help you discover the hottest emerging topics of discussion on Twitter.

key is to ‘spike’, rather than long build up:

BTW, Twitter trends

How are Trends determined?

Trends are determined by an algorithm and, by default, are tailored for you based on who you follow, your interests, and your location. This algorithm identifies topics that are popular now, rather than topics that have been popular for a while or on a daily basis, to help you discover the hottest emerging topics of discussion on Twitter.

key is to 'spike', rather than long build up: thus #FreddieGray and #OccupyWallStreet did not trend.

BTW, Twitter trends

How are Trends determined?

Trends are determined by an algorithm and, by default, are tailored for you based on who you follow, your interests, and your location. This algorithm identifies topics that are popular now, rather than topics that have been popular for a while or on a daily basis, to help you discover the hottest emerging topics of discussion on Twitter.

key is to 'spike', rather than long build up: thus #FreddieGray and #OccupyWallStreet did not trend.

Twitter seems to hand-curate also (unsurprisingly)

"Trumping Hate on Twitter", Siegal et al, 2019

"Trumping Hate on Twitter", Siegal et al, 2019

Strong prior that Trump's rhetoric up to 2016 election lead to permanent increase in racist language on Twitter.

“Trumping Hate on Twitter”, Siegal et al, 2019

Strong prior that Trump’s rhetoric up to 2016 election lead to permanent increase in racist language on Twitter.

Authors gather $\sim$ 750M election tweets, and use dictionary and non-dictionary approaches.

"Trumping Hate on Twitter", Siegal et al, 2019

Strong prior that Trump's rhetoric up to 2016 election lead to permanent increase in racist language on Twitter.

Authors gather $\sim$ 750M election tweets, and use dictionary and non-dictionary approaches.

Find no evidence of sustained increase:

“Trumping Hate on Twitter”, Siegal et al, 2019

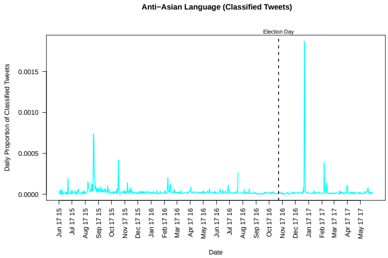
Strong prior that Trump’s rhetoric up to 2016 election lead to permanent increase in racist language on Twitter.

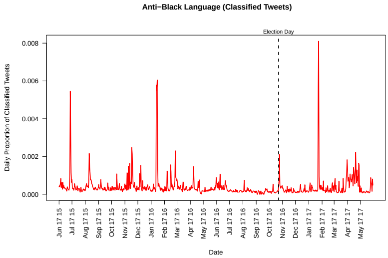
Authors gather $\sim$ 750M election tweets, and use dictionary and non-dictionary approaches.

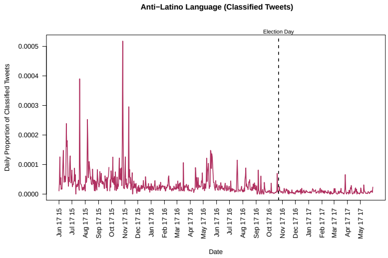
Find no evidence of sustained increase: “[W]hile key campaign events and policy announcements produced brief spikes in hateful language, these bursts quickly dissipated”

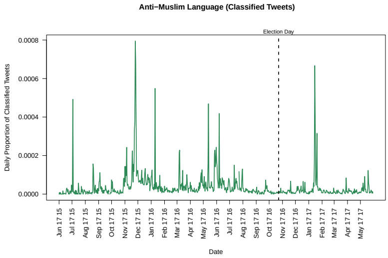
Anti-Asian, Anti-Black, Anti-Latino, Anti-Muslim tweets

Anti-Asian, Anti-Black, Anti-Latino, Anti-Muslim tweets









Recap: “Meme-tracking and the Dynamics of the News Cycle”

Recap: "Meme-tracking and the Dynamics of the News Cycle"

What is a meme for these authors?

Recap: "Meme-tracking and the Dynamics of the News Cycle"

What is a meme for these authors?

Why are they hard to track?

Recap: "Meme-tracking and the Dynamics of the News Cycle"

What is a meme for these authors?

Why are they hard to track?

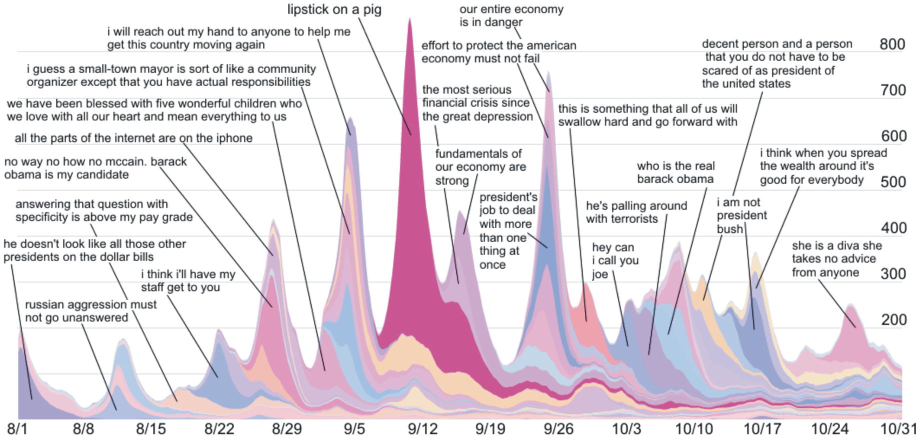
Authors find peak and decay of memes is symmetric: what does this suggest?

McCain: “The fundamentals of our economy are strong”

Length	Qoute text	Vo
7	the fundamentals of our economy are strong	365
7	the fundamentals of the economy are strong	988
6	fundamentals of our economy are strong	645
6	fundamentals of the economy are strong	557
41	if john mccain hadn't said that the fundamentals of our economy are strong on the day of one of our nation's worst financial crises the claim that he invented the blackberry would have been the most preposterous thing said all week	224
4	fundamentals of the economy	172
7	the fundamentals of the economy are sound	119
18	i promise you we will never put america in this position again we will clean up wall street	83
7	the fundamentals of our economy are sound	81
4	clean up wall street	78
12	our economy i think still the fundamentals of our economy are strong	75
6	fundamentals of the economy are sound	72
27	the fundamentals of our economy are strong but these are very very difficult times and i promise you we will never put america in this position again	68
5	the economy is in crisis	66
6	these are very very difficult times	63

Top 50 threads in 2008/9

Top 50 threads in 2008/9



Quotus: Bias

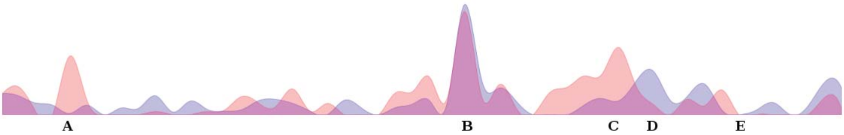


Figure 1: Volume of quotations for each word from a fragment of the 2010 State of the Union Address split by political leaning: conservative outlets shown in red and liberal outlets shown in blue. Quotes from the marked positions are reproduced in Table 1 and shown in the QUOTUS visualization in Figure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

Exercise

Exercise

1. Why don't the authors simply study how media outlets differ in terms of the stories they report?

Exercise

1. Why don't the authors simply study how media outlets differ in terms of the stories they report? Wouldn't that tell us something about ideological bias?

Exercise

1. Why don't the authors simply study how media outlets differ in terms of the stories they report? Wouldn't that tell us something about ideological bias?

2. In US politics, how (substantively) do you expect ideological bias to manifest itself in terms of reporting Biden's (current) speeches? (assuming he is not quoted out of context)

Exercise

1. Why don't the authors simply study how media outlets differ in terms of the stories they report? Wouldn't that tell us something about ideological bias?

2. In US politics, how (substantively) do you expect ideological bias to manifest itself in terms of reporting Biden's (current) speeches? (assuming he is not quoted out of context)

Questions and Answers

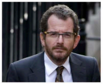
Questions and Answers

What do you say is permissible and **what do** you say is impermissible on this issue [of boundaries for privacy]?

J.K. Rowling, phone hacking victim

… it is not normal for anyone, famous or not famous, for their address to be known to millions of newspaper readers or users of the Internet. So **that is where I would draw the line**. …

Mr. Jay, lead counsel



… I paraphrase: he [the bin-man] turned up with sackfuls of Elton [John]'s documents, including the bank statements. **Did you have any qualms** about that, Mr Morgan?

Piers Morgan, tabloid editor

Yes, slightly. I mean, it clearly is, you know, a strange thing to be doing. … **Did I think he was doing anything illegal?** No. **Did I think it was on the cusp of unethical?** Yes.

… **Would you agree** that the magnitude of legal risk to a company is merely a function of the magnitude of **ethical misbehaviour** within a company?

Rupert Murdoch, tabloid owner

No. Clearly it may be. Serious breaches of the law are certainly unethical, but I think **I can think of other unethical things** which I would call unethical and extremely serious, but -- **which are not criminal**.

Questions and Answers

Questions and Answers

We want to understand the space of questions: some are accusatory, some seek information etc.

Questions and Answers

We want to understand the space of questions: some are accusatory, some seek information etc.

But we don't have labels for the questions:

Questions and Answers

We want to understand the space of questions: some are accusatory, some seek information etc.

But we don't have labels for the questions: may be we can use the answers.

Questions and Answers

We want to understand the space of questions: some are accusatory, some seek information etc.

But we don't have labels for the questions: may be we can use the answers.

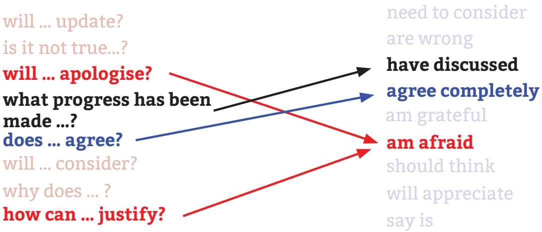
Questions with similar rhetorical functions will map to similar answers. . .

Questions and Answers

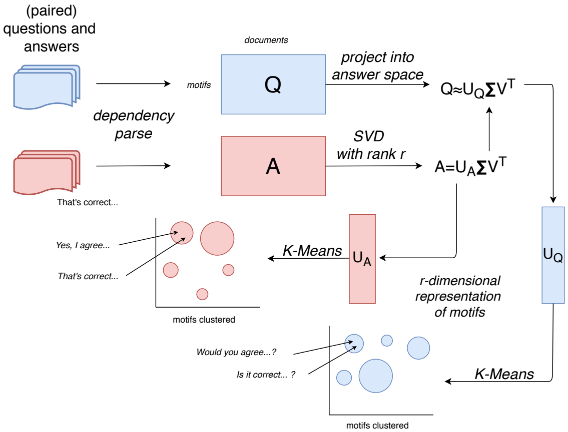
We want to understand the space of questions: some are accusatory, some seek information etc.

But we don't have labels for the questions: may be we can use the answers.

Questions with similar rhetorical functions will map to similar answers. . .



Map Questions into Answer Space



Zhang et al "Asking Too Much", 2018

Types of Questions that Prime Ministers face. . .

Zhang et al "Asking Too Much", 2018

Types of Questions that Prime Ministers face...

0. Issue update
1. Shared concerns
2. Narrow factual
3. Prompt for comment
4. Agreement
5. Self promotion
6. Concede & accept
7. Condemnatory

Zhang et al "Asking Too Much", 2018

Types of Questions that Prime Ministers face. . .

0. Issue update

Will you update us on Afghanistan?

1. Shared concerns

2. Narrow factual

What funding will be given to the College of Dentistry?

3. Prompt for comment
4. Agreement
5. Self promotion
6. Concede & accept
7. Condemnatory

Zhang et al "Asking Too Much", 2018

Types of Questions that Prime Ministers face...

0. Issue update
1. Shared concerns
2. Narrow factual
3. Prompt for comment
4. Agreement
5. Self promotion
6. **Concede & accept**
7. **Condemnatory**

Will you accept that [Brexit] will undermine our security?

Will you apologise for leading us into the Iraq disaster?

Huang (2021) on Leveson

Huang (2021) on Leveson

Considers various technical improvements to framework,

Huang (2021) on Leveson

Considers various technical improvements to framework, especially in terms of model selection.

Huang (2021) on Leveson

Considers various technical improvements to framework, especially in terms of model selection.

Looks at respondent types in Leveson:

Huang (2021) on Leveson

Considers various technical improvements to framework, especially in terms of model selection.

Looks at respondent types in Leveson:

Utterance type	Victim (compared to media, police, govt.)
Receive deferential question	**More** likely
Receive leading question	*Less* likely
Receive formal or hypothetical question	*Less* likely
Give evasive answer	*Less* likely