Entropy and Information-Theoretic Methods for Text Analysis

Ryan J. Gallagher
@ryanjgallag

netsi
Northeastern University
Network Science Institute

How do we quantify the diversity of language and topics expressed in #BlackLivesMatter and #AllLivesMatter tweets?

Entropy 😱

How do we measure to what extent an individual has adopted the language of a subreddit like r/TheRedPill?

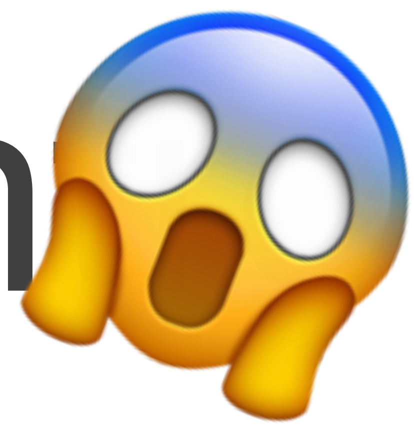
😱 Entropy 😱

How do we measure how particular words and phrases drove the divergence in language between #BlackLivesMatter and #AllLivesMatter?

Entropy

How do we extract topical frames around a particular word or set of words with no assumptions on how those frames were generated?

Entropy

Enpy

What is entropy?

If we wanted to measure surprise, what properties would we want from that measure?

What is entropy?

If we wanted to measure surprise, what properties would we want from that measure?

1. **Continuous:** we want continuity because jumps are 😨😰😭

What is entropy?

If we wanted to measure surprise, what properties would we want from that measure?

1. **Continuous:** we want continuity because jumps are 😨😰😭

2. **Additive:** we want to be able to add up how surprising events are ➕

What is entropy?

If we wanted to measure surprise, what properties would we want from that measure?

1. **Continuous:** we want continuity because jumps are 😨😰😭

2. **Additive:** we want to be able to add up how surprising events are ➕

3. **Symmetric:** we want to be able to add up events in any order

What is entropy?

If we wanted to measure surprise, what properties would we want from that measure?

1. **Continuous:** we want continuity because jumps are 😨😰😭

2. **Additive:** we want to be able to add up how surprising events are ➕

3. **Symmetric:** we want to be able to add up events in any order

4. **Maximal:** we want the surprise of a collection of events to be at its maximum when all of the events are equally likely 🏔️

What is entropy?

If we wanted to measure surprise, what properties would we want from that measure?

1. **Continuous:** we want continuity because jumps are 😨😰😭

2. **Additive:** we want to be able to add up how surprising events are ➕

3. **Symmetric:** we want to be able to add up events in any order

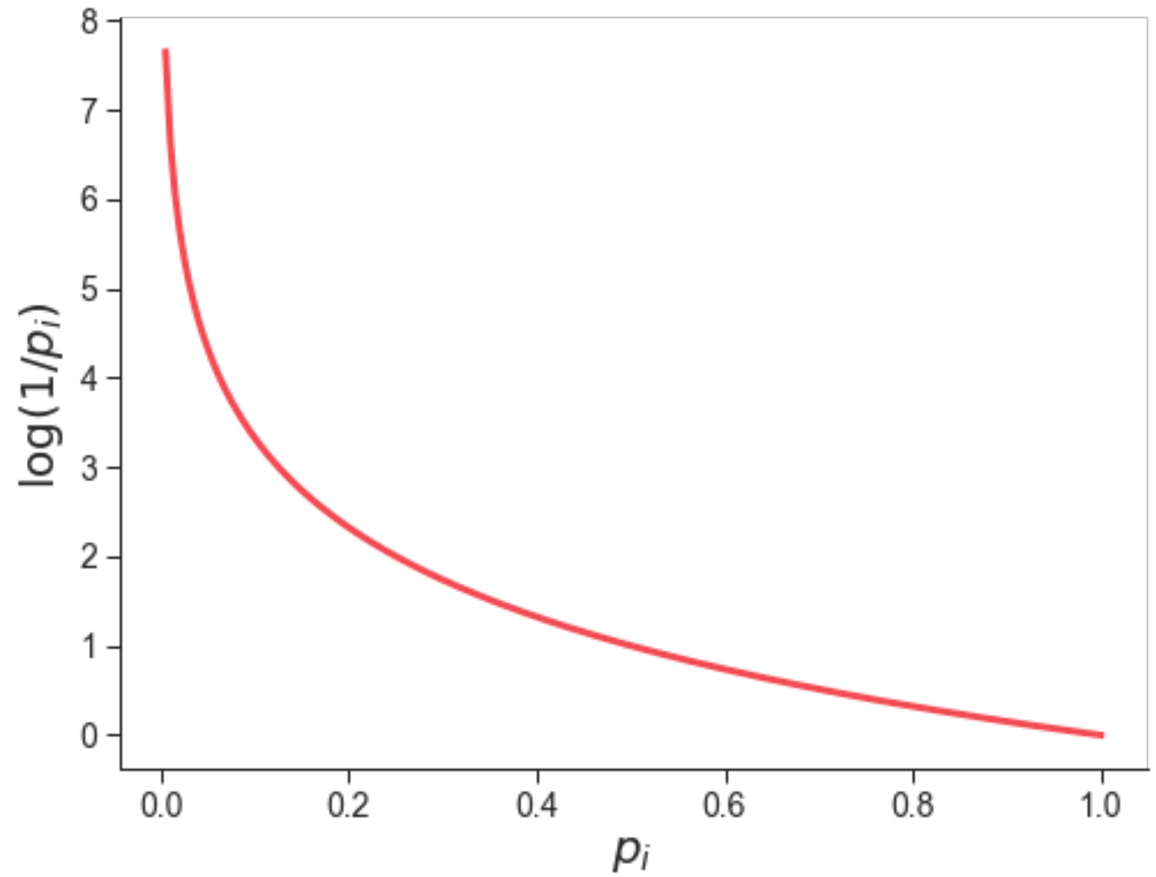
4. **Maximal:** we want the surprise of a collection of events to be at its maximum when all of the events are equally likely ⛰️

5. **Minimal:** we want the surprise of a collection of events to be at its minimum when only one event can occur ☝🏼

What is entropy?

There is only one function that satisfies all of these properties:

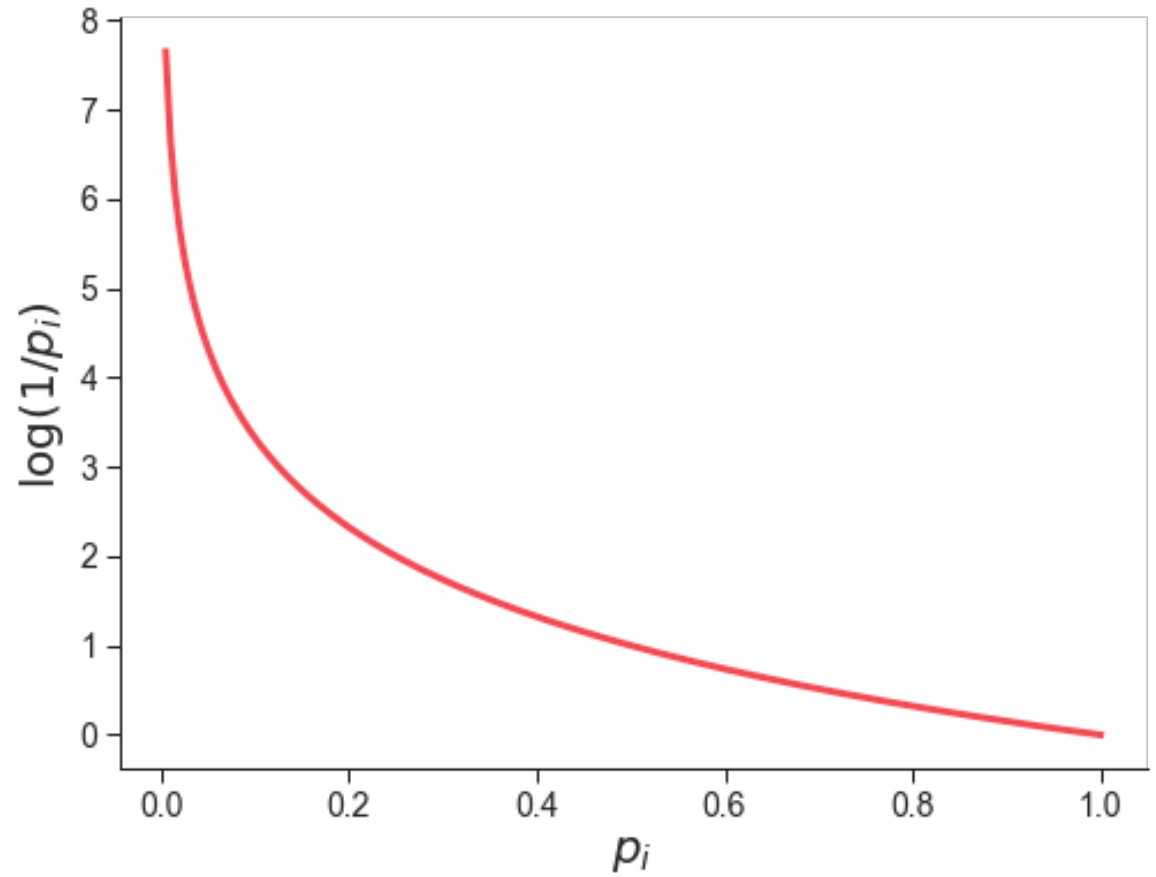
$$\log \frac{1}{p_i}$$



What is entropy?

There is only one function that satisfies all of these properties:

$$\log \frac{1}{p_i}$$



Entropy is the average surprise across a collection of events

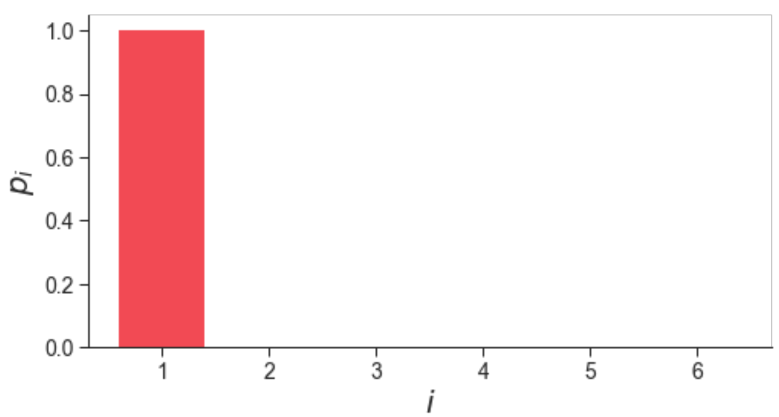
$$H(P) = -\sum_{i=1}^{n} p_i \log p_i$$

😱

Entropy as Diversity

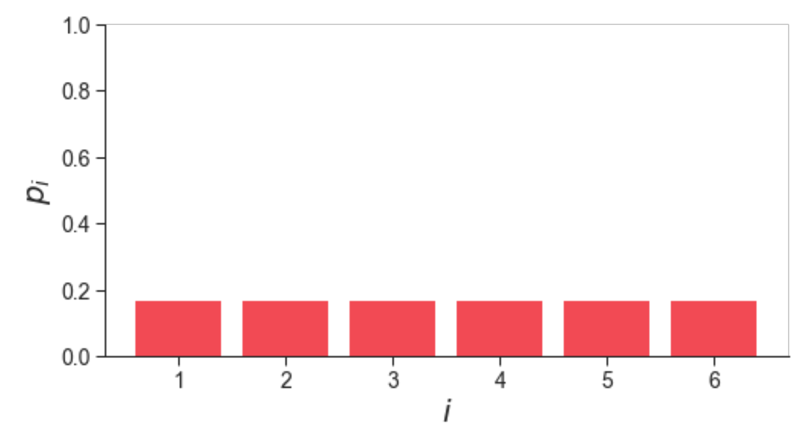
Entropy is at its minimum if the same event is always guaranteed to occur

$$H(P) = 1 \cdot \log 1 = 0$$



Entropy is at its maximum if it is equally likely any given event occurs

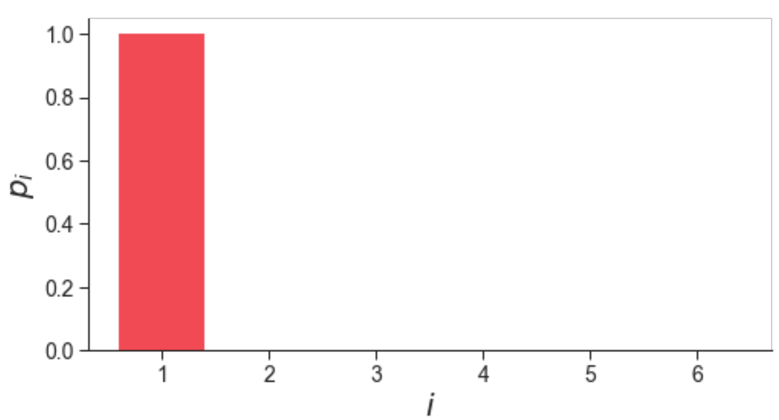
$$H(P) = \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n}$$



Entropy as Diversity

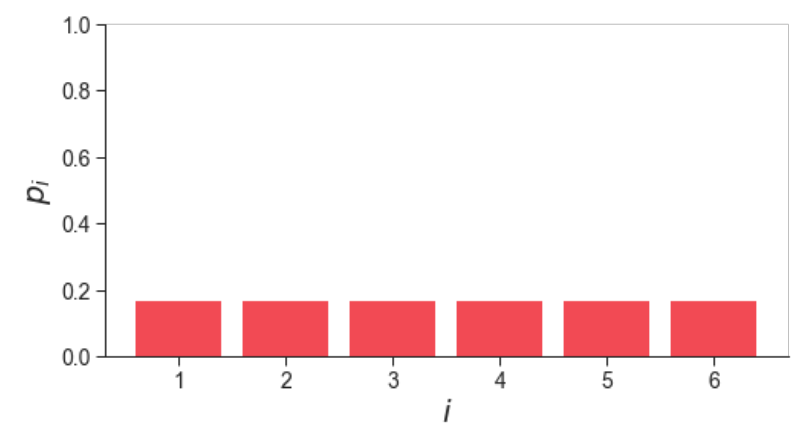
Entropy is at its minimum if the same event is always guaranteed to occur

$$H(P) = 1 \cdot \log 1 = 0$$

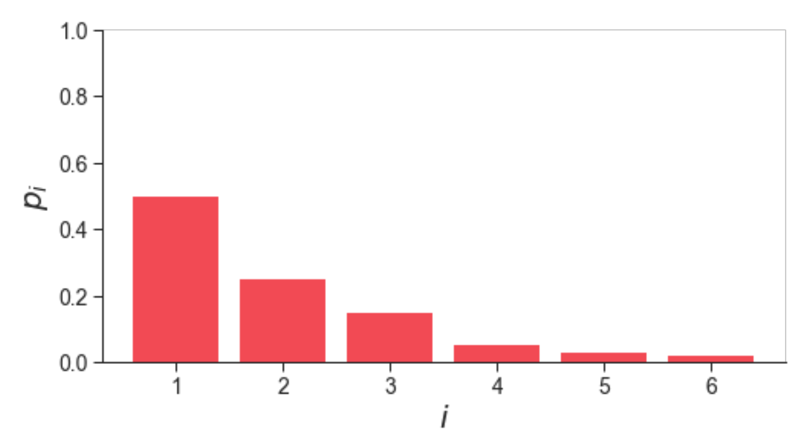


Entropy is at its maximum if it is equally likely any given event occurs

$$H(P) = \sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n}$$



We can think of entropy as the “skew” or “diversity” of a system



Entropy as Diversity

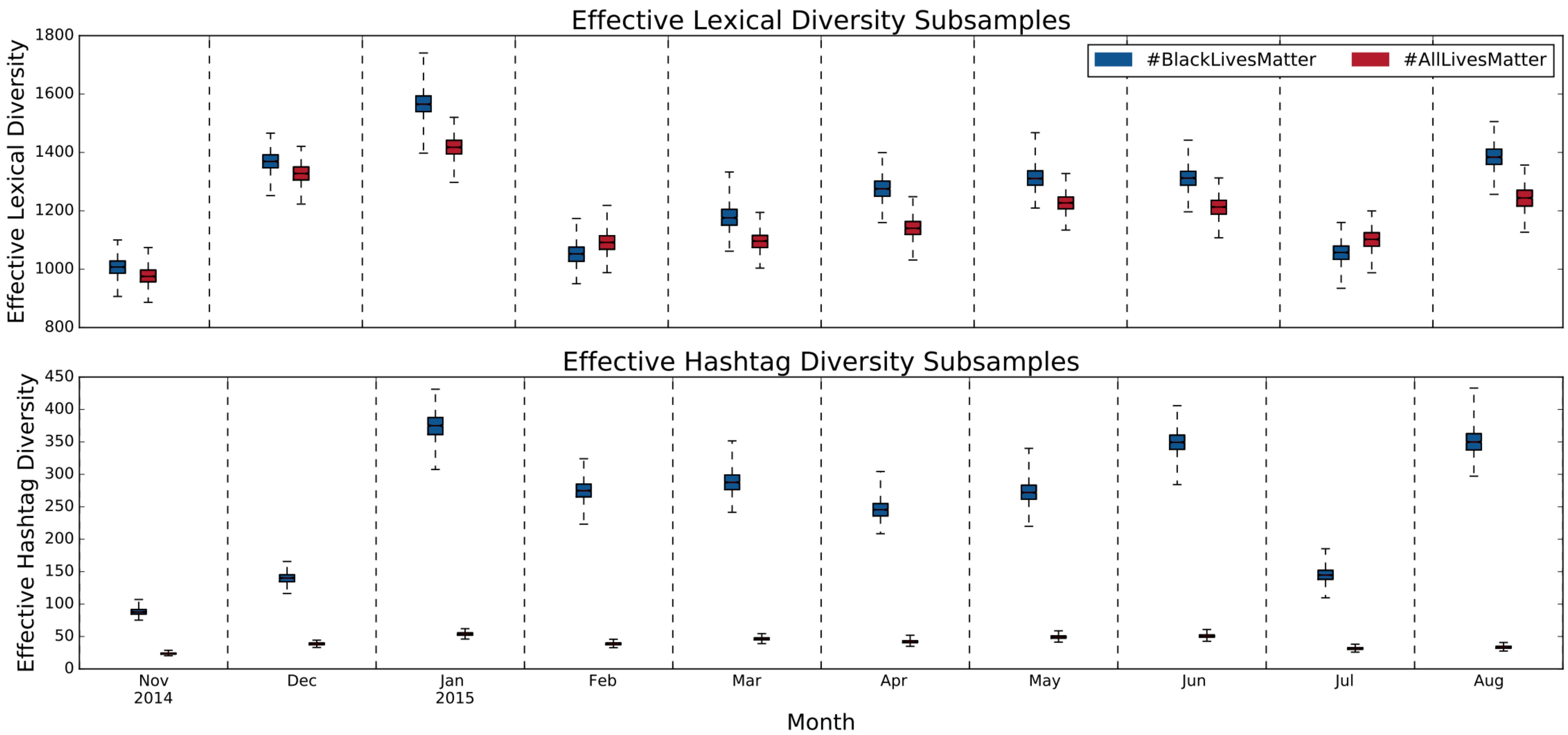
Suppose we have #BlackLivesMatter and #AllLivesMatter tweets and we are interested in understanding how diverse the language is within each hashtag, controlling for the number of tweets in each hashtag

1. Subsample the same number of tweets from both hashtags

2. Calculate the probability distribution over words and hashtags

3. Calculate the entropy of the distributions

4. Repeat steps 1—3 to bootstrap a distribution of language diversities



"Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter." Gallagher et al. *PLoS ONE*, 2018.

Divergence of Distributions

We can also compare two distributions by looking at their *divergence*

$$D_{KL}(P\,||\,Q) = \sum_{i=1}^{n} p_i \left(\log \frac{1}{q_i} - \log \frac{1}{p_i} \right)$$

Divergence of Distributions

We can also compare two distributions by looking at their *divergence*

$$D_{KL}(P\,||\,Q) = \sum_{i=1}^{n} p_i \left(\log \frac{1}{q_i} - \log \frac{1}{p_i} \right)$$

Kullback-Leibler 🧐

Divergence of Distributions

We can also compare two distributions by looking at their *divergence*

$$D_{KL}(\underline{P\,||\,Q}) = \sum_{i=1}^{n} p_i \left(\log \frac{1}{q_i} - \log \frac{1}{p_i} \right)$$

Divergence of distribution Q from distribution P

Divergence of Distributions

We can also compare two distributions by looking at their *divergence*

$$\underbrace{D_{KL}(P\,||\,Q)}_{\text{Divergence of distribution Q from distribution P}} = \sum_{i=1}^{n} p_i \underbrace{\left(\log \frac{1}{q_i} - \log \frac{1}{p_i} \right)}_{\text{How much the surprise of event } i \text{ in Q differs from the surprise in P}}$$

Divergence of Distributions

We can also compare two distributions by looking at their *divergence*

$$\underbrace{D_{KL}(P\,||\,Q)}_{\text{Divergence of distribution Q from distribution P}} = \sum_{i=1}^{n} p_i \underbrace{\left(\log \frac{1}{q_i} - \log \frac{1}{p_i} \right)}_{\text{How much the surprise of event } i \text{ in Q differs from the surprise in P}}$$

For example, *i* could represent the *i*th word in a corpus, where q_i is the probability of seeing that word in text *Q* and p_i is the probability of seeing that word in text *P*

Divergence of Language

Suppose we have an online community and we are interested in how the language of users varies over time with respect to the community

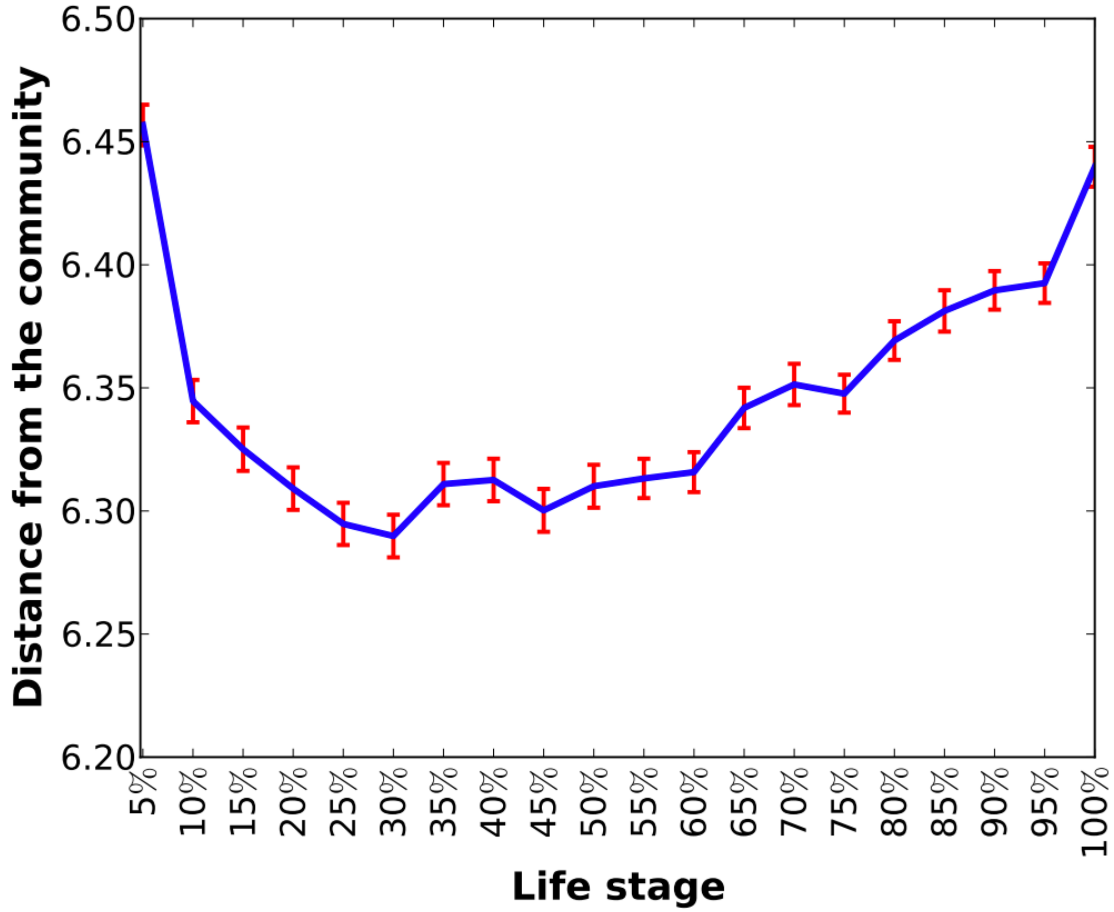
1. Calculate the distribution of language of the community

2. Build a language model over the community, which assigns a probability to seeing a given word or phrase

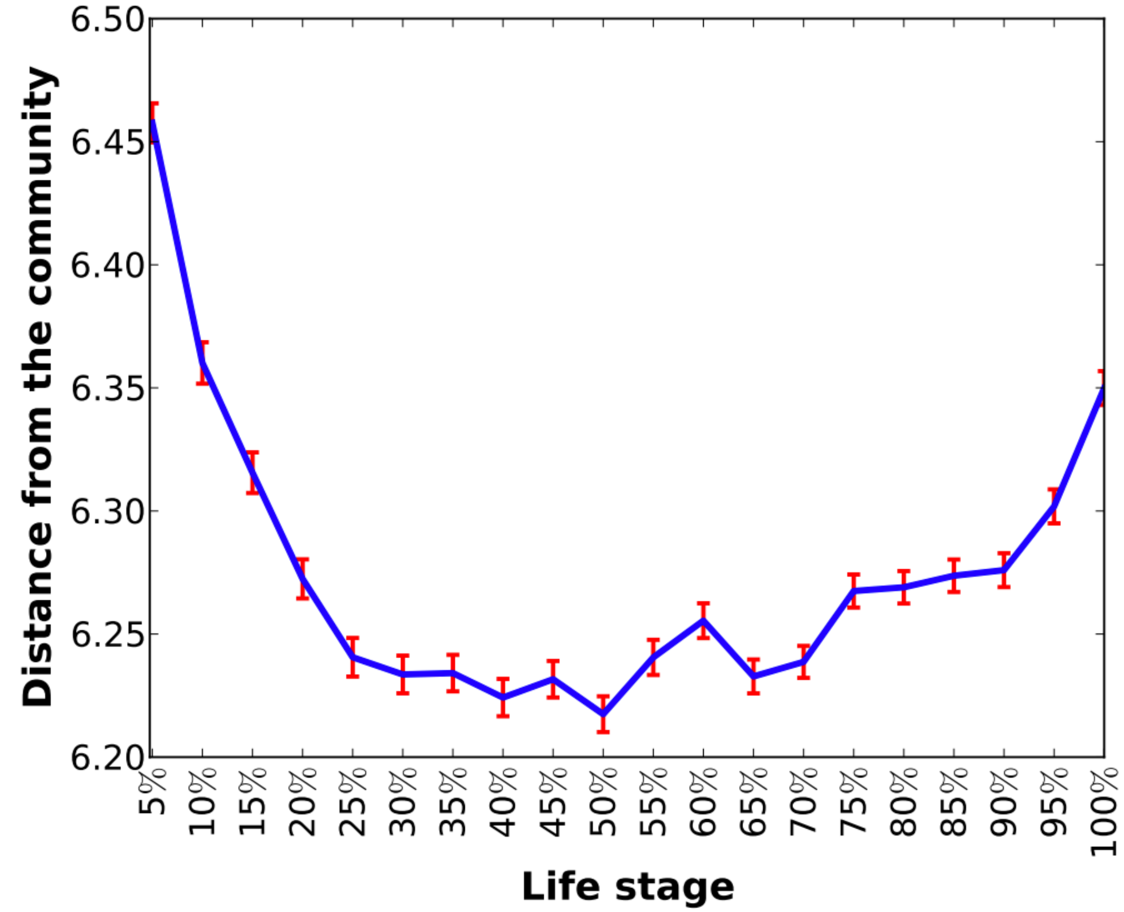
3. Apply the language model to a user's language to calculate their distribution of language

4. Calculate the KL divergence of the user's language from the community's language

5. Average this divergence across users as a function of user "lifetime"



(a) BeerAdvocate



(b) RateBeer

“No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities.” Danescu-Niculescu-Mizil et al. *WWW*, 2013.

Jensen-Shannon Divergence

The Kullback-Leibler divergence is not an ideal measure for text analysis

$$D_{KL}(P\,||\,Q) = \sum_{i=1}^{n} p_i \left(\log \frac{1}{q_i} - \log \frac{1}{p_i} \right)$$

Jensen-Shannon Divergence

The Kullback-Leibler divergence is not an ideal measure for text analysis

$$D_{KL}(P\,||\,Q) = \sum_{i=1}^{n} p_i \left(\underbrace{\log \frac{1}{q_i} - \log \frac{1}{p_i}}_{} \right)$$

If a single word in P is not in Q (i.e. $q_i = 0$), then the log explodes 💣💥

Jensen-Shannon Divergence

The Kullback-Leibler divergence is not an ideal measure for text analysis

$$D_{KL}(P\,||\,Q) = \sum_{i=1}^{n} p_i \left(\underline{\log \frac{1}{q_i} - \log \frac{1}{p_i}} \right)$$

If a single word in P is not in Q (i.e. $q_i = 0$), then the log explodes 💣💥

Instead, we can look at the Jensen-Shannon divergence

$$D_{JS}(P\,||\,Q) = \frac{1}{2} D_{KL}(P\,||\,M) + \frac{1}{2} D_{KL}(Q\,||\,M)$$

Jensen-Shannon Divergence

The Kullback-Leibler divergence is not an ideal measure for text analysis

$$D_{KL}(P\,||\,Q) = \sum_{i=1}^{n} p_i \left(\underline{\log \frac{1}{q_i} - \log \frac{1}{p_i}} \right)$$

If a single word in P is not in Q (i.e. $q_i = 0$), then the log explodes 💣💥

Instead, we can look at the Jensen-Shannon divergence

$$D_{JS}(P\,||\,Q) = \frac{1}{2} D_{KL}(P\,||\,\underline{M}) + \frac{1}{2} D_{KL}(Q\,||\,M)$$

Mixed distribution
$M = 1/2\, P + 1/2\, Q$

Jensen-Shannon Divergence

The Kullback-Leibler divergence is not an ideal measure for text analysis

$$D_{KL}(P\,||\,Q) = \sum_{i=1}^{n} p_i \underbrace{\left(\log \frac{1}{q_i} - \log \frac{1}{p_i} \right)}$$

If a single word in P is not in Q (i.e. $q_i = 0$), then the log explodes 💣💥

Instead, we can look at the Jensen-Shannon divergence

$$D_{JS}(P\,||\,Q) = \frac{1}{2} D_{KL}(P\,||\,M) + \frac{1}{2} D_{KL}(Q\,||\,M)$$

Advantages:

1. Does not explode (unlike KL divergence)

2. Symmetric (unlike KL divergence)

3. 0 if both texts are exactly the same, 1 if they have no words in common

Interpretability of JSD

The Jensen-Shannon divergence is interpretable at the word level

$$D_{JS}(P\,||\,Q) = \frac{1}{2}D_{KL}(P\,||\,M) + \frac{1}{2}D_{KL}(Q\,||\,M)$$

Interpretability of JSD

The Jensen-Shannon divergence is interpretable at the word level

$$D_{JS}(P\,||\,Q) = \frac{1}{2}D_{KL}(P\,||\,M) + \frac{1}{2}D_{KL}(Q\,||\,M)$$

$$= \sum_{i=1}^{n} \underline{-\,m_i \log m_i + \frac{1}{2}\left(p_i \log p_i + q_i \log q_i\right)}$$

Contribution of each word to the divergence

Interpretability of JSD

The Jensen-Shannon divergence is interpretable at the word level

$$D_{JS}(P\,||\,Q) = \frac{1}{2}D_{KL}(P\,||\,M) + \frac{1}{2}D_{KL}(Q\,||\,M)$$

$$= \sum_{i=1}^{n} \underbrace{-\,m_i \log m_i + \frac{1}{2}\left(p_i \log p_i + q_i \log q_i\right)}_{\text{Contribution of each word to the divergence}}$$

This means that we can not only measure *how much* two texts diverge,
but also *why* two texts diverge

Interpretability of JSD

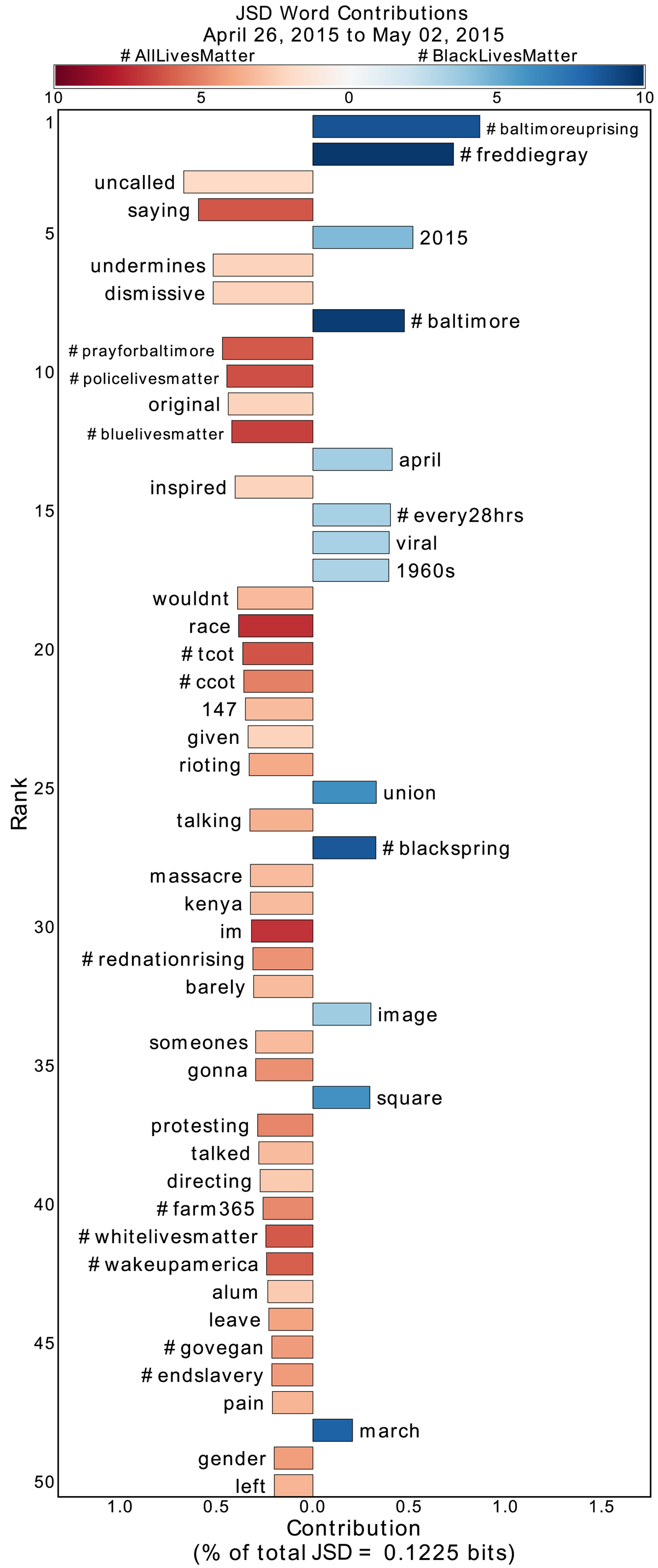
Suppose we want to understand in what ways #BlackLivesMatter and #AllLivesMatter diverged in terms of their language during the Baltimore protests following the death of Freddie Gray

1. Calculate the word distributions of #BlackLivesMatter and #AllLivesMatter

2. Calculate the Jensen-Shannon divergence between the distributions

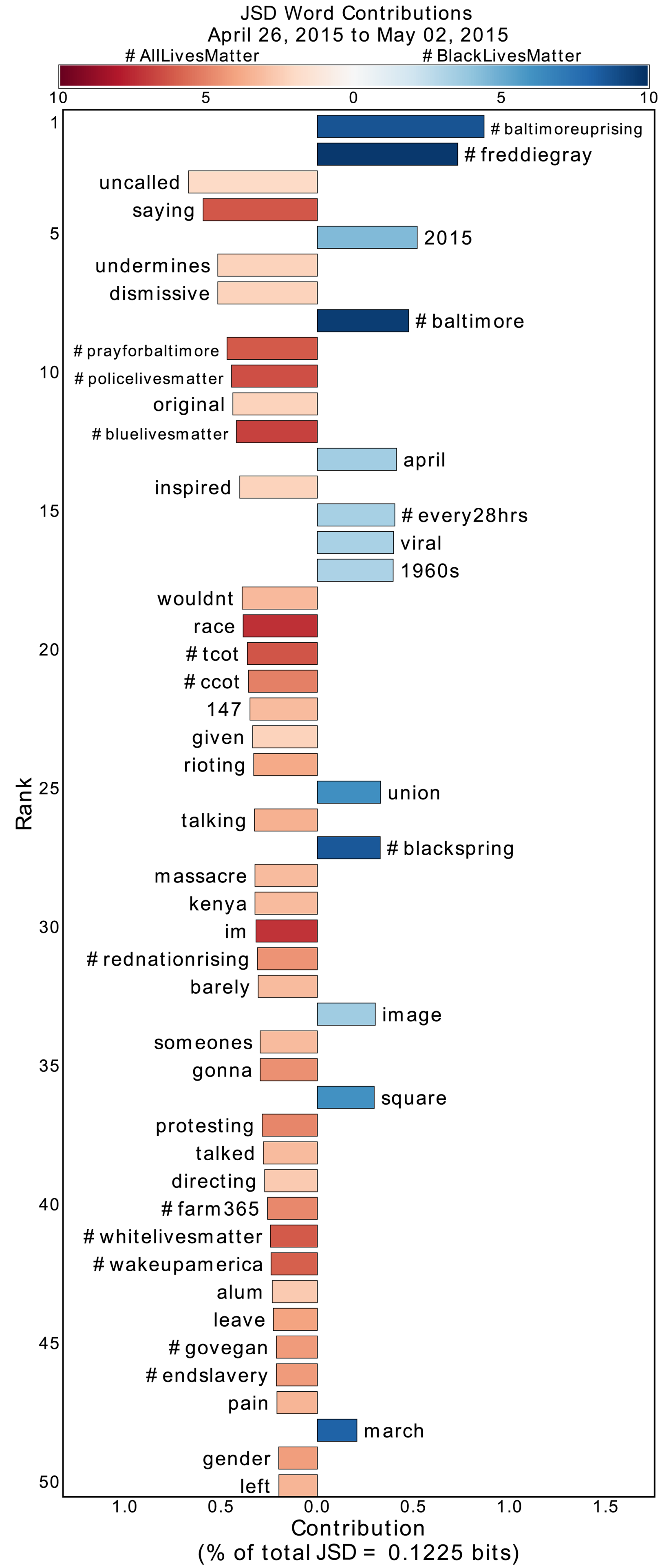
3. Calculate and attribute each word's contribution to the divergence

4. Provide context to the contribution by calculating the diversity of language around each word



"Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter." Gallagher et al. *PLoS ONE*, 2018.

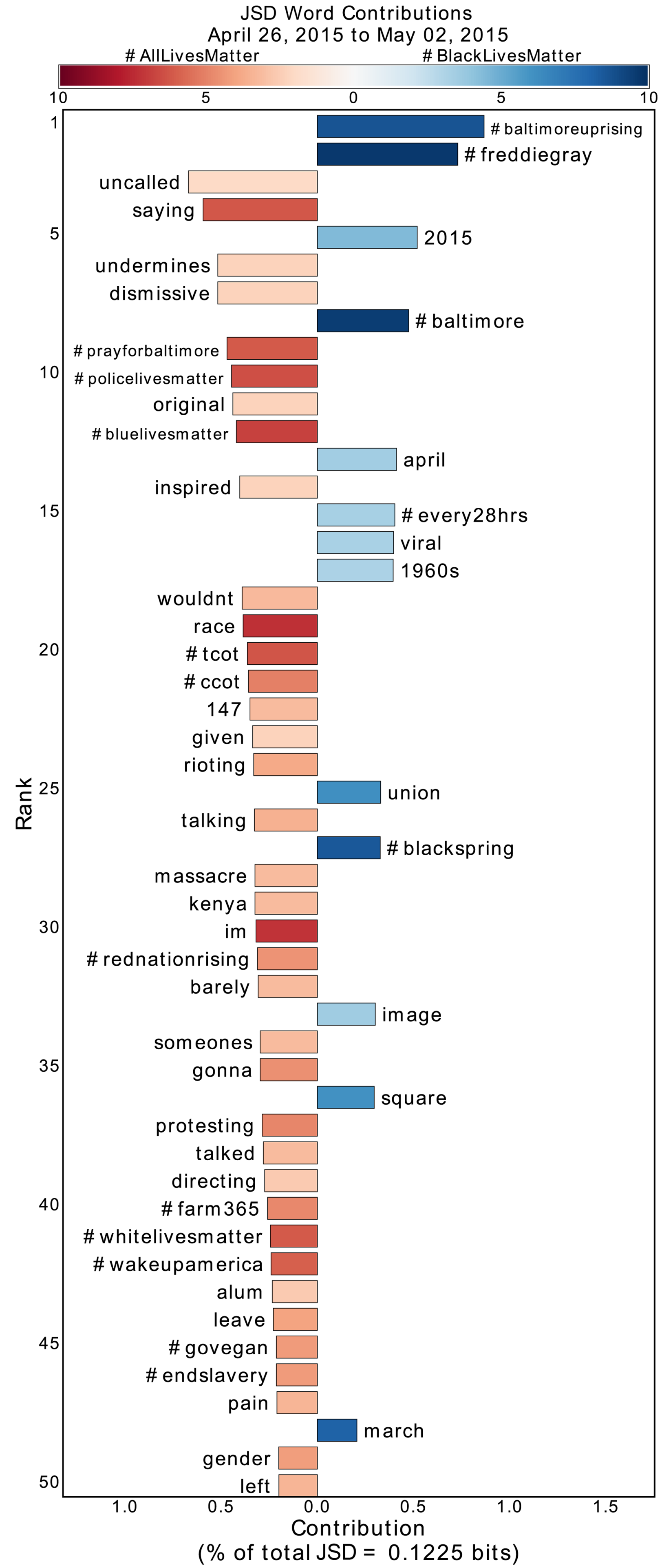
Length of bar indicates the contribution to the divergence



"Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter." Gallagher et al. *PLoS ONE*, 2018.

Length of bar indicates the contribution to the divergence

Word is more frequent in #AllLivesMatter

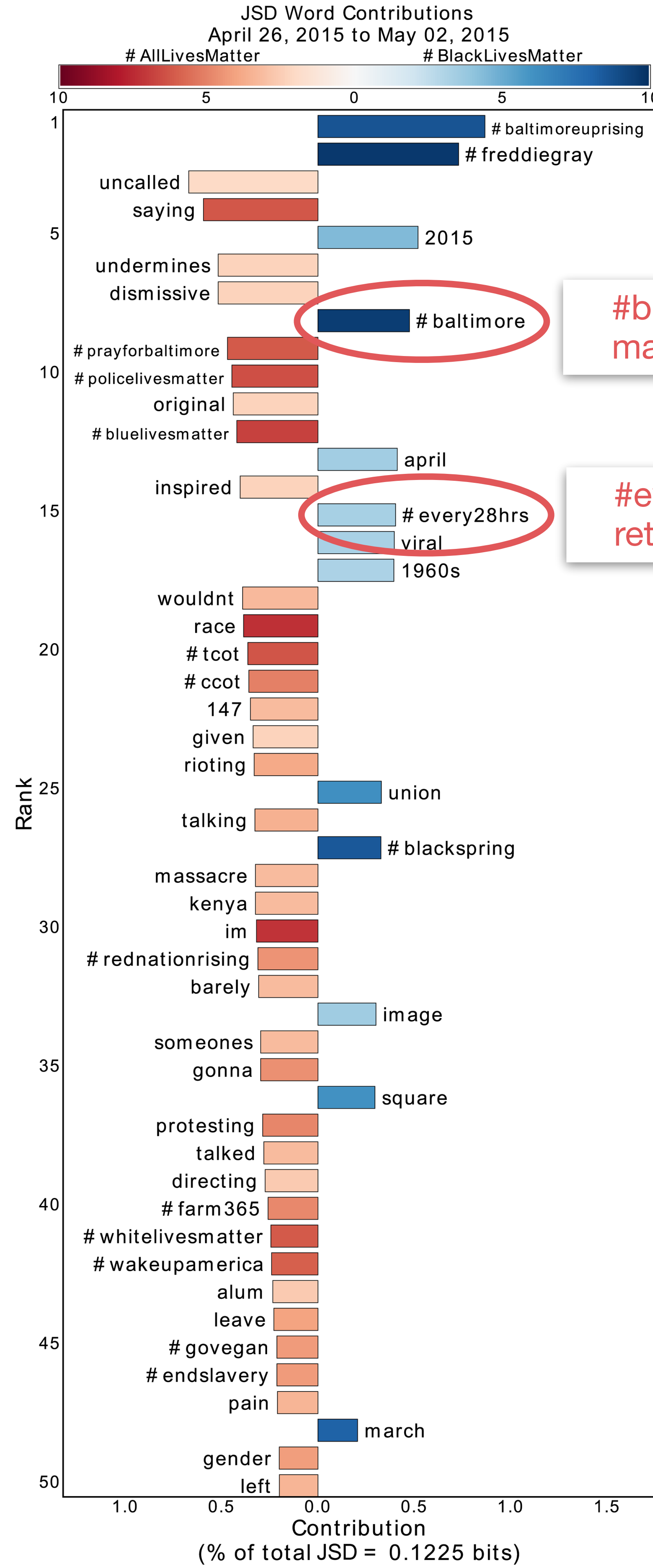


Word is more frequent in #BlackLivesMatter

"Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter." Gallagher et al. *PLoS ONE*, 2018.

Color indicates the diversity of language used around a particular word (darker = more diversity)

Word is more frequent in #AllLivesMatter

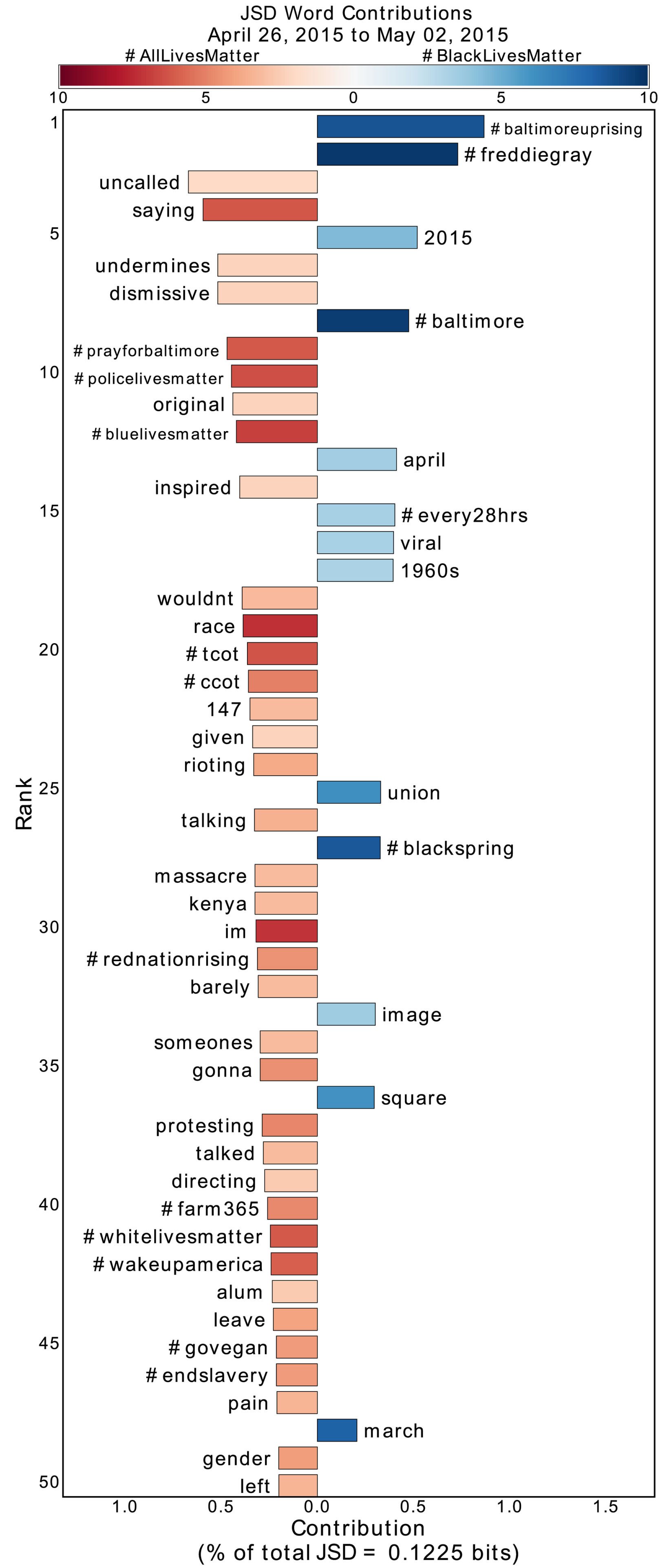


#baltimore was used in many different tweets

#every28hours was primarily used in retweets of a single tweet

Word is more frequent in #BlackLivesMatter

"Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter." Gallagher et al. *PLoS ONE*, 2018.



"Divergent Discourse Between Protests and Counter-Protests: #BlackLivesMatter and #AllLivesMatter." Gallagher et al. *PLoS ONE*, 2018.

Information-Theoretic Topics (example)

Documents

$$d_1 \quad \boxed{x_1 \quad x_2} \quad (1,1,0,0,0)$$

$$d_2 \quad \boxed{x_3 \quad x_4} \quad (0,0,1,1,0)$$

Probability table

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	1/2	0
$X_1 = 1$	0	1/2

Information-Theoretic Topics (example)

Documents

$$d_1 \quad \boxed{x_1 \quad x_2} \quad (1,1,0,0,0)$$

$$d_2 \quad \boxed{x_3 \quad x_4} \quad (0,0,1,1,0)$$

Probability table

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	1/2	0
$X_1 = 1$	0	1/2

Words 1 and 2 are related:

$$I(X_1 : X_2) = D_{KL}\big(p(x_1, x_2) \,||\, p(x_1)p(x_2)\big) = 1 \text{ bit}$$

Information-Theoretic Topics (example)

Documents

d_1: $\boxed{x_1 \quad x_2}$ $(1, 1, 0, 0, 0)$

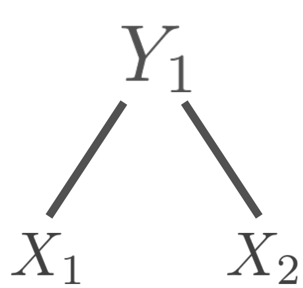
d_2: $\boxed{x_3 \quad x_4}$ $(0, 0, 1, 1, 0)$

Probability table

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	1/2	0
$X_1 = 1$	0	1/2

Words 1 and 2 are related:

$$I(X_1 : X_2) = D_{KL}\big(p(x_1, x_2) \,||\, p(x_1)p(x_2)\big) = 1 \text{ bit}$$



Hypothesize a latent factor: $Y_1 = X_1 = X_2$

Information-Theoretic Topics (example)

Documents

d_1: $\boxed{x_1 \quad x_2}$ $(1,1,0,0,0)$

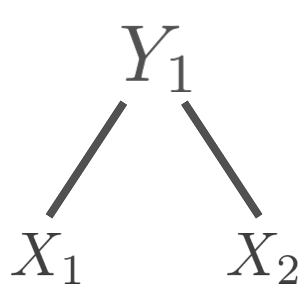
d_2: $\boxed{x_3 \quad x_4}$ $(0,0,1,1,0)$

Probability table

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	1/2	0
$X_1 = 1$	0	1/2

Words 1 and 2 are related:

$$I(X_1 : X_2) = D_{KL}\big(p(x_1, x_2) \,||\, p(x_1)p(x_2)\big) = 1 \text{ bit}$$



Hypothesize a latent factor: $Y_1 = X_1 = X_2$

Then conditioned on Y_1, words 1 and 2 are independent

$$D_{KL}\big(p(x_1, x_2 \mid y_1) \,||\, p(x_1 \mid y_1)p(x_2 \mid y_1)\big) = 0 \text{ bits}$$

Information-Theoretic Topics (example)

Documents

d_1: $x_1 \quad x_2$ — $(1, 1, 0, 0, 0)$

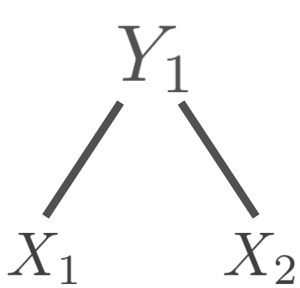
d_2: $x_3 \quad x_4$ — $(0, 0, 1, 1, 0)$

Probability table

	$X_2 = 0$	$X_2 = 1$
$X_1 = 0$	1/2	0
$X_1 = 1$	0	1/2

Words 1 and 2 are related:

$$I(X_1 : X_2) = D_{KL}\big(p(x_1, x_2) \,||\, p(x_1)p(x_2)\big) = 1 \text{ bit}$$



Hypothesize a latent factor: $Y_1 = X_1 = X_2$

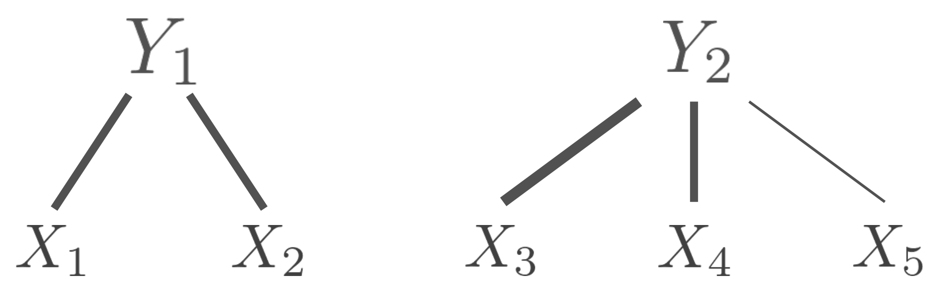
Then conditioned on Y_1, words 1 and 2 are independent

$$D_{KL}\big(p(x_1, x_2 \mid y_1) \,||\, p(x_1 \mid y_1)p(x_2 \mid y_1)\big) = 0 \text{ bits}$$

Goal: find latent factors (topics) that make words conditionally independent

CorEx Topic Model

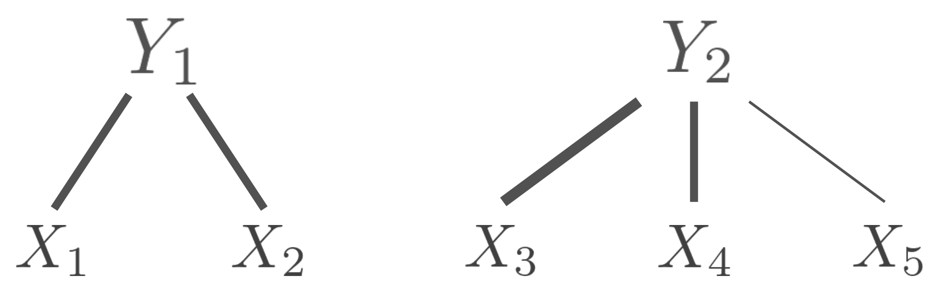
Goal: find latent factors (topics) that make words conditionally independent



$$\min_{Y} D_{KL}\left(p(x_1, x_2, \ldots x_n \mid y) \,||\, \prod_i p(x_i \mid y)\right)$$

CorEx Topic Model

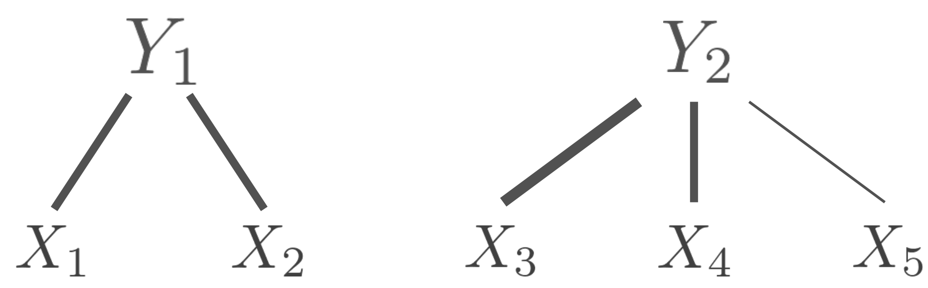
Goal: find latent factors (topics) that make words conditionally independent



$$\min_Y D_{KL}\left(p(x_1, x_2, \ldots x_n \mid y) \,||\, \prod_i p(x_i \mid y)\right) = \min_Y \underbrace{TC(X_1, X_2, \ldots, X_N \mid Y)}_{\text{Total correlation conditioned on } Y}$$

CorEx Topic Model

Goal: find latent factors (topics) that make words conditionally independent



$$\min_{Y} D_{KL}\left(p(x_1, x_2, \dots x_n \mid y) \,||\, \prod_i p(x_i \mid y) \right) = \min_{Y} TC(X_1, X_2, \dots, X_N \mid Y)$$

$TC(X \mid Y) = 0$ if and only if the topic “explains” all the dependencies (total correlation)

Hence, “Total **Cor**relation **Ex**planation” (CorEx)

“Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge.” Gallagher et al. *TACL*, 2017.

Why CorEx?

As a computational social scientist, why might you use CorEx instead of LDA?

"Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Gallagher et al. *TACL*, 2017.

Why CorEx?

As a computational social scientist, why might you use CorEx instead of LDA?

1. CorEx does not assume anything about the data generating process. There are no priors, the only parameter is the number of topics

"Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Gallagher et al. *TACL*, 2017.

Why CorEx?

As a computational social scientist, why might you use CorEx instead of LDA?

1. CorEx does not assume anything about the data generating process. There are no priors, the only parameter is the number of topics

2. There is a principled method for choosing the number of topics because each topic explains a certain amount of information

"Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Gallagher et al. *TACL*, 2017.

Why CorEx?

As a computational social scientist, why might you use CorEx instead of LDA?

1. CorEx does not assume anything about the data generating process. There are no priors, the only parameter is the number of topics

2. There is a principled method for choosing the number of topics because each topic explains a certain amount of information

3. CorEx allows the user to guide the topic model through *anchor words*

"Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Gallagher et al. *TACL*, 2017.

Why CorEx?

As a computational social scientist, why might you use CorEx instead of LDA?

1. CorEx does not assume anything about the data generating process. There are no priors, the only parameter is the number of topics

2. There is a principled method for choosing the number of topics because each topic explains a certain amount of information

3. CorEx allows the user to guide the topic model through *anchor words*

And for many other fun reasons under the hood that I can't fit in a 15 minute talk

"Anchored Correlation Explanation: Topic Modeling with Minimal Domain Knowledge." Gallagher et al. *TACL*, 2017.

Anchoring Strategies

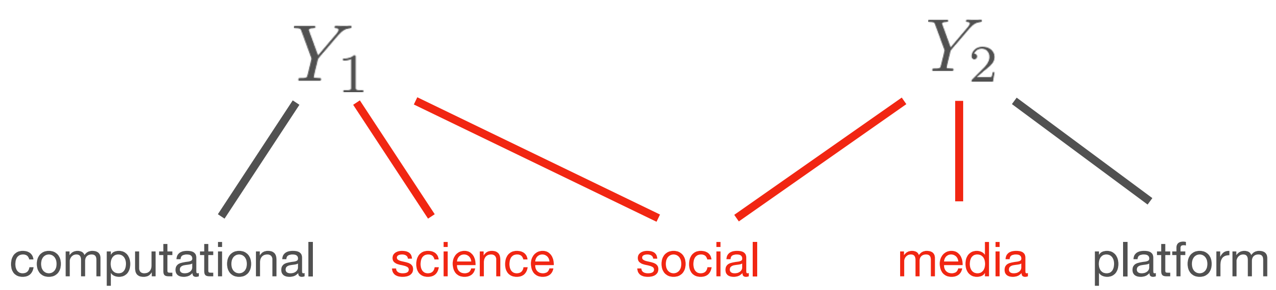
Topic Representation

Anchoring to unveil topics that do not naturally emerge

Anchoring Strategies

Topic Representation
Anchoring to unveil topics that do not naturally emerge

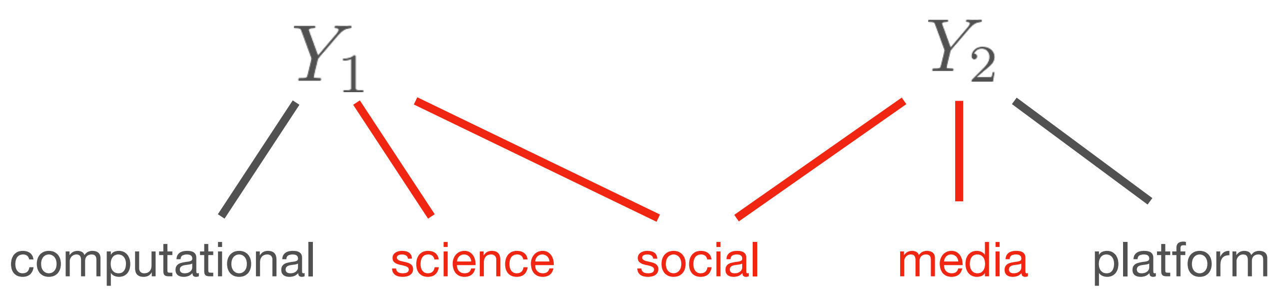
Topic Separability
Anchoring to help enforce separation between topics



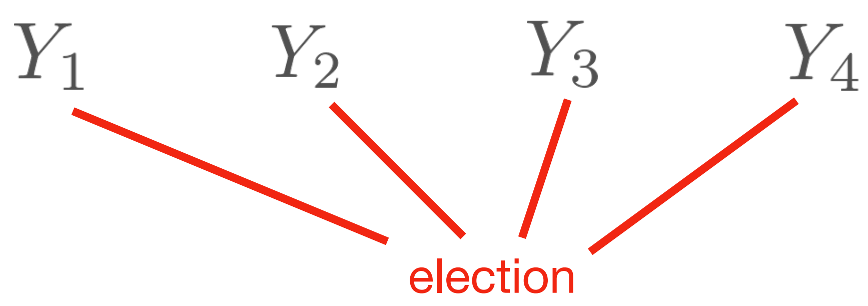
Anchoring Strategies

Topic Representation
Anchoring to unveil topics that do not naturally emerge

Topic Separability
Anchoring to help enforce separation between topics

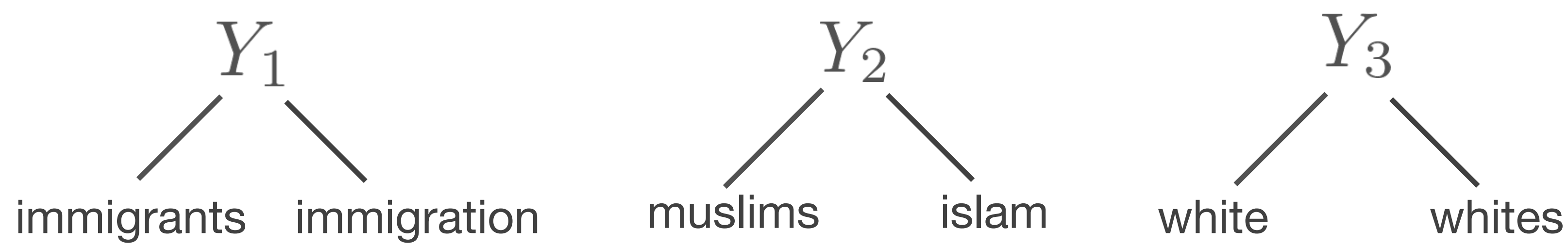


Topic Frames
Anchoring to disambiguate different frames around a word



Anchoring for Topic Representation

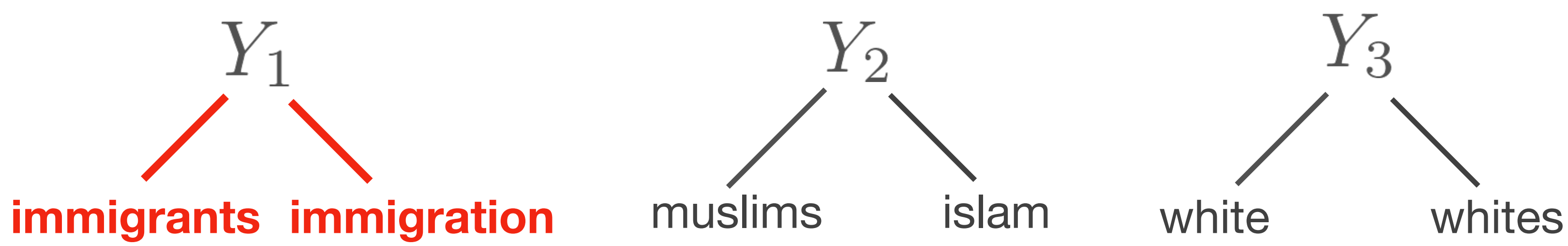
Data: news articles about the campaigns of Clinton and Trump, up to August 2016
Method: train one CorEx topic model for each corpus, anchor words for comparison



Work by Abigail Ross and the Computational Story Lab, University of Vermont

Anchoring for Topic Representation

Data: news articles about the campaigns of Clinton and Trump, up to August 2016
Method: train one CorEx topic model for each corpus, anchor words for comparison



Clinton Topic

1: **immigration**, **immigrants**, jobs, economic, trade, health, tax, wall, care, economy

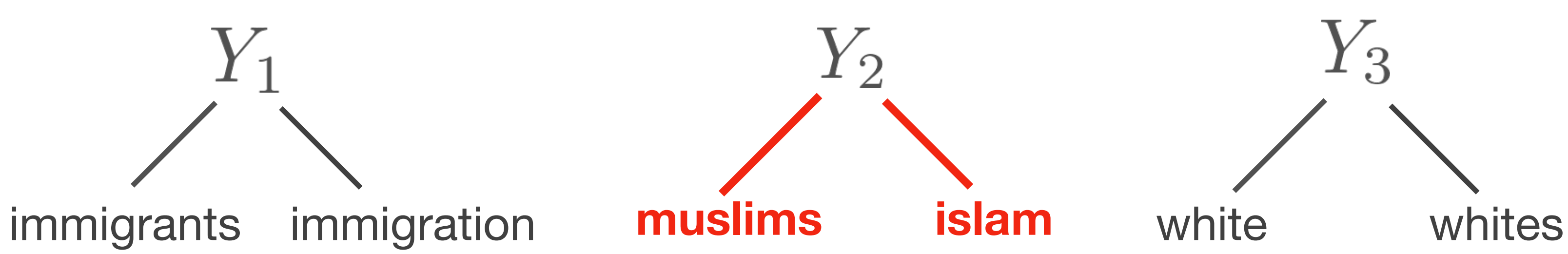
Trump Topic

1: **immigration**, **immigrants**, illegal, border, mexican, undocumented, rapists, mexico, wall, illegally

Work by Abigail Ross and the Computational Story Lab, University of Vermont

Anchoring for Topic Representation

Data: news articles about the campaigns of Clinton and Trump, up to August 2016
Method: train one CorEx topic model for each corpus, anchor words for comparison



Clinton Topic

2: muslims, **islam**, islamic, gun, terrorism, war, military, iraq, terrorist, syria

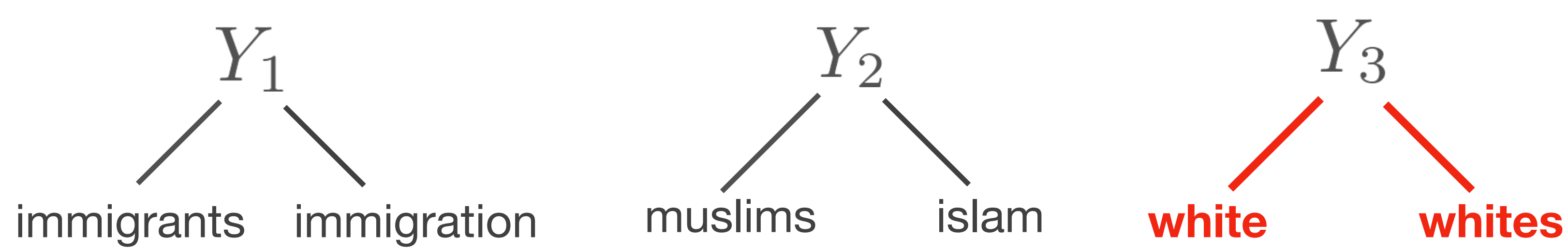
Trump Topic

2: muslims, **islam**, united, ban, entering, islamic, muslim, terrorism, terrorist, terrorists

Work by Abigail Ross and the Computational Story Lab, University of Vermont

Anchoring for Topic Representation

Data: news articles about the campaigns of Clinton and Trump, up to August 2016
Method: train one CorEx topic model for each corpus, anchor words for comparison



Clinton Topic

3: **white**, i, you, what, do, if, we, it’s, like, people

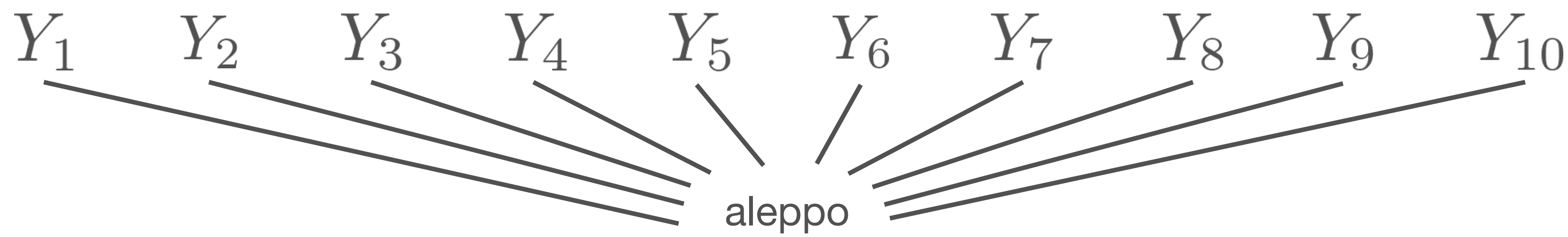
Trump Topic

3: **white**, house, **whites**, supremacists, supremacist, duke, klan, klux, ku, supremacy

Work by Abigail Ross and the Computational Story Lab, University of Vermont

Anchoring for Topic Frames

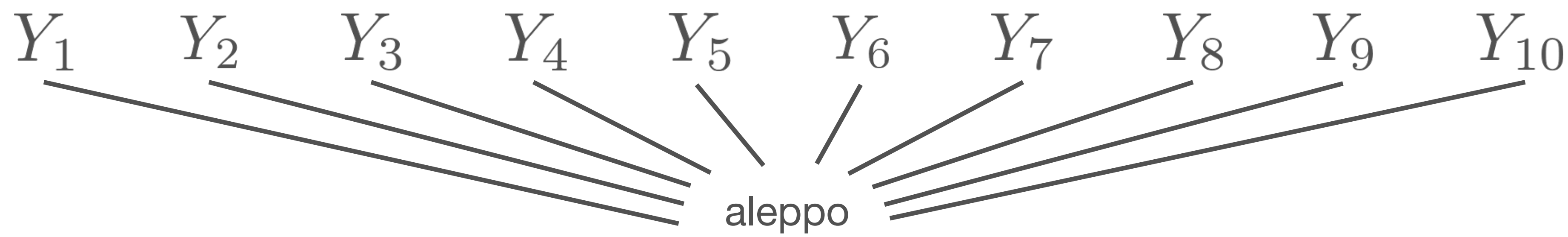
Data: ~1 million English newswire articles since June 2015 from countries in the Middle East



Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Anchoring for Topic Frames

Data: ~1 million English newswire articles since June 2015 from countries in the Middle East

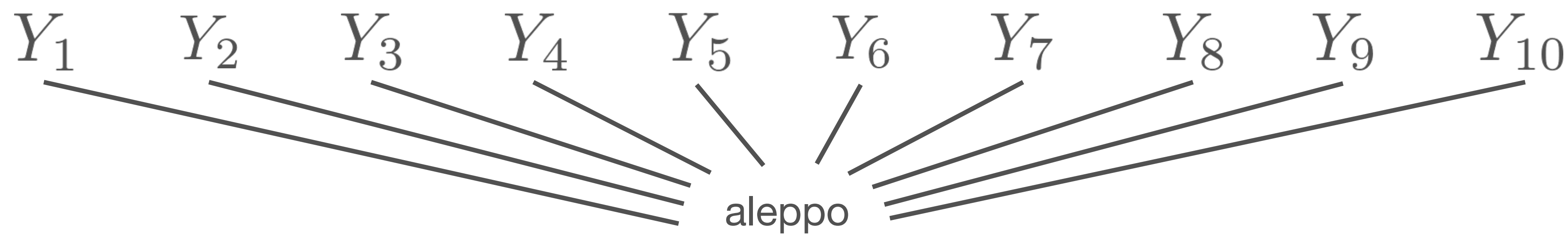


Note: this data broadly covers the Middle East and a priori we do not expect 10 topics to emerge about Aleppo

Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Anchoring for Topic Frames

Data: ~1 million English newswire articles since June 2015 from countries in the Middle East



1: **aleppo**, killed, police, security, attack, state, arrested, authorities

2: **aleppo**, forces, syria, military, war, army, civilians, iraq, militants

3: **aleppo**, health, medical, food, care, water, small, conditions, treatment, patients

4: country, **aleppo**, east, across, group, region, middle

5: two, **aleppo**, took, another, place, taking, leaders

6: **aleppo**, russia, iran, barack, obama, moscow, washington, putin

7: **aleppo**, political, court, part, accused, opposition, called, saying, parliament, democratic

8: government, **aleppo**, minister, foreign, states, united, prime, UN, law, nations

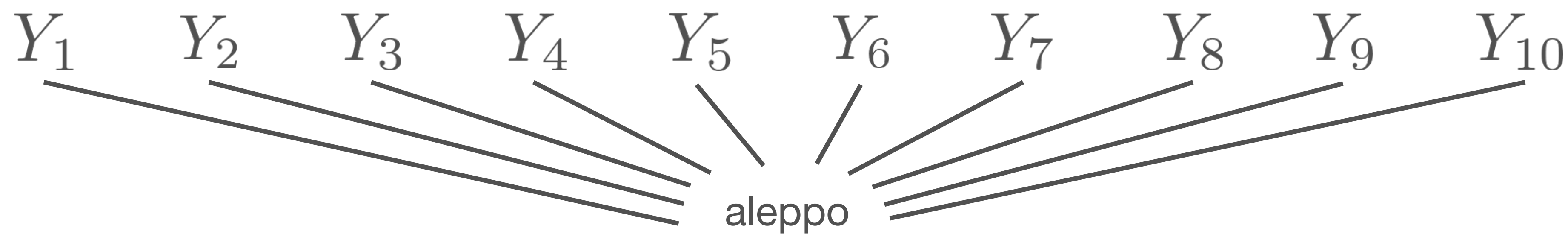
9: **aleppo**, city, area, near, air, northern, least, town, eastern, injured

10: **aleppo**, people, children, human, rights, women, social, school, society, lives

Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Anchoring for Topic Frames

Data: ~1 million English newswire articles since June 2015 from countries in the Middle East



1: **aleppo**, killed, police, security, attack, state, arrested, authorities

2: **aleppo**, forces, syria, military, war, army, civilians, iraq, militants

3: **aleppo**, health, medical, food, care, water, small, conditions, treatment, patients

4: country, **aleppo**, east, across, group, region, middle

5: two, **aleppo**, took, another, place, taking, leaders

6: **aleppo**, russia, iran, barack, obama, moscow, washington, putin

7: **aleppo**, political, court, part, accused, opposition, called, saying, parliament, democratic

8: government, **aleppo**, minister, foreign, states, united, prime, UN, law, nations

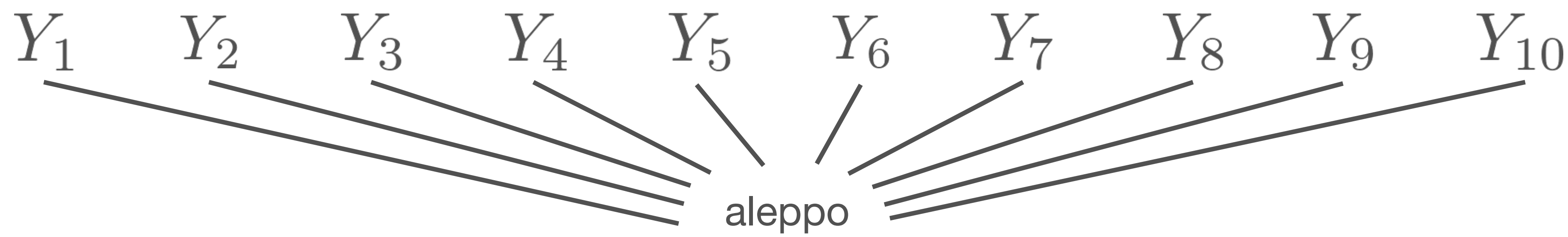
9: **aleppo**, city, area, near, air, northern, least, town, eastern, injured

10: **aleppo**, people, children, human, rights, women, social, school, society, lives

Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Anchoring for Topic Frames

Data: ~1 million English newswire articles since June 2015 from countries in the Middle East



1: aleppo, killed, police, security, attack, state, arrested, authorities

2: aleppo, forces, syria, military, war, army, civilians, iraq, militants

3: **aleppo**, health, medical, food, care, water, small, conditions, treatment, patients

4: country, aleppo, east, across, group, region, middle

5: two, aleppo, took, another, place, taking, leaders

6: aleppo, russia, iran, barack, obama, moscow, washington, putin

7: aleppo, political, court, part, accused, opposition, called, saying, parliament, democratic

8: government, aleppo, minister, foreign, states, united, prime, UN, law, nations

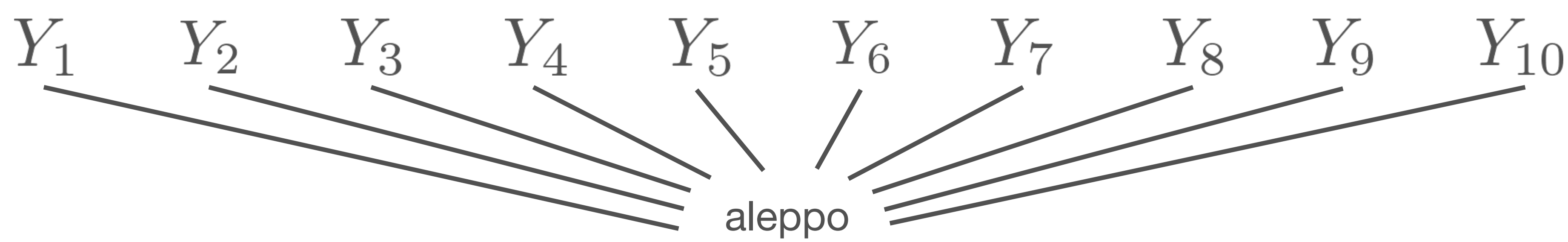
9: aleppo, city, area, near, air, northern, least, town, eastern, injured

10: aleppo, people, children, human, rights, women, social, school, society, lives

Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Anchoring for Topic Frames

Data: ~1 million English newswire articles since June 2015 from countries in the Middle East



1: aleppo, killed, police, security, attack, state, arrested, authorities

2: aleppo, forces, syria, military, war, army, civilians, iraq, militants

3: aleppo, health, medical, food, care, water, small, conditions, treatment, patients

4: country, **aleppo**, east, across, group, region, middle

5: two, aleppo, took, another, place, taking, leaders

6: aleppo, russia, iran, barack, obama, moscow, washington, putin

7: aleppo, political, court, part, accused, opposition, called, saying, parliament, democratic

8: government, aleppo, minister, foreign, states, united, prime, UN, law, nations

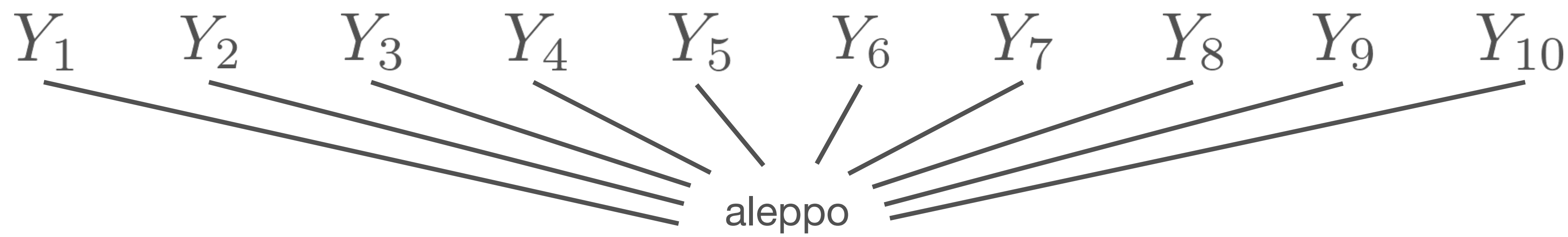
9: **aleppo**, city, area, near, air, northern, least, town, eastern, injured

10: aleppo, people, children, human, rights, women, social, school, society, lives

Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Anchoring for Topic Frames

Data: ~1 million English newswire articles since June 2015 from countries in the Middle East



1: **aleppo**, killed, police, security, attack, state, arrested, authorities

2: **aleppo**, forces, syria, military, war, army, civilians, iraq, militants

3: **aleppo**, health, medical, food, care, water, small, conditions, treatment, patients

4: country, **aleppo**, east, across, group, region, middle

5: two, **aleppo**, took, another, place, taking, leaders

6: **aleppo**, russia, iran, barack, obama, moscow, washington, putin

7: **aleppo**, political, court, part, accused, opposition, called, saying, parliament, democratic

8: government, **aleppo**, minister, foreign, states, united, prime, UN, law, nations

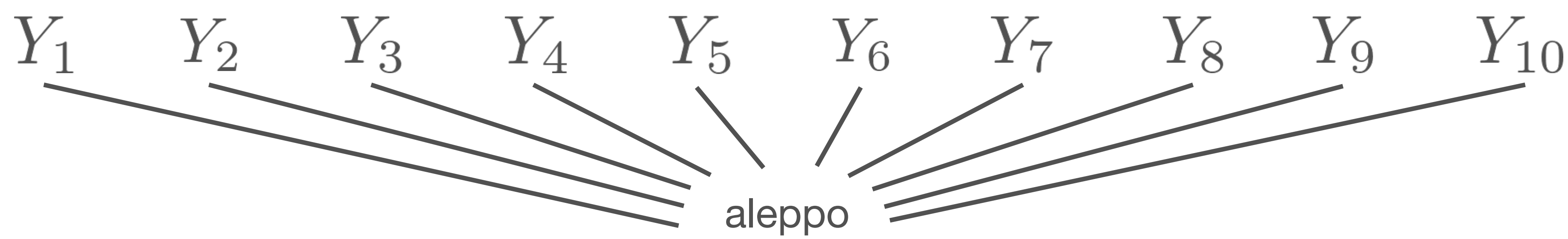
9: **aleppo**, city, area, near, air, northern, least, town, eastern, injured

10: **aleppo**, people, children, human, rights, women, social, school, society, lives

Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Anchoring for Topic Frames

Data: ~1 million English newswire articles since June 2015 from countries in the Middle East



1: **aleppo**, killed, police, security, attack, state, arrested, authorities

2: **aleppo**, forces, syria, military, war, army, civilians, iraq, militants

3: **aleppo**, health, medical, food, care, water, small, conditions, treatment, patients

4: country, **aleppo**, east, across, group, region, middle

5: two, **aleppo**, took, another, place, taking, leaders

6: **aleppo**, russia, iran, barack, obama, moscow, washington, putin

7: **aleppo**, political, court, part, accused, opposition, called, saying, parliament, democratic

8: government, **aleppo**, minister, foreign, states, united, prime, UN, law, nations

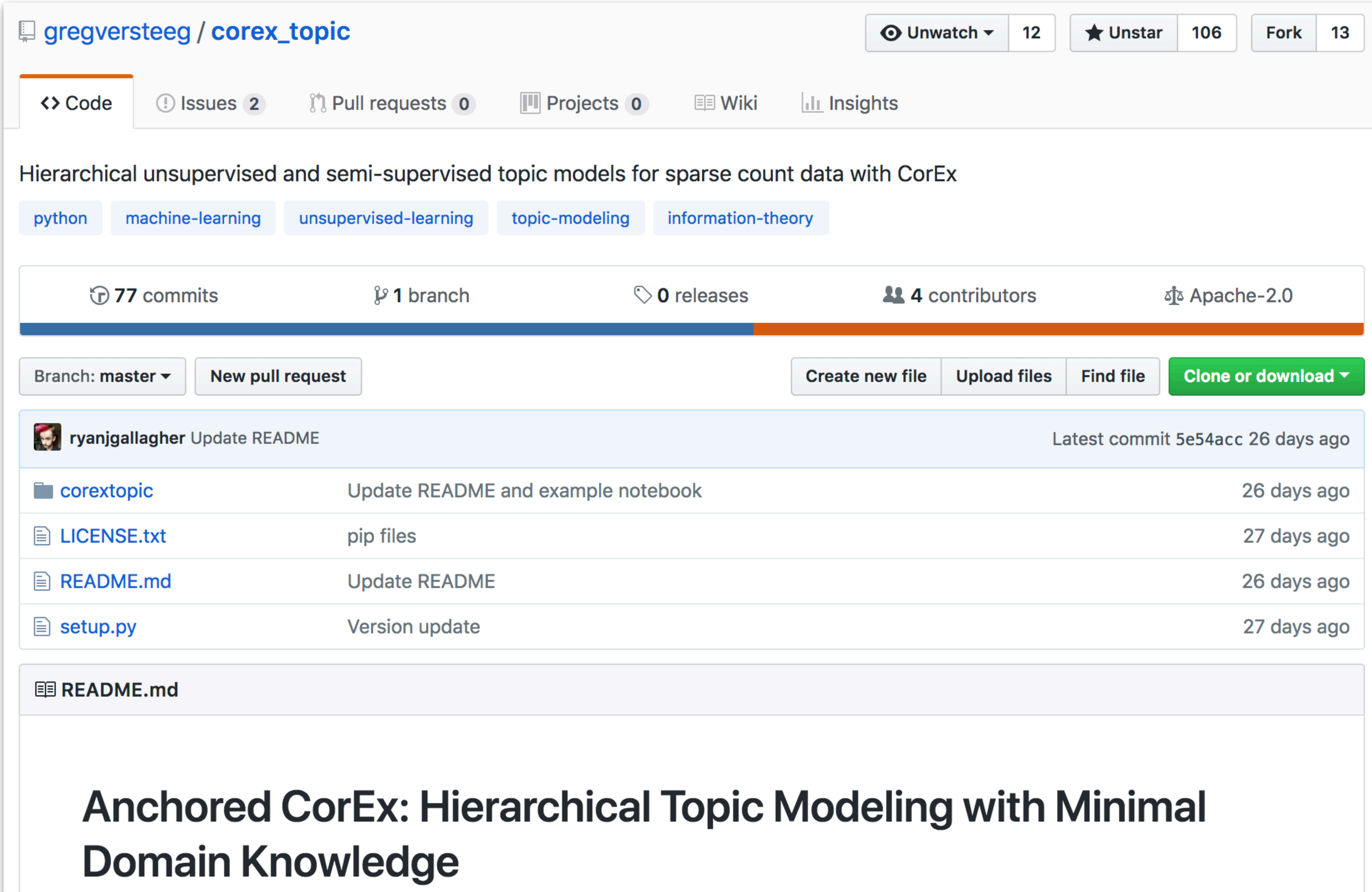
9: **aleppo**, city, area, near, air, northern, least, town, eastern, injured

10: **aleppo**, people, children, human, rights, women, social, school, society, lives

Work by Brendan Kennedy and Greg Ver Steeg, Information Sciences Institute

Code is open source and documented

github.com/gregversteeg/corex_topic



Detailed Jupyter notebook working through how to use and understand the CorEx topic model

Anchoring for Semi-Supervised Topic Modeling

Anchored CorEx is an extension of CorEx that allows the "anchoring" of words to topics. When anchoring a word to a topic, CorEx is trying to maximize the mutual information between that word and the anchored topic. So, anchoring provides a way to guide the topic model towards specific subsets of words that the user would like to explore.

The anchoring mechanism is flexible, and so there are many possibilities of anchoring. We explored the following types of anchoring in our TACL paper:

1. Anchoring a single set of words to a single topic. This can help promote a topic that did not naturally emerge when running an unsupervised instance of the CorEx topic model. For example, one might anchor words like "snow," "cold," and "avalanche" to a topic if one suspects there should be a snow avalanche topic within a set of disaster relief articles.
2. Anchoring single sets of words to multiple topics. This can help find different aspects of a topic that may be discussed in several different contexts. For example, one might anchor "protest" to three topics and "riot" to three other topics to understand different framings that arise from tweets about political protests.
3. Anchoring different sets of words to multiple topics. This can help enforce topic separability if there appear to be chimera topics. For example, one might anchor "mountain," "Bernese," and "dog" to one topic and "mountain," "rocky," and "colorado" to another topic to help separate topics that merge discussion of Bernese Mountain Dogs and the Rocky Mountains.

We'll demonstrate how to anchor words to the the CorEx topic model and how to develop other anchoring strategies.

]:
```
# Anchor one word to the first topic
anchor_words = ['nasa']
```

]:
```
# Anchor the word 'nasa' to the first topic
anchored_topic_model = ct.Corex(n_hidden=50, seed=2)
anchored_topic_model.fit(doc_word, words=words, anchors=anchor_words, anchor_strength=6);
```

This anchors the single word "nasa" to the first topic.

]:
```
topic_words,_ = zip(*anchored_topic_model.get_topics(topic=0))
print('0: ' + ','.join(topic_words))
```

```
0: nasa,gov,ames,institute,jpl,station,propulsion,jsc,arc,shafer
```

We can anchor multiple groups of words to multiple topics as well.

Information Theory for Text Analysis

1. We can measure the diversity of language through application of entropy

2. We can quantify *how much* texts diverge from one another

3. We can quantify *why* texts diverge from one another

4. We can learn topics with more control and less assumptions than LDA

Thank you for your time!

@ryanjgallag
ryanjgallag@gmail.com

github.com/gregversteeg/corex_topic