

# An Introduction to Analyzing Political Texts

## Part II

Arthur Spirling

New York University

November 15, 2019

# Where Are We?

# Where Are We?



# Where Are We?

We've covered the basics of **document** representation and characterization.



# Where Are We?

We've covered the basics of **document** representation and characterization.

**Now** begin to think about documents as members of **categories** or **classes**



# Where Are We?

We've covered the basics of **document** representation and characterization.

Now begin to think about documents as members of **categories** or **classes**

→ simple, fast **dictionary based** ways to classify/categorize



# Where Are We?



We've covered the basics of **document** representation and characterization.

Now begin to think about documents as members of **categories** or **classes**

→ simple, fast **dictionary based** ways to classify/categorize

**and** move on to supervised and unsupervised learning problems.

# Terminology



# Terminology

Unsupervised techniques:

# Terminology

Unsupervised techniques: learning  
(hidden or latent) structure in  
unlabeled data.

e.g. PCA of legislators's votes:

# Terminology

Unsupervised techniques: learning  
(hidden or latent) structure in  
unlabeled data.

e.g. PCA of legislators's votes: want to see  
how they are organized—

# Terminology

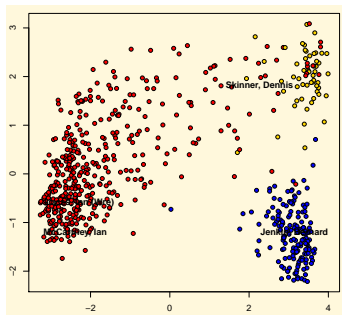
Unsupervised techniques: learning  
(hidden or latent) structure in  
unlabeled data.

e.g. PCA of legislators's votes: want to see  
how they are organized—by party? by  
ideology? by race?

# Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

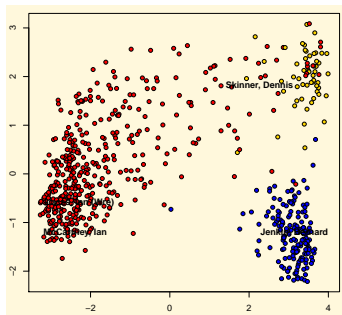


# Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

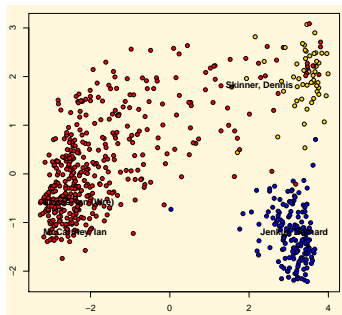
Supervised techniques:



# Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?

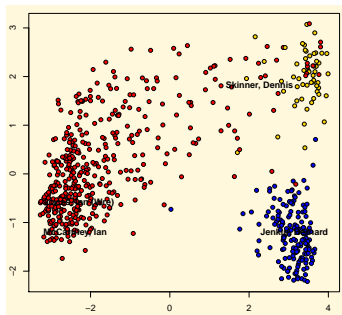


Supervised techniques: learning relationship between inputs and a labeled set of outputs.

# Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



Supervised techniques: learning relationship between inputs and a labeled set of outputs.

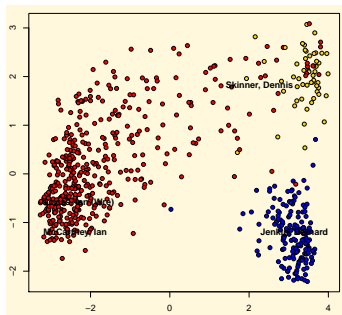
e.g. opinion mining:



# Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



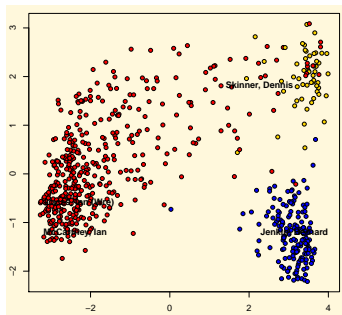
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ( $y \in \{0, 1\}$ )?

# Terminology

Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



Supervised techniques: learning relationship between inputs and a labeled set of outputs.

e.g. opinion mining: what makes a critic like or dislike a movie ( $y \in \{0, 1\}$ )?

**CRITIC REVIEWS FOR STAR WARS: EPISODE VII - THE FORCE AWAKENS**

All Critics (313) | Top Critics (48) | My Critics | Fresh (293) | Rotten (20)

The new movie, as an act of pure storytelling, streams by with fluency and zip.

[Full Review...](#) | December 21, 2015

**Anthony Lane**  
New Yorker  
★ Top Critic

At the end The Force Awakens looks more like a nostalgic film that will work as a transition to the new Star Wars' age. [Full Review in Spanish]

[Full Review...](#) | December 29, 2015

**Salvador Franco Reyes**

While Star Wars: The Force Awakens gets temporarily bogged down taking us back to the world that we left in 1983, it introduces us to the new and exciting torch-bearers of the franchise.

[Full Review...](#) | December 30, 2015

**Blake Howard**  
Graffiti With Punctuation

This film is a well-planned product that balances nostalgia with the capacity to attract new generations into the Star Wars universe. [Full Review in Spanish]

[Full Review...](#) | December 29, 2015

# Overview: Supervised Learning

# Overview: Supervised Learning

label some examples of each category

# Overview: Supervised Learning

label some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ )

# Overview: Supervised Learning

label some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ );  
some statements that were liberal,

# Overview: Supervised Learning

label some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ );  
some statements that were liberal, some that were conservative.

# Overview: Supervised Learning

**label** some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ );  
some statements that were liberal, some that were conservative.

**train** a 'machine' on these examples (e.g. logistic regression),



# Overview: Supervised Learning

**label** some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ );  
some statements that were liberal, some that were conservative.

**train** a 'machine' on these examples (e.g. logistic regression), using the  
**features** (DTM, other stuff) as the 'independent' variables.

# Overview: Supervised Learning

**label** some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ );  
some statements that were liberal, some that were conservative.

**train** a 'machine' on these examples (e.g. logistic regression), using the **features** (DTM, other stuff) as the 'independent' variables.

e.g. does the commentator use the word 'fetus' or 'baby' in discussing abortion law?

# Overview: Supervised Learning

**label** some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ );  
some statements that were liberal, some that were conservative.

**train** a 'machine' on these examples (e.g. logistic regression), using the **features** (DTM, other stuff) as the 'independent' variables.

e.g. does the commentator use the word 'fetus' or 'baby' in discussing abortion law?

**classify** use the learned relationship to predict the outcomes of documents ( $y \in \{0, 1\}$ , review sentiment)

# Overview: Supervised Learning

**label** some examples of each category

e.g. some reviews that were positive ( $y = 1$ ), some that were negative ( $y = 0$ );  
some statements that were liberal, some that were conservative.

**train** a 'machine' on these examples (e.g. logistic regression), using the **features** (DTM, other stuff) as the 'independent' variables.

e.g. does the commentator use the word 'fetus' or 'baby' in discussing abortion law?

**classify** use the learned relationship to predict the outcomes of documents ( $y \in \{0, 1\}$ , review sentiment) not in the training set.

# Overview of Dictionaries

# Overview of Dictionaries

idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents

# Overview of Dictionaries

idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.

# Overview of Dictionaries

- idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.
- common in opinion mining/sentiment analysis,



# Overview of Dictionaries

- idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.
- common in opinion mining/sentiment analysis, and in coding events or manifestos.

# Overview of Dictionaries

- idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.
- common in opinion mining/sentiment analysis, and in coding events or manifestos.

Often **derived from** supervised learning techniques

# Overview of Dictionaries

- idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.
- common in opinion mining/sentiment analysis, and in coding events or manifestos.

Often **derived from** supervised learning techniques  
and often **used in** supervised learning problems, as a starting point.

# Overview of Dictionaries

- idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.
- common in opinion mining/sentiment analysis, and in coding events or manifestos.

Often **derived from** supervised learning techniques  
and often **used in** supervised learning problems, as a starting point.  
so we'll cover them here in that context.

# Overview of Dictionaries

- idea: set of **pre-defined words** with specific connotations that allow us to **classify** documents automatically, quickly and accurately.
- common in opinion mining/sentiment analysis, and in coding events or manifestos.

Often **derived from** supervised learning techniques  
and often **used in** supervised learning problems, as a starting point.  
so we'll cover them here in that context.

# Classification with Dictionary Methods

# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

- 1 Categorize documents as belonging to a certain class (mutually exclusive? jointly exhaustive?)



# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

- 1 Categorize documents as belonging to a certain class (mutually exclusive? jointly exhaustive?)

e.g. this review is 'positive',

# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

- 1 Categorize documents as belonging to a certain class (mutually exclusive? jointly exhaustive?)

e.g. this review is 'positive', this speech is 'liberal'

# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

- 1 Categorize documents as belonging to a certain class (mutually exclusive? jointly exhaustive?)

e.g. this review is 'positive', this speech is 'liberal'

- 2 Measure **extent to which** document is associated with given category

# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

- 1 Categorize documents as belonging to a certain class (mutually exclusive? jointly exhaustive?)

e.g. this review is 'positive', this speech is 'liberal'

- 2 Measure **extent to which** document is associated with given category

e.g. this review is generally 'positive', but has some negative elements.

# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

- 1 Categorize documents as belonging to a certain class (mutually exclusive? jointly exhaustive?)

e.g. this review is 'positive', this speech is 'liberal'

- 2 Measure **extent to which** document is associated with given category

e.g. this review is generally 'positive', but has some negative elements.

We have a **pre-determined** list of words, the (weighted) presence of which helps us with (1) and (2).

# Classification with Dictionary Methods

**Aim** Typically we are trying to do one of two closely related things:

- 1 Categorize documents as belonging to a certain class (mutually exclusive? jointly exhaustive?)

e.g. this review is 'positive', this speech is 'liberal'

- 2 Measure **extent to which** document is associated with given category

e.g. this review is generally 'positive', but has some negative elements.

We have a **pre-determined** list of words, the (weighted) presence of which helps us with (1) and (2).

# More Specifically

# More Specifically

We have a set of **key words**, with attendant scores,



# More Specifically

We have a set of **key words**, with attendant scores,  
e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$

# More Specifically

We have a set of **key words**, with attendant scores,

e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$

→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

## More Specifically

We have a set of **key words**, with attendant scores,

- e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$   
→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

i.e. for document  $i$  and words  $m = 1, \dots, M$  in the dictionary,

## More Specifically

We have a set of **key words**, with attendant scores,

e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$

→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

i.e. for document  $i$  and words  $m = 1, \dots, M$  in the dictionary,

$$\text{tone of document } i = \sum_{m=1}^M \frac{s_m w_{im}}{N_i}$$

## More Specifically

We have a set of **key words**, with attendant scores,

e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$

→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

i.e. for document  $i$  and words  $m = 1, \dots, M$  in the dictionary,

$$\text{tone of document } i = \sum_{m=1}^M \frac{s_m w_{im}}{N_i}$$

where  $s_m$  is the score of word  $m$

## More Specifically

We have a set of **key words**, with attendant scores,

e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$

→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

i.e. for document  $i$  and words  $m = 1, \dots, M$  in the dictionary,

$$\text{tone of document } i = \sum_{m=1}^M \frac{s_m w_{im}}{N_i}$$

where  $s_m$  is the score of word  $m$

and  $w_{im}$  is the number of occurrences of the  $m$ th dictionary word in the document  $i$

## More Specifically

We have a set of **key words**, with attendant scores,

e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$

→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

i.e. for document  $i$  and words  $m = 1, \dots, M$  in the dictionary,

$$\text{tone of document } i = \sum_{m=1}^M \frac{s_m w_{im}}{N_i}$$

where  $s_m$  is the score of word  $m$

and  $w_{im}$  is the number of occurrences of the  $m$ th dictionary word in the document  $i$

and  $N_i$  is the total number of all dictionary words in the document.

## More Specifically

We have a set of **key words**, with attendant scores,

e.g. for movie reviews: 'terrible' is scored as  $-1$ ; 'fantastic' as  $+1$

→ the **relative rate** of occurrence of these terms tells us about the overall **tone** or category that the document should be placed in.

i.e. for document  $i$  and words  $m = 1, \dots, M$  in the dictionary,

$$\text{tone of document } i = \sum_{m=1}^M \frac{s_m w_{im}}{N_i}$$

where  $s_m$  is the score of word  $m$

and  $w_{im}$  is the number of occurrences of the  $m$ th dictionary word in the document  $i$

and  $N_i$  is the total number of all dictionary words in the document.

→ just add up the number of times the words appear and multiply by the score (normalizing by doc dictionary presence)



# (Simple) Example: Barnes' review of *The Big Short*

## (Simple) Example: Barnes' review of *The Big Short*

*Director and co-screenwriter Adam McKay (Step Brothers) bungles a great opportunity to savage the architects of the 2008 financial crisis in The Big Short, wasting an A-list ensemble cast in the process. Steve Carell, Brad Pitt, Christian Bale and Ryan Gosling play various tenuously related members of the finance industry, men who made a killing by betting against the housing market, which at that point had superficially swelled to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is bad, from the U-Turn-era Oliver Stone visuals to Carell's sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain complex financial concepts. After a brutal opening half, it finally settles into a groove, and there's a queasy charge in watching a credit-drunk America walking towards that cliff's edge, but not enough to save the film.*

## Retain words in Hu & Liu Dictionary...

*Director and co-screenwriter Adam McKay (Step Brothers) bungles a **great** opportunity to **savage** the architects of the 2008 financial **crisis** in The Big Short, **wasting** an A-list ensemble cast in the process. Steve Carell, Brad Pitt, Christian Bale and Ryan Gosling play various **tenuously** related members of the finance industry, men who made made a **killing** by betting against the housing market, which at that point had **superficially swelled** to record highs. All of the elements are in place for a lacerating satire, but almost every aesthetic choice in the film is **bad**, from the U-Turn-era Oliver Stone visuals to Carell's sketch-comedy performance to the cheeky cutaways where Selena Gomez and Anthony Bourdain explain **complex** financial concepts. After a **brutal** opening half, it finally settles into a groove, and there's a queasy charge in watching a credit-**drunk** America walking towards that cliff's edge, but not **enough** to save the film.*

# Retain words in Hu & Liu Dictionary...

*great*  
*crisis*

*savage*  
*wasting*

*tenuously*

*killing*

*superficially swelled*

*bad*

*complex*

*brutal*

*drunk*

*enough*

# Simple math...

# Simple math...

negative 11

# Simple math...

negative 11

positive 2

# Simple math...

negative 11

positive 2

total 13



# Simple math...

negative 11

positive 2

total 13

$$\text{tone} = \frac{2-11}{13} = \frac{-9}{13}$$

# Simple math...

negative 11

positive 2

total 13

$$\text{tone} = \frac{2-11}{13} = \frac{-9}{13}$$



# Partner Exercise

# Partner Exercise



You are working for `rottentomatoes.com`, and want to automatically code (written) movie reviews as being between 1 and 5 stars.

**MOVIES OPENING THIS WEEK** [Get Tickets](#)

No Score Yet	Gods Of Egypt	FEB 26
58%	Triple 9	FEB 26
78%	Eddie The Eagle	FEB 26
No Score Yet	Crouching Tiger, Hidden Dragon	
100%	Only Yesterday	

**TOP BOX OFFICE**

83%	Deadpool	
82%	Kung Fu Panda 3	
60%	Risen	
88%	The Witch	\$8.8M
49%	How To Be Single	\$8.2M
60%	Race	\$7.4M
23%	Zoolander 2	\$5.5M

**Grandfathered**  
68% 51%  
Christina Milian, Daniel Chun

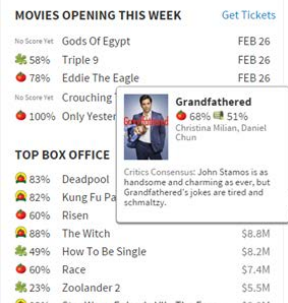
Critics Consensus: John Stamos is as handsome and charming as ever, but Grandfathered's jokes are tired and schmalzy.

# Partner Exercise



You are working for `rottentomatoes.com`, and want to automatically code (written) movie reviews as being between 1 and 5 stars.

- 1 Would the Hu & Liu approach work better for distinguishing a 1 star review from a 5 star review, or a 4 from a 5 star review? Why? How could you improve upon this?



# Partner Exercise

The screenshot shows the Rotten Tomatoes homepage. At the top is the 'Rotten Tomatoes' logo and a search bar. Below the logo is a 'TRENDING ON RT' section with links to 'Oscars Personality Quiz', 'Deadpool', and 'Winter T'. A large featured image shows characters from 'The Walking Dead'. Below this is a 'TUMBLR PICKS' section with the text 'Our Favorite Richonne Moments From Last Night's The'. The 'MOVIES OPENING THIS WEEK' section lists movies with their scores and release dates: 'Gods Of Egypt' (No Score Yet, FEB 26), 'Triple 9' (58%, FEB 26), 'Eddie The Eagle' (78%, FEB 26), 'Crouching' (No Score Yet), and 'Only Yesterday' (100%). The 'TOP BOX OFFICE' section lists movies with their scores and box office numbers: 'Deadpool' (83%, \$8.8M), 'Kung Fu Panda 3' (82%, \$8.2M), 'Risen' (60%, \$7.4M), 'The Witch' (88%, \$5.5M), 'How To Be Single' (49%, \$5.5M), 'Race' (60%, \$5.5M), and 'Zoolander 2' (23%, \$5.5M). A tooltip for 'Grandfathered' is visible, showing a score of 68% and a critics consensus: 'Critics Consensus: John Stamos is as handsome and charming as ever, but Grandfathered's jokes are tired and schmalzy.'

You are working for `rottentomatoes.com`, and want to automatically code (written) movie reviews as being between 1 and 5 stars.

- 1 Would the Hu & Liu approach work better for distinguishing a 1 star review from a 5 star review, or a 4 from a 5 star review? Why? How could you improve upon this?
- 2 Why does sarcasm cause problems, and what should we do about it?

# Partner Exercise

The screenshot shows the Rotten Tomatoes homepage. At the top is the 'Rotten Tomatoes' logo and a search bar. Below the logo are links for 'TRENDING ON RT', 'Oscars Personality Quiz', 'Deadpool', and 'Winter T'. A large featured image shows characters from 'The Walking Dead'. Below this is a 'TUMBLR PICKS' section with the text 'Our Favorite Richonne Moments From Last Night's The'. The 'MOVIES OPENING THIS WEEK' section lists movies with their scores and release dates. A 'TOP BOX OFFICE' section lists movies with their scores and box office numbers. A 'Grandfathered' movie is highlighted with a critics consensus.

**MOVIES OPENING THIS WEEK** [Get Tickets](#)

Score	Movie	Release Date
No Score Yet	Gods Of Egypt	FEB 26
58%	Triple 9	FEB 26
78%	Eddie The Eagle	FEB 26
No Score Yet	Crouching	
100%	Only Yesterday	

**TOP BOX OFFICE**

Score	Movie	Box Office
83%	Deadpool	\$8.8M
82%	Kung Fu Panda 3	\$8.2M
60%	Risen	\$7.4M
88%	The Witch	\$5.5M
49%	How To Be Single	\$5.5M
60%	Race	\$5.5M
23%	Zoolander 2	\$5.5M

**Grandfathered**  
68% 51%  
Christina Milian, Daniel Chun

Critics Consensus: John Stamos is as handsome and charming as ever, but Grandfathered's jokes are tired and schmalzy.

You are working for `rottentomatoes.com`, and want to automatically code (written) movie reviews as being between 1 and 5 stars.

- 1 Would the Hu & Liu approach work better for distinguishing a 1 star review from a 5 star review, or a 4 from a 5 star review? Why? How could you improve upon this?
- 2 Why does sarcasm cause problems, and what should we do about it?
- 3 Why might be generally nervous about BOW approaches?

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)



# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al,

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

80 categories,

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

80 categories, organized **hierarchically** into 4 larger groups.

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

80 categories, organized **hierarchically** into 4 larger groups.

e.g. all anger words (e.g. hate)  $\subset$  negative emotion  $\subset$  affective processes  $\subset$  psychological processes

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

80 categories, organized **hierarchically** into 4 larger groups.

e.g. all anger words (e.g. hate)  $\subset$  negative emotion  $\subset$  affective processes  $\subset$  psychological processes

NB words can be in **multiple** categories,

# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

80 categories, organized **hierarchically** into 4 larger groups.

e.g. all anger words (e.g. hate)  $\subset$  negative emotion  $\subset$  affective processes  $\subset$  psychological processes

**NB** words can be in **multiple** categories, and each subdictionary score is incremented as such words appear.



# Dictionaries: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al, <http://liwc.wpengine.com/>

LIWC2007 dictionary contains 2290 words and word stems (see also LIWC2015)

80 categories, organized **hierarchically** into 4 larger groups.

e.g. all anger words (e.g. hate)  $\subset$  negative emotion  $\subset$  affective processes  $\subset$  psychological processes

**NB** words can be in **multiple** categories, and each subdictionary score is incremented as such words appear.

Based on somewhat involved human coding/judgement and **proprietary**.

# Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

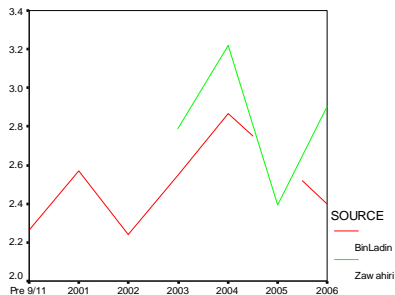
# Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

“The LIWC analyses suggest that Bin Ladin has been increasing in his cognitively complexity and emotionality since 9/11, as reflected by his increased use of exclusive, positive emotion, and negative emotion word use. ”

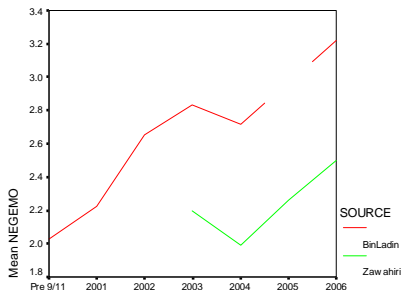
# Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

“The LIWC analyses suggest that Bin Ladin has been increasing in his cognitively complexity and emotionality since 9/11, as reflected by his increased use of exclusive, positive emotion, and negative emotion word use. ”

C. Positive emotion (happy, love)



D. Negative emotion (hate, sad)



# Making Dictionaries from Scratch

# Making Dictionaries from Scratch

Not trivial,

# Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.

# Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.  
Generally,



# Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.

Generally, need to ensure that we get **all relevant content** (no false negatives) and **only** that content (no false positives)

# Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.

Generally, need to ensure that we get **all relevant content** (no false negatives) and **only** that content (no false positives)

Works best when the contrasts are binary/obvious

# Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.

Generally, need to ensure that we get **all relevant content** (no false negatives) and **only** that content (no false positives)

Works best when the contrasts are binary/obvious

e.g. obviously 'for' vs 'against'

# Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.

Generally, need to ensure that we get **all relevant content** (no false negatives) and **only** that content (no false positives)

Works best when the contrasts are binary/obvious

e.g. obviously 'for' vs 'against'

**NB** Typically start with distinct **types** of documents (classified by hand),

# Making Dictionaries from Scratch

Not trivial, extending pre-existing is the norm.

Generally, need to ensure that we get **all relevant content** (no false negatives) and **only** that content (no false positives)

Works best when the contrasts are binary/obvious

e.g. obviously 'for' vs 'against'

**NB** Typically start with distinct **types** of documents (classified by hand), and learn which words are important for **discriminating** between them.

# Supervised Learning

# Wordscores (Laver, Benoit & Garry, 2003)

# Wordscores (Laver, Benoit & Garry, 2003)





# Wordscores (Laver, Benoit & Garry, 2003)



# Wordscores (Laver, Benoit & Garry, 2003)



Long standing interest in scaling **political texts** relative to one another:



# Wordscores (Laver, Benoit & Garry, 2003)



Long standing interest in scaling **political texts** relative to one another:

e.g. are parties moving together over time, such that manifestos are converging?



# Wordscores (Laver, Benoit & Garry, 2003)



Long standing interest in scaling **political texts** relative to one another:

- e.g. are parties moving together over time, such that manifestos are converging?
- e.g. do members of parliament speak in line with their constituency's ideology (roll calls typically uninformative)?

# Wordscores (Laver, Benoit & Garry, 2003)



Long standing interest in scaling **political texts** relative to one another:

- e.g. are parties moving together over time, such that manifestos are converging?
- e.g. do members of parliament speak in line with their constituency's ideology (roll calls typically uninformative)?

→ LBG suggest a way of scoring documents in a “naive Bayes” style, so that we can answer such questions.



- 1 Begin with a **reference set** (training set) of texts that have **known positions**.

1 Begin with a **reference set** (training set) of texts that have **known positions**.

e.g. we find a 'left' document and give it score  $-1$ ; and a 'right' document and give it score  $1$



1 Begin with a **reference set** (training set) of texts that have **known positions**.

e.g. we find a 'left' document and give it score  $-1$ ; and a 'right' document and give it score  $1$

2 Generate **word scores** from these reference texts

- 1 Begin with a **reference set** (training set) of texts that have **known positions**.

e.g. we find a 'left' document and give it score  $-1$ ; and a 'right' document and give it score  $1$

- 2 Generate **word scores** from these reference texts
- 3 Score the **virgin texts** (test set) of texts using those word scores, possibly transform virgin scores to original metric.

# Scoring the words

# Scoring the words

Suppose we have a given reference document  $R$ ,

# Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

# Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

In document  $R$ , count the number of times word  $i$  occurs, denote as  $f_{iR}$ .

# Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

In document  $R$ , count the number of times word  $i$  occurs, denote as  $f_{iR}$ . Also record the **total** number of words in document  $R$ , and denote as  $W_R$ .

# Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

In document  $R$ , count the number of times word  $i$  occurs, denote as  $f_{iR}$ . Also record the **total** number of words in document  $R$ , and denote as  $W_R$ .

Do the same for Communist party manifesto  $L$ , which we score as  $A_L = -1$ .



# Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

In document  $R$ , count the number of times word  $i$  occurs, denote as  $f_{iR}$ . Also record the **total** number of words in document  $R$ , and denote as  $W_R$ .

Do the same for Communist party manifesto  $L$ , which we score as  $A_L = -1$ . Then calculate  $f_{iL}$  and  $W_L$ .

# Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

In document  $R$ , count the number of times word  $i$  occurs, denote as  $f_{iR}$ . Also record the **total** number of words in document  $R$ , and denote as  $W_R$ .

Do the same for Communist party manifesto  $L$ , which we score as  $A_L = -1$ . Then calculate  $f_{iL}$  and  $W_L$ .

Define  $P_{iR}$  as (approximately the probability of word  $i$  given we are in document  $R$ ),

## Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

In document  $R$ , count the number of times word  $i$  occurs, denote as  $f_{iR}$ . Also record the **total** number of words in document  $R$ , and denote as  $W_R$ .

Do the same for Communist party manifesto  $L$ , which we score as  $A_L = -1$ . Then calculate  $f_{iL}$  and  $W_L$ .

Define  $P_{iR}$  as (approximately the probability of word  $i$  given we are in document  $R$ ),

$$P_{iR} = \frac{\frac{f_{iR}}{W_R}}{\frac{f_{iR}}{W_R} + \frac{f_{iL}}{W_L}}$$

## Scoring the words

Suppose we have a given reference document  $R$ , which is scored as  $A_R = 1$ . E.g. Neo-Nazi manifesto.

In document  $R$ , count the number of times word  $i$  occurs, denote as  $f_{iR}$ . Also record the **total** number of words in document  $R$ , and denote as  $W_R$ .

Do the same for Communist party manifesto  $L$ , which we score as  $A_L = -1$ . Then calculate  $f_{iL}$  and  $W_L$ .

Define  $P_{iR}$  as (approximately the probability of word  $i$  given we are in document  $R$ ),

$$P_{iR} = \frac{\frac{f_{iR}}{W_R}}{\frac{f_{iR}}{W_R} + \frac{f_{iL}}{W_L}}$$

and define  $P_{iL}$  in similar way.

# Score of a given word $i$

# Score of a given word $i$

is then

# Score of a given word $i$

is then

$$S_i = A_L P_{iL} + A_R P_{iR},$$

## Score of a given word $i$

is then

$$S_i = A_L P_{iL} + A_R P_{iR},$$

which in our simple case is  $S_i = P_{iR} - P_{iL}$ .



## Score of a given word $i$

is then

$$S_i = A_L P_{iL} + A_R P_{iR},$$

which in our simple case is  $S_i = P_{iR} - P_{iL}$ .

and the score of a **virgin** document is then

$$S_V = \sum_i \frac{f_{iV}}{W_V} \cdot S_i$$

## Score of a given word $i$

is then

$$S_i = A_L P_{iL} + A_R P_{iR},$$

which in our simple case is  $S_i = P_{iR} - P_{iL}$ .

and the score of a **virgin** document is then

$$S_V = \sum_i \frac{f_{iV}}{W_V} \cdot S_i$$

**NB**  $S_V$  is the mean of the scores of the words in  $V$  weighted by their term frequency.

## Score of a given word $i$

is then

$$S_i = A_L P_{iL} + A_R P_{iR},$$

which in our simple case is  $S_i = P_{iR} - P_{iL}$ .

and the score of a **virgin** document is then

$$S_V = \sum_i \frac{f_{iV}}{W_V} \cdot S_i$$

**NB**  $S_V$  is the mean of the scores of the words in  $V$  weighted by their term frequency.

**NB** any **new** words in the virgin document that were *not* in the reference texts are **ignored**: the sum is only over the words we've seen in the reference texts.

# Example

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005}$

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83$ .

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005}$



## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so  $S_i = 0.83 - 0.16 = 0.66$

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so  $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party,

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so  $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions immigrant 20 times in a thousand words.

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so  $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions immigrant 20 times in a thousand words.

well the relevant calculation for that word is  $0.02 \times 0.66 = 0.0132.$

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so  $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions immigrant 20 times in a thousand words.

well the relevant calculation for that word is  $0.02 \times 0.66 = 0.0132.$

but virgin manifesto, from Labour party,

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so  $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions immigrant 20 times in a thousand words.

well the relevant calculation for that word is  $0.02 \times 0.66 = 0.0132.$

but virgin manifesto, from Labour party, mentions it 10 times in a thousand words:  $0.01 \times 0.66 = 0.006$

## Example

Neo-Nazi manifesto uses 'immigrant' 25 times in 1000 words, while Communists use it only 5 times.

then  $P_{iR} = \frac{0.025}{0.025+0.005} = 0.83.$

and  $P_{iL} = \frac{0.005}{0.025+0.005} = 0.16.$

so  $S_i = 0.83 - 0.16 = 0.66$

we see a virgin manifesto, from the Conservative party, and it mentions immigrant 20 times in a thousand words.

well the relevant calculation for that word is  $0.02 \times 0.66 = 0.0132.$

but virgin manifesto, from Labour party, mentions it 10 times in a thousand words:  $0.01 \times 0.66 = 0.006$

→ can rescale these back to original  $(-1, 1)$  dimension.

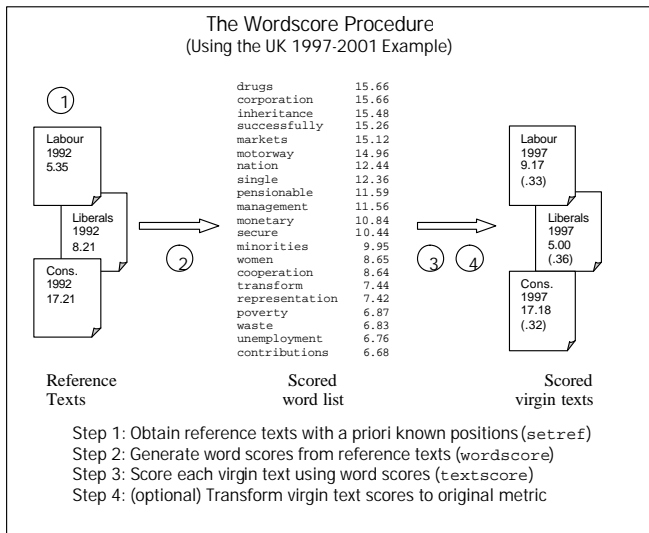


# New Labour Moderates its Economic Policy

# New Labour Moderates its Economic Policy



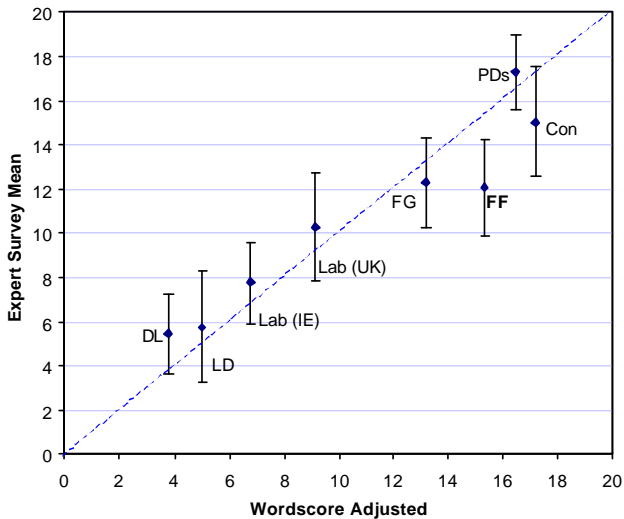
# New Labour Moderates its Economic Policy



# Compared to Expert Surveys

# Compared to Expert Surveys

(a) Economic Scale



# Comments

Extremely influential approach:

Extremely influential approach: avoids having to pick features of interest



Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice,

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot.

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot.  
very important to obtain [extreme](#) and appropriate [reference](#),

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot. very important to obtain [extreme](#) and appropriate [reference](#), and [score](#) them appropriately.

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot. very important to obtain [extreme](#) and appropriate [reference](#), and [score](#) them appropriately. Need to be from [domain](#) of virgin texts,

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot. very important to obtain [extreme](#) and appropriate [reference](#), and [score](#) them appropriately. Need to be from [domain](#) of virgin texts, and have [lots](#) of words.

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot.  
very important to obtain [extreme](#) and appropriate [reference](#), and [score](#) them appropriately. Need to be from [domain](#) of virgin texts, and have [lots](#) of words.

but Lowe (2008):



# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot.  
very important to obtain [extreme](#) and appropriate [reference](#), and [score](#) them appropriately. Need to be from [domain](#) of virgin texts, and have [lots](#) of words.

but Lowe (2008): no statistical model,

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have [uncertainty](#) estimates to boot.  
very important to obtain [extreme](#) and appropriate [reference](#), and [score](#) them appropriately. Need to be from [domain](#) of virgin texts, and have [lots](#) of words.

but Lowe (2008): no statistical model, inconsistent scoring assumptions,

# Comments

Extremely influential approach: avoids having to pick features of interest (features that don't distinguish between reference texts have  $S_i = 0$ )

and helpful/valid in practice, and can have **uncertainty** estimates to boot.  
very important to obtain **extreme** and appropriate **reference**, and **score** them appropriately. Need to be from **domain** of virgin texts, and have **lots** of words.

but Lowe (2008): no statistical model, inconsistent scoring assumptions, and difficult to pick up 'centrist language' (is equivalent to any language used commonly by all parties for linguistic reasons).

# Unsupervised Learning

# Overview: Unsupervised Learning

# Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

# Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties,

# Overview: Unsupervised Learning

Now our data—humans, documents, observations—are not pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its latent properties, what 'kind' of speech it is,



# Overview: Unsupervised Learning

Now our data—humans, documents, observations—are **not** pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its **latent** properties, what 'kind' of speech it is, what 'topics' it covers,

# Overview: Unsupervised Learning

Now our data—humans, documents, observations—are **not** pre-labeled in terms of some underlying concept.

e.g. while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its **latent** properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

# Overview: Unsupervised Learning

**Now** our data—humans, documents, observations—are **not** pre-labeled in terms of some underlying concept.

**e.g.** while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its **latent** properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

**Goal** is to take the observations and find hidden **structure** and **meaning** in them.

# Overview: Unsupervised Learning

**Now** our data—humans, documents, observations—are **not** pre-labeled in terms of some underlying concept.

**e.g.** while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its **latent** properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

**Goal** is to take the observations and find hidden **structure** and **meaning** in them.

→ look for (dis)similarities between documents (or observations):

# Overview: Unsupervised Learning

**Now** our data—humans, documents, observations—are **not** pre-labeled in terms of some underlying concept.

**e.g.** while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its **latent** properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

**Goal** is to take the observations and find hidden **structure** and **meaning** in them.

→ look for (dis)similarities between documents (or observations): it will be up to us to **interpret** what the groups/dimensions/concepts represent **after** the technique has been used.

# Overview: Unsupervised Learning

**Now** our data—humans, documents, observations—are **not** pre-labeled in terms of some underlying concept.

**e.g.** while we know who gave a particular speech, we don't yet know what that speech 'represents' in terms of its **latent** properties, what 'kind' of speech it is, what 'topics' it covers, what speeches it is similar to conceptually, etc.

**Goal** is to take the observations and find hidden **structure** and **meaning** in them.

→ look for (dis)similarities between documents (or observations): it will be up to us to **interpret** what the groups/dimensions/concepts represent **after** the technique has been used.

So...

So...

in contrast to supervised approaches,



# So...

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

# So...

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

# So...

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible?

# So...

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible? do somewhat similar (e.g. 2, 3, 4 clusters) specifications imply the same thing?

# So...

in contrast to supervised approaches, we won't know 'how correct' the output is in a simple statistical sense

While there *are* ways to compare unsupervised models, we are generally judging utility/fit in a different way

So, does it tell us something interesting/plausible? do somewhat similar (e.g. 2, 3, 4 clusters) specifications imply the same thing?

(not "what is the recall/precision/accuracy?")

# Topic Models

# Goal

# Goal

*Topic models are algorithms for discovering the **main themes** that pervade a large and otherwise **unstructured** collection of documents.*



# Goal

*Topic models are algorithms for discovering the **main themes** that pervade a large and otherwise **unstructured** collection of documents. Topic models can **organize** the collection according to the discovered themes.*

Blei, 2012

# Goal

*Topic models are algorithms for discovering the **main themes** that pervade a large and otherwise **unstructured** collection of documents. Topic models can **organize** the collection according to the discovered themes.*

Blei, 2012

Note that in **social science** we often use the outputs from topic models as a **measurement** strategy:

# Goal

*Topic models are algorithms for discovering the **main themes** that pervade a large and otherwise **unstructured** collection of documents. Topic models can **organize** the collection according to the discovered themes.*

Blei, 2012

Note that in **social science** we often use the outputs from topic models as a **measurement** strategy:

“who pays more attention to education policy, conservatives or liberals?”

# Topic Modeling

# Topic Modeling

Document 1

Document 2

Document 3

⋮

Document  $N$

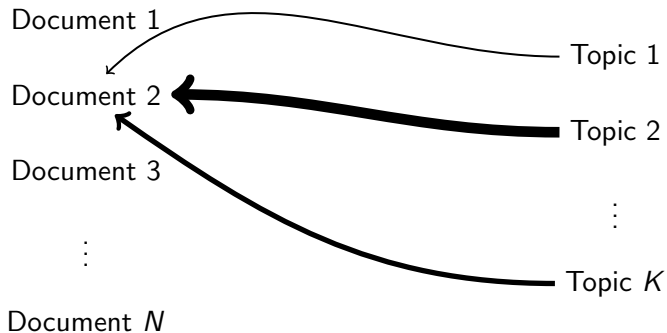
Topic 1

Topic 2

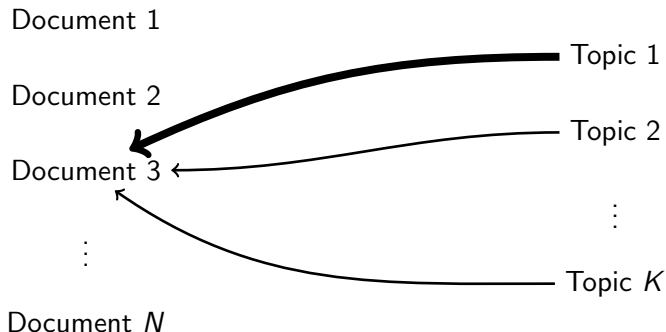
⋮

Topic  $K$

# Topic Modeling



# Topic Modeling



# DGP: intuition



# DGP: intuition

Documents exhibit different topics,

# DGP: intuition

Documents exhibit different topics, and in different **proportions**.

# DGP: intuition

Documents exhibit different topics, and in different **proportions**.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

# DGP: intuition

Documents exhibit different topics, and in different **proportions**.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a **topic** as a **distribution** over a **fixed vocabulary**.

# DGP: intuition

Documents exhibit different topics, and in different **proportions**.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a **topic** as a **distribution** over a **fixed vocabulary**.

E.g. the `trade` topic will have words like `import` and `tariff` with high probability.

# DGP: intuition

Documents exhibit different topics, and in different **proportions**.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a **topic** as a **distribution** over a **fixed vocabulary**.

E.g. the `trade` topic will have words like `import` and `tariff` with high probability.

Technically we assume the topics are generated **first**,

# DGP: intuition

Documents exhibit different topics, and in different **proportions**.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a **topic** as a **distribution** over a **fixed vocabulary**.

E.g. the `trade` topic will have words like `import` and `tariff` with high probability.

Technically we assume the topics are generated **first**, and the documents are generated second (from those topics).

# DGP: intuition

Documents exhibit different topics, and in different **proportions**.

E.g. a speech by the Finance minister might be 50% drawn from the `trade` topic, 40% from the `spending` topic, 9.9% from the `taxation` topic, 0.1% from the `health` topic.

Think of a **topic** as a **distribution** over a **fixed vocabulary**.

E.g. the `trade` topic will have words like `import` and `tariff` with high probability.

Technically we assume the topics are generated **first**, and the documents are generated second (from those topics).

Now, where do the **words** in the documents come from?



# Intuition: Generating Words

# Intuition: Generating Words

For each document. . .

# Intuition: Generating Words

For each document. . .

- 1 Randomly choose a **distribution** over topics.

# Intuition: Generating Words

For each document. . .

- 1 Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.

# Intuition: Generating Words

For each document. . .

- 1 Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.
- 2 Then, for every **word** in the document. . .

# Intuition: Generating Words

For each document. . .

- ① Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.
- ② Then, for every **word** in the document. . .
  - ① Randomly choose a topic from the distribution over topics from step 1.

# Intuition: Generating Words

For each document. . .

- ① Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.
- ② Then, for every **word** in the document. . .
  - ① Randomly choose a topic from the distribution over topics from step 1.
  - ② Randomly choose a word from the distribution over the vocabulary that the topic implies.

# First Part



# First Part

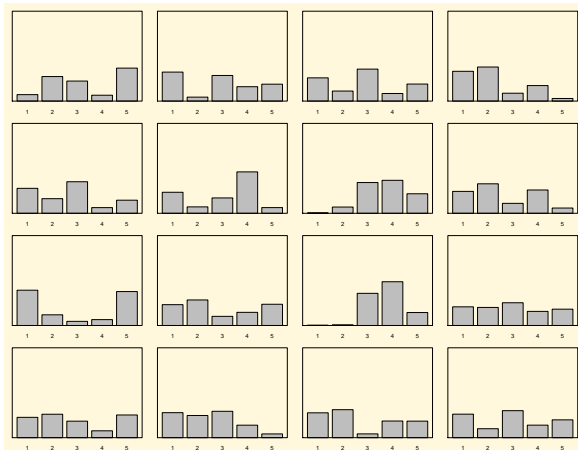
Randomly choose a **distribution** over topics.

# First Part

Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.

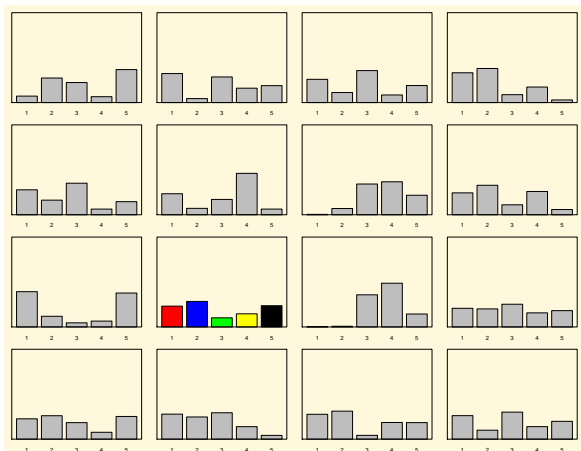
# First Part

Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.



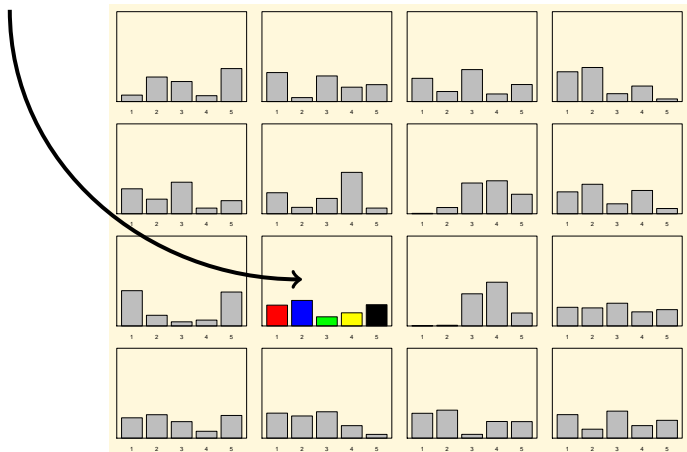
# First Part

Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.



# First Part

Randomly choose a **distribution** over topics. That is, choose one of many **multinomial** distributions, each which mixes the topics in different proportions.



# Second Part

## Second Part

Then, for every **word** in the document. . .

## Second Part

Then, for every **word** in the document. . .

- 1 Randomly choose a topic from the distribution over topics from step 1.



## Second Part

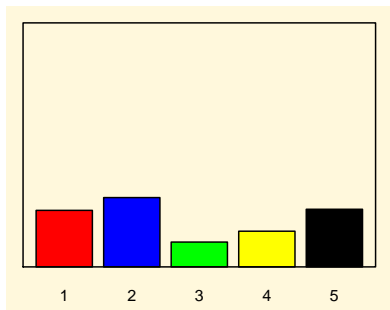
Then, for every **word** in the document. . .

- 1 Randomly choose a topic from the distribution over topics from step 1.
- 2 Randomly choose a word from the distribution over the vocabulary that the topic implies.

## Second Part

Then, for every **word** in the document...

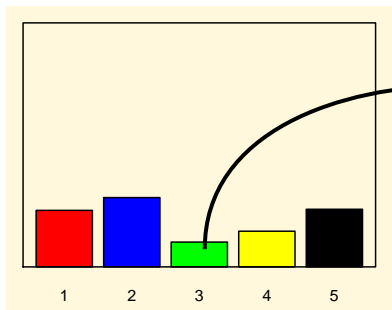
- 1 Randomly choose a topic from the distribution over topics from step 1.
- 2 Randomly choose a word from the distribution over the vocabulary that the topic implies.



## Second Part

Then, for every **word** in the document. . .

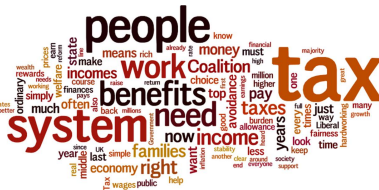
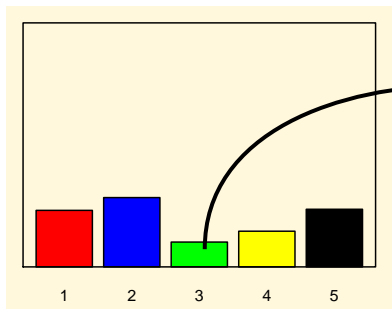
- 1 Randomly choose a topic from the distribution over topics from step 1.
- 2 Randomly choose a word from the distribution over the vocabulary that the topic implies.



## Second Part

Then, for every **word** in the document...

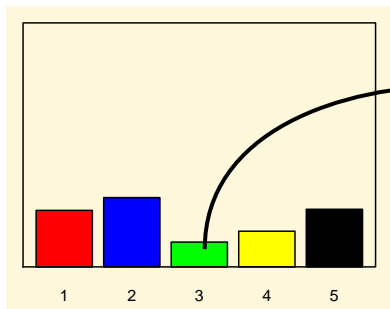
- 1 Randomly choose a topic from the distribution over topics from step 1.
- 2 Randomly choose a word from the distribution over the vocabulary that the topic implies.



## Second Part

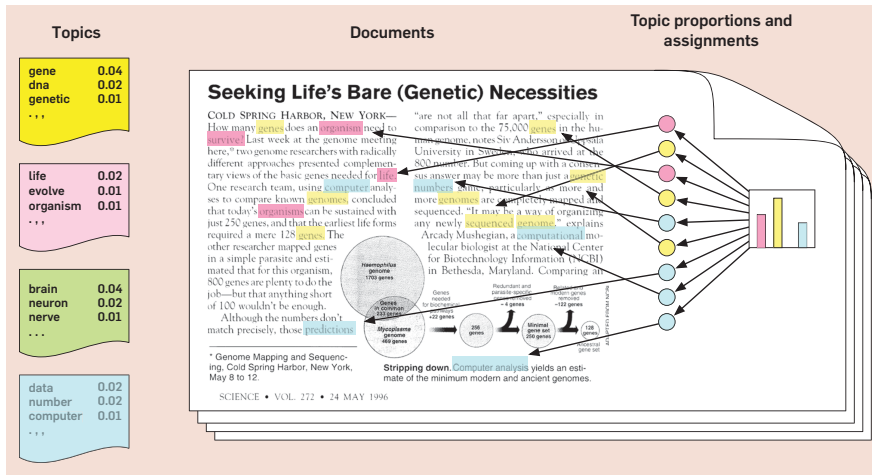
Then, for every **word** in the document...

- 1 Randomly choose a topic from the distribution over topics from step 1.
- 2 Randomly choose a word from the distribution over the vocabulary that the topic implies.



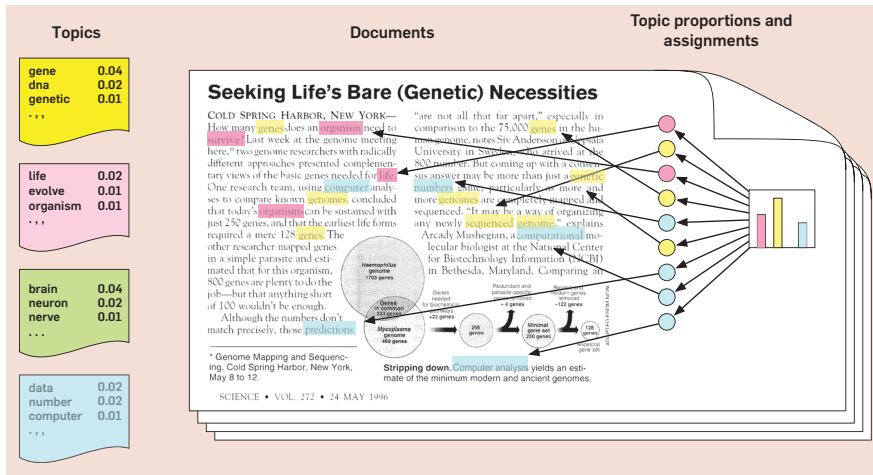
# Topic Modeling a Document (Blei, 2012)

# Topic Modeling a Document (Blei, 2012)



Note that all documents share **same** set of topics:

# Topic Modeling a Document (Blei, 2012)



Note that all documents share **same** set of topics: but some (e.g. **neuro**) may be (basically) absent in a given document.





Some of our variables—the documents which contain the words—are observable.

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are latent.

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are latent.

We need a distribution from which to draw the per-document topic distribution. We use a Dirichlet distribution as a prior for that.

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are **latent**.

We need a distribution from which to draw the per-document topic distribution. We use a **Dirichlet** distribution as a prior for that.

And Dirichlet is used for the **allocation** of the words in the documents to different topics:

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are **latent**.

We need a distribution from which to draw the per-document topic distribution. We use a **Dirichlet** distribution as a prior for that.

And Dirichlet is used for the **allocation** of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are **latent**.

We need a distribution from which to draw the per-document topic distribution. We use a **Dirichlet** distribution as a prior for that.

And Dirichlet is used for the **allocation** of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

→ **Latent**

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are **latent**.

We need a distribution from which to draw the per-document topic distribution. We use a **Dirichlet** distribution as a prior for that.

And Dirichlet is used for the **allocation** of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

→ **Latent Dirichlet**



Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are **latent**.

We need a distribution from which to draw the per-document topic distribution. We use a **Dirichlet** distribution as a prior for that.

And Dirichlet is used for the **allocation** of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

→ **Latent Dirichlet Allocation**.

Some of our variables—the documents which contain the words—are observable. But, topic structure—topics themselves, per-document topic distributions, per-document per-word topic assignments—are **latent**.

We need a distribution from which to draw the per-document topic distribution. We use a **Dirichlet** distribution as a prior for that.

And Dirichlet is used for the **allocation** of the words in the documents to different topics: it is used as a prior over the distribution of words (which define the topics).

→ **Latent Dirichlet Allocation**. **LDA**.

# Estimation

# Estimation

Ultimately,

# Estimation

Ultimately, we will use the observed data, the **words**,

# Estimation

Ultimately, we will use the observed data, the **words**, to make an inference about the **latent** parameters:

# Estimation

Ultimately, we will use the observed data, the **words**, to make an inference about the **latent** parameters: the  $\beta$ s, the  $z$ s, the  $\theta$ s.

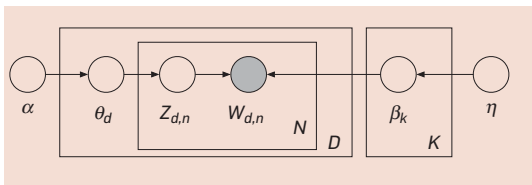
# Estimation

Ultimately, we will use the observed data, the **words**, to make an inference about the **latent** parameters: the  $\beta$ s, the  $z$ s, the  $\theta$ s. That will be a **conditional** probability.



# Estimation

Ultimately, we will use the observed data, the **words**, to make an inference about the **latent** parameters: the  $\beta$ s, the  $z$ s, the  $\theta$ s. That will be a **conditional** probability.



This is a complicated estimation problem, so we typically simulate/**approximate** the solution.

# Results

# Results

For a user-selected  $k$ , a typical implementation of LDA will return...

# Results

For a user-selected  $k$ , a typical implementation of LDA will return...

The **word distribution** for each topic.

# Results

For a user-selected  $k$ , a typical implementation of LDA will return...

The **word distribution** for each topic.

The **topic distribution** for each document.

# Results

For a user-selected  $k$ , a typical implementation of LDA will return...

The **word distribution** for each topic.

The **topic distribution** for each document.

Some implementations allow you to estimate e.g.  $\alpha$  (concentration parameter), in which case this is also returned.

# Results

For a user-selected  $k$ , a typical implementation of LDA will return...

The **word distribution** for each topic.

The **topic distribution** for each document.

Some implementations allow you to estimate e.g.  $\alpha$  (concentration parameter), in which case this is also returned. And perhaps some kind of fit statistic(s).

# A Manifesto Example



# A Manifesto Example

69 UK manifestos.

# A Manifesto Example

69 UK manifestos. Some preprocessing.

# A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics.

# A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

# A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

The (some selected) word distributions for each topic.

# A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

The (some selected) word distributions for each topic. Sum down the columns is one.

# A Manifesto Example

69 UK manifestos. Some preprocessing. Used `topicmodels` to fit five topics. Has Gibbs sampling and variational options.

The (some selected) word distributions for each topic. Sum down the columns is one.

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
conservative	0.00188	0.00088	0.00185	0.00221	0.00168
party	0.00145	0.00067	0.00066	0.00577	0.00093
general	0.00073	0.00033	0.00018	0.00192	0.00040
election	0.00079	0.00053	0.00022	0.00235	0.00076
manifesto	0.00059	0.00078	0.00032	0.00099	0.00048
:	:	:	:	:	:

Continued...



## Continued...

'Top' 6 most frequent words in each topic:

## Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

## Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Up to [analyst](#) to label the topics!

## Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Up to [analyst](#) to label the topics!

Meaningless 'junk' topics not unusual:

## Continued...

'Top' 6 most frequent words in each topic: might help interpretation (!)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	people	new	[markup]	new	must
2	local	government	people	labour	government
3	government	people	new	government	labour
4	new	continue	work	people	shall
5	tax	can	[markup]	shall	can
6	liberal	conservative	support	britain	policy

Up to [analyst](#) to label the topics!

Meaningless 'junk' topics not unusual: debate as to whether one has to interpret [every](#) topic.

# Continued

The topic distribution for each document. . .

The topic distribution for each document. . .



The topic distribution for each document. . .

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
doc 1	0.00009	0.00009	0.00009	0.00009	0.99965
doc 2	0.00011	0.00011	0.00011	0.00011	0.99954
doc 3	0.00010	0.00010	0.00010	0.00010	0.99959
doc 4	0.00006	0.00006	0.00006	0.00006	0.99978
doc 5	0.00002	0.00002	0.00002	0.00002	0.99991
doc 6	0.00019	0.00019	0.00019	0.00019	0.99924
⋮	⋮	⋮	⋮	⋮	⋮

The topic distribution for each document. . .

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
doc 1	0.00009	0.00009	0.00009	0.00009	0.99965
doc 2	0.00011	0.00011	0.00011	0.00011	0.99954
doc 3	0.00010	0.00010	0.00010	0.00010	0.99959
doc 4	0.00006	0.00006	0.00006	0.00006	0.99978
doc 5	0.00002	0.00002	0.00002	0.00002	0.99991
doc 6	0.00019	0.00019	0.00019	0.00019	0.99924
⋮	⋮	⋮	⋮	⋮	⋮

# Practical Notes I

# Practical Notes I

Texts are usually **preprocessed**:

# Practical Notes I

Texts are usually **preprocessed**: stop words removed,

# Practical Notes I

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed.

# Practical Notes I

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed. Punctuation often removed.

# Practical Notes I

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.



# Practical Notes I

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the **number of topics**,  $K$ , is not picked automatically.

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the **number of topics**,  $K$ , is not picked automatically. Analysts select various  $K$ s and check that their results are 'robust'. But see over.

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the **number of topics**,  $K$ , is not picked automatically. Analysts select various  $K$ s and check that their results are 'robust'. But see over.

As with all **unsupervised** learning,

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the **number of topics**,  $K$ , is not picked automatically. Analysts select various  $K$ s and check that their results are 'robust'. But see over.

As with all **unsupervised** learning, interpretation is non-trivial, and requires a lot of validation. Rant: 'just-so' stories abound. Lazy analysts conclude whatever they want.

Texts are usually **preprocessed**: stop words removed, (very) rare tokens removed. Punctuation often removed. Stemming seems less common.

In most social science examples, the **number of topics**,  $K$ , is not picked automatically. Analysts select various  $K$ s and check that their results are 'robust'. But see over.

As with all **unsupervised** learning, interpretation is non-trivial, and requires a lot of validation. Rant: 'just-so' stories abound. Lazy analysts conclude whatever they want.

## Practical Notes II: Picking $k$

## Practical Notes II: Picking $k$

Crudely: in social science,

## Practical Notes II: Picking $k$

Crudely: in social science, researchers fit 'enough' topics until they see what they think they should.



## Practical Notes II: Picking $k$

Crudely: in social science, researchers fit ‘enough’ topics until they see what they think they should. E.g. a certain topic—like finance suddenly peels off—so stop there.

## Practical Notes II: Picking $k$

Crudely: in social science, researchers fit ‘enough’ topics until they see what they think they should. E.g. a certain topic—like finance suddenly peels off—so stop there.

→ Check findings are robust in the neighborhood: if best model has  $k = 35$ , check  $k = 30 - 40$  yields similar inferences.

## Practical Notes II: Picking $k$

Crudely: in social science, researchers fit ‘enough’ topics until they see what they think they should. E.g. a certain topic—like *finance* suddenly peels off—so stop there.

→ Check findings are robust in the neighborhood: if best model has  $k = 35$ , check  $k = 30 - 40$  yields similar inferences.

NB: social scientists typically fit far fewer topics than CS, even to same data.

## Practical Notes II: Picking $k$

Crudely: in social science, researchers fit ‘enough’ topics until they see what they think they should. E.g. a certain topic—like *finance* suddenly peels off—so stop there.

→ Check findings are robust in the neighborhood: if best model has  $k = 35$ , check  $k = 30 - 40$  yields similar inferences.

NB: social scientists typically fit far fewer topics than CS, even to same data.

## Picking $k$ , continued...

CS: split into training and test sets.

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.
- 2 record value of  $\alpha$  (hyperparameter on document specific topic distributions) and word distributions for the topics (the  $\beta$ s)



## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.
- 2 record value of  $\alpha$  (hyperparameter on document specific topic distributions) and word distributions for the topics (the  $\beta$ s)

We'll write the  $\beta$ s as  $\beta$ , then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\beta, \alpha) = \sum_d \log p(w_d|\beta, \alpha)$$

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.
- 2 record value of  $\alpha$  (hyperparameter on document specific topic distributions) and word distributions for the topics (the  $\beta$ s)

We'll write the  $\beta$ s as  $\beta$ , then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\beta, \alpha) = \sum_d \log p(w_d|\beta, \alpha)$$

where  $\mathbf{w}$  are the words in the **test** set.

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.
- 2 record value of  $\alpha$  (hyperparameter on document specific topic distributions) and word distributions for the topics (the  $\beta$ s)

We'll write the  $\beta$ s as  $\beta$ , then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\beta, \alpha) = \sum_d \log p(w_d|\beta, \alpha)$$

where  $\mathbf{w}$  are the words in the **test** set. Higher  $\mathcal{L}$  implies better model.

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.
- 2 record value of  $\alpha$  (hyperparameter on document specific topic distributions) and word distributions for the topics (the  $\beta$ s)

We'll write the  $\beta$ s as  $\beta$ , then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\beta, \alpha) = \sum_d \log p(w_d|\beta, \alpha)$$

where  $\mathbf{w}$  are the words in the **test** set. Higher  $\mathcal{L}$  implies better model. Intuition is to calculate likelihood of seeing the test words, given what we know produced the training set.

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.
- 2 record value of  $\alpha$  (hyperparameter on document specific topic distributions) and word distributions for the topics (the  $\beta$ s)

We'll write the  $\beta$ s as  $\beta$ , then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\beta, \alpha) = \sum_d \log p(w_d|\beta, \alpha)$$

where  $\mathbf{w}$  are the words in the **test** set. Higher  $\mathcal{L}$  implies better model. Intuition is to calculate likelihood of seeing the test words, given what we know produced the training set.

Do this for all  $k$ .

## Picking $k$ , continued...

CS: split into training and test sets. In the **training** set,

- 1 pick some value of  $k$  and fit a topic model.
- 2 record value of  $\alpha$  (hyperparameter on document specific topic distributions) and word distributions for the topics (the  $\beta$ s)

We'll write the  $\beta$ s as  $\beta$ , then we want

$$\mathcal{L}(\mathbf{w}) = \log p(\mathbf{w}|\beta, \alpha) = \sum_d \log p(w_d|\beta, \alpha)$$

where  $\mathbf{w}$  are the words in the **test** set. Higher  $\mathcal{L}$  implies better model. Intuition is to calculate likelihood of seeing the test words, given what we know produced the training set.

Do this for all  $k$ .

# In practice...

# In practice...

Perplexity is popular option



## In practice...

Perplexity is popular option

$$\text{perplexity} = \exp \left( -\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}} \right),$$

## In practice...

Perplexity is popular option

$$\text{perplexity} = \exp \left( -\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}} \right),$$

where lower is better.

# In practice...

Perplexity is popular option

$$\text{perplexity} = \exp \left( -\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}} \right),$$

where lower is better.

In general,  $\mathcal{L}(\mathbf{w})$  is intractable,

# In practice...

Perplexity is popular option

$$\text{perplexity} = \exp \left( -\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}} \right),$$

where lower is better.

In general,  $\mathcal{L}(\mathbf{w})$  is *intractable*, but there are ways to approximate it.

## In practice...

Perplexity is popular option

$$\text{perplexity} = \exp \left( - \frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}} \right),$$

where lower is better.

In general,  $\mathcal{L}(\mathbf{w})$  is *intractable*, but there are ways to approximate it.

But:

## In practice...

Perplexity is popular option

$$\text{perplexity} = \exp \left( -\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}} \right),$$

where lower is better.

In general,  $\mathcal{L}(\mathbf{w})$  is *intractable*, but there are ways to approximate it.

But: the topic models that hold-out calculations suggest are optimal and not much liked by humans!

## In practice...

Perplexity is popular option

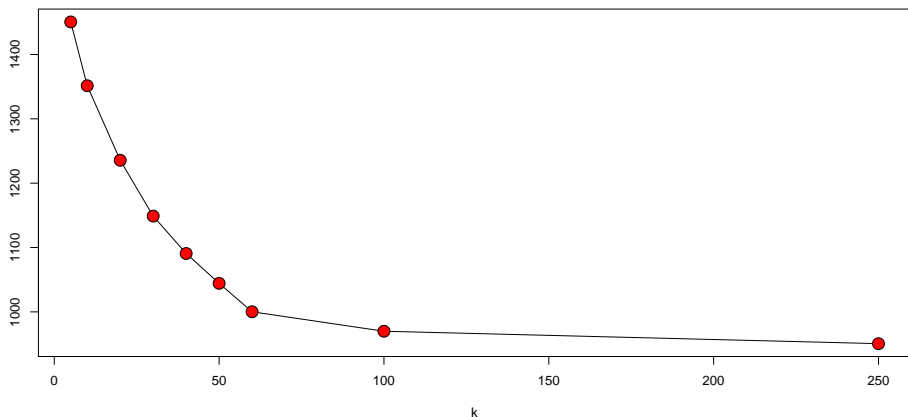
$$\text{perplexity} = \exp \left( -\frac{\mathcal{L}(\mathbf{w})}{\text{count of tokens}} \right),$$

where lower is better.

In general,  $\mathcal{L}(\mathbf{w})$  is **intractable**, but there are ways to approximate it.

But: the topic models that hold-out calculations suggest are optimal and not much liked by humans! “Reading Tea Leaves: How Humans Interpret Topic Models” by Chang et al.

# Perplexity Likes a Lot of Topics (manifestos)





# Pork to Policy (Catalinac, 2016)

# Pork to Policy (Catalinac, 2016)



# Pork to Policy (Catalinac, 2016)

Japan is a curious IR case:



# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy.



# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area.

# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

- ① Rise of China?



# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

- ① Rise of China? Need to focus on security.



# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

① Rise of China? Need to focus on security.

vs.

② Change in Electoral System?

# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

- ① Rise of China? Need to focus on security.

vs.

- ② Change in Electoral System? Moved from promising **pork** to having to deliver **policy** as part of Westminster-style polity.

# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

① Rise of China? Need to focus on security.

vs.

② Change in Electoral System? Moved from promising **pork** to having to deliver **policy** as part of Westminster-style polity.

To decide, we need data source that covers all lower house **legislators**

# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?

① Rise of China? Need to focus on security.

vs.

② Change in Electoral System? Moved from promising **pork** to having to deliver **policy** as part of Westminster-style polity.

To decide, we need data source that covers all lower house **legislators** where they set out their **policy priorities** over time.

# Pork to Policy (Catalinac, 2016)



Japan is a curious IR case: wealthy post-war not very interested in foreign policy. Recent times have seen a (re-)emergence in this area. Why?


- ① Rise of China? Need to focus on security.

vs.

- ② Change in Electoral System? Moved from promising **pork** to having to deliver **policy** as part of Westminster-style polity.

To decide, we need data source that covers all lower house **legislators** where they set out their **policy priorities** over time. See if/when they shift priorities.

# Manifestos



自由民主党公認  
のろた 芳成  
ほうせい よしなり  
五十六歳

青年に働く場を  
ふるさと秋田に活力を

**意** 我が郷の誇り、嵐山が月夜不況の犠牲とな  
なっている。緊急融資や救済土木事業  
等の実施を強く迫る。

**活** 公共住宅や公共建築物の木造化を推進。  
木材産業の活性化を図る。

**増** 建設費の急激拡大のため、公共事業のい  
づその増額確保にはすみをつける。

**守** 輸入米の阻止、やる気の出る米価確保は  
のろたに課せられた使命。

**希** 高速交通体系の速やかな整備と、それに  
伴う先端産業の誘致で若者の働く場所を  
確保。親・子・孫が希望を持って生活で  
きる豊かな郷土づくりを目指す。

**福** 高齢化社会を速に、老人・母と子の健康  
と幸せを守る福祉の充実を図る。

**望** 心の原点、ふるさと創りに全力。政治の  
眼を秋田へ向けさせ、二十一世紀の豊か  
な秋田を目指す。

「実行の二文字を胸に刻んで」

## のろた

主な結団と役職  
昭和4年 郷土会に生まれる  
昭和13年 秋田県議員  
昭和14年 青森県議員  
昭和15年 青森県議員  
昭和16年 秋田県議員  
昭和17年 秋田県議員  
昭和18年 秋田県議員  
昭和19年 秋田県議員  
昭和20年 秋田県議員  
昭和21年 秋田県議員  
昭和22年 秋田県議員  
昭和23年 秋田県議員  
昭和24年 秋田県議員  
昭和25年 秋田県議員  
昭和26年 秋田県議員  
昭和27年 秋田県議員  
昭和28年 秋田県議員  
昭和29年 秋田県議員  
昭和30年 秋田県議員  
昭和31年 秋田県議員  
昭和32年 秋田県議員  
昭和33年 秋田県議員  
昭和34年 秋田県議員  
昭和35年 秋田県議員  
昭和36年 秋田県議員  
昭和37年 秋田県議員  
昭和38年 秋田県議員  
昭和39年 秋田県議員  
昭和40年 秋田県議員  
昭和41年 秋田県議員  
昭和42年 秋田県議員  
昭和43年 秋田県議員  
昭和44年 秋田県議員  
昭和45年 秋田県議員  
昭和46年 秋田県議員  
昭和47年 秋田県議員  
昭和48年 秋田県議員  
昭和49年 秋田県議員  
昭和50年 秋田県議員  
昭和51年 秋田県議員  
昭和52年 秋田県議員  
昭和53年 秋田県議員  
昭和54年 秋田県議員  
昭和55年 秋田県議員  
昭和56年 秋田県議員  
昭和57年 秋田県議員  
昭和58年 秋田県議員  
昭和59年 秋田県議員  
昭和60年 秋田県議員  
昭和61年 秋田県議員  
昭和62年 秋田県議員  
昭和63年 秋田県議員  
昭和64年 秋田県議員  
昭和65年 秋田県議員  
昭和66年 秋田県議員  
昭和67年 秋田県議員  
昭和68年 秋田県議員  
昭和69年 秋田県議員  
昭和70年 秋田県議員  
昭和71年 秋田県議員  
昭和72年 秋田県議員  
昭和73年 秋田県議員  
昭和74年 秋田県議員  
昭和75年 秋田県議員  
昭和76年 秋田県議員  
昭和77年 秋田県議員  
昭和78年 秋田県議員  
昭和79年 秋田県議員  
昭和80年 秋田県議員  
昭和81年 秋田県議員  
昭和82年 秋田県議員  
昭和83年 秋田県議員  
昭和84年 秋田県議員  
昭和85年 秋田県議員  
昭和86年 秋田県議員  
昭和87年 秋田県議員  
昭和88年 秋田県議員  
昭和89年 秋田県議員  
昭和90年 秋田県議員  
昭和91年 秋田県議員  
昭和92年 秋田県議員  
昭和93年 秋田県議員  
昭和94年 秋田県議員  
昭和95年 秋田県議員  
昭和96年 秋田県議員  
昭和97年 秋田県議員  
昭和98年 秋田県議員  
昭和99年 秋田県議員  
昭和100年 秋田県議員

**主な結団と就職**

昭和4年 創立者に含まれる  
 昭和11年 労働問題員  
 昭和12年 労働問題員  
 昭和13年 労働問題員  
 昭和14年 労働問題員  
 昭和15年 労働問題員  
 昭和16年 労働問題員  
 昭和17年 労働問題員  
 昭和18年 労働問題員  
 昭和19年 労働問題員  
 昭和20年 労働問題員  
 昭和21年 労働問題員  
 昭和22年 労働問題員  
 昭和23年 労働問題員  
 昭和24年 労働問題員  
 昭和25年 労働問題員  
 昭和26年 労働問題員  
 昭和27年 労働問題員  
 昭和28年 労働問題員  
 昭和29年 労働問題員  
 昭和30年 労働問題員  
 昭和31年 労働問題員  
 昭和32年 労働問題員  
 昭和33年 労働問題員  
 昭和34年 労働問題員  
 昭和35年 労働問題員  
 昭和36年 労働問題員  
 昭和37年 労働問題員  
 昭和38年 労働問題員  
 昭和39年 労働問題員  
 昭和40年 労働問題員  
 昭和41年 労働問題員  
 昭和42年 労働問題員  
 昭和43年 労働問題員  
 昭和44年 労働問題員  
 昭和45年 労働問題員  
 昭和46年 労働問題員  
 昭和47年 労働問題員  
 昭和48年 労働問題員  
 昭和49年 労働問題員  
 昭和50年 労働問題員  
 昭和51年 労働問題員  
 昭和52年 労働問題員  
 昭和53年 労働問題員  
 昭和54年 労働問題員  
 昭和55年 労働問題員  
 昭和56年 労働問題員  
 昭和57年 労働問題員  
 昭和58年 労働問題員  
 昭和59年 労働問題員  
 昭和60年 労働問題員  
 昭和61年 労働問題員  
 昭和62年 労働問題員  
 昭和63年 労働問題員  
 昭和64年 労働問題員  
 昭和65年 労働問題員  
 昭和66年 労働問題員  
 昭和67年 労働問題員  
 昭和68年 労働問題員  
 昭和69年 労働問題員  
 昭和70年 労働問題員  
 昭和71年 労働問題員  
 昭和72年 労働問題員  
 昭和73年 労働問題員  
 昭和74年 労働問題員  
 昭和75年 労働問題員  
 昭和76年 労働問題員  
 昭和77年 労働問題員  
 昭和78年 労働問題員  
 昭和79年 労働問題員  
 昭和80年 労働問題員  
 昭和81年 労働問題員  
 昭和82年 労働問題員  
 昭和83年 労働問題員  
 昭和84年 労働問題員  
 昭和85年 労働問題員  
 昭和86年 労働問題員  
 昭和87年 労働問題員  
 昭和88年 労働問題員  
 昭和89年 労働問題員  
 昭和90年 労働問題員  
 昭和91年 労働問題員  
 昭和92年 労働問題員  
 昭和93年 労働問題員  
 昭和94年 労働問題員  
 昭和95年 労働問題員  
 昭和96年 労働問題員  
 昭和97年 労働問題員  
 昭和98年 労働問題員  
 昭和99年 労働問題員  
 昭和100年 労働問題員

## のろた

「実行の二文字を胸に刻んで」

**急** 我が国の将来、富山が高度不況の犠牲となつていく。緊急融資や救済土木事業等の実施を強く迫る。

**活** 公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。


**増** 建設費の急激拡大のため、公共事業のいつその増額確保にはすゝみをつける。

**守** 輸入米の阻止、やる気の出る米価確保ののろたに課せられた使命。

**希** 高度交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親、子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

**福** 高齢化社会を速に、老人、母と子の健康と幸せを守る福祉の充実を図る。

**心** 心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。



自由民主党公認  
**のろた 芳成**  
 まほう さい  
 五十六歳

7,497.



[illegible]

7,497. 1986–2009.

# Manifestos

**主な結団と就職**

昭和4年 鹿児島に生まれる  
昭和13年 参議院議員  
昭和14年 参議院議員  
昭和15年 参議院議員  
昭和16年 参議院議員  
昭和17年 参議院議員  
昭和18年 参議院議員  
昭和19年 参議院議員  
昭和20年 参議院議員  
昭和21年 参議院議員  
昭和22年 参議院議員  
昭和23年 参議院議員  
昭和24年 参議院議員  
昭和25年 参議院議員  
昭和26年 参議院議員  
昭和27年 参議院議員  
昭和28年 参議院議員  
昭和29年 参議院議員  
昭和30年 参議院議員  
昭和31年 参議院議員  
昭和32年 参議院議員  
昭和33年 参議院議員  
昭和34年 参議院議員  
昭和35年 参議院議員  
昭和36年 参議院議員  
昭和37年 参議院議員  
昭和38年 参議院議員  
昭和39年 参議院議員  
昭和40年 参議院議員  
昭和41年 参議院議員  
昭和42年 参議院議員  
昭和43年 参議院議員  
昭和44年 参議院議員  
昭和45年 参議院議員  
昭和46年 参議院議員  
昭和47年 参議院議員  
昭和48年 参議院議員  
昭和49年 参議院議員  
昭和50年 参議院議員  
昭和51年 参議院議員  
昭和52年 参議院議員  
昭和53年 参議院議員  
昭和54年 参議院議員  
昭和55年 参議院議員  
昭和56年 参議院議員  
昭和57年 参議院議員  
昭和58年 参議院議員  
昭和59年 参議院議員  
昭和60年 参議院議員  
昭和61年 参議院議員  
昭和62年 参議院議員  
昭和63年 参議院議員  
昭和64年 参議院議員  
昭和65年 参議院議員  
昭和66年 参議院議員  
昭和67年 参議院議員  
昭和68年 参議院議員  
昭和69年 参議院議員  
昭和70年 参議院議員  
昭和71年 参議院議員  
昭和72年 参議院議員  
昭和73年 参議院議員  
昭和74年 参議院議員  
昭和75年 参議院議員  
昭和76年 参議院議員  
昭和77年 参議院議員  
昭和78年 参議院議員  
昭和79年 参議院議員  
昭和80年 参議院議員  
昭和81年 参議院議員  
昭和82年 参議院議員  
昭和83年 参議院議員  
昭和84年 参議院議員  
昭和85年 参議院議員  
昭和86年 参議院議員  
昭和87年 参議院議員  
昭和88年 参議院議員  
昭和89年 参議院議員  
昭和90年 参議院議員  
昭和91年 参議院議員  
昭和92年 参議院議員  
昭和93年 参議院議員  
昭和94年 参議院議員  
昭和95年 参議院議員  
昭和96年 参議院議員  
昭和97年 参議院議員  
昭和98年 参議院議員  
昭和99年 参議院議員  
平成元年 参議院議員  
平成2年 参議院議員  
平成3年 参議院議員  
平成4年 参議院議員  
平成5年 参議院議員  
平成6年 参議院議員  
平成7年 参議院議員  
平成8年 参議院議員  
平成9年 参議院議員  
平成10年 参議院議員  
平成11年 参議院議員  
平成12年 参議院議員  
平成13年 参議院議員  
平成14年 参議院議員  
平成15年 参議院議員  
平成16年 参議院議員  
平成17年 参議院議員  
平成18年 参議院議員  
平成19年 参議院議員  
平成20年 参議院議員  
平成21年 参議院議員  
平成22年 参議院議員  
平成23年 参議院議員  
平成24年 参議院議員  
平成25年 参議院議員  
平成26年 参議院議員  
平成27年 参議院議員  
平成28年 参議院議員  
平成29年 参議院議員  
平成30年 参議院議員  
平成31年 参議院議員  
平成32年 参議院議員  
平成33年 参議院議員  
平成34年 参議院議員  
平成35年 参議院議員  
平成36年 参議院議員  
平成37年 参議院議員  
平成38年 参議院議員  
平成39年 参議院議員  
平成40年 参議院議員  
平成41年 参議院議員  
平成42年 参議院議員  
平成43年 参議院議員  
平成44年 参議院議員  
平成45年 参議院議員  
平成46年 参議院議員  
平成47年 参議院議員  
平成48年 参議院議員  
平成49年 参議院議員  
平成50年 参議院議員  
平成51年 参議院議員  
平成52年 参議院議員  
平成53年 参議院議員  
平成54年 参議院議員  
平成55年 参議院議員  
平成56年 参議院議員  
平成57年 参議院議員  
平成58年 参議院議員  
平成59年 参議院議員  
平成60年 参議院議員  
平成61年 参議院議員  
平成62年 参議院議員  
平成63年 参議院議員  
平成64年 参議院議員  
平成65年 参議院議員  
平成66年 参議院議員  
平成67年 参議院議員  
平成68年 参議院議員  
平成69年 参議院議員  
平成70年 参議院議員  
平成71年 参議院議員  
平成72年 参議院議員  
平成73年 参議院議員  
平成74年 参議院議員  
平成75年 参議院議員  
平成76年 参議院議員  
平成77年 参議院議員  
平成78年 参議院議員  
平成79年 参議院議員  
平成80年 参議院議員  
平成81年 参議院議員  
平成82年 参議院議員  
平成83年 参議院議員  
平成84年 参議院議員  
平成85年 参議院議員  
平成86年 参議院議員  
平成87年 参議院議員  
平成88年 参議院議員  
平成89年 参議院議員  
平成90年 参議院議員  
平成91年 参議院議員  
平成92年 参議院議員  
平成93年 参議院議員  
平成94年 参議院議員  
平成95年 参議院議員  
平成96年 参議院議員  
平成97年 参議院議員  
平成98年 参議院議員  
平成99年 参議院議員  
令和元年 参議院議員  
令和2年 参議院議員  
令和3年 参議院議員  
令和4年 参議院議員  
令和5年 参議院議員  
令和6年 参議院議員  
令和7年 参議院議員  
令和8年 参議院議員  
令和9年 参議院議員  
令和10年 参議院議員  
令和11年 参議院議員  
令和12年 参議院議員  
令和13年 参議院議員  
令和14年 参議院議員  
令和15年 参議院議員  
令和16年 参議院議員  
令和17年 参議院議員  
令和18年 参議院議員  
令和19年 参議院議員  
令和20年 参議院議員  
令和21年 参議院議員  
令和22年 参議院議員  
令和23年 参議院議員  
令和24年 参議院議員  
令和25年 参議院議員  
令和26年 参議院議員  
令和27年 参議院議員  
令和28年 参議院議員  
令和29年 参議院議員  
令和30年 参議院議員  
令和31年 参議院議員  
令和32年 参議院議員  
令和33年 参議院議員  
令和34年 参議院議員  
令和35年 参議院議員  
令和36年 参議院議員  
令和37年 参議院議員  
令和38年 参議院議員  
令和39年 参議院議員  
令和40年 参議院議員  
令和41年 参議院議員  
令和42年 参議院議員  
令和43年 参議院議員  
令和44年 参議院議員  
令和45年 参議院議員  
令和46年 参議院議員  
令和47年 参議院議員  
令和48年 参議院議員  
令和49年 参議院議員  
令和50年 参議院議員  
令和51年 参議院議員  
令和52年 参議院議員  
令和53年 参議院議員  
令和54年 参議院議員  
令和55年 参議院議員  
令和56年 参議院議員  
令和57年 参議院議員  
令和58年 参議院議員  
令和59年 参議院議員  
令和60年 参議院議員  
令和61年 参議院議員  
令和62年 参議院議員  
令和63年 参議院議員  
令和64年 参議院議員  
令和65年 参議院議員  
令和66年 参議院議員  
令和67年 参議院議員  
令和68年 参議院議員  
令和69年 参議院議員  
令和70年 参議院議員  
令和71年 参議院議員  
令和72年 参議院議員  
令和73年 参議院議員  
令和74年 参議院議員  
令和75年 参議院議員  
令和76年 参議院議員  
令和77年 参議院議員  
令和78年 参議院議員  
令和79年 参議院議員  
令和80年 参議院議員  
令和81年 参議院議員  
令和82年 参議院議員  
令和83年 参議院議員  
令和84年 参議院議員  
令和85年 参議院議員  
令和86年 参議院議員  
令和87年 参議院議員  
令和88年 参議院議員  
令和89年 参議院議員  
令和90年 参議院議員  
令和91年 参議院議員  
令和92年 参議院議員  
令和93年 参議院議員  
令和94年 参議院議員  
令和95年 参議院議員  
令和96年 参議院議員  
令和97年 参議院議員  
令和98年 参議院議員  
令和99年 参議院議員  
令和100年 参議院議員

## のろた

「実行」の二字を胸に刻んで

**急** 我が国の将来、富山が国富不況の犠牲となつていく。緊急融資や救済土木事業等の実施を強く迫る。

**活** 公共住宅や公共建築物の木造化を推進。木材産業の活性化を図る。

**増** 建設費の急激な拡大のため、公共事業のいつその増額確保にはすまいをつける。

**守** 輸入米の阻止、やる気の出る米価確保のろたに課せられた使命。

**希** 高度交通網の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親・子・孫が希望を持って生活できる豊かな郷土づくりを目指す。

**福** 高齢化社会を速に、老人・母と子の健康と幸せを守る福祉の充実を図る。

**心** 心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

自由民主党公認

五十六歳

のろた ほうた

芳成

青年に働く場を  
ふるさと秋田に活力を

7,497. 1986–2009. Standardized form.

# Manifestos

**主な結団と役職**

昭和4年 衆議院に選ばれる  
昭和13年 参議院議員  
昭和15年 参議院議員  
昭和16年 参議院議員  
昭和17年 参議院議員  
昭和18年 参議院議員  
昭和19年 参議院議員  
昭和20年 参議院議員  
昭和21年 参議院議員  
昭和22年 参議院議員  
昭和23年 参議院議員  
昭和24年 参議院議員  
昭和25年 参議院議員  
昭和26年 参議院議員  
昭和27年 参議院議員  
昭和28年 参議院議員  
昭和29年 参議院議員  
昭和30年 参議院議員  
昭和31年 参議院議員  
昭和32年 参議院議員  
昭和33年 参議院議員  
昭和34年 参議院議員  
昭和35年 参議院議員  
昭和36年 参議院議員  
昭和37年 参議院議員  
昭和38年 参議院議員  
昭和39年 参議院議員  
昭和40年 参議院議員  
昭和41年 参議院議員  
昭和42年 参議院議員  
昭和43年 参議院議員  
昭和44年 参議院議員  
昭和45年 参議院議員  
昭和46年 参議院議員  
昭和47年 参議院議員  
昭和48年 参議院議員  
昭和49年 参議院議員  
昭和50年 参議院議員  
昭和51年 参議院議員  
昭和52年 参議院議員  
昭和53年 参議院議員  
昭和54年 参議院議員  
昭和55年 参議院議員  
昭和56年 参議院議員  
昭和57年 参議院議員  
昭和58年 参議院議員  
昭和59年 参議院議員  
昭和60年 参議院議員  
昭和61年 参議院議員  
昭和62年 参議院議員  
昭和63年 参議院議員  
昭和64年 参議院議員  
昭和65年 参議院議員  
昭和66年 参議院議員  
昭和67年 参議院議員  
昭和68年 参議院議員  
昭和69年 参議院議員  
昭和70年 参議院議員  
昭和71年 参議院議員  
昭和72年 参議院議員  
昭和73年 参議院議員  
昭和74年 参議院議員  
昭和75年 参議院議員  
昭和76年 参議院議員  
昭和77年 参議院議員  
昭和78年 参議院議員  
昭和79年 参議院議員  
昭和80年 参議院議員  
昭和81年 参議院議員  
昭和82年 参議院議員  
昭和83年 参議院議員  
昭和84年 参議院議員  
昭和85年 参議院議員  
昭和86年 参議院議員  
昭和87年 参議院議員  
昭和88年 参議院議員  
昭和89年 参議院議員  
昭和90年 参議院議員  
昭和91年 参議院議員  
昭和92年 参議院議員  
昭和93年 参議院議員  
昭和94年 参議院議員  
昭和95年 参議院議員  
昭和96年 参議院議員  
昭和97年 参議院議員  
昭和98年 参議院議員  
昭和99年 参議院議員  
昭和100年 参議院議員

# のろた

「実行」の文字を胸に刻んで

**希望**

高齡化社会を速に、老人、母と子の健康と幸せを守る福祉の充実を図る。

**守**

心の原点、ふるさと創りに全力、政治の眼を秋田へ向けさせ、二十一世紀の豊かな秋田を目指す。

**増**

高齡化社会を速に、老人、母と子の健康と幸せを守る福祉の充実を図る。

**活**

高齡化社会を速に、老人、母と子の健康と幸せを守る福祉の充実を図る。

**意**

我が国の将来、秋田が国高不況の犠牲となっていない、緊急融資や救済土木事業等の実施を強く迫る。

公共住宅や公共建築物の老朽化を推進、木材産業の活性化を図る。

建設費の急激な拡大のため、公共事業のいっそうの増額確保に努める。

輸入米の阻止、やる気の出る米価確保のめざすに課せられた使命。

高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保、親子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

自由民主党公認

**のろた 芳成**

五十六歳

7,497. 1986–2009. Standardized form.

*"...instructed to write whatever they want in the form and return it before 5 PM of the first day of the campaign. At least two days before the election, local electoral commissions are required to distribute the forms of all candidates running in the district to all registered voters"*

# Manifestos

主な経歴と役職

昭和4年 鹿児島に生まれる

昭和13年 参議院議員

昭和15年 参議院議員

昭和17年 参議院議員

昭和19年 参議院議員

昭和21年 参議院議員

昭和23年 参議院議員

昭和25年 参議院議員

昭和27年 参議院議員

昭和29年 参議院議員

昭和31年 参議院議員

昭和33年 参議院議員

昭和35年 参議院議員

昭和37年 参議院議員

昭和39年 参議院議員

昭和41年 参議院議員

昭和43年 参議院議員

昭和45年 参議院議員

昭和47年 参議院議員

昭和49年 参議院議員

昭和51年 参議院議員

昭和53年 参議院議員

昭和55年 参議院議員

昭和57年 参議院議員

昭和59年 参議院議員

昭和61年 参議院議員

昭和63年 参議院議員

昭和65年 参議院議員

昭和67年 参議院議員

昭和69年 参議院議員

昭和71年 参議院議員

昭和73年 参議院議員

昭和75年 参議院議員

昭和77年 参議院議員

昭和79年 参議院議員

昭和81年 参議院議員

昭和83年 参議院議員

昭和85年 参議院議員

昭和87年 参議院議員

昭和89年 参議院議員

昭和91年 参議院議員

昭和93年 参議院議員

昭和95年 参議院議員

昭和97年 参議院議員

昭和99年 参議院議員

平成1年 参議院議員

平成3年 参議院議員

平成5年 参議院議員

平成7年 参議院議員

平成9年 参議院議員

平成11年 参議院議員

平成13年 参議院議員

平成15年 参議院議員

平成17年 参議院議員

平成19年 参議院議員

平成21年 参議院議員

平成23年 参議院議員

平成25年 参議院議員

平成27年 参議院議員

平成29年 参議院議員

平成31年 参議院議員

令和1年 参議院議員

令和3年 参議院議員

令和5年 参議院議員

令和7年 参議院議員

令和9年 参議院議員

令和11年 参議院議員

令和13年 参議院議員

令和15年 参議院議員

令和17年 参議院議員

令和19年 参議院議員

令和21年 参議院議員

令和23年 参議院議員

令和25年 参議院議員

令和27年 参議院議員

令和29年 参議院議員

令和31年 参議院議員

令和33年 参議院議員

令和35年 参議院議員

令和37年 参議院議員

令和39年 参議院議員

令和41年 参議院議員

令和43年 参議院議員

令和45年 参議院議員

令和47年 参議院議員

令和49年 参議院議員

令和51年 参議院議員

令和53年 参議院議員

令和55年 参議院議員

令和57年 参議院議員

令和59年 参議院議員

令和61年 参議院議員

令和63年 参議院議員

令和65年 参議院議員

令和67年 参議院議員

令和69年 参議院議員

令和71年 参議院議員

令和73年 参議院議員

令和75年 参議院議員

令和77年 参議院議員

令和79年 参議院議員

令和81年 参議院議員

令和83年 参議院議員

令和85年 参議院議員

令和87年 参議院議員

令和89年 参議院議員

令和91年 参議院議員

令和93年 参議院議員

令和95年 参議院議員

令和97年 参議院議員

令和99年 参議院議員

## のろた

「実行」の「文字」を胸に刻んで

自由民主党公認

のろた 芳成

五十六歳

### 青年に働く場を ふるさと秋田に活力を

我が国の将来、秋田が国高不況の犠牲となっていく。緊急融資や教職土木事業等の実施を強く迫る。

公共住宅や公共建築物の造成を推進。木材産業の活性化を図る。

建設費の急激な拡大のため、公共事業のいつそうの増額確保にはすみを付ける。

輸入米の阻止、やる気の出る米価確保のみに止まらず、米の活用を推進。

高速交通体系の速やかな整備と、それに伴う先端産業の誘致で若者の働く場所を確保。親子、孫が希望を持って生活できる豊かな郷土づくりを目指す。

高齢化社会を速に、老人、母と子の健康と幸せを守る福祉の充実を図る。

心の原点、ふるさと創りに全力。政治の眼を秋田へ向けさせ、「二十一世紀の豊かな秋田」を目指す。

7,497. 1986–2009. Standardized form.

*"...instructed to write whatever they want in the form and return it before 5 PM of the first day of the campaign. At least two days before the election, local electoral commissions are required to distribute the forms of all candidates running in the district to all registered voters"*

Manifestos were [hand transcribed](#) from microfilm.

# Manifestos

[illegible]

7,497. 1986–2009. Standardized form.

*"...instructed to write whatever they want in the form and return it before 5 PM of the first day of the campaign. At least two days before the election, local electoral commissions are required to distribute the forms of all candidates running in the district to all registered voters"*

Manifestos were **hand transcribed** from microfilm. Japanese install of Windows/R used to fit LDA.

# Topic Distribution over Words

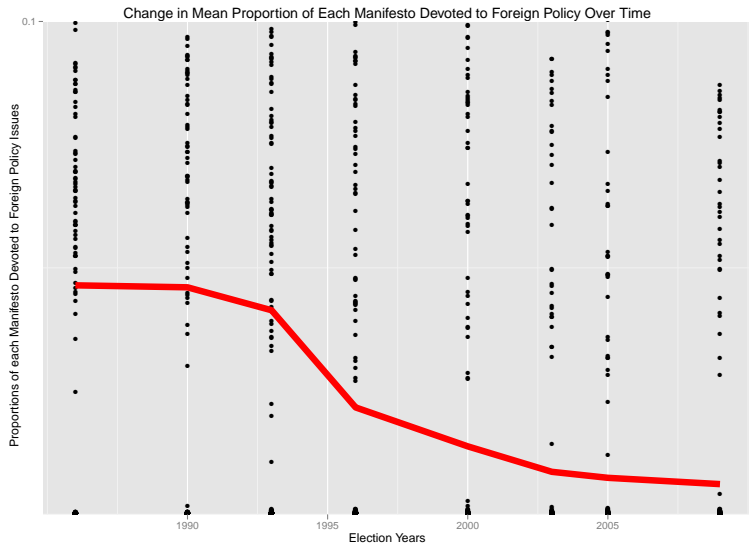
# Topic Distribution over Words

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1 改革	年金	推進	区	政治	日本
2 郵政	円	整備	政策	改革	国
3 民営	廃止	回る	地域	国民	外交
4 小泉	改革	つとめる	まち	企業	国家
5 構造	兆	社会	鹿児島	自民党	社会
6 政府	実現	対策	全力	日本	国民
7 官	無駄	振興	選挙	共産党	保障
8 推進	日本	充実	国政	献金	安全
9 民	増税	促進	作り	金権	地域
10 自民党	削減	安定	横浜	党	拉致
11 日本	一元化	確立	対策	選挙	経済
12 制度	政権	企業	中小	禁止	守る
13 民間	子供	実現	発電	憲法	問題
14 年金	地域	中小	推進	腐敗	北朝鮮
15 実現	ひと	育成	エネルギー	団体	教育
16 進める	サラリーマン	制度	企業	区	責任
17 断行	制度	政治	声	ソ連	力
18 地方	議員	地域	実現	守る	創る
19 止める	金	福祉	活性	平和	安心
20 保障	民主党	事業	自民党	円	目指す
21 財政	年間	改革	地方	反対	調り
22 作る	一掃	確保	尽くす	真	憲法
23 賛成	郵政	強化	商店	是正	可能
24 社会	道路	教育	いかす	一掃	道
25 国民	交代	施設	全国	憲政	未来
26 公務員	社会保険庁	生活	政党	抜本	ひと
27 力	月額	支援	ひと	定数	再生
28 経済	手当	環境	支援	政党	将来
29 国	談合	発展	経済	金丸	解決
30 安心	支援	協賛	福祉	政策	基土

# Change in proportion of 'Pork' Topic

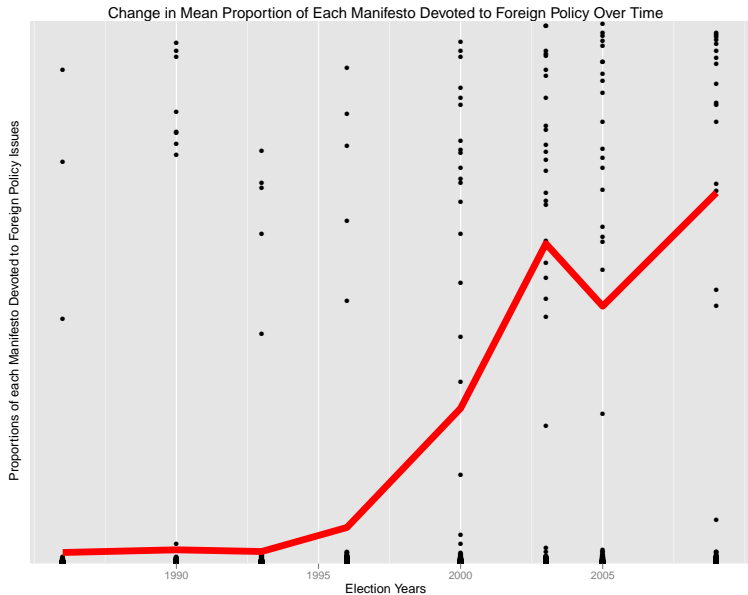


# Change in proportion of 'Pork' Topic



# Change in proportion of 'Foreign Policy' Topic

# Change in proportion of 'Foreign Policy' Topic



# Special Topics: Structural Topic Model

# Structural Topic Model

# Structural Topic Model

In general, we have lots of **metadata**:

# Structural Topic Model

In general, we have lots of **metadata**: e.g. author covariates, like gender or party membership.

# Structural Topic Model

In general, we have lots of **metadata**: e.g. author covariates, like gender or party membership.

But this it non-trivial to include in LDA.



# Structural Topic Model

In general, we have lots of **metadata**: e.g. author covariates, like gender or party membership.

But this it non-trivial to include in LDA.

→  $STM = LDA + \text{contextual information}$

# Structural Topic Model

In general, we have lots of **metadata**: e.g. author covariates, like gender or party membership.

But this it non-trivial to include in LDA.

→  $STM = LDA + \text{contextual information}$

This allows **more accurate estimation** and

# Structural Topic Model

In general, we have lots of **metadata**: e.g. author covariates, like gender or party membership.

But this is non-trivial to include in LDA.

→  $STM = LDA + \text{contextual information}$

This allows **more accurate estimation** and **more interpretable results**.

# Structural Topic Model

In general, we have lots of **metadata**: e.g. author covariates, like gender or party membership.

But this is non-trivial to include in LDA.

→  $STM = LDA + \text{contextual information}$

This allows **more accurate estimation** and **more interpretable results**.

Also allows us to 'test' hypothesis in more sensible way (though be careful!)

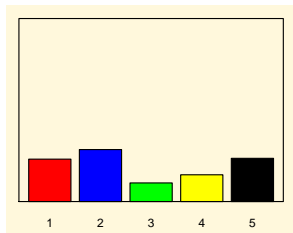
# Compare: Per Document Topic Distribution ( $\theta$ )

# Compare: Per Document Topic Distribution ( $\theta$ )

LDA: each document  
has some topic  
distribution.

# Compare: Per Document Topic Distribution ( $\theta$ )

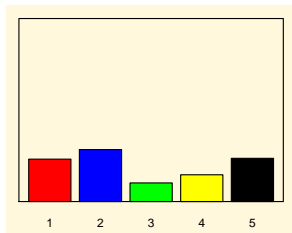
LDA: each document  
has some topic  
distribution.



# Compare: Per Document Topic Distribution ( $\theta$ )

**LDA**: each document has some topic distribution.

**STM**, that topic distribution is a function of the document metadata.



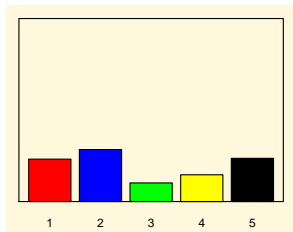


# Compare: Per Document Topic Distribution ( $\theta$ )

**LDA**: each document has some topic distribution.

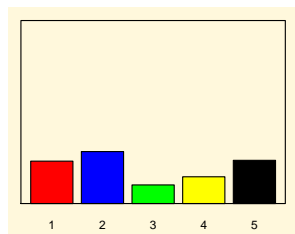
**STM**, that topic distribution is a function of the document metadata.

e.g. perhaps male author ( $X = 0$ ) documents have different topics relative to female ( $X = 1$ ) author docs.



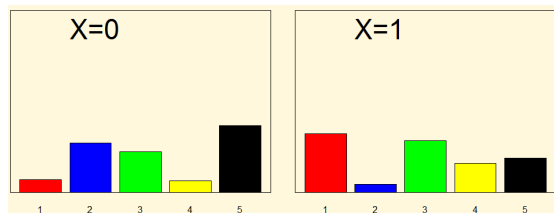
# Compare: Per Document Topic Distribution ( $\theta$ )

**LDA**: each document has some topic distribution.



**STM**, that topic distribution is a function of the document metadata.

e.g. perhaps male author ( $X = 0$ ) documents have different topics relative to female ( $X = 1$ ) author docs.



# Compare: Per Topic Word Distribution ( $\beta$ )

# Compare: Per Topic Word Distribution ( $\beta$ )

LDA: topic ('immigration') has a given distribution over words.

# Compare: Per Topic Word Distribution ( $\beta$ )

LDA: topic ('immigration') has a given distribution over words.



# Compare: Per Topic Word Distribution ( $\beta$ )

LDA: topic ('immigration') has a given distribution over words.



STM: that word distribution is a function of the document metadata.

STM: that word distribution is a function of the document metadata.

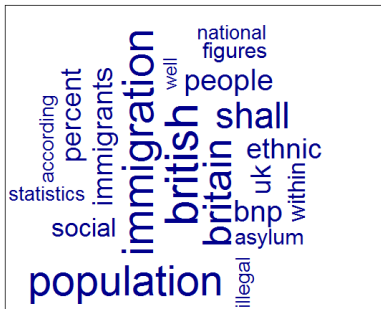
e.g. perhaps right parties ( $X = 0$ ) talk about a given topic differently to left ( $X = 1$ ) parties.



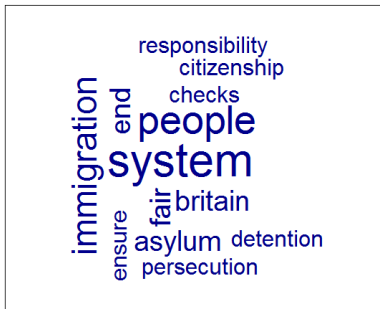
**STM**: that word distribution is a function of the document metadata.

e.g. perhaps right parties ( $X = 0$ ) talk about a given topic differently to left ( $X = 1$ ) parties.

$X=0$

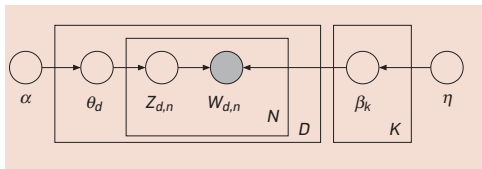


$X=1$

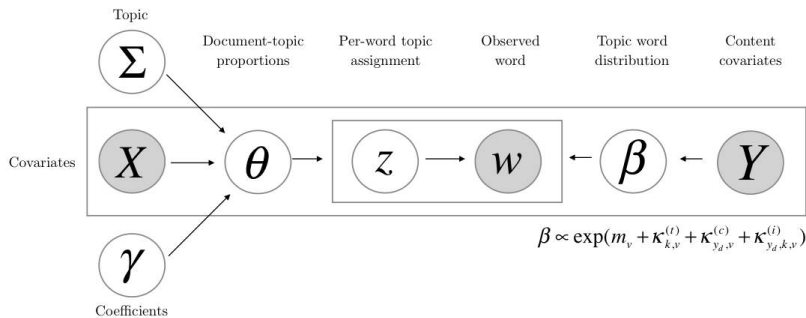
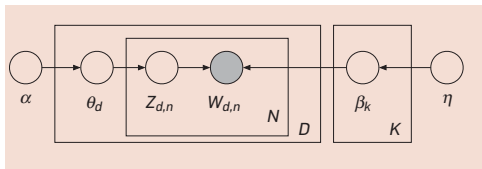


# Compare: Plate Diagram

# Compare: Plate Diagram



# Compare: Plate Diagram



# More Slides: Naive Bayes

# Naive Bayes Classification

# Naive Bayes Classification

**Motivation:** emails  $d$  arrive and must be classified as belonging to one of two classes  $c \in \{\text{spam}, \text{ham}\}$ .

# Naive Bayes Classification

**Motivation:** emails  $d$  arrive and must be classified as belonging to one of two classes  $c \in \{\text{spam}, \text{ham}\}$ .

by using the words/features frequencies the emails contain.



# Naive Bayes Classification

**Motivation:** emails  $d$  arrive and must be classified as belonging to one of two classes  $c \in \{\text{spam}, \text{ham}\}$ .

by using the words/features frequencies the emails contain.

use Naive Bayes,

# Naive Bayes Classification

**Motivation:** emails  $d$  arrive and must be classified as belonging to one of two classes  $c \in \{\text{spam}, \text{ham}\}$ .

by using the words/features frequencies the emails contain.

use Naive Bayes, also **simple Bayes**, or **independence Bayes**,

# Naive Bayes Classification

**Motivation:** emails  $d$  arrive and must be classified as belonging to one of two classes  $c \in \{\text{spam}, \text{ham}\}$ .

by using the words/features frequencies the emails contain.

use Naive Bayes, also **simple Bayes**, or **independence Bayes**,  
is a family of **classifiers** which apply **Bayes's theorem** and make 'naive' assumptions about **independence** between the features of a document.

# Naive Bayes Classification

**Motivation:** emails  $d$  arrive and must be classified as belonging to one of two classes  $c \in \{\text{spam}, \text{ham}\}$ .

by using the words/features frequencies the emails contain.

use Naive Bayes, also **simple Bayes**, or **independence Bayes**,  
is a family of **classifiers** which apply **Bayes's theorem** and make 'naive' assumptions about **independence** between the features of a document.

→ fast, simple, accurate, efficient and therefore **popular**.

# Set up

# Set up

We're interested in the probability that an email is in a given category,

## Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

# Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is



## Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

## Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9,

# Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9, because a lot of spam emails use that term.

# Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9, because a lot of spam emails use that term.

NB we are assuming terms basically occur randomly throughout the document/no position effects

# Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9, because a lot of spam emails use that term.

NB we are assuming terms basically occur randomly throughout the document/no position effects

We can write the probability that a given email  $d$  contains **all** the terms,

# Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9, because a lot of spam emails use that term.

NB we are assuming terms basically occur randomly throughout the document/no position effects

We can write the probability that a given email  $d$  contains **all** the terms, if it's from a class  $c$ , as

# Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9, because a lot of spam emails use that term.

NB we are assuming terms basically occur randomly throughout the document/no position effects

We can write the probability that a given email  $d$  contains **all** the terms, if it's from a class  $c$ , as

$$\Pr(d|c) = \prod_{k=1}^K \Pr(t_k|c)$$

## Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9, because a lot of spam emails use that term.

NB we are assuming terms basically occur randomly throughout the document/no position effects

We can write the probability that a given email  $d$  contains **all** the terms, if it's from a class  $c$ , as

$$\Pr(d|c) = \prod_{k=1}^K \Pr(t_k|c)$$

but this is not what we want:



## Set up

We're interested in the probability that an email is in a given **category**, given its features—i.e. frequency of terms.

The conditional probability of a term  $t_k$  occurring in a document, given that document is of class  $c$ , is  $= \Pr(t_k|c)$

e.g. probability of seeing 'beneficiary' in a spam email might be 0.9, because a lot of spam emails use that term.

NB we are assuming terms basically occur randomly throughout the document/no position effects

We can write the probability that a given email  $d$  contains **all** the terms, if it's from a class  $c$ , as

$$\Pr(d|c) = \prod_{k=1}^K \Pr(t_k|c)$$

but this is not what we want: we want  $\Pr(c|d)$ .

# Reminder: Bayes' Theorem

# Reminder: Bayes' Theorem

# Reminder: Bayes' Theorem

Recall that:

# Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

# Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.

## Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.
- e.g. you know a die shows an odd number, what is the probability that this odd number is 3?

## Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.
- e.g. you know a die shows an odd number, what is the probability that this odd number is 3?  $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}}$



## Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.
- e.g. you know a die shows an odd number, what is the probability that this odd number is 3?  $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$ .

## Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.
- e.g. you know a die shows an odd number, what is the probability that this odd number is 3?  $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$ .
- of course, it is also true that  $\Pr(B|A) = \frac{\Pr(B, A)}{\Pr(A)}$ .

## Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.
- e.g. you know a die shows an odd number, what is the probability that this odd number is 3?  $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$ .
- of course, it is also true that  $\Pr(B|A) = \frac{\Pr(B, A)}{\Pr(A)}$ .
  - but then, since  $\Pr(A, B) = \Pr(B, A)$ , we must have  $\Pr(A|B) \Pr(B) = \Pr(B|A) \Pr(A)$ , and thus...

## Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.
- e.g. you know a die shows an odd number, what is the probability that this odd number is 3?  $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$ .
- of course, it is also true that  $\Pr(B|A) = \frac{\Pr(B, A)}{\Pr(A)}$ .
  - but then, since  $\Pr(A, B) = \Pr(B, A)$ , we must have  $\Pr(A|B) \Pr(B) = \Pr(B|A) \Pr(A)$ , and thus... **Bayes' law**

## Reminder: Bayes' Theorem

Recall that:

$$\Pr(A|B) = \frac{\Pr(A, B)}{\Pr(B)}$$

- the probability that  $A$  occurs given that  $B$  occurred = the probability of both  $A$  and  $B$  occurring, divided by the probability that  $B$  occurs.

e.g. you know a die shows an odd number, what is the probability that this odd number is 3?  $\Pr(3|\text{odd}) = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{1}{3}$ .

- of course, it is also true that  $\Pr(B|A) = \frac{\Pr(B, A)}{\Pr(A)}$ .
- but then, since  $\Pr(A, B) = \Pr(B, A)$ , we must have  $\Pr(A|B) \Pr(B) = \Pr(B|A) \Pr(A)$ , and thus... **Bayes' law**

$$\Pr(A|B) = \frac{\Pr(A) \Pr(B|A)}{\Pr(B)}.$$

And...

And...

- interest is in  $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$ .

# And...

- interest is in  $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$ .
- Notice that  $\Pr(B)$  itself does not tell us whether a particular value of  $A$  is more or less likely to be observed,



# And...

- interest is in  $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$ .
- Notice that  $\Pr(B)$  itself does not tell us whether a particular value of  $A$  is more or less likely to be observed, so drop it and rewrite:

And...

- interest is in  $\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(B)}$ .
- Notice that  $\Pr(B)$  itself does not tell us whether a particular value of  $A$  is more or less likely to be observed, so drop it and rewrite:

$$\Pr(A|B) \propto \Pr(A) \Pr(B|A)$$

Here,  $\Pr(A)$  is our **prior** for  $A$ , while  $\Pr(B|A)$  will be the **likelihood** for the data we saw.

# Partner Exercise

# Partner Exercise

- 1 We know  $\Pr(A, B) = \Pr(B, A)$ . Can we conclude  $\Pr(A|B) = \Pr(B|A)$ ?

# Partner Exercise

- 1 We know  $\Pr(A, B) = \Pr(B, A)$ . Can we conclude  $\Pr(A|B) = \Pr(B|A)$ ?
- 2 If  $\Pr(A|B) = \Pr(A)$ ,

## Partner Exercise

- 1 We know  $\Pr(A, B) = \Pr(B, A)$ . Can we conclude  $\Pr(A|B) = \Pr(B|A)$ ?
- 2 If  $\Pr(A|B) = \Pr(A)$ , what does that tell us about events  $A$  and  $B$ ?

## Partner Exercise

- 1 We know  $\Pr(A, B) = \Pr(B, A)$ . Can we conclude  $\Pr(A|B) = \Pr(B|A)$ ?
- 2 If  $\Pr(A|B) = \Pr(A)$ , what does that tell us about events  $A$  and  $B$ ?
- 3 A subject claims to have psychic abilities—he can tell you how a (fair) coin will come down in nine tosses. He has less than a  $\frac{1}{500}$  chance of being correct by chance, but he succeeds in the task! Do you 'update' that he has psychic abilities? Why or why not?

So...



So...

We can express our quantity of interest as:

So...

We can express our quantity of interest as:

$$\Pr(c|d) = \frac{\Pr(c) \Pr(d|c)}{\Pr(d)}$$

So...

We can express our quantity of interest as:

$$\Pr(c|d) = \frac{\Pr(c) \Pr(d|c)}{\Pr(d)}$$

and

$$\Pr(c|d) \propto \underbrace{\Pr(c)}_{\text{prior}}$$

So...

We can express our quantity of interest as:

$$\Pr(c|d) = \frac{\Pr(c) \Pr(d|c)}{\Pr(d)}$$

and

$$\Pr(c|d) \propto \underbrace{\Pr(c)}_{\text{prior}} \underbrace{\prod_{k=1}^K \Pr(t_k|c)}_{\text{likelihood}}$$

So...

We can express our quantity of interest as:

$$\Pr(c|d) = \frac{\Pr(c) \Pr(d|c)}{\Pr(d)}$$

and

$$\Pr(c|d) \propto \underbrace{\Pr(c)}_{\text{prior}} \underbrace{\prod_{k=1}^K \Pr(t_k|c)}_{\text{likelihood}}$$

where  $\Pr(c)$  is the **prior probability** of a document occurring in class  $c$ ;

So...

We can express our quantity of interest as:

$$\Pr(c|d) = \frac{\Pr(c) \Pr(d|c)}{\Pr(d)}$$

and

$$\Pr(c|d) \propto \underbrace{\Pr(c)}_{\text{prior}} \underbrace{\prod_{k=1}^K \Pr(t_k|c)}_{\text{likelihood}}$$

where  $\Pr(c)$  is the **prior probability** of a document occurring in class  $c$ ; and  $\Pr(t_k|c)$  is interpreted as “measure of the how much evidence  $t_k$  contributes that  $c$  is the correct class”

# Goal

# Goal

We want to classify **new data**,



# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

e.g. We look at 10,000 emails to this point in time, and classify them as  $c \in \{\text{spam}, \text{ham}\}$ .

# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

- e.g. We look at 10,000 emails to this point in time, and classify them as  $c \in \{\text{spam}, \text{ham}\}$ . We use that information, and the terms associated with the two classes,

# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

- e.g. We look at 10,000 emails to this point in time, and classify them as  $c \in \{\text{spam}, \text{ham}\}$ . We use that information, and the terms associated with the two classes, to categorize **tomorrow's** email.

# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

e.g. We look at 10,000 emails to this point in time, and classify them as  $c \in \{\text{spam}, \text{ham}\}$ . We use that information, and the terms associated with the two classes, to categorize **tomorrow's** email.

In particular, we typically want to assign the document to a single **best** class.

# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

e.g. We look at 10,000 emails to this point in time, and classify them as  $c \in \{\text{spam}, \text{ham}\}$ . We use that information, and the terms associated with the two classes, to categorize **tomorrow's** email.

In particular, we typically want to assign the document to a single **best** class.

→ The 'best' class is the **maximum a posteriori** class,  $c_{map}$ :

# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

e.g. We look at 10,000 emails to this point in time, and classify them as  $c \in \{\text{spam}, \text{ham}\}$ . We use that information, and the terms associated with the two classes, to categorize **tomorrow's** email.

In particular, we typically want to assign the document to a single **best** class.

→ The 'best' class is the **maximum a posteriori** class,  $c_{map}$ :

$$c_{map} = \arg \max_c \widehat{\Pr(c|d)}$$

# Goal

We want to classify **new data**, based on patterns we observe in our **training** set (which we will classify by hand).

e.g. We look at 10,000 emails to this point in time, and classify them as  $c \in \{\text{spam}, \text{ham}\}$ . We use that information, and the terms associated with the two classes, to categorize **tomorrow's** email.

In particular, we typically want to assign the document to a single **best** class.

→ The 'best' class is the **maximum a posteriori** class,  $c_{map}$ :

$$c_{map} = \arg \max_c \widehat{\Pr(c|d)} = \arg \max_c \widehat{\Pr(c)} \prod_{k=1}^K \widehat{\Pr(t_k|c)}$$



# Example

# Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam

# Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam
	3	inherit plan money	ham
	4	cost amount amazon	ham
	5	prince william news	ham

# Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam
	3	inherit plan money	ham
	4	cost amount amazon	ham
	5	prince william news	ham
test	6	prince prince money	?

# Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam
	3	inherit plan money	ham
	4	cost amount amazon	ham
	5	prince william news	ham
test	6	prince prince money	?

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{money}|\text{ham}) = \frac{1}{9}$$

# Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam
	3	inherit plan money	ham
	4	cost amount amazon	ham
	5	prince william news	ham
test	6	prince prince money	?

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{money}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5} \frac{1}{9} \frac{1}{9} \frac{1}{9} = 0.00082$$

## Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam
	3	inherit plan money	ham
	4	cost amount amazon	ham
	5	prince william news	ham
test	6	prince prince money	?

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{money}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{ham}|\text{d}) \propto \frac{3}{5} \frac{1}{9} \frac{1}{9} \frac{1}{9} = 0.00082$$

$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$

$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$

$$\Pr(\text{money}|\text{spam}) = \frac{1}{6}$$

# Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam
	3	inherit plan money	ham
	4	cost amount amazon	ham
	5	prince william news	ham
test	6	prince prince money	?

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{money}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{ham}|d) \propto \frac{3}{5} \frac{1}{9} \frac{1}{9} \frac{1}{9} = 0.00082$$

$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$

$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$

$$\Pr(\text{money}|\text{spam}) = \frac{1}{6}$$

$$\Pr(\text{spam}|d) \propto \frac{2}{5} \frac{2}{6} \frac{2}{6} \frac{1}{6} = 0.0074$$



# Example

	email	words	classification
training	1	money inherit prince	spam
	2	prince inherit amount	spam
	3	inherit plan money	ham
	4	cost amount amazon	ham
	5	prince william news	ham
test	6	prince prince money	?

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{prince}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{money}|\text{ham}) = \frac{1}{9}$$

$$\Pr(\text{ham}|d) \propto \frac{3}{5} \frac{1}{9} \frac{1}{9} \frac{1}{9} = 0.00082$$

$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$

$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$

$$\Pr(\text{money}|\text{spam}) = \frac{1}{6}$$

$$\Pr(\text{spam}|d) \propto \frac{2}{5} \frac{2}{6} \frac{2}{6} \frac{1}{6} = 0.0074$$

→  $C_{map} = \text{spam}$

# Classifier is 'Naive'...

# Classifier is 'Naive'...

- 1 we assume conditional independence:

# Classifier is 'Naive'...

- 1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

# Classifier is 'Naive'...

- 1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.
- e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam.

# Classifier is 'Naive'...

1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars}),$$

# Classifier is 'Naive'...

1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars})$ , enables us to write everything as a simple product,

# Classifier is 'Naive'...

1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars})$ , enables us to write everything as a simple product,  $\prod_{k=1}^K \widehat{\Pr(t_k|c)}$ .



# Classifier is 'Naive'...

- 1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars})$ , enables us to write everything as a simple product,  $\prod_{k=1}^K \widehat{\Pr(t_k|c)}$ .

- 2 we assume **positional independence**:

# Classifier is 'Naive'...

- 1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars})$ , enables us to write everything as a simple product,  $\prod_{k=1}^K \widehat{\Pr(t_k|c)}$ .

- 2 we assume **positional independence**: probability that a term occurs in a particular place is constant for the entire document.

# Classifier is 'Naive'...

- 1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars})$ , enables us to write everything as a simple product,  $\prod_{k=1}^K \widehat{\Pr(t_k|c)}$ .

- 2 we assume **positional independence**: probability that a term occurs in a particular place is constant for the entire document. This implies we only need one probability distribution of terms, and that it's valid for every position.

# Classifier is 'Naive'...

- 1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars})$ , enables us to write everything as a simple product,  $\prod_{k=1}^K \widehat{\Pr(t_k|c)}$ .

- 2 we assume **positional independence**: probability that a term occurs in a particular place is constant for the entire document. This implies we only need one probability distribution of terms, and that it's valid for every position.

e.g. probability of observing 'dear' is the same regardless of which word in the document we are considering (1st, 2nd, 3rd etc). Equivalent to **bag of words**. (not an issue for Bernoulli)

# Classifier is 'Naive'...

- 1 we assume **conditional independence**: probability that a particular feature occurs is independent of any other feature occurring, once we condition on a given category.

e.g. probability of observing 'money' is independent of probability of observing 'dollars' given the emails are spam. This implies

$\Pr(\text{money}|c) = \Pr(\text{money}|c, \text{dollars})$ , enables us to write everything as a simple product,  $\prod_{k=1}^K \widehat{\Pr(t_k|c)}$ .

- 2 we assume **positional independence**: probability that a term occurs in a particular place is constant for the entire document. This implies we only need one probability distribution of terms, and that it's valid for every position.

e.g. probability of observing 'dear' is the same regardless of which word in the document we are considering (1st, 2nd, 3rd etc). Equivalent to **bag of words**. (not an issue for Bernoulli)

# Partner Exercise

## Partner Exercise

A feature of NB classification is that while the estimated probabilities can be **wildly wrong**, the classification decisions (the classes to which the documents are assigned) are **correct**.

# Partner Exercise

A feature of NB classification is that while the estimated probabilities can be **wildly wrong**, the classification decisions (the classes to which the documents are assigned) are **correct**.

- 1 Why does this happen?



# Partner Exercise

A feature of NB classification is that while the estimated probabilities can be **wildly wrong**, the classification decisions (the classes to which the documents are assigned) are **correct**.

- 1 Why does this happen?
- 2 What does this imply about the relationship between **estimation** ('modeling') and **accuracy**?

# Example: Jihadi Clerics

# Example: Jihadi Clerics

## Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10:14pm EDT

**Noor Huda Ismail**

PhD Candidate in Politics and International Relations , Monash University



# Example: Jihadi Clerics

## Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10:14pm EDT

**Noor Huda Ismail**

PhD Candidate in Politics and International Relations , Monash University



Nielsen (2012) investigates why certain scholars of Islam become Jihadi:

# Example: Jihadi Clerics

## Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10:14pm EDT

**Noor Huda Ismail**

PhD Candidate in Politics and International Relations , Monash University



Nielsen (2012) investigates why certain scholars of Islam become **Jihadi**: i.e. why they encourage armed struggle (especially against the west)

# Example: Jihadi Clerics

## Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10:14pm EDT

**Noor Huda Ismail**

PhD Candidate in Politics and International Relations, Monash University



Nielsen (2012) investigates why certain scholars of Islam become **Jihadi**: i.e. why they encourage armed struggle (especially against the west)

Requires that he first classifies scholars as **Jihadi** and  $\neg$  **Jihadi**:

# Example: Jihadi Clerics

## Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10:14pm EDT

**Noor Huda Ismail**

PhD Candidate in Politics and International Relations, Monash University



Nielsen (2012) investigates why certain scholars of Islam become **Jihadi**: i.e. why they encourage armed struggle (especially against the west)

Requires that he first classifies scholars as **Jihadi** and  $\neg$  **Jihadi**: has 27,142 texts from 101 clerics,

# Example: Jihadi Clerics

## Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10:14pm EDT

**Noor Huda Ismail**

PhD Candidate in Politics and International Relations, Monash University



Nielsen (2012) investigates why certain scholars of Islam become **Jihadi**: i.e. why they encourage armed struggle (especially against the west)

Requires that he first classifies scholars as **Jihadi** and  $\neg$  **Jihadi**: has 27,142 texts from 101 clerics, and difficult to do by hand.



# Example: Jihadi Clerics

## Indonesian cleric's support for ISIS increases the security threat

July 20, 2014 10:14pm EDT

Noor Huda Ismail

PhD Candidate in Politics and International Relations, Monash University



Nielsen (2012) investigates why certain scholars of Islam become **Jihadi**: i.e. why they encourage armed struggle (especially against the west)

Requires that he first classifies scholars as **Jihadi** and  $\neg$  **Jihadi**: has 27,142 texts from 101 clerics, and difficult to do by hand.

# Jihadi Clerics

# Jihadi Clerics

Training set:

# Jihadi Clerics

Training set: self-identified Jihadi texts (765),

# Jihadi Clerics

**Training** set: self-identified Jihadi texts (765), and sample from Islamic website as  $\neg$  Jihadi (1951)

# Jihadi Clerics

**Training** set: self-identified Jihadi texts (765), and sample from Islamic website as  $\neg$  Jihadi (1951)

**Preprocess**: drops terms occurring in less than 10%, or more than 40% of documents,

# Jihadi Clerics

**Training** set: self-identified Jihadi texts (765), and sample from Islamic website as  $\neg$  Jihadi (1951)

**Preprocess**: drops terms occurring in less than 10%, or more than 40% of documents, and uses 'light' stemmer for Arabic

# Jihadi Clerics

**Training** set: self-identified Jihadi texts (765), and sample from Islamic website as  $\neg$  Jihadi (1951)

**Preprocess**: drops terms occurring in less than 10%, or more than 40% of documents, and uses 'light' stemmer for Arabic

Can assign a *Jihad Score* to each document: basically the logged likelihood ratio,  $\sum_i \log \frac{\Pr(t_k | \text{Jihad})}{\Pr(t_k | \neg \text{Jihad})}$  (note: doesn't know what 'real world' priors are, so drops them here)



# Jihadi Clerics

**Training** set: self-identified Jihadi texts (765), and sample from Islamic website as  $\neg$  Jihadi (1951)

**Preprocess**: drops terms occurring in less than 10%, or more than 40% of documents, and uses 'light' stemmer for Arabic

Can assign a *Jihad Score* to each document: basically the logged likelihood ratio,  $\sum_i \log \frac{\Pr(t_k | \text{Jihad})}{\Pr(t_k | \neg \text{Jihad})}$  (note: doesn't know what 'real world' priors are, so drops them here)

Then for each cleric,

# Jihadi Clerics

**Training** set: self-identified Jihadi texts (765), and sample from Islamic website as  $\neg$  Jihadi (1951)

**Preprocess**: drops terms occurring in less than 10%, or more than 40% of documents, and uses 'light' stemmer for Arabic

Can assign a *Jihad Score* to each document: basically the logged likelihood ratio,  $\sum_i \log \frac{\Pr(t_k | \text{Jihad})}{\Pr(t_k | \neg \text{Jihad})}$  (note: doesn't know what 'real world' priors are, so drops them here)

Then for each cleric, **concatenate all works** into **one** and give this 'document'/cleric a score.

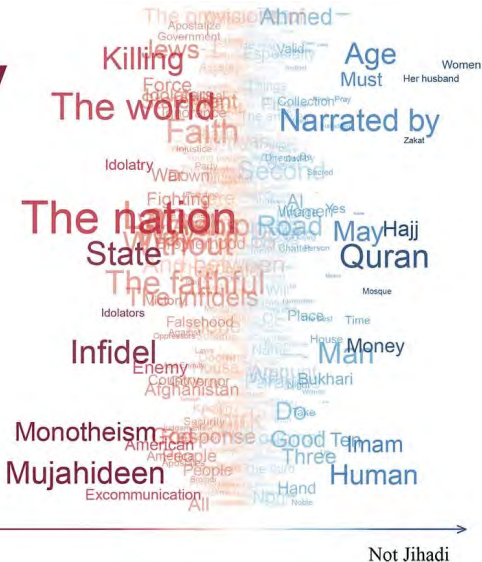
# Discriminating Words

## Discriminating Words

# Apostasy

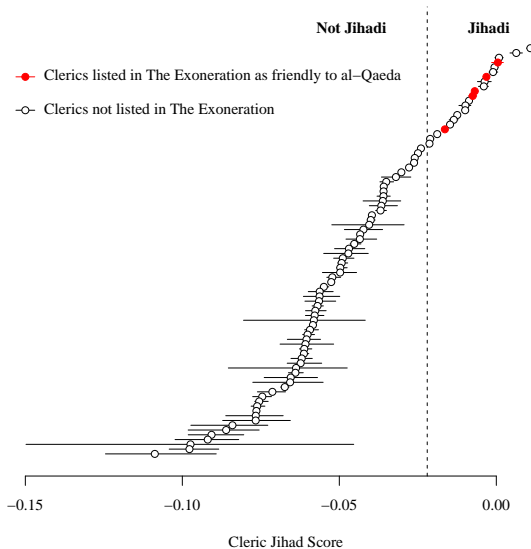
# Jihad

## Word Frequency

$$a = 1/250$$
$$a = 1/500$$
$$a = 1/1000$$
$$a = 1/2000$$


# Validation: *Exoneration*

# Validation: *Exoneration*



**Figure 4.9:** Jihad Scores Predict Inclusion in *The Exoneration*