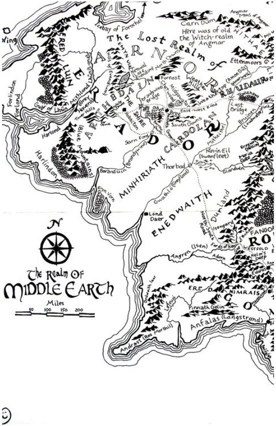
University of Tokyo: Text-as-Data
Day 2, Part I

Arthur Spirling

June 4, 2017

Where Are We?

Where Are We?


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

Where Are We?

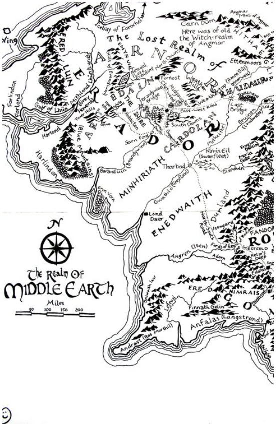
We've covered the basics of document representation and characterization.

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?

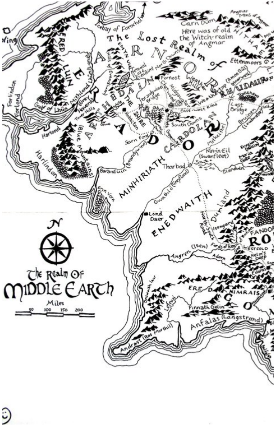
We've covered the basics of document representation and characterization.

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

→ simple, fast dictionary based ways to classify/categorize

cover some 'major' dictionaries in social science

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

$\rightarrow$ simple, fast dictionary based ways to classify/categorize

cover some 'major' dictionaries in social science

and move on to supervised 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

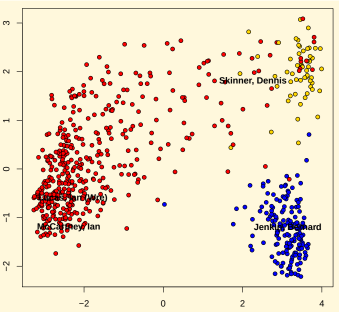
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

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.

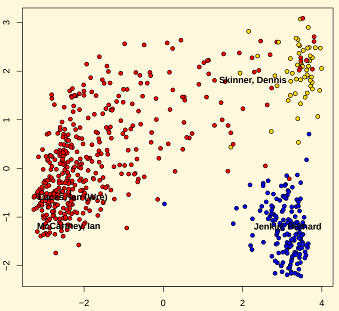
e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



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?

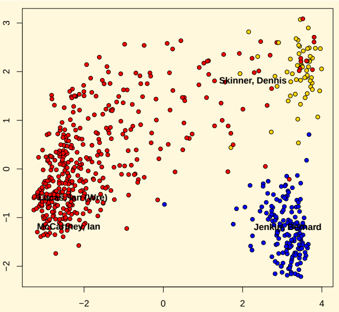


Supervised techniques:

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?

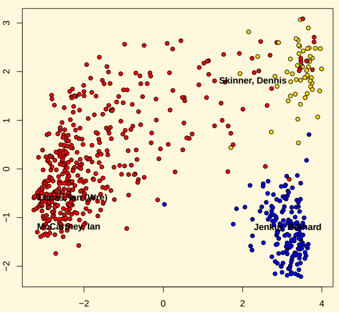


Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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?



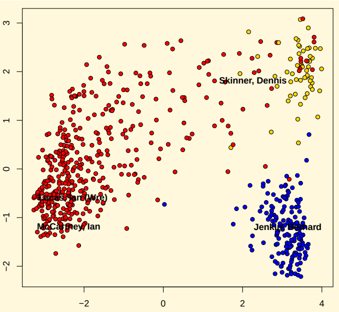
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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?



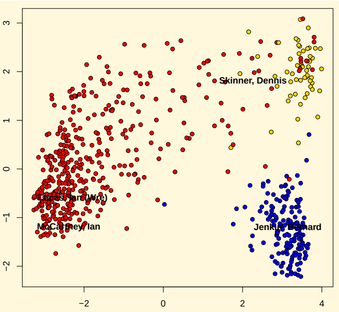
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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

Terminology

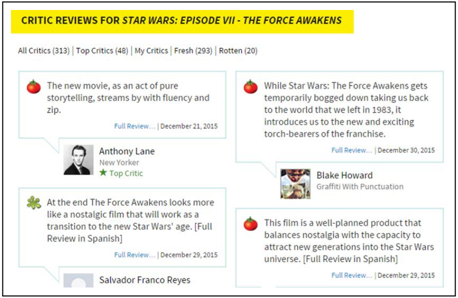
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

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 ($y \in \{0, 1\}$)?



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

Overview

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

Overview

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

Overview

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

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

$\rightarrow$ common in opinion mining/sentiment analysis, and in coding events or manifestos.

Overview

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

$\rightarrow$ common in opinion mining/sentiment analysis, and in coding events or manifestos.

Often derived from supervised learning techniques

Overview

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

$\rightarrow$ 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

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

$\rightarrow$ 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

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

$\rightarrow$ 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 $+1$

$\rightarrow$ 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 $+1$

$\rightarrow$ 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 $m = 1, \ldots, M$ 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 $+1$

$\rightarrow$ 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 $m = 1, \ldots, M$ in the dictionary,

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

More Specifically

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

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

$\rightarrow$ 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 $m = 1, \ldots, M$ in the dictionary,

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

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 $+1$

$\rightarrow$ 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 $m = 1, \ldots, M$ in the dictionary,

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

where s_m is the score of word m

and w_{im} is the number of occurrences of the mth 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 $+1$

$\rightarrow$ 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 $m = 1, \ldots, M$ in the dictionary,

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

where s_m is the score of word m

and w_{im} is the number of occurrences of the mth 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 $+1$

$\rightarrow$ 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 $m = 1, \ldots, M$ in the dictionary,

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

where s_m is the score of word m

and w_{im} is the number of occurrences of the mth dictionary word in the document i

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

$\rightarrow$ 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 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 *savage*
crisis *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

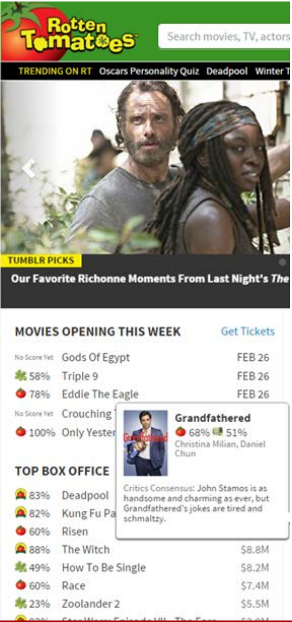
positive 2

total 13

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

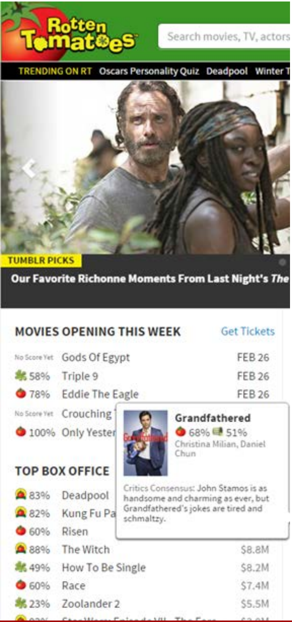
Partner Exercise

Partner Exercise



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

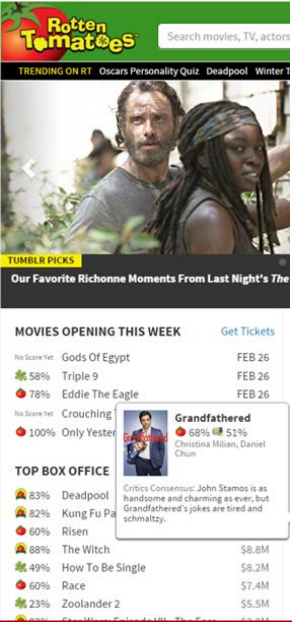
Partner Exercise



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

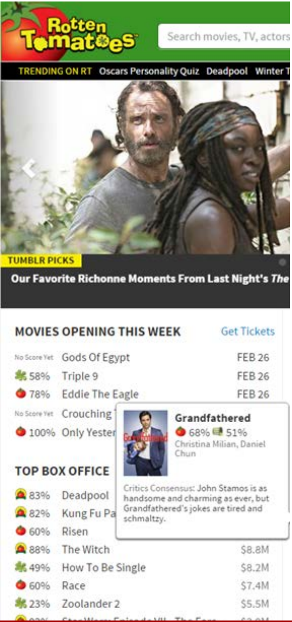


You are working for `rottentomatoes.com`, and want to automatically code (written) movie reviews as being between 1 and 5 stars.

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



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?

Notes

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka):

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued:

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued: could be continuous.

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued: could be continuous.

e.g. might want to differentiate '`good`' from '`great`' from '`best`'. Hard to come up with rules!

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued: could be continuous.

e.g. might want to differentiate 'good' from 'great' from 'best'. Hard to come up with rules!

NB Tone of the document can be presented as a continuous value,

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued: could be continuous.

e.g. might want to differentiate 'good' from 'great' from 'best'. Hard to come up with rules!

NB Tone of the document can be presented as a continuous value, or used to put documents in categories via some cutoff rule.

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued: could be continuous.

e.g. might want to differentiate 'good' from 'great' from 'best'. Hard to come up with rules!

NB Tone of the document can be presented as a continuous value, or used to put documents in categories via some cutoff rule.

e.g. all documents with tone>0 are deemed 'positive'

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued: could be continuous.

e.g. might want to differentiate 'good' from 'great' from 'best'. Hard to come up with rules!

NB Tone of the document can be presented as a continuous value, or used to put documents in categories via some cutoff rule.

e.g. all documents with tone>0 are deemed 'positive'

NB Bag-of-words assn may be especially dubious for some dictionary tasks

Notes

Typically assume that "every word contributes isomorphically" (Young & Saroka): each word in dictionary has one of two values and sum totals matter.

But no requirement that s_m be dichotomous or integer valued: could be continuous.

e.g. might want to differentiate 'good' from 'great' from 'best'. Hard to come up with rules!

NB Tone of the document can be presented as a continuous value, or used to put documents in categories via some cutoff rule.

e.g. all documents with tone>0 are deemed 'positive'

NB Bag-of-words assn may be especially dubious for some dictionary tasks

e.g. context matters: "was `not` good" gets $+1$!

Dictionaries I: General Inquirer

Dictionaries I: General Inquirer

Stone (1965) begins efforts to automatically analyze psychological states of authors

Dictionaries I: General Inquirer

Stone (1965) begins efforts to automatically analyze psychological states of authors

'General Inquirer' combines several dictionaries to make total of 182 categories:

Dictionaries I: General Inquirer

Stone (1965) begins efforts to automatically analyze psychological states of authors

'General Inquirer' combines several dictionaries to make total of 182 categories:

- Harvard IV-4 dictionary: psychology, themes, topics

Dictionaries I: General Inquirer

Stone (1965) begins efforts to automatically analyze psychological states of authors

'General Inquirer' combines several dictionaries to make total of 182 categories:

- Harvard IV-4 dictionary: psychology, themes, topics
- Lasswell dictionary: "commonsense categories of meaning", 8 basic value categories

Dictionaries I: General Inquirer

Stone (1965) begins efforts to automatically analyze psychological states of authors

'General Inquirer' combines several dictionaries to make total of 182 categories:

- Harvard IV-4 dictionary: psychology, themes, topics
- Lasswell dictionary: "commonsense categories of meaning", 8 basic value categories
- Semin and Fielder categories: interpersonal/pyschological properties of words

General Inquirer (selected)

General Inquirer (selected)

Entry	Source	Positiv	Negativ	Pstv	Affil	Ngtv	Hostile	Strong	Power
ABILITY	H4Lvd	Positiv						Strong	
ABJECT	H4		Negativ						
ABLE	H4Lvd	Positiv		Pstv				Strong	
ABNORMAL	H4Lvd		Negativ			Ngtv			
ABOARD	H4Lvd								
ABOLISH	H4Lvd		Negativ			Ngtv	Hostile	Strong	Power
ABOLITION	Lvd								
ABOMINABLE	H4		Negativ					Strong	
ABRASIVE	H4		Negativ				Hostile	Strong	
ABROAD	H4Lvd								
ABRUPT	H4Lvd		Negativ			Ngtv			
ABSCOND	H4		Negativ				Hostile		
ABSENCE	H4Lvd		Negativ						
ABSENT#1	H4Lvd		Negativ						
ABSENT#2	H4Lvd								
ABSENT-MINDED	H4		Negativ						
ABSENTEE	H4		Negativ				Hostile		
ABSOLUTE#1	H4Lvd							Strong	
ABSOLUTE#2	H4Lvd							Strong	

General Inquirer (selected)

Entry	Source	Positiv	Negativ	Pstv	Affil	Ngtv	Hostile	Strong	Power
ABILITY	H4Lvd	Positiv						Strong	
ABJECT	H4		Negativ						
ABLE	H4Lvd	Positiv		Pstv				Strong	
ABNORMAL	H4Lvd		Negativ			Ngtv			
ABOARD	H4Lvd								
ABOLISH	H4Lvd		Negativ			Ngtv	Hostile	Strong	Power
ABOLITION	Lvd								
ABOMINABLE	H4		Negativ					Strong	
ABRASIVE	H4		Negativ				Hostile	Strong	
ABROAD	H4Lvd								
ABRUPT	H4Lvd		Negativ			Ngtv			
ABSCOND	H4		Negativ				Hostile		
ABSENCE	H4Lvd		Negativ						
ABSENT#1	H4Lvd		Negativ						
ABSENT#2	H4Lvd								
ABSENT-MINDED	H4		Negativ						
ABSENTEE	H4		Negativ				Hostile		
ABSOLUTE#1	H4Lvd							Strong	
ABSOLUTE#2	H4Lvd							Strong	

provides dictionaries and software,

General Inquirer (selected)

Entry	Source	Positiv	Negativ	Pstv	Affil	Ngtv	Hostile	Strong	Power
ABILITY	H4Lvd	Positiv						Strong	
ABJECT	H4		Negativ						
ABLE	H4Lvd	Positiv		Pstv				Strong	
ABNORMAL	H4Lvd		Negativ			Ngtv			
ABOARD	H4Lvd								
ABOLISH	H4Lvd		Negativ			Ngtv	Hostile	Strong	Power
ABOLITION	Lvd								
ABOMINABLE	H4		Negativ					Strong	
ABRASIVE	H4		Negativ				Hostile	Strong	
ABROAD	H4Lvd								
ABRUPT	H4Lvd		Negativ			Ngtv			
ABSCOND	H4		Negativ				Hostile		
ABSENCE	H4Lvd		Negativ						
ABSENT#1	H4Lvd		Negativ						
ABSENT#2	H4Lvd								
ABSENT-MINDED	H4		Negativ						
ABSENTEE	H4		Negativ				Hostile		
ABSOLUTE#1	H4Lvd							Strong	
ABSOLUTE#2	H4Lvd							Strong	

provides dictionaries and software, which performs some stemming and disambiguation in terms of context

General Inquirer (selected)

Entry	Source	Positiv	Negativ	Pstv	Affil	Ngtv	Hostile	Strong	Power
ABILITY	H4Lvd	Positiv						Strong	
ABJECT	H4		Negativ						
ABLE	H4Lvd	Positiv		Pstv				Strong	
ABNORMAL	H4Lvd		Negativ			Ngtv			
ABOARD	H4Lvd								
ABOLISH	H4Lvd		Negativ			Ngtv	Hostile	Strong	Power
ABOLITION	Lvd								
ABOMINABLE	H4		Negativ					Strong	
ABRASIVE	H4		Negativ				Hostile	Strong	
ABROAD	H4Lvd								
ABRUPT	H4Lvd		Negativ			Ngtv			
ABSCOND	H4		Negativ				Hostile		
ABSENCE	H4Lvd		Negativ						
ABSENT#1	H4Lvd		Negativ						
ABSENT#2	H4Lvd								
ABSENT-MINDED	H4		Negativ						
ABSENTEE	H4		Negativ				Hostile		
ABSOLUTE#1	H4Lvd							Strong	
ABSOLUTE#2	H4Lvd							Strong	

provides dictionaries and software, which performs some stemming and disambiguation in terms of context

e.g. ADULT has two meanings: one is a ‘virtue’, one is a ‘role’

Bainbridge, “Personality Capture” (2014)

Bainbridge, “Personality Capture” (2014)

	Declaration of Independence	**‘Plymouth Rock and the Pilgrims’**
	Jefferson et al	**Mark Twain**

Bainbridge, "Personality Capture" (2014)

	Declaration of Independence	'Plymouth Rock and the Pilgrims'
	Jefferson et al	Mark Twain
Affiliation	4.7%	2.1%
Hostile	3.6%	1.1%
Power	8.5%	1.8%
Submission	2.1%	1.0%

Bainbridge, “Personality Capture” (2014)

	Declaration of Independence	‘Plymouth Rock and the Pilgrims’
	Jefferson et al	**Mark Twain**
Affiliation	4.7%	2.1%
Hostile	3.6%	1.1%
Power	8.5%	1.8%
Submission	2.1%	1.0%
Virtue	3.9%	2.7%
Vice	1.7%	1.1%
Overstated	5.6%	3.9%
Understated	0.6%	2.5%

Dictionaries II: Linguistic Inquiry and Word Count (LIWC)

Dictionaries II: Linguistic Inquiry and Word Count (LIWC)

Pennebaker et al,

Dictionaries II: Linguistic Inquiry and Word Count (LIWC)

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

Dictionaries II: 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 II: 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 II: 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 II: 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 II: 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 II: 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 II: 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.

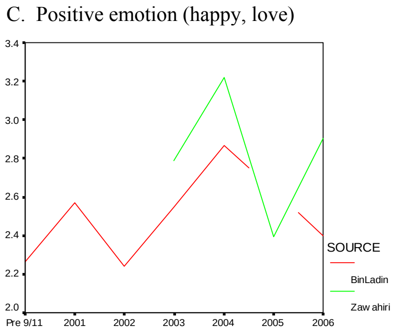
Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

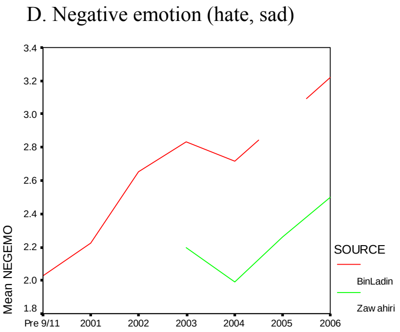
Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

“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. "





Application: Ramey, Klinger & Hollibaugh

Application: Ramey, Klinger & Hollibaugh

Mairesse et al. (2007) provide estimates of 'big 5' personality traits from LIWC categories

Application: Ramey, Klinger & Hollibaugh

Mairesse et al. (2007) provide estimates of 'big 5' personality traits from LIWC categories

Application: Ramey, Klinger & Hollibaugh

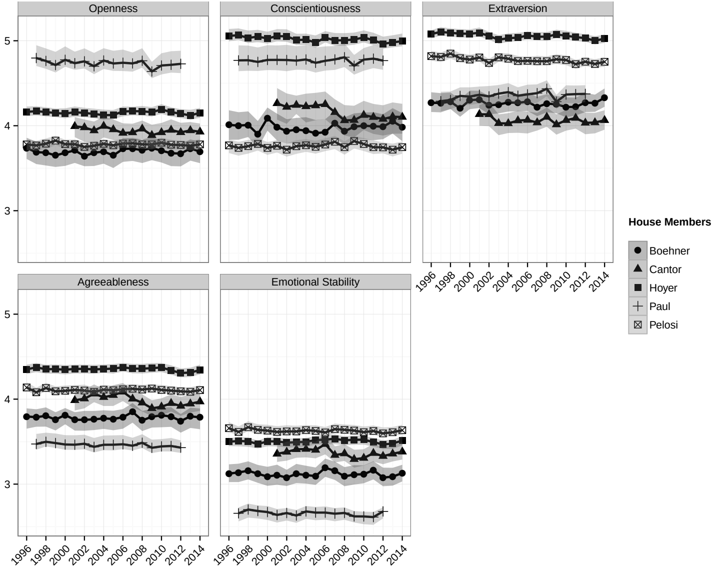
Mairesse et al. (2007) provide estimates of 'big 5' personality traits from LIWC categories

Ramey et al apply to Congressional speech.

Application: Ramey, Klinger & Hollibaugh

Mairesse et al. (2007) provide estimates of 'big 5' personality traits from LIWC categories

Ramey et al apply to Congressional speech.



Dictionaries III: Laver & Garry

Dictionaries III: Laver & Garry

2000 Laver and Garry create dictionary for manifestos where basic unit is strings of ~ 10 words in length.

Dictionaries III: Laver & Garry

2000 Laver and Garry create dictionary for manifestos where basic unit is strings of ~ 10 words in length.

$\rightarrow$ hierarchical, with topmost level pertaining to five policy domains: economy, political system, social system, external relations, 'other' (waffle)

Dictionaries III: Laver & Garry

2000 Laver and Garry create dictionary for manifestos where basic unit is strings of $\sim$ 10 words in length.

$\rightarrow$ hierarchical, with topmost level pertaining to five policy domains: economy, political system, social system, external relations, 'other' (waffle)

get good/valid results and high correlation with expert surveys.

Dictionaries III: Laver & Garry

2000 Laver and Garry create dictionary for manifestos where basic unit is strings of ∼ 10 words in length.

→ hierarchical, with topmost level pertaining to five policy domains: economy, political system, social system, external relations, 'other' (waffle)

get good/valid results and high correlation with expert surveys.

1 1 1	ECONOMY/+State+/Budget Budget
1 1 1 1	ECONOMY/+State+/Budget/Spending Increase public spending
1 1 1 1 1	ECONOMY/+State+/Budget/Spending/Health
1 1 1 1 2	ECONOMY/+State+/Budget/Spending/Educ. and training

Dictionaries IV: Hu & Liu

Dictionaries IV: Hu & Liu

2004 Hu and Liu ("Mining and Summarizing Customer Reviews")

Dictionaries IV: Hu & Liu

2004 Hu and Liu ("Mining and Summarizing Customer Reviews") provide 6800 words which are positive and negative derived from `amazon.com` and others.

Dictionaries IV: Hu & Liu

2004 Hu and Liu ("Mining and Summarizing Customer Reviews") provide 6800 words which are positive and negative derived from `amazon.com` and others.

Dictionaries IV: Hu & Liu

2004 Hu and Liu ("Mining and Summarizing Customer Reviews") provide 6800 words which are positive and negative derived from `amazon.com` and others.

1,036 of 1,144 people found the following review helpful

★★★★★ **With Great Powers Comes Great Responsibility**

By Tommy H. on July 17, 2009

I admit it, I'm a ladies' man. And when you put this shirt on a ladies' man, it's like giving an AK-47 to a ninja. Sure it looks cool and probably would make for a good movie, but you know somebody is probably going to get hurt in the end (no pun intended). That's what almost happened to me, this is my story...

Being Careful...

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

$\rightarrow$ matter of adding extra words in the various categories.

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

$\rightarrow$ matter of adding extra words in the various categories.

But much care is needed when a dictionary designed for one context is applied to another.

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

→ matter of adding extra words in the various categories.

But much care is needed when a dictionary designed for one context is applied to another.

e.g. Loughran & MacDonald, 2011: common dictionaries fail badly when applied to financial texts

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

→ matter of adding extra words in the various categories.

But much care is needed when a dictionary designed for one context is applied to another.

e.g. Loughran & MacDonald, 2011: common dictionaries fail badly when applied to financial texts—e.g. `cost` is a neutral term in reports, but negative in Harvard IV

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

→ matter of adding extra words in the various categories.

But much care is needed when a dictionary designed for one context is applied to another.

e.g. Loughran & MacDonald, 2011: common dictionaries fail badly when applied to financial texts—e.g. `cost` is a neutral term in reports, but negative in Harvard IV

plus virtually impossible to validate dictionaries: very expensive,

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

→ matter of adding extra words in the various categories.

But much care is needed when a dictionary designed for one context is applied to another.

e.g. Loughran & MacDonald, 2011: common dictionaries fail badly when applied to financial texts—e.g. `cost` is a neutral term in reports, but negative in Harvard IV

plus virtually impossible to validate dictionaries: very expensive, at least.

Being Careful...

In principle, it is straightforward to extend dictionary from one domain to another

→ matter of adding extra words in the various categories.

But much care is needed when a dictionary designed for one context is applied to another.

e.g. Loughran & MacDonald, 2011: common dictionaries fail badly when applied to financial texts—e.g. `cost` is a neutral term in reports, but negative in Harvard IV

plus virtually impossible to validate dictionaries: very expensive, at least.

btw humans *not* very good at producing discriminating terms for e.g. opinion mining (Pang et al, 2002)

Events, dear boy...

Events, dear boy…

Scholars of International Relations need access to events

Events, dear boy. . .

Scholars of International Relations need access to events

Real time media reports are obvious source. . .

Events, dear boy. . .

Scholars of International Relations need access to events

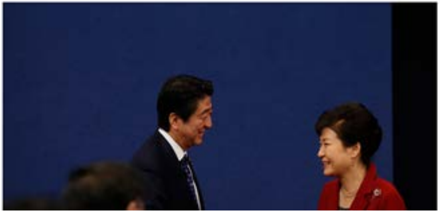
Real time media reports are obvious source. . .

ASIA PACIFIC

Leaders of South Korea and Japan Meet in Effort to Mend Ties

点击查看本文中文版 | Read in Chinese

By CHOE SANG-HUN NOV. 1, 2015



Events, dear boy...

Scholars of International Relations need access to events

Real time media reports are obvious source...

Yet need to be coded automatically to be helpful.

Events, dear boy. . .

Scholars of International Relations need access to events

Real time media reports are obvious source. . .


ASIA PACIFIC

Leaders of South Korea and Japan Meet in Effort to Mend Ties

点击查看本文中文版 | Read in Chinese

By CHOE SANG-HUN NOV. 1, 2015

Yet need to be coded automatically to be helpful.

Partner Exercise

Partner Exercise

Consider compiling an events data set (by hand or by machine) from press reports.

Partner Exercise

Consider compiling an events data set (by hand or by machine) from press reports.

1 Would you prefer to have the universe of press reports from the United States, or the United Kingdom or North Korea? Why, and what does that tell you about journalistic agendas?

Partner Exercise

Consider compiling an events data set (by hand or by machine) from press reports.

1 Would you prefer to have the universe of press reports from the United States, or the United Kingdom or North Korea? Why, and what does that tell you about journalistic agendas?

2 Sports reports cause problems for automatic event extraction. Why?

Partner Exercise

Consider compiling an events data set (by hand or by machine) from press reports.

1. Would you prefer to have the universe of press reports from the United States, or the United Kingdom or North Korea? Why, and what does that tell you about journalistic agendas?
2. Sports reports cause problems for automatic event extraction. Why?

Premise and Resources

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 TABARI —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed (‘lead’) contains. . .

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 TABARI —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed ('lead') contains. . .

source of event,

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed (‘lead’) contains. . .
source of event, subject of sentence

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 TABARI —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed ('lead') contains...

source of event, subject of sentence

target of event,

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed ('lead') contains...

source of event, subject of sentence

target of event, object of sentence (direct or indirect)

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed ('lead') contains...

source of event, subject of sentence

target of event, object of sentence (direct or indirect)

type of event,

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed ('lead') contains...

source of event, subject of sentence

target of event, object of sentence (direct or indirect)

type of event, transitive verb of sentence

Premise and Resources

1994 Philip Schrodt develops Kansas Event Data System

2000 `TABARI` —Textual Analysis by Augmeted Replacement Instructions—open source.

also many related products, including CAMEO dealing specifically with mediation

while Virtual Research Associates Reader VRA is proprietary version.

idea first sentence of Reuters news feed ('lead') contains...

source of event, subject of sentence

target of event, object of sentence (direct or indirect)

type of event, transitive verb of sentence

Use and Example (Lowe & King, 2003)

Use and Example (Lowe & King, 2003)

Russian artillery^{S} south of the Chechen capital Grozny blasted^{223} Chechen positions^{T} overnight before falling silent at dawn, witnesses said on Tuesday

Use and Example (Lowe & King, 2003)

Russian artillery^{S} south of the Chechen capital Grozny blasted^{223} Chechen positions^{T} overnight before falling silent at dawn, witnesses said on Tuesday

Use and Example (Lowe & King, 2003)

```
Russian artillery^S south of the Chechen capital
Grozny blasted^223 Chechen positions^T overnight
before falling silent at dawn, witnesses said on
Tuesday
```

S is the source

Use and Example (Lowe & King, 2003)

Russian artillery^{S} south of the Chechen capital Grozny blasted^{223} Chechen positions^{T} overnight before falling silent at dawn, witnesses said on Tuesday

S is the source

T is the target

Use and Example (Lowe & King, 2003)

> `Russian artillery`^{S} `south of the Chechen capital Grozny` `blasted`^{223} `Chechen positions`^{T} `overnight before falling silent at dawn, witnesses said on Tuesday`

S is the source

T is the target

223 is the code of the event between them

Hierarchical Coding Scheme (CAMEO)/Dictionary

Hierarchical Coding Scheme (CAMEO)/Dictionary

12: REJECT

120: Reject, not specified below

121: Reject material cooperation

- 1211: Reject economic cooperation
- 1212: Reject military cooperation

122: Reject request or demand for material aid, not specified below

- 1221: Reject request for economic aid
- 1222: Reject request for military aid
- 1223: Reject request for humanitarian aid
- 1224: Reject request for military protection or peacekeeping

Hierarchical Coding Scheme (CAMEO)/Dictionary

12: REJECT

120: Reject, not specified below

121: Reject material cooperation

- 1211: Reject economic cooperation
- 1212: Reject military cooperation

122: Reject request or demand for material aid, not specified below

- 1221: Reject request for economic aid
- 1222: Reject request for military aid
- 1223: Reject request for humanitarian aid
- 1224: Reject request for military protection or peacekeeping

CAMEO	**1222**
Name	**Reject request for military aid**
Description	Refuse to extend military assistance.
Example	The Turkish government has refused to commit to any direct assistance to the US-led war against Iraq, citing domestic opposition.

Actors (CAMEO)/Dictionary

Actors (CAMEO)/Dictionary

UGAREBLRA	Lord's Resistance Army
UIG	Uighur (Chinese ethnic minority)
UIS	Unidentified state actors
UKR	Ukraine
URY	Uruguay
USA	United States
USR	Union of Soviet Socialist Republics (USSR)
UZB	Uzbekistan
VAT	Holy See (Vatican City)
VCT	Saint Vincent and the Grenadines
VEN	Venezuela
VGB	British Virgin Islands

Actors (CAMEO)/Dictionary

UGAREBLRA	Lord's Resistance Army
UIG	Uighur (Chinese ethnic minority)
UIS	Unidentified state actors
UKR	Ukraine
URY	Uruguay
USA	United States
USR	Union of Soviet Socialist Republics (USSR)
UZB	Uzbekistan
VAT	Holy See (Vatican City)
VCT	Saint Vincent and the Grenadines
VEN	Venezuela
VGB	British Virgin Islands

Delving More Deeply

Delving More Deeply

- Begins with basic parsing:

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.
- Much effort to disambiguate:

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as 'he' in subsequent sentences

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as '`he`' in subsequent sentences

  Synonyms (and metonyms!) also require dictionaries (`WordNet`).

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as '`he`' in subsequent sentences

  Synonyms (and metonyms!) also require dictionaries (`WordNet`).

  e.g. '`US`', '`American`'

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as '`he`' in subsequent sentences

  Synonyms (and metonyms!) also require dictionaries (`WordNet`).

  e.g. `'US'`, `'American'` (`'US'`, `'Washington'`)

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as '`he`' in subsequent sentences

  Synonyms (and metonyms!) also require dictionaries (`WordNet`).

  e.g. `'US'`, `'American'` (`'US'`, `'Washington'`)

  Care over verb/noun problems.

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as '`he`' in subsequent sentences

  Synonyms (and metonyms!) also require dictionaries (`WordNet`).

  e.g. '`US`', '`American`' ('`US`', '`Washington`')

  Care over verb/noun problems.

  e.g. '`attack`' as noun and verb

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as '`he`' in subsequent sentences

  Synonyms (and metonyms!) also require dictionaries (`WordNet`).

  e.g. '`US`', '`American`' ('`US`', '`Washington`')

  Care over verb/noun problems.

  e.g. '`attack`' as noun and verb

- Excellent performance relative to human coders (Lowe & King, 2003):

Delving More Deeply

- Begins with basic parsing: POS, stemming, stop words etc.

- Much effort to disambiguate:

  Use of pronouns causes problems.

  e.g. President is referred to as '`he`' in subsequent sentences

  Synonyms (and metonyms!) also require dictionaries (`WordNet`).

  e.g. '`US`', '`American`' ('`US`', '`Washington`')

  Care over verb/noun problems.

  e.g. '`attack`' as noun and verb

- Excellent performance relative to human coders (Lowe & King, 2003): both in terms of reliability and validity.

Example: Dayton Peace Accords

Example: Dayton Peace Accords

Yugoslavia breaking up; Bosnian War

Example: Dayton Peace Accords

Yugoslavia breaking up; Bosnian War

$\rightarrow$ multiple peace attempts failed,

Example: Dayton Peace Accords

Yugoslavia breaking up; Bosnian War

→ multiple peace attempts failed,

Until US put intense pressure on parties.

Example: Dayton Peace Accords

Yugoslavia breaking up; Bosnian War

→ multiple peace attempts failed,

Until US put intense pressure on parties.

→ can we see this in automatic mediation estimates?

Example: Dayton Peace Accords

Yugoslavia breaking up; Bosnian War

→ multiple peace attempts failed,

Until US put intense pressure on parties.

→ can we see this in automatic mediation estimates?

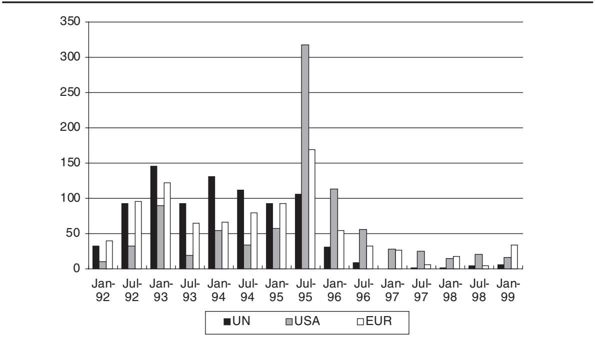


Figure 3: Six-Month Totals of Mediation Events in the Balkans by Mediator
NOTE: UN = United Nations; USA = United States; EUR = major European states, plus the European Union.

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.

Discrimination

Discrimination

So Once researcher has *extreme* examples of text,

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them...

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them...

$\rightarrow$ these words then become scored as part of the dictionary/thesaurus.

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them. . .

$\rightarrow$ these words then become scored as part of the dictionary/thesaurus. Can use `WordNet` to find synonyms.

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them...

→ these words then become scored as part of the dictionary/thesaurus. Can use `WordNet` to find synonyms.

2013 Taddy provides *Multinomial Inverse Regression* to dimension reduce text, and make outcomes a product of that (reduced) set of Xs

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them...

→ these words then become scored as part of the dictionary/thesaurus. Can use `WordNet` to find synonyms.

2013 Taddy provides *Multinomial Inverse Regression* to dimension reduce text, and make outcomes a product of that (reduced) set of Xs

→ can be used to produce key predictors/keywords that discriminate in terms of categories.

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them...

→ these words then become scored as part of the dictionary/thesaurus. Can use `WordNet` to find synonyms.

2013 Taddy provides *Multinomial Inverse Regression* to dimension reduce text, and make outcomes a product of that (reduced) set of Xs

→ can be used to produce key predictors/keywords that discriminate in terms of categories.

2009 Monroe, Colaresi & Quinn consider ways to capture partisan differences in speech,

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them...

→ these words then become scored as part of the dictionary/thesaurus. Can use `WordNet` to find synonyms.

2013 Taddy provides *Multinomial Inverse Regression* to dimension reduce text, and make outcomes a product of that (reduced) set of Xs

→ can be used to produce key predictors/keywords that discriminate in terms of categories.

2009 Monroe, Colaresi & Quinn consider ways to capture partisan differences in speech, and suggest Bayesian shrinkage estimator approach.

Discrimination

So Once researcher has *extreme* examples of text, various methods to identify the words that discriminate between them...

→ these words then become scored as part of the dictionary/thesaurus. Can use `WordNet` to find synonyms.

2013 Taddy provides *Multinomial Inverse Regression* to dimension reduce text, and make outcomes a product of that (reduced) set of Xs

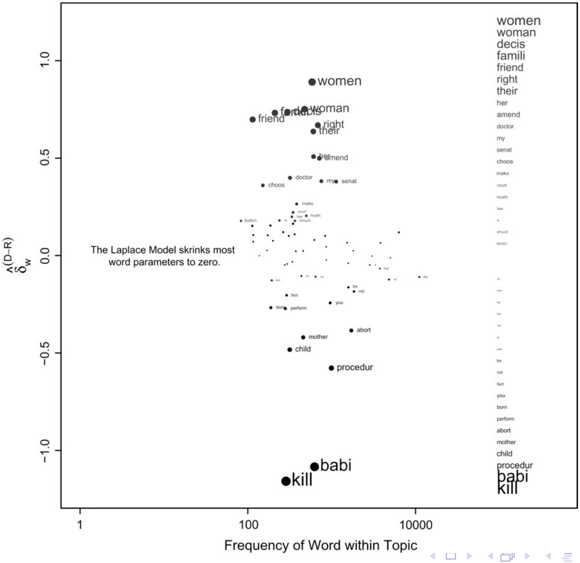
→ can be used to produce key predictors/keywords that discriminate in terms of categories.

2009 Monroe, Colaresi & Quinn consider ways to capture partisan differences in speech, and suggest Bayesian shrinkage estimator approach.

→ previous approaches tend to overfit to obscure words or groups that don't have much validity in context.

Most Democratic and Republican Words on Abortion (106th, Laplace prior)

Most Democratic and Republican Words on Abortion (106th, Laplace prior)



Supervised Learning

Naive Bayes Classification

Naive Bayes Classification

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

Naive Bayes Classification

Motivation: emails d arrive and must 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 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,

Motivation: emails d arrive and must 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 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 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.

$\rightarrow$ 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

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

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}:

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

Estimation Notes I

Estimation Notes I

The ‘hats’ appear because neither $\widehat{\Pr(c)}$ nor $\widehat{\Pr(t_k|c)}$ are known.

Estimation Notes I

The 'hats' appear because neither $\widehat{\Pr(c)}$ nor $\widehat{\Pr(t_k|c)}$ are known.

$\rightarrow$ they are (can be) estimated from the training set.

Estimation Notes I

The 'hats' appear because neither $\widehat{\Pr(c)}$ nor $\widehat{\Pr(t_k|c)}$ are known.

$\rightarrow$ they are (can be) estimated from the training set.

We can use $\frac{N_c}{N}$ for $\widehat{\Pr(c)}$,

Estimation Notes I

The 'hats' appear because neither $\widehat{\Pr(c)}$ nor $\widehat{\Pr(t_k|c)}$ are known.

$\rightarrow$ they are (can be) estimated from the training set.

We can use $\frac{N_c}{N}$ for $\widehat{\Pr(c)}$, where N_c is the number of documents in class c in our training set (MLE).

Estimation Notes I

The 'hats' appear because neither $\widehat{\Pr(c)}$ nor $\widehat{\Pr(t_k|c)}$ are known.

$\rightarrow$ they are (can be) estimated from the training set.

We can use $\frac{N_c}{N}$ for $\widehat{\Pr(c)}$, where N_c is the number of documents in class c in our training set (MLE).

We can use $\frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$ for $\widehat{\Pr(t_k|c)}$ (MLE).

Estimation Notes I

The 'hats' appear because neither $\widehat{\Pr(c)}$ nor $\widehat{\Pr(t_k|c)}$ are known.

$\rightarrow$ they are (can be) estimated from the training set.

We can use $\frac{N_c}{N}$ for $\widehat{\Pr(c)}$, where N_c is the number of documents in class c in our training set (MLE).

We can use $\frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$ for $\widehat{\Pr(t_k|c)}$ (MLE).

here T_{ct} is the number of occurrences of t in training documents that come from class c, including multiple occurrences.

Estimation Notes I

The 'hats' appear because neither $\widehat{\Pr(c)}$ nor $\widehat{\Pr(t_k|c)}$ are known.

$\rightarrow$ they are (can be) estimated from the training set.

We can use $\frac{N_c}{N}$ for $\widehat{\Pr(c)}$, where N_c is the number of documents in class c in our training set (MLE).

We can use $\frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$ for $\widehat{\Pr(t_k|c)}$ (MLE).

here T_{ct} is the number of occurrences of t in training documents that come from class c, including multiple occurrences.

and denominator is the total number all terms in the training documents in c.

Example

Example

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

Example

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

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

$$\Pr(\text{prince}|\text{spam}) = \tfrac{2}{6}$$
$$\Pr(\text{prince}|\text{spam}) = \tfrac{2}{6}$$
$$\Pr(\text{money}|\text{spam}) = \tfrac{1}{6}$$
$$\Pr(\text{spam}|\text{d}) \propto \tfrac{2}{5}\tfrac{2}{6}\tfrac{2}{6}\tfrac{1}{6} = 0.0074$$

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	?

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

$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$
$$\Pr(\text{prince}|\text{spam}) = \frac{2}{6}$$
$$\Pr(\text{money}|\text{spam}) = \frac{1}{6}$$
$$\Pr(\text{spam}|\text{d}) \propto \frac{2}{5}\frac{2}{6}\frac{2}{6}\frac{1}{6} = 0.0074$$

$$\rightarrow \boxed{c_{map} = \text{spam}}$$

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t.

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t.

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t. This is a multinomial NB model.

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t. This is a multinomial NB model.

aside Could have $\widehat{\Pr(t_k|c)}$ as fraction of documents containing t,

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t. This is a multinomial NB model.

aside Could have $\widehat{\Pr(t_k|c)}$ as fraction of documents containing t, which implies Bernoulli model (ignores number of occurrences).

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t. This is a multinomial NB model.

aside Could have $\widehat{\Pr(t_k|c)}$ as fraction of documents containing t, which implies Bernoulli model (ignores number of occurrences).

As usual,

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t. This is a multinomial NB model.

aside Could have $\widehat{\Pr(t_k|c)}$ as fraction of documents containing t, which implies Bernoulli model (ignores number of occurrences).

As usual, working with products of probabilities is difficult computationally

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t. This is a multinomial NB model.

aside Could have $\widehat{\Pr(t_k|c)}$ as fraction of documents containing t, which implies Bernoulli model (ignores number of occurrences).

As usual, working with products of probabilities is difficult computationally

$\rightarrow$ take logs:

Estimation Notes II

$\widehat{\Pr(t_k|c)}$ is the fraction of tokens in documents from class c that are t. Can also see it as fraction of positions in documents from class c that contain term t. This is a multinomial NB model.

aside Could have $\widehat{\Pr(t_k|c)}$ as fraction of documents containing t, which implies Bernoulli model (ignores number of occurrences).

As usual, working with products of probabilities is difficult computationally

$\rightarrow$ take logs:

$$c_{map} = \arg\max_c \left[\log \widehat{\Pr(c)} + \sum \log \widehat{\Pr(t_k|c)}\right]$$

Estimation Notes III

Estimation Notes III

Sparsity can be a problem in the training set.

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails,

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word '`cost`' (but it does occur in the ham set),

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word '`cost`' (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word '`cost`' (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word `cost` (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

$\rightarrow$ well, $\widehat{\Pr(t_k|c)} = \widehat{\Pr(\text{`cost'}|\text{spam})} = 0$.

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word ‘`cost`’ (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

$\rightarrow$ well, $\widehat{\Pr(t_k|c)} = \widehat{\Pr(\text{`cost'}|\text{spam})} = 0$. And that will be multiplied into the product.

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word ‘`cost`’ (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

→ well, $\widehat{\Pr(t_k|c)} = \Pr(\widehat{\text{`cost'}|}\text{spam}) = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word ‘cost’ (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

→ well, $\widehat{\Pr(t_k|c)} = \Pr(\text{`}\widehat{\text{cost'}|}\text{spam}) = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

So may want to add one to each count:

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word '`cost`' (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

→ well, $\widehat{\Pr(t_k|c)} = \Pr(\text{`}\widehat{\text{cost'}|}\text{spam}) = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

So may want to add one to each count: $\frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)}$

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word '`cost`' (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

$\rightarrow$ well, $\widehat{\Pr(t_k|c)} = \widehat{\Pr(\text{'cost'}|}\text{spam}) = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

So may want to add one to each count: $\frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)}$ to avoid wiping out the products (or causing problems for taking logs).

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word '`cost`' (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

$\rightarrow$ well, $\widehat{\Pr(t_k|c)} = \widehat{\Pr(\text{`cost'}|\text{spam})} = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

So may want to add one to each count: $\frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)}$ to avoid wiping out the products (or causing problems for taking logs). Equivalent to adding size of the vocabulary to the counts within the class.

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word '`cost`' (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What's the probability that email is spam?

$\rightarrow$ well, $\widehat{\Pr(t_k|c)} = \widehat{\Pr(\text{'cost'}|\text{spam})} = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

So may want to add one to each count: $\frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)}$ to avoid wiping out the products (or causing problems for taking logs). Equivalent to adding size of the vocabulary to the counts within the class.

$\rightarrow$ Laplace smoothing,

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word ‘`cost`’ (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What’s the probability that email is spam?

→ well, $\widehat{\Pr(t_k|c)} = \widehat{\Pr(\text{‘cost’}|\text{spam})} = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

So may want to add one to each count: $\frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)}$ to avoid wiping out the products (or causing problems for taking logs). Equivalent to adding size of the vocabulary to the counts within the class.

→ Laplace smoothing, equivalent to a uniform prior on term (each term occurs once for each class).

Estimation Notes III

Sparsity can be a problem in the training set. Suppose, in our training set of spam emails, we never see the word ‘`cost`’ (but it does occur in the ham set), and it shows up in our actual email tomorrow.

Q What’s the probability that email is spam?

→ well, $\widehat{\Pr(t_k|c)} = \widehat{\Pr(\text{‘cost’}|\text{spam})} = 0$. And that will be multiplied into the product. So, $\Pr(\text{spam}|d) = 0$.

So may want to add one to each count: $\frac{T_{ct}+1}{\sum_{t' \in V}(T_{ct'}+1)}$ to avoid wiping out the products (or causing problems for taking logs). Equivalent to adding size of the vocabulary to the counts within the class.

→ Laplace smoothing, equivalent to a uniform prior on term (each term occurs once for each class). Use slightly different smoother for Bernoulli case.

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 as 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 as 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 as 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 as 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 as 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 as 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 as 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 ¬ 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 ¬ 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 ¬ 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)

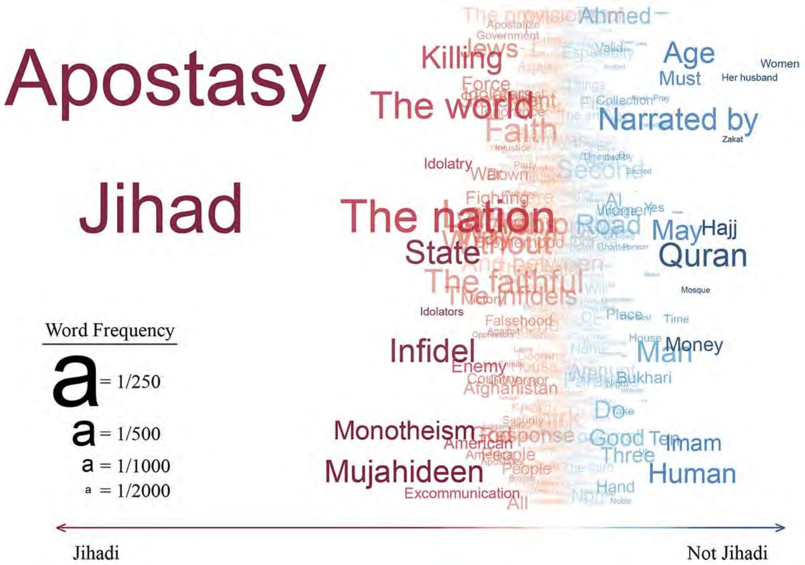
Preprocess: drops terms occurring in less than 10%, or more than 40% of documents, and uses 'light' stemmer for Arabic

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



Validation: *Exoneration*

Validation: *Exoneration*

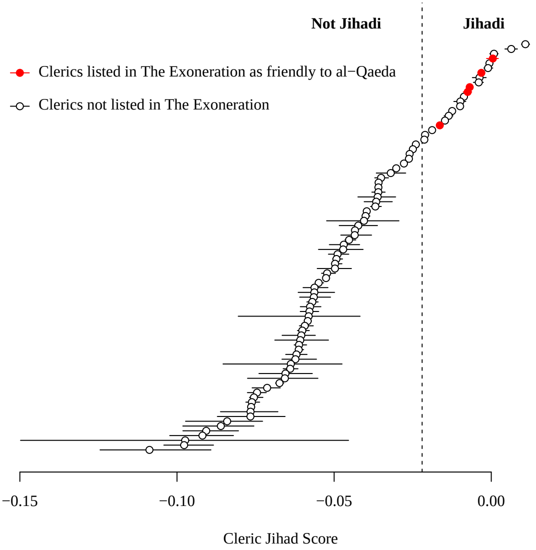


Figure 4.9: *Jihad Scores Predict Inclusion in The Exoneration*

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 NB style, so that we can answer such questions.

Basics

Basics

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

Basics

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

Basics

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

Basics

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 $P_{iR} = \frac{0.025}{0.025+0.05}$

Example

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

then $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

Example

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

then $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05}$

Example

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

then $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

Example

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

then $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

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 $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

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 $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

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 $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

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 $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

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 $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

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 $P_{iR} = \frac{0.025}{0.025+0.05} = 0.83.$

and $P_{iL} = \frac{0.005}{0.025+0.05} = 0.16.$

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 $0.02 \times 0.66 = 0.0132$.

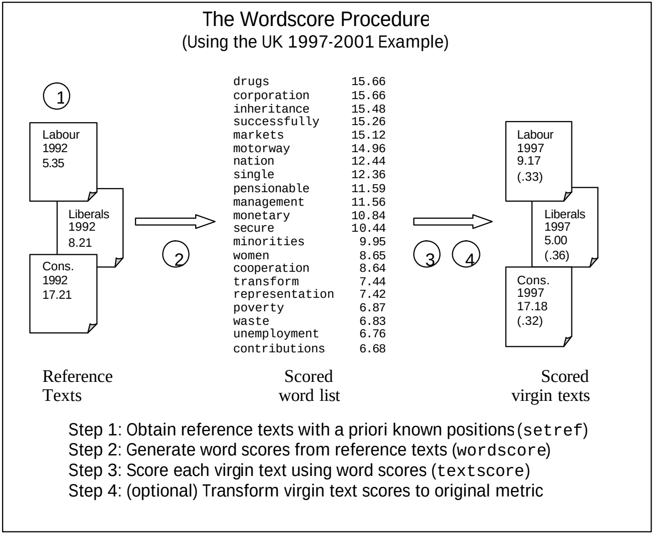
but virgin manifesto, from Labour party, mentions it 10 times in a thousand words: $0.01 \times 0.66 = 0.006$

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

New Labour Moderates its Economic Policy

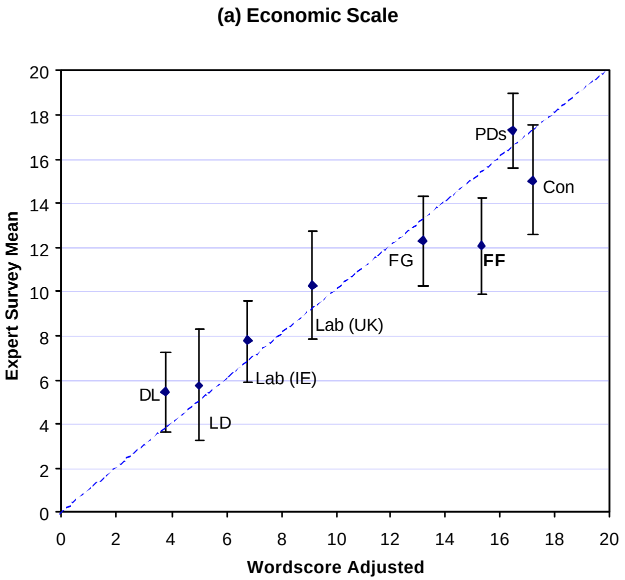
New Labour Moderates its Economic Policy

New Labour Moderates its Economic Policy



Compared to Expert Surveys

Compared to Expert Surveys



Comments

Comments

Extremely influential approach:

Comments

Extremely influential approach: avoids having to pick features of interest

Comments

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,

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.

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 (typically?) unhappy (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 (typically?) unhappy (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 (typically?) unhappy (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 (typically?) unhappy (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).

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 (typically?) unhappy (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).

while Beauchamp (2011) provides comparison and extension to more purely Bayesian approach.

Special Topic: Estimating Proportions

Estimating Proportions, Hopkins & King (2010)

Estimating Proportions, Hopkins & King (2010)

Suppose you want to know the proportion of e.g. blog posts or Facebook updates that are sympathetic to Trump.

Estimating Proportions, Hopkins & King (2010)

Suppose you want to know the proportion of e.g. blog posts or Facebook updates that are sympathetic to Trump.

$\rightarrow$ could train a (Naive Bayes) classifier on documents,

Estimating Proportions, Hopkins & King (2010)

Suppose you want to know the proportion of e.g. blog posts or Facebook updates that are sympathetic to Trump.

→ could train a (Naive Bayes) classifier on documents, and then calculate the proportion of the test set that fits into the class of interest.

Estimating Proportions, Hopkins & King (2010)

Suppose you want to know the proportion of e.g. blog posts or Facebook updates that are sympathetic to Trump.

→ could train a (Naive Bayes) classifier on documents, and then calculate the proportion of the test set that fits into the class of interest.

But problematic when labeled set is not random sample of population

Estimating Proportions, Hopkins & King (2010)

Suppose you want to know the proportion of e.g. blog posts or Facebook updates that are sympathetic to Trump.

→ could train a (Naive Bayes) classifier on documents, and then calculate the proportion of the test set that fits into the class of interest.

But problematic when labeled set is not random sample of population (perhaps due to 'drift'—sample collected once, and population moves on)

Estimating Proportions, Hopkins & King (2010)

Suppose you want to know the proportion of e.g. blog posts or Facebook updates that are sympathetic to Trump.

$\rightarrow$ could train a (Naive Bayes) classifier on documents, and then calculate the proportion of the test set that fits into the class of interest.

But problematic when labeled set is not random sample of population (perhaps due to 'drift'—sample collected once, and population moves on)

and DGP is typically $\Pr(t_k|c)$ not $\Pr(c|t_k)$, which is what aggregating would imply (causes some problems for inference, though H&K are v vague here)

Estimating Proportions, Hopkins & King (2010)

Suppose you want to know the proportion of e.g. blog posts or Facebook updates that are sympathetic to Trump.

→ could train a (Naive Bayes) classifier on documents, and then calculate the proportion of the test set that fits into the class of interest.

But problematic when labeled set is not random sample of population (perhaps due to 'drift'—sample collected once, and population moves on)

and DGP is typically $\Pr(t_k|c)$ not $\Pr(c|t_k)$, which is what aggregating would imply (causes some problems for inference, though H&K are v vague here)

→ would like unbiased approach (and be nice if non-parametric), that avoids the intermediate step of document classification.

What to do

What to do

Convert all features to K stems S, and count binary instances only

What to do

Convert all features to K stems S, and count binary instances only (stem occurs, or not, in document)

What to do

Convert all features to K stems S, and count binary instances only (stem occurs, or not, in document)

Hand code say, 500, texts into the J classes.

What to do

Convert all features to K stems S, and count binary instances only (stem occurs, or not, in document)

Hand code say, 500, texts into the J classes.

Notice that

$$\underbrace{\Pr(\mathbf{S})}_{2^K \times 1} = \underbrace{\Pr(\mathbf{S}|c)}_{2^K \times J} \underbrace{\Pr(c)}_{J \times 1}$$

What to do

Convert all features to K stems S, and count binary instances only (stem occurs, or not, in document)

Hand code say, 500, texts into the J classes.

Notice that

$$\underbrace{\Pr(\mathbf{S})}_{2^K \times 1} = \underbrace{\Pr(\mathbf{S}|c)}_{2^K \times J} \underbrace{\Pr(c)}_{J \times 1}$$

where $\Pr(\mathbf{S})$ is the probability of each of the word stem (binary) combinations occurring, which we can tabulate from the target texts (test set).

What to do

Convert all features to K stems S, and count binary instances only (stem occurs, or not, in document)

Hand code say, 500, texts into the J classes.

Notice that

$$\underbrace{\Pr(\mathbf{S})}_{2^K \times 1} = \underbrace{\Pr(\mathbf{S}|c)}_{2^K \times J} \underbrace{\Pr(c)}_{J \times 1}$$

where $\Pr(\mathbf{S})$ is the probability of each of the word stem (binary) combinations occurring, which we can tabulate from the target texts (test set).

and $\Pr(\mathbf{S}|c)$ is the probability of each of the word stem combinations occurring within class c: which we assume is identical to the same quantity in the labeled set (and then tabulate).

What to do

Convert all features to K stems S, and count binary instances only (stem occurs, or not, in document)

Hand code say, 500, texts into the J classes.

Notice that

$$\underbrace{\Pr(\mathbf{S})}_{2^K \times 1} = \underbrace{\Pr(\mathbf{S}|c)}_{2^K \times J}\underbrace{\Pr(c)}_{J \times 1}$$

where $\Pr(\mathbf{S})$ is the probability of each of the word stem (binary) combinations occurring, which we can tabulate from the target texts (test set).

and $\Pr(\mathbf{S}|c)$ is the probability of each of the word stem combinations occurring within class c: which we assume is identical to the same quantity in the labeled set (and then tabulate).

while $\Pr(c)$ is the proportion of documents in class c,

What to do

Convert all features to K stems S, and count binary instances only (stem occurs, or not, in document)

Hand code say, 500, texts into the J classes.

Notice that

$$\underbrace{\Pr(\mathbf{S})}_{2^K \times 1} = \underbrace{\Pr(\mathbf{S}|c)}_{2^K \times J}\underbrace{\Pr(c)}_{J \times 1}$$

where $\Pr(\mathbf{S})$ is the probability of each of the word stem (binary) combinations occurring, which we can tabulate from the target texts (test set).

and $\Pr(\mathbf{S}|c)$ is the probability of each of the word stem combinations occurring within class c: which we assume is identical to the same quantity in the labeled set (and then tabulate).

while $\Pr(c)$ is the proportion of documents in class c, which is what we want to know.

Estimation Notes I

Estimation Notes I

e.g. If there are $K = 3$ stems of interest,

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles:

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0, 0, 0], [0, 0, 1], [0, 1, 0], [1, 0, 0], [1, 1, 0], [1, 0, 1], [0, 1, 1], [1, 1, 1]$.

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0],[0,0,1],[0,1,0],[1,0,0],[1,1,0],[1,0,1],[0,1,1],[1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large,

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0],[0,0,1],[0,1,0],[1,0,0],[1,1,0],[1,0,1],[0,1,1],[1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable:

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}.

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0],[0,0,1],[0,1,0],[1,0,0],[1,1,0],[1,0,1],[0,1,1],[1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}. Plus, number of possible stem profiles (y) is much larger than number of observations,

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\Pr(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\Pr(\mathbf{S})}_{y} = \underbrace{\Pr(\mathbf{S}|c)}_{X}\underbrace{\Pr(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}. Plus, number of possible stem profiles (y) is much larger than number of observations, meaning that many of the profile combinations are never observed (we have no information about them).

Estimation Notes I

e.g. If there are $K = 3$ stems of interest, then $\text{Pr}(\mathbf{S})$ gives the probability (proportion of documents in the target set) of the $2^3 = 8$ profiles: $[0,0,0], [0,0,1], [0,1,0], [1,0,0], [1,1,0], [1,0,1], [0,1,1], [1,1,1]$.

then set up a linear regression and report $\hat{\beta}$:

$$\underbrace{\text{Pr}(\mathbf{S})}_{y} = \underbrace{\text{Pr}(\mathbf{S}|c)}_{X}\underbrace{\text{Pr}(c)}_{\beta}$$

but given K is large, problem is clearly intractable: try having y of length 2^{300}. Plus, number of possible stem profiles (y) is much larger than number of observations, meaning that many of the profile combinations are never observed (we have no information about them).

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation),

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Judge *relative* performance via mean absolute proportion error.

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Judge *relative* performance via mean absolute proportion error.

NB "*among all documents in a given category, the prevalence of particular word profiles in the labeled set should be the same in expectation as in the population set*".

Estimation Notes II

so choose subset of 5–25 stems and estimate $\Pr(c)$,

then repeat process with different subsets (number determined by cross-validation), before averaging results across subsets. Bootstrap for CIs.

$\rightsquigarrow$ kernel smoothing of sparse matrices.

Judge *relative* performance via mean absolute proportion error.

NB "*among all documents in a given category, the prevalence of particular word profiles in the labeled set should be the same in expectation as in the population set*". This is key assumption. btw, what happened to the danger of drift?!

Performance: Congress, Editorials, Enron

Figure 4 Additional Out-of-Sample Validation