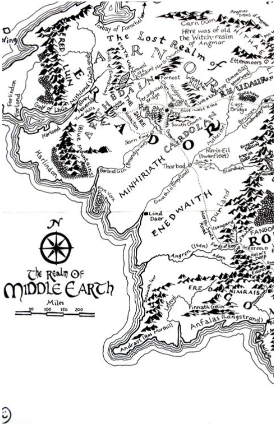
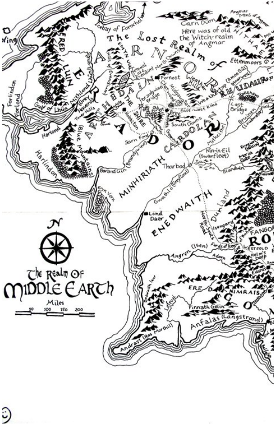
Where Are We?

Where Are We?



Carn Dûm
Here was of old the Witch-realm of Angmar
The Lost Realm of
ARNOR
Ettenmoors
Fornost
Weather Hills
Bree
East-West Road
Last Bridge
South Downs
Sarn Ford
Tharbad
MINHIRIATH
CARDOLAN
ENEDWAITH
Dunland
Isengard
Lond Daer
Forlindon
Forlond
Gulf of Lhûn
Harlond
Harlindon
Blue Mountains
Pinnath Gelin
Anfalas (Langstrand)
The Realm of
MIDDLE EARTH
Miles

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

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

Where Are We?

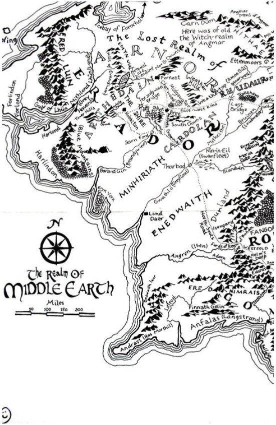
We've covered the basics of document representation and characterization.

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.

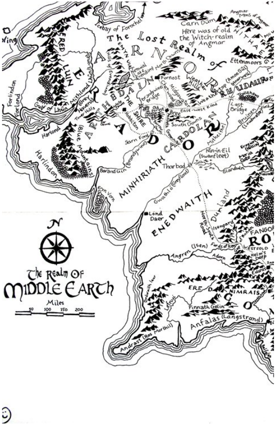
Now begin to think about documents as members of categories or classes

→ simple, fast dictionary based ways to classify/categorize

cover some 'major' dictionaries in social science

and demonstrate challenges that emerge in constructing and using dictionaries,

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

and demonstrate challenges that emerge in constructing and using dictionaries, especially for novel tasks.

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

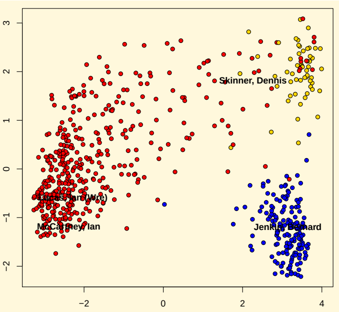
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

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.

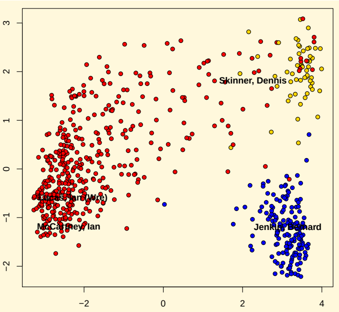
e.g. PCA of legislators's votes: want to see how they are organized—by party? by ideology? by race?



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?

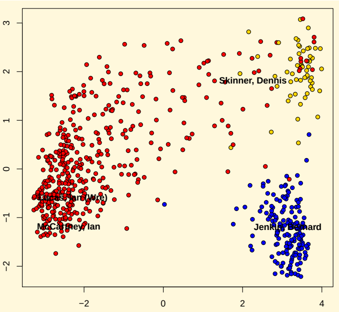


Supervised techniques:

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?

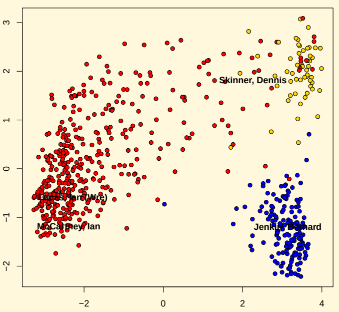


Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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?



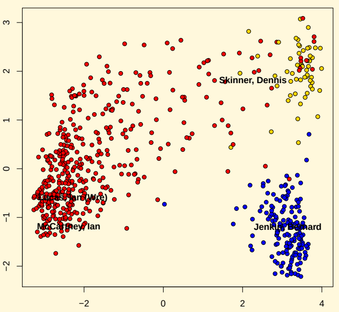
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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?



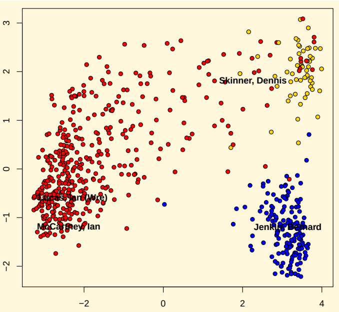
Supervised techniques: learning relationship between inputs and a labeled set of outputs.

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

Terminology

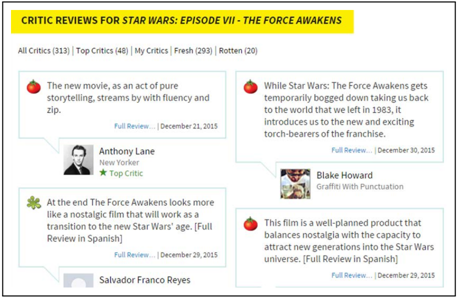
Unsupervised techniques: learning (hidden or latent) structure in unlabeled data.

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—some $f(x)$—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—some $f(x)$—to predict the outcomes of documents ($y \in \{0, 1\}$, review sentiment) not in the training set.

Overview: Dictionary

Overview: Dictionary

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

Overview: Dictionary

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

Overview: Dictionary

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,

Overview: Dictionary

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

→ common in opinion mining/sentiment analysis, and in coding events or manifestos.

Overview: Dictionary

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: Dictionary

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: Dictionary

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

→ common in opinion mining/sentiment analysis, and in coding events or manifestos.

Often derived from supervised learning techniques

and often used in supervised learning problems, as a starting point.

so we'll cover them here.

Overview: Dictionary

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.

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

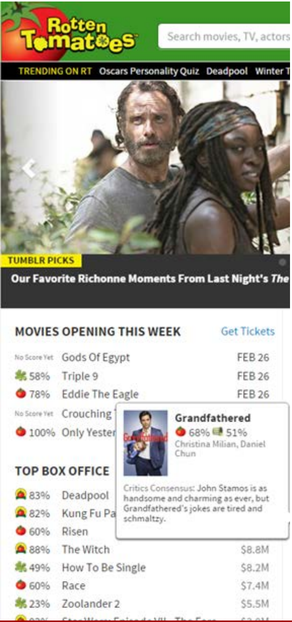
positive 2

total 13

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

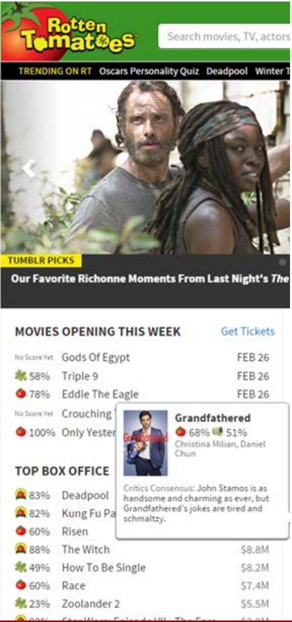
Partner Exercise

Exercise



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

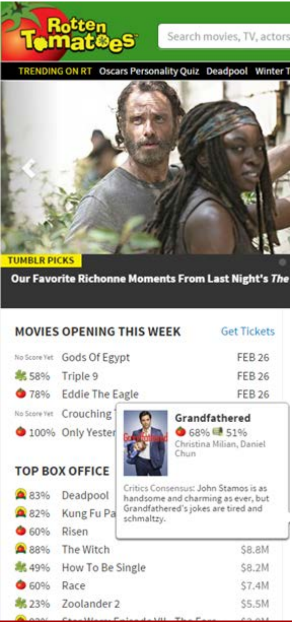
Exercise



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?

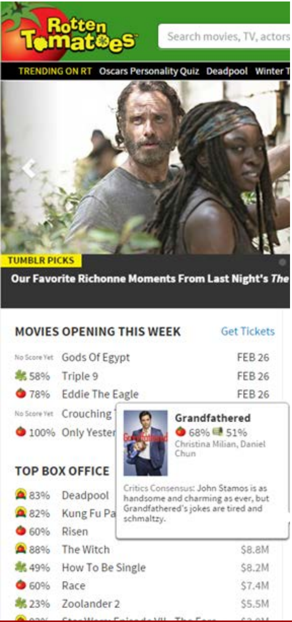
Exercise



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?

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 $+1$!

Sarcasm Detection

Sarcasm Detection

Notoriously difficult problem (see Pang & Lee, 2008)

Sarcasm Detection

Notoriously difficult problem (see Pang & Lee, 2008)

e.g. "Great for insomniacs!" (book review)

Sarcasm Detection

Notoriously difficult problem (see Pang & Lee, 2008)

e.g. "Great for insomniacs!" (book review)

Early approaches simply used presence of tags e.g. `[laughter]` in proceedings to make inferences about what was sarcastic.

Sarcasm Detection

Notoriously difficult problem (see Pang & Lee, 2008)

e.g. "Great for insomniacs!" (book review)

Early approaches simply used presence of tags e.g. `[laughter]` in proceedings to make inferences about what was sarcastic. Or prosody in spoken word (e.g. lower pitch).

Sarcasm Detection

Notoriously difficult problem (see Pang & Lee, 2008)

e.g. "Great for insomniacs!" (book review)

Early approaches simply used presence of tags e.g. `[laughter]` in proceedings to make inferences about what was sarcastic. Or prosody in spoken word (e.g. lower pitch).

Davidov et al ('Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon') get good results by looking for particular pattern:

Sarcasm Detection

Notoriously difficult problem (see Pang & Lee, 2008)

e.g. "Great for insomniacs!" (book review)

Early approaches simply used presence of tags e.g. `[laughter]` in proceedings to make inferences about what was sarcastic. Or prosody in spoken word (e.g. lower pitch).

Davidov et al ('Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon') get good results by looking for particular pattern: [high frequency word] [content word] [high frequency word] in reviews

Sarcasm Detection

Notoriously difficult problem (see Pang & Lee, 2008)

e.g. "Great for insomniacs!" (book review)

Early approaches simply used presence of tags e.g. `[laughter]` in proceedings to make inferences about what was sarcastic. Or prosody in spoken word (e.g. lower pitch).

Davidov et al ('Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon') get good results by looking for particular pattern: [high frequency word] [content word] [high frequency word] in reviews

btw punctuation adds relatively little to accuracy.

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’

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.

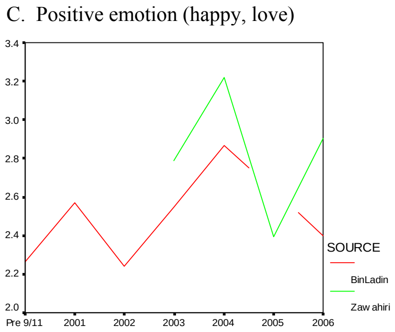
Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

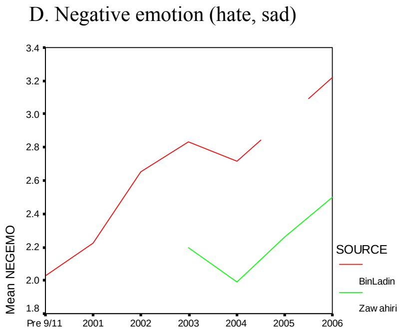
Pennebaker & Chung, 2007: Computerized Analysis of Al-Qaeda Transcripts

"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. "





Dictionaries V: Hu & Liu

Dictionaries V: Hu & Liu

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

Dictionaries V: 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 V: 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 V: 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

→ 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. . .

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.

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.

Exercise

Exercise

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

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?

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?

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?

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.

Word embeddings may offer automatic way forward here (Hamilton et al, "Inducing Domain-Specific Sentiment Lexicons from Unlabeled Corpora")

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.

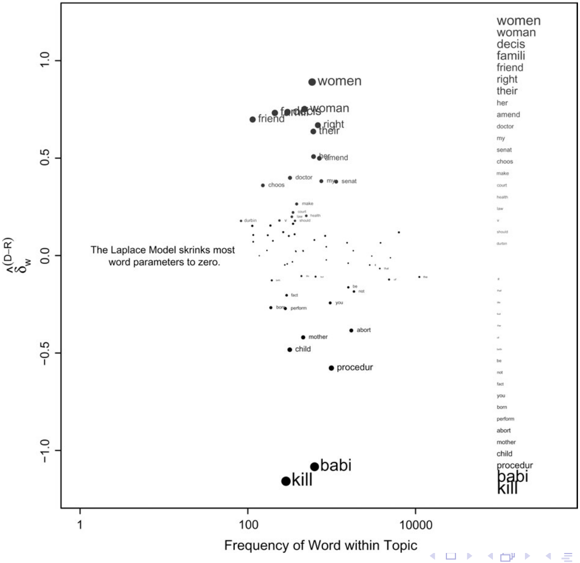
Discrimination

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.

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

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



Goldman-Sachs Case Study

Goldman-Sachs Case Study

Goldman-Sachs Case Study

GS world's largest investment bank,

Goldman-Sachs Case Study

GS world's largest investment bank, $\sim \$90B$ revenue.

Goldman-Sachs Case Study

GS world's largest investment bank, $\sim \$90B$ revenue.

Jari Stehn, GS Managing Director, senior economist in US Economics Group,

Goldman-Sachs Case Study

GS world's largest investment bank, $\sim \$90B$ revenue.

[Jari Stehn](), GS Managing Director, senior economist in US Economics Group, Econ PhD from Oxford (macro focus),

Goldman-Sachs Case Study

GS world's largest investment bank, $\sim \$90B$ revenue.

Jari Stehn, GS Managing Director, senior economist in US Economics Group, Econ PhD from Oxford (macro focus), previously employed at IMF.

Goldman-Sachs Case Study

GS world's largest investment bank, $\sim \$90B$ revenue.

Jari Stehn, GS Managing Director, senior economist in US Economics Group, Econ PhD from Oxford (macro focus), previously employed at IMF.

Works alone or in pair to produce weekly report on investment topic of interest to clients, public, academics and policy makers.

Goldman-Sachs Case Study

GS world's largest investment bank, $\sim \$90B$ revenue.

Jari Stehn, GS Managing Director, senior economist in US Economics Group, Econ PhD from Oxford (macro focus), previously employed at IMF.

Works alone or in pair to produce weekly report on investment topic of interest to clients, public, academics and policy makers. Not peer-reviewed, but not 'journalism'.

Background

Background

Background

Federal Reserve Board oversees federal reserve system,

Background

Federal Reserve Board oversees federal reserve system, and sets monetary policy via Federal Open Market Committee decisions on interest rates.

Background

Federal Reserve Board oversees federal reserve system, and sets monetary policy via Federal Open Market Committee decisions on interest rates.

Meet at least four times a year,

Background

Federal Reserve Board oversees federal reserve system, and sets monetary policy via Federal Open Market Committee decisions on interest rates.

Meet at least four times a year, and release statement after meeting (followed by minutes, also release 'Beige book' of economic indicators 8 times a year)

Background

Federal Reserve Board oversees federal reserve system, and sets monetary policy via Federal Open Market Committee decisions on interest rates.

Meet at least four times a year, and release statement after meeting (followed by minutes, also release 'Beige book' of economic indicators 8 times a year)

FOMC perception of economic situation has extremely influential effects on financial markets,

Background

Federal Reserve Board oversees federal reserve system, and sets monetary policy via Federal Open Market Committee decisions on interest rates.

Meet at least four times a year, and release statement after meeting (followed by minutes, also release 'Beige book' of economic indicators 8 times a year)

FOMC perception of economic situation has extremely influential effects on financial markets, so are made with great care via 'Fedspeak'—

Background

Federal Reserve Board oversees federal reserve system, and sets monetary policy via Federal Open Market Committee decisions on interest rates.

Meet at least four times a year, and release statement after meeting (followed by minutes, also release 'Beige book' of economic indicators 8 times a year)

FOMC perception of economic situation has extremely influential effects on financial markets, so are made with great care via 'Fedspeak'—which is ambiguous in tone.

Example from Sept 2015

Example from Sept 2015

Information received since the Federal Open Market Committee met in July suggests that economic activity is expanding at a moderate pace. Household spending and business fixed investment have been increasing moderately, and the housing sector has improved further; however, net exports have been soft.

Example from Sept 2015

Information received since the Federal Open Market Committee met in July suggests that economic activity is expanding at a moderate pace. Household spending and business fixed investment have been increasing moderately, and the housing sector has improved further; however, net exports have been soft.

The Committee continues to see the risks to the outlook for economic activity and the labor market as nearly balanced but is monitoring developments abroad. Inflation is anticipated to remain near its recent low level in the near term but the Committee expects inflation to rise gradually toward 2 percent over the medium term as the labor market improves further and the transitory effects of declines in energy and import prices dissipate.

Exercise

Exercise

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting?

Exercise

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

Exercise

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

$\rightarrow$ How?

Exercise

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

$\rightarrow$ How?

Problem and Approach

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting?

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

$\rightarrow$ How?

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

$\rightarrow$ How?

Use 'hawkish' and 'dovish' dictionary,

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

→ How?

Use 'hawkish' and 'dovish' dictionary, with key measure being ratio of those terms.

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

$\rightarrow$ How?

Use 'hawkish' and 'dovish' dictionary, with key measure being ratio of those terms.

$\rightarrow$ nouns (`strength` vs `recession`) and adjectives (`healthy`, vs `weak`)

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

→ How?

Use 'hawkish' and 'dovish' dictionary, with key measure being ratio of those terms.

→ nouns ('`strength`' vs '`recession`') and adjectives ('`healthy`', vs '`weak`')

NB statements generally most marginally informative (as expected, since they come first),

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

→ How?

Use 'hawkish' and 'dovish' dictionary, with key measure being ratio of those terms.

→ nouns ('`strength`' vs '`recession`') and adjectives ('`healthy`', vs '`weak`')

NB statements generally most marginally informative (as expected, since they come first), with pseudo-R^2 (from ordered logit/probit?) ~ 0.15

Problem and Approach

Can we predict interest rate decisions $\{-1, 0, +1\}$ at next meeting from prior FOMC statements, minutes, books since last meeting? (range is 1994–2013)

→ How?

Use 'hawkish' and 'dovish' dictionary, with key measure being ratio of those terms.

→ nouns ('`strength`' vs '`recession`') and adjectives ('`healthy`', vs '`weak`')

NB statements generally most marginally informative (as expected, since they come first), with pseudo-R^2 (from ordered logit/probit?) ~ 0.15

→ rising to ~ 0.25 when all sources included (NB: speeches generally uninformative)

More Results

More Results

Next, study themes of documents: pairs of hawkish/dovish adjectives that pertain to nouns of various topics—growth, inflation, monetary policy.

More Results

Next, study themes of documents: pairs of hawkish/dovish adjectives that pertain to nouns of various topics—growth, inflation, monetary policy.

+ count terms pertaining to (un)certainty

More Results

Next, study themes of documents: pairs of hawkish/dovish adjectives that pertain to nouns of various topics—growth, inflation, monetary policy.

+ count terms pertaining to (un)certainty

Themes about growth are most important for monetary policy

More Results

Next, study themes of documents: pairs of hawkish/dovish adjectives that pertain to nouns of various topics—growth, inflation, monetary policy.

+ count terms pertaining to (un)certainty

Themes about growth are most important for monetary policy

Show predictions from this model do a good job of tracking actual policy over time.

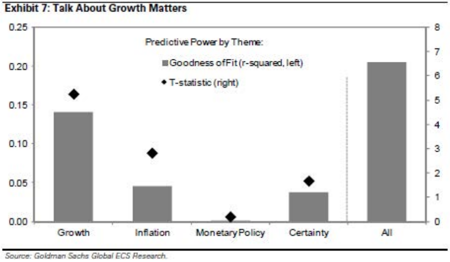
More Results

Next, study themes of documents: pairs of hawkish/dovish adjectives that pertain to nouns of various topics—growth, inflation, monetary policy.

+ count terms pertaining to (un)certainty

Themes about growth are most important for monetary policy

Show predictions from this model do a good job of tracking actual policy over time.



More Results

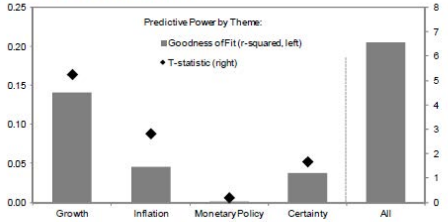
Next, study themes of documents: pairs of hawkish/dovish adjectives that pertain to nouns of various topics—growth, inflation, monetary policy.

+ count terms pertaining to (un)certainty

Themes about growth are most important for monetary policy

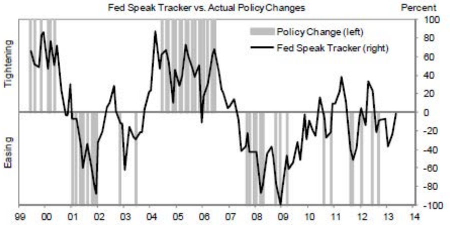
Show predictions from this model do a good job of tracking actual policy over time.

Exhibit 7: Talk About Growth Matters



Source: Goldman Sachs Global ECS Research.

Exhibit 8: The Fed Speak Tracker



Source: Goldman Sachs Global ECS Research.

Lie Detection

Lie Detection

Zhou et al (2004) investigate whether deception can be detected in Text-based Asynchronous Computer-Mediated Communication

Lie Detection

Zhou et al (2004) investigate whether deception can be detected in Text-based Asynchronous Computer-Mediated Communication—email.

Lie Detection

Zhou et al (2004) investigate whether deception can be detected in Text-based Asynchronous Computer-Mediated Communication—email.

So Recruit undergrad students,

Lie Detection

Zhou et al (2004) investigate whether deception can be detected in Text-based Asynchronous Computer-Mediated Communication—email.

So Recruit undergrad students, and have them play roles in Desert Survival Problem where they need to agree on 12 items to take from crashed jeep.

Lie Detection

Zhou et al (2004) investigate whether deception can be detected in Text-based Asynchronous Computer-Mediated Communication—email.

So Recruit undergrad students, and have them play roles in Desert Survival Problem where they need to agree on 12 items to take from crashed jeep.

\+ one person in dyad told to deceive other and not represent true preferences about rank order.

Lie Detection

Zhou et al (2004) investigate whether deception can be detected in Text-based Asynchronous Computer-Mediated Communication—email.

So Recruit undergrad students, and have them play roles in Desert Survival Problem where they need to agree on 12 items to take from crashed jeep.

+ one person in dyad told to deceive other and not represent true preferences about rank order.

Item	Your Rank	Actual Rank	Team Rank	Team Difference	Your Difference
A ball of steel wool					
A small ax					
A loaded .45-caliber pistol					
Can of Crisco shortening					
Newspapers (one per person)					
Cigarette lighter (without fluid)					
Extra shirt and pants for each survivor					
20 x 20 ft. piece of heavy-duty canvas					
A sectional air map made of plastic					
One quart of 100-proof whiskey					
A compass					
Family-size chocolate bars (one per person)					
Score					

Exercise

Exercise

Exercise

Q how would you expect deceiver's messages differ from truth tellers?

Exercise

Q how would you expect deceiver's messages differ from truth tellers?

→ in terms of e.g. number of words, informality, uncertainty, complexity, pausality etc.

Results

Results

Q how do deceiver's messages differ from truth tellers?

Results

Q how do deceiver's messages differ from truth tellers?

A used more words, verbs, noun phrases, sentences

Results

Q how do deceiver's messages differ from truth tellers?

A used more words, verbs, noun phrases, sentences

more informality:

Results

Q how do deceiver's messages differ from truth tellers?

A used more words, verbs, noun phrases, sentences

more informality: more typos.

Results

Q how do deceiver's messages differ from truth tellers?

A used more words, verbs, noun phrases, sentences

more informality: more typos.

more uncertain (weak modifiers e.g. 'about') and non-immediate (passive voice, externalizing)

Results

- Q how do deceiver's messages differ from truth tellers?
- A used more words, verbs, noun phrases, sentences

  more informality: more typos.

  more uncertain (weak modifiers e.g. 'about') and non-immediate (passive voice, externalizing)

  less complex and less diverse (∼TTR)

Results

Q how do deceiver's messages differ from truth tellers?

A used more words, verbs, noun phrases, sentences

more informality: more typos.

more uncertain (weak modifiers e.g. 'about') and non-immediate (passive voice, externalizing)

less complex and less diverse (∼TTR)

less pausality (more punctuation), more group references ('we')

Results

Q how do deceiver's messages differ from truth tellers?

A used more words, verbs, noun phrases, sentences

more informality: more typos.

more uncertain (weak modifiers e.g. 'about') and non-immediate (passive voice, externalizing)

less complex and less diverse ($\sim$TTR)

less pausality (more punctuation), more group references ('we')

btw, passive voice means subject and object of sentence are switched:

Results

Q how do deceiver's messages differ from truth tellers?

A used more words, verbs, noun phrases, sentences

more informality: more typos.

more uncertain (weak modifiers e.g. 'about') and non-immediate (passive voice, externalizing)

less complex and less diverse ($\sim$TTR)

less pausality (more punctuation), more group references ('we')

btw, passive voice means subject and object of sentence are switched: "I am packing my bag" $\rightarrow$ "My bag is being packed by me."

I will **definitely** see you next time, when I intend to forego persiflage and conduct a profound lucubration, skirring over new topics in a way that could never be described as prolix.