Computer-Assisted Content Analysis: Assigning Categories to Documents

Will Lowe (University of Mannheim)

Menu

Session 1: Classical Content Analysis

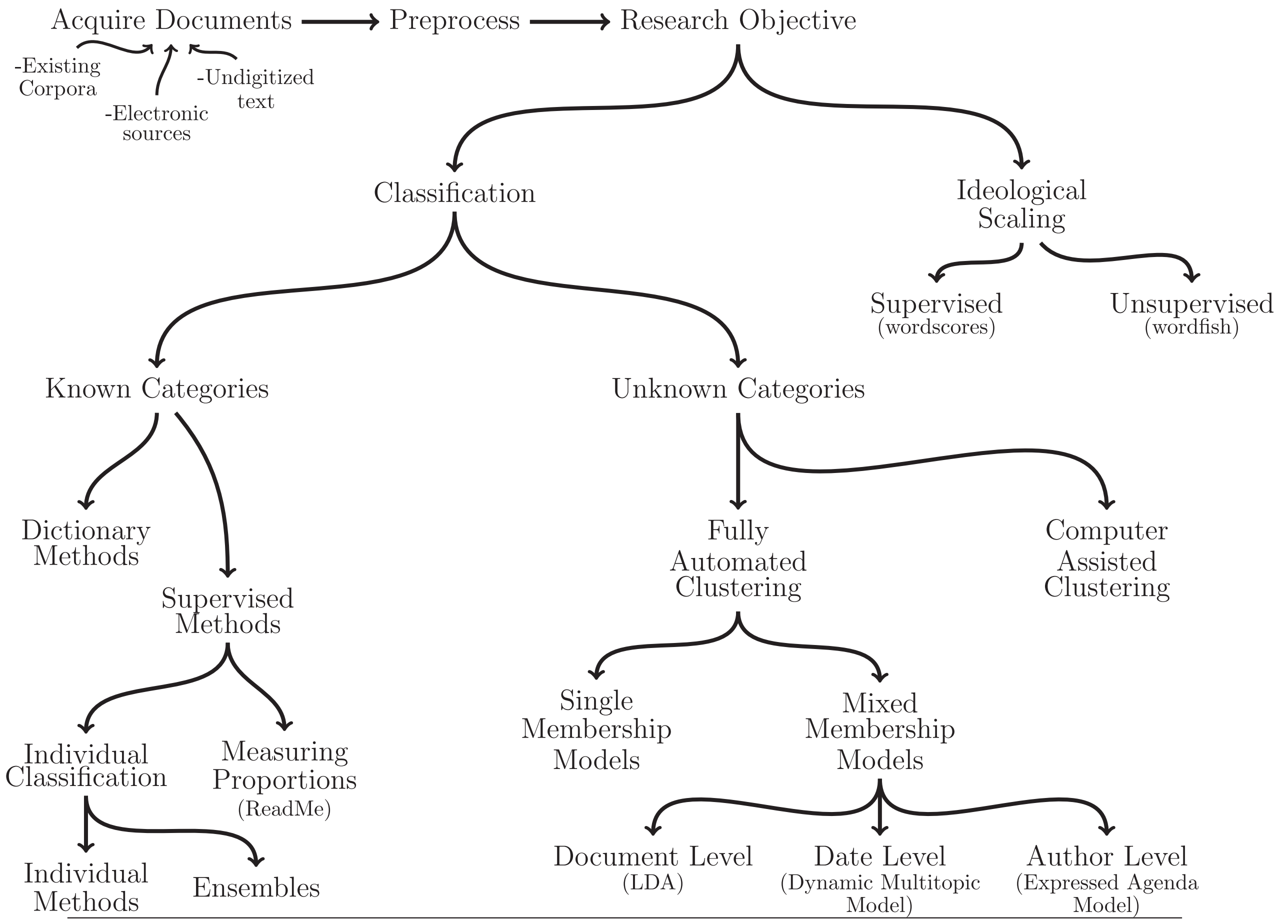
Session 2:

Document classification

The indirect approach to classification
Evaluation and interpretation
Evaluation for the lazy
The direct approach to classification

Session 3: Scaling Models

Text as Data



Source: Grimmer and Stewart (2013)

Classification Approaches when Categories are Known

Examples:

Are campaign advertisements positive or negative?

What policy areas do newspaper editorials cover?

Are international statements belligerent or peaceful?

Do court letters represent liberal or conservative positions?

What language is this article written in?

Is this email spam?

Classification Approaches when Categories are Known

Yesterday we talked how to do this using a dictionary approach.

An alternative is supervised machine learning methods:

1. coders categorize a set of documents by hand
2. the algorithm “learns” how to sort the documents in categories
3. characteristics of training set are used to assign new documents to categories.

Classification Approaches when Categories are Known

Assume that each document has a *single* topic Z

Let θ_k be the *probability* that Z=k for each document

Assume that (some) topic labels are observed

Classification Approaches

Much of machine learning, computational linguistics, and AI deals with classification problems

Methods:

Naive Bayes, Maximum Entropy, Support Vector Machines, Neural Networks, Bagging, Boosting, . . .

We only touch on the issues here. . .

Classification approaches

In the simple framework of yesterday

Z is the true category of a document

θ is be the posterior probability that a document is a particular category

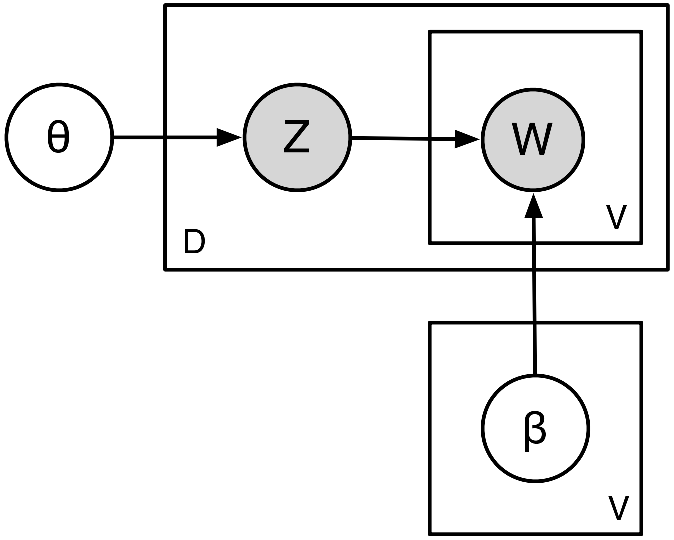
Classification approaches

As before, we have two approaches

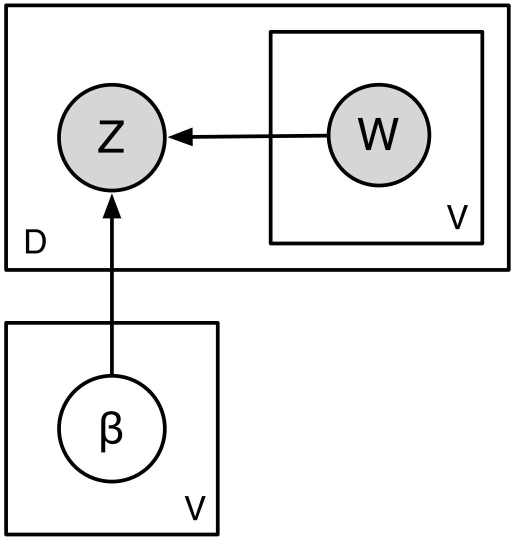
Discriminative: Model $P(Z \mid \{W\}, \beta) = \theta_{z|w}$ directly

Generative: Model $P(\{W\} \mid Z, \beta)$ and $P(Z) = \theta_z$, then get $P(Z \mid \{W\}, \beta) = \theta_{z|w}$ via Bayes theorem

Generative vs discriminative training

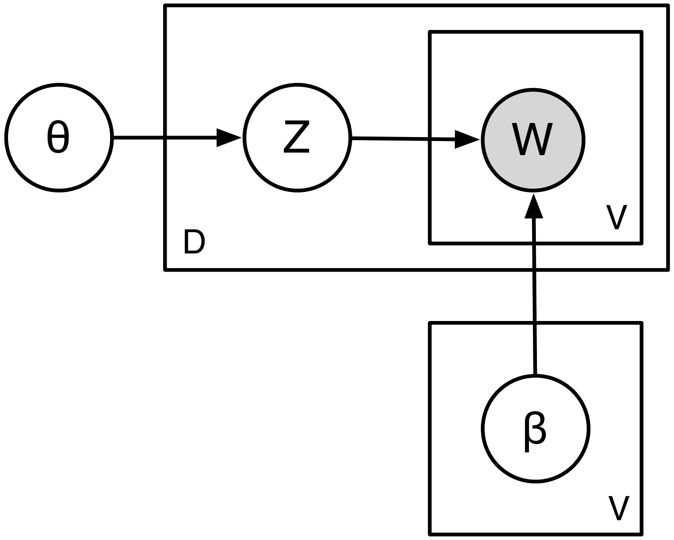


Naive Bayes

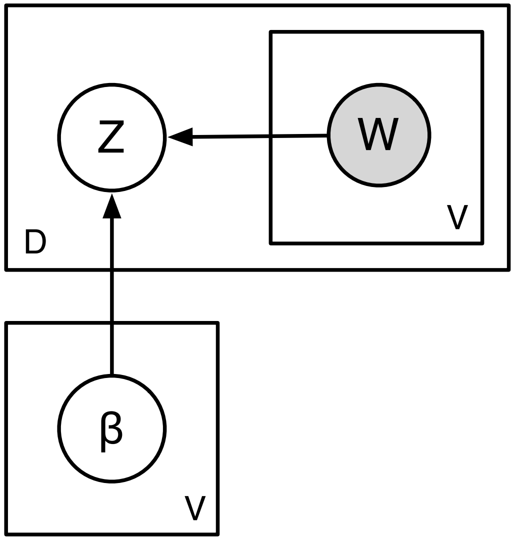


Maximum Entropy, etc...

Generative vs discriminative testing



Naive Bayes



Maximum Entropy, etc...

Either way. . .

Desirable classification outcome:

	$P(Z = \text{'Domestic'} \mid \{W\}_d)$	$P(Z = \text{'Foreign'} \mid \{W\}_d)$
D_1	0.75	0.25
D_2	0.82	0.18
. . .	. . .	. . .
D_{N-1}	0.02	0.98
D_N	0.45	0.55

where $\theta_{z|w} = P(Z = z \mid \{W\})$

The Basic Steps

1. Construct a training set

    (a) create a coding scheme
    (b) select documents (ideally randomly sampled)

2. Apply the supervised learning method to learn features of a training set and infer labels

3. Validate and classify remaining documents

Indirect approach: Naive Bayes

Background:

Amicus Curiae (friend of the court) briefs are submitted to an appellate court

They usually present a legal argument for or against one of the parties to a case

Amicus Curiae can be submitted by any group that feels that it has a stake in the case

Affirmative action

Evans et al. use cases about the constitutionality of 'affirmative action' programs at university level

Regents of the University of California vs. Bakke (1978)

Grutter vs Bollinger, Lehman, Shields, and the Regents of the University of Michigan (2003)

Gratz and Hamacher vs Bollinger, Lehman, Shields and the Regents of the University of Michigan (2003)

The arguments are as much *political* (state vs federal rights, social welfare, constitutional interpretation) as they are *legal*

Affirmative Action

This work uses document classification to answer two questions

To what extent can the *direction* of an AC brief be predicted on the basis of its words?

What can we learn about *language* of each side of the case?

Naive Bayes

Naive Bayes is a relatively old (∼1975) classification method

Suppose you had to guess whether document j is liberal (Z=‘Lib’) or conservative (Z=‘Con’) based on its word profile $\{W\}_j$.

Probability can be derived by applying Bayes theorem:

$$P(Z = \text{‘Lib’} \mid \{W\}_j) \;=\; \frac{P(\{W\}_j \mid Z = \text{‘Lib’})\; P(Z = \text{‘Lib’})}{P(\{W\}_j)}$$

Naive Bayes

$$P(Z = \text{`Lib'} \mid \{W\}_j) \;\propto\; P(\{W\}_j \mid Z = \text{`Lib'})\; P(Z = \text{`Lib'})$$

We can drop $P(\{W\}_j)$ since it is constant across categories.

Given a representative training set, estimating $P(Z = L)$ is easy:

$$\hat{P}(Z = \text{`Lib'}) \;=\; \frac{\#\text{ training docs that are liberal}}{\#\text{ of training docs}}$$

Naive Bayes

Estimating the probability that a word profile $\{W\}_j$ occurs given that the document is liberal $P(\{W\}_j \mid Z = \text{‘Lib’})$ is more challenging, because any one word profile is likely to occur only once.

Solution:

words are assumed to be generated *independently* given the category Z (the ‘naive’ and wrong assumption).

$$P(\{W\}_j \mid Z = \text{‘Lib’}) \;=\; \Pi_i P(W_i \mid Z = \text{‘Lib’})$$

Naive Bayes

With this assumption, we can estimate the probability of observing a word i given that the document is liberal: proportion of word i in liberal training set.

The classifier then chooses the class Z (Liberal or Conservative) with the highest aggregate probability.

Note that every new word adds a bit of information that re-adjusts the conditional probabilities.

Naive Bayes

Note that with two classes (here: liberal and conservative) this has a rather neat interpretation:

$$\frac{P(Z = \text{`Lib'} \mid \{W\}_j)}{P(Z = \text{`Con'} \mid \{W\}_j)} = \prod_i \frac{P(W_i \mid Z = \textit{`Lib'})}{P(W_i \mid Z = \textit{`Con'})} \times \frac{P(Z = \textit{`Lib'})}{P(Z = \textit{`Con'})}$$

Logging this probability ratio, every new word *adds* a bit of information that pushes the ratio above or below 0

Naive Bayes

Example: Naive Bayes with only word class ‘discriminat*’.

$$P(W = \text{‘discriminat*’} \mid Z = \text{‘Lib’}) = (26 + 13)/(20002 + 18722) \approx 0.001$$

$$P(W = \text{‘discriminat*’} \mid Z = \text{‘Con’}) = (70 + 48)/(17368 + 17698) \approx 0.003$$

Assume that liberal and conservative supporting briefs are equally likely (true in the training set)

$$\frac{P(Z = \text{‘Lib’})}{P(Z = \text{‘Con’})} = 1$$

Last step: calculate posterior classification probabilities for a new document (based on occurrence of this word).

Naive Bayes

Amicus brief from ‘King County Bar Association’ containing 3667 words (File 6019_al18-utf8.txt) and 4 matches to disciminat*.

```
      that "the state shall not [discriminate] against, or grant preferential treatment
the lingering effects of racial [discrimination] against minority groups in this
 remedy the effects of societal [discrimination]. Another four Justices (Stevens
      that "the state shall not [discriminate] against, or grant preferential treatment
```

Naive Bayes

A priori, the probabilities are. . .

Probability that we observe the word discriminat* 4 out of 3667 times if the document is liberal:

```
> dbinom(4, size=3667, prob=0.001007127)
[1] 0.1930602
```

Probability that we observe the word discriminat* 4 out of 3667 times if the document is conservative:

```
> dbinom(4, size=3667, prob=0.003365083)
[1] 0.004188261
```

Logged probability ratio = 3.83

Naive Bayes

Conclusion: Seeing 4 instances of discriminat* gives the posterior classification probabilities

$$\theta_{\text{liberal}} = 0.979$$

$$\theta_{\text{conservative}} = \text{1-0.979=0.021}$$

This is *quite* confident

. . . but other words will be less loaded or push the other way

Evaluating Classifiers: Accuracy, Precision and Recall

How can we evaluate how good a supervised classifier works?

Use the confusion matrix (training set versus machine)

Evaluating Classifiers: Accuracy, Precision and Recall

		Machine	
		Liberal	Conservative
Training Data	Liberal	40	10
	Conservative	40	60

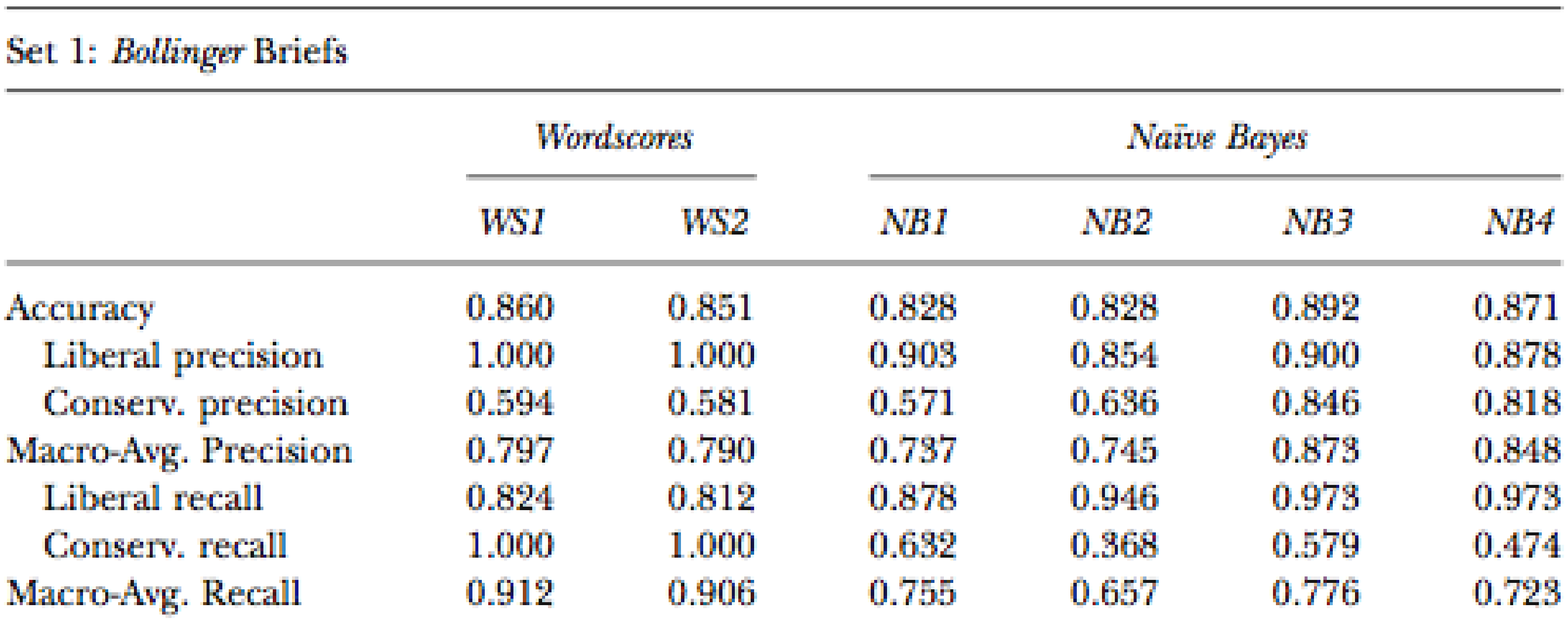
Accuracy = (40+60)/150 = .66

Precision ($Z_{machine}$=Liberal) = 40/(40+40) = 0.5

Precision ($Z_{machine}$=Cons) = 60/(10+60) = 0.86

Recall ($Z_{training}$=Liberal) = 40/(40+10) = 0.80

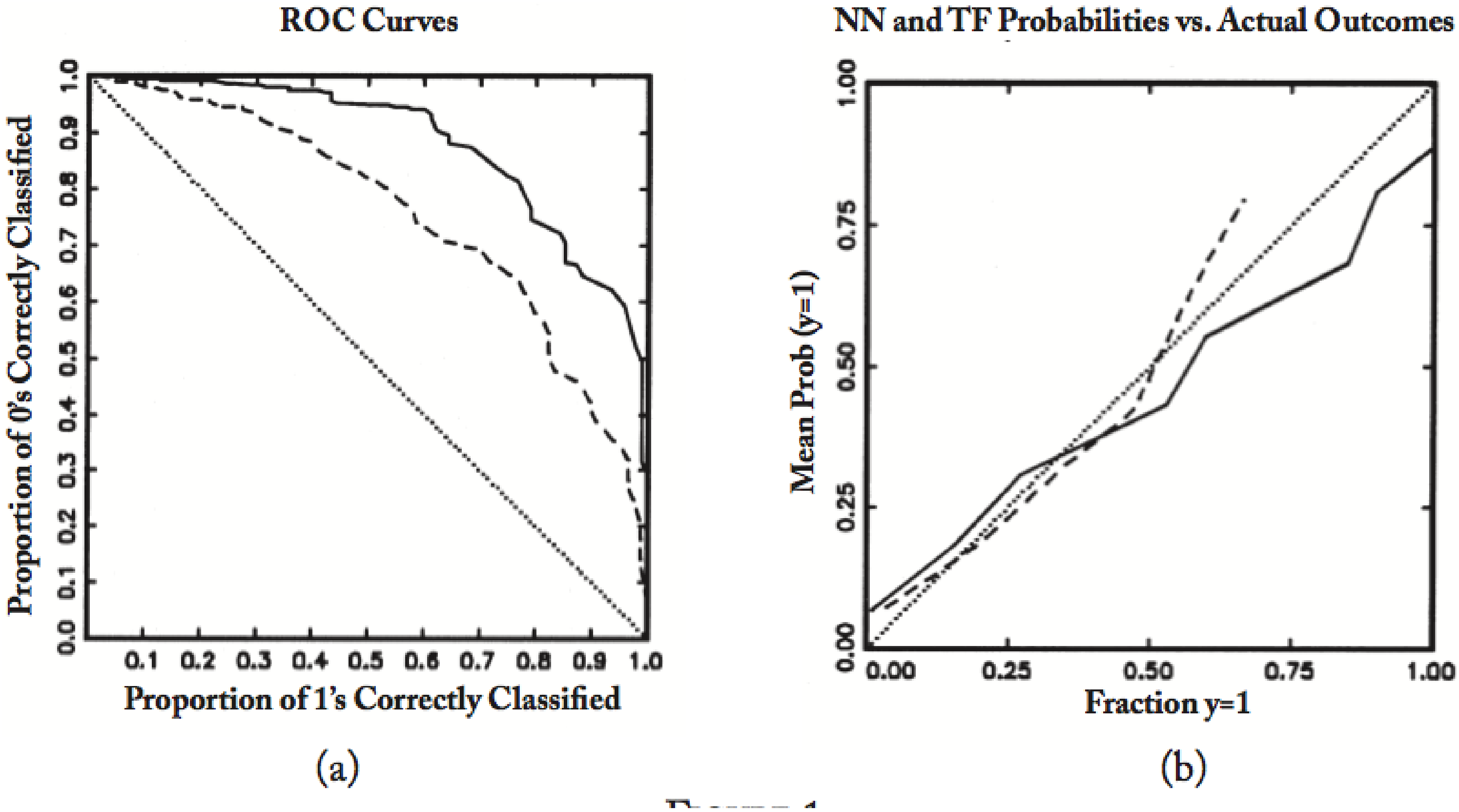
Recall ($Z_{training}$=Cons) = 60/(40+60) = 0.60



Set 1: *Bollinger* Briefs

	Wordscores		*Naïve Bayes*			
	WS1	*WS2*	*NB1*	*NB2*	*NB3*	*NB4*
Accuracy	0.860	0.851	0.828	0.828	0.892	0.871
Liberal precision	1.000	1.000	0.903	0.854	0.900	0.878
Conserv. precision	0.594	0.581	0.571	0.636	0.846	0.818
Macro-Avg. Precision	0.797	0.790	0.737	0.745	0.873	0.848
Liberal recall	0.824	0.812	0.878	0.946	0.973	0.973
Conserv. recall	1.000	1.000	0.632	0.368	0.579	0.474
Macro-Avg. Recall	0.912	0.906	0.755	0.657	0.776	0.723

Trading off precision and recall



(a)

(b)

(King and Zeng, 2001)

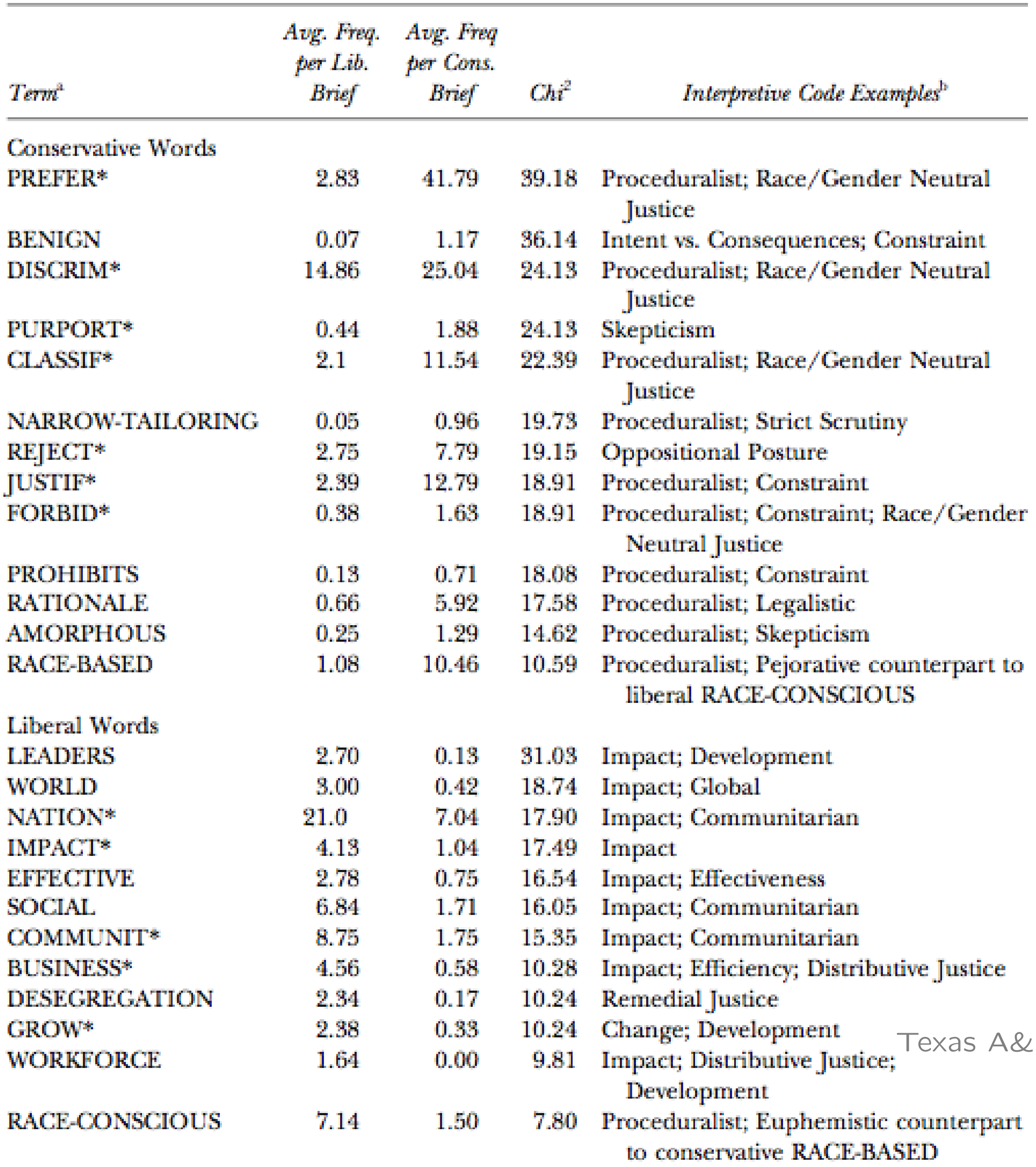
Evaluation

All classification models have a secret extra parameter: the *threshold*

Distinctive Words

Detecting different rhetorical styles of liberal and conservative groups:

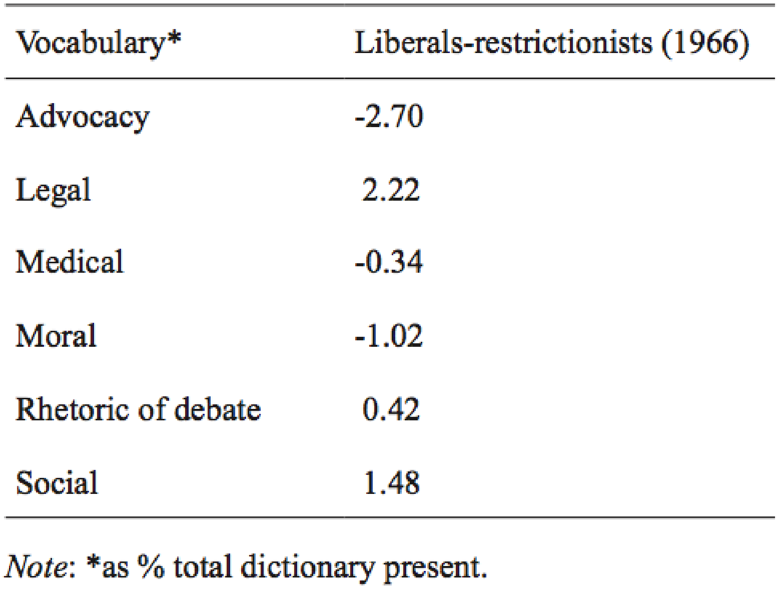
“Liberal groups use language emphasizing the impact of affirmative action polices, while conservative words indicate concern over legal-constitutional limits on administrative procedure” (Evans et al. 2007, p. 1029)



Term[a]	Avg. Freq. per Lib. Brief	Avg. Freq per Cons. Brief	Chi²	Interpretive Code Examples[b]
Conservative Words				
PREFER*	2.83	41.79	39.18	Proceduralist; Race/Gender Neutral Justice
BENIGN	0.07	1.17	36.14	Intent vs. Consequences; Constraint
DISCRIM*	14.86	25.04	24.13	Proceduralist; Race/Gender Neutral Justice
PURPORT*	0.44	1.88	24.13	Skepticism
CLASSIF*	2.1	11.54	22.39	Proceduralist; Race/Gender Neutral Justice
NARROW-TAILORING	0.05	0.96	19.73	Proceduralist; Strict Scrutiny
REJECT*	2.75	7.79	19.15	Oppositional Posture
JUSTIF*	2.39	12.79	18.91	Proceduralist; Constraint
FORBID*	0.38	1.63	18.91	Proceduralist; Constraint; Race/Gender Neutral Justice
PROHIBITS	0.13	0.71	18.08	Proceduralist; Constraint
RATIONALE	0.66	5.92	17.58	Proceduralist; Legalistic
AMORPHOUS	0.25	1.29	14.62	Proceduralist; Skepticism
RACE-BASED	1.08	10.46	10.59	Proceduralist; Pejorative counterpart to liberal RACE-CONSCIOUS
Liberal Words				
LEADERS	2.70	0.13	31.03	Impact; Development
WORLD	3.00	0.42	18.74	Impact; Global
NATION*	21.0	7.04	17.90	Impact; Communitarian
IMPACT*	4.13	1.04	17.49	Impact
EFFECTIVE	2.78	0.75	16.54	Impact; Effectiveness
SOCIAL	6.84	1.71	16.05	Impact; Communitarian
COMMUNIT*	8.75	1.75	15.35	Impact; Communitarian
BUSINESS*	4.56	0.58	10.28	Impact; Efficiency; Distributive Justice
DESEGREGATION	2.34	0.17	10.24	Remedial Justice
GROW*	2.38	0.33	10.24	Change; Development
WORKFORCE	1.64	0.00	9.81	Impact; Distributive Justice; Development
RACE-CONSCIOUS	7.14	1.50	7.80	Proceduralist; Euphemistic counterpart to conservative RACE-BASED

Compare and Contrast

Bara et al. (2007) abortion debate with a thematic dictionary



Vocabulary*	Liberals-restrictionists (1966)
Advocacy	-2.70
Legal	2.22
Medical	-0.34
Moral	-1.02
Rhetoric of debate	0.42
Social	1.48

Note: *as % total dictionary present.

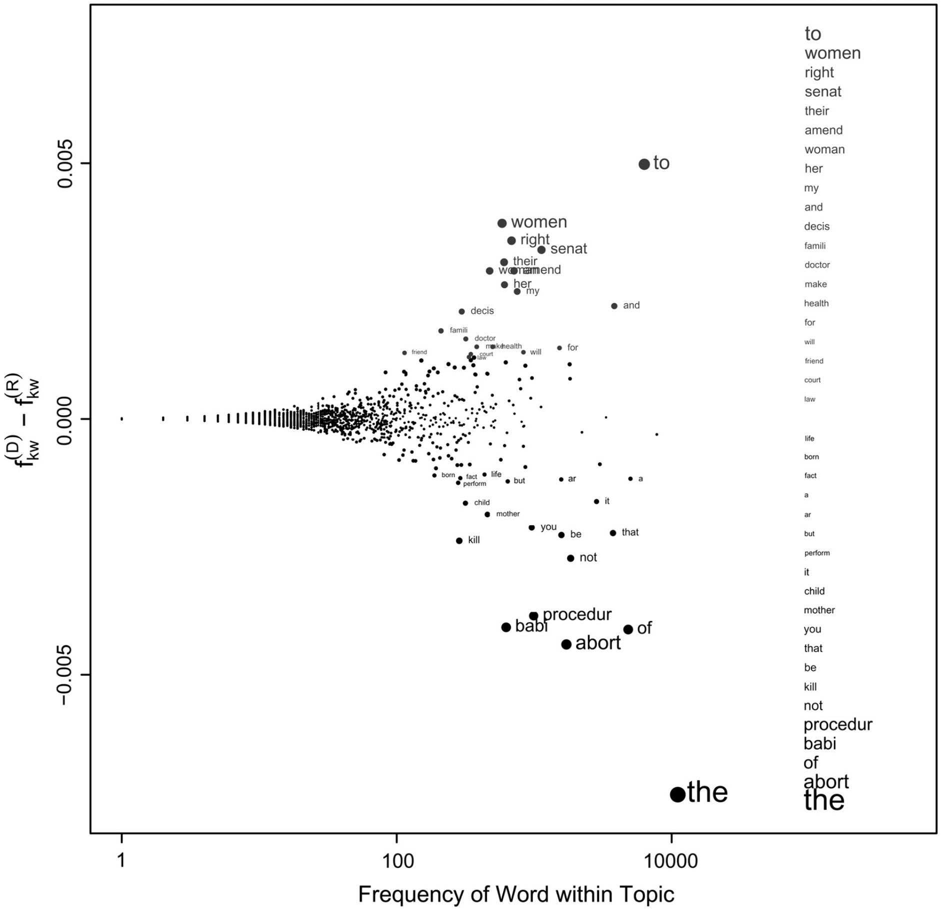
Vocabulary Usage

We can use known Z to characterize vocabulary usage directly, if we're careful. . .

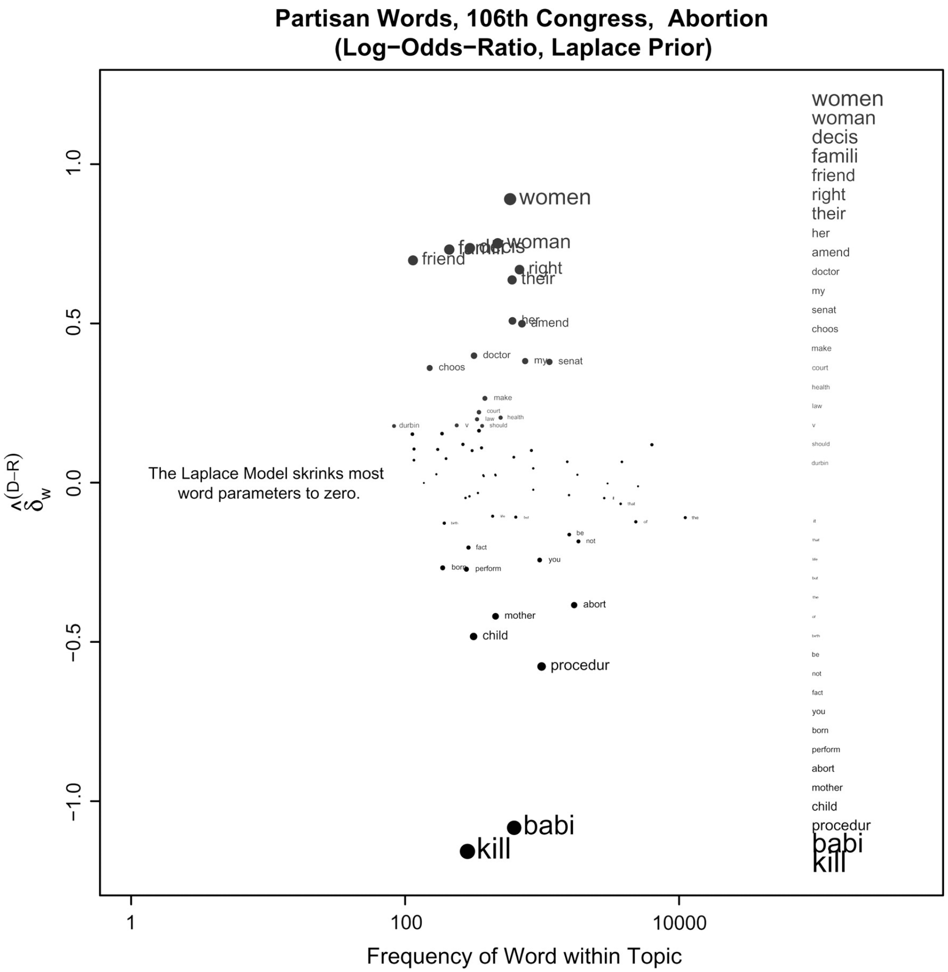
Monroe et al. (2008) compare different measures of 'partisan vocabulary' in abortion debates

- Simple frequencies will be misleading
- They settle for Laplace regularized odds-ratios

Partisan Words, 106th Congress, Abortion
(Difference of Proportions)



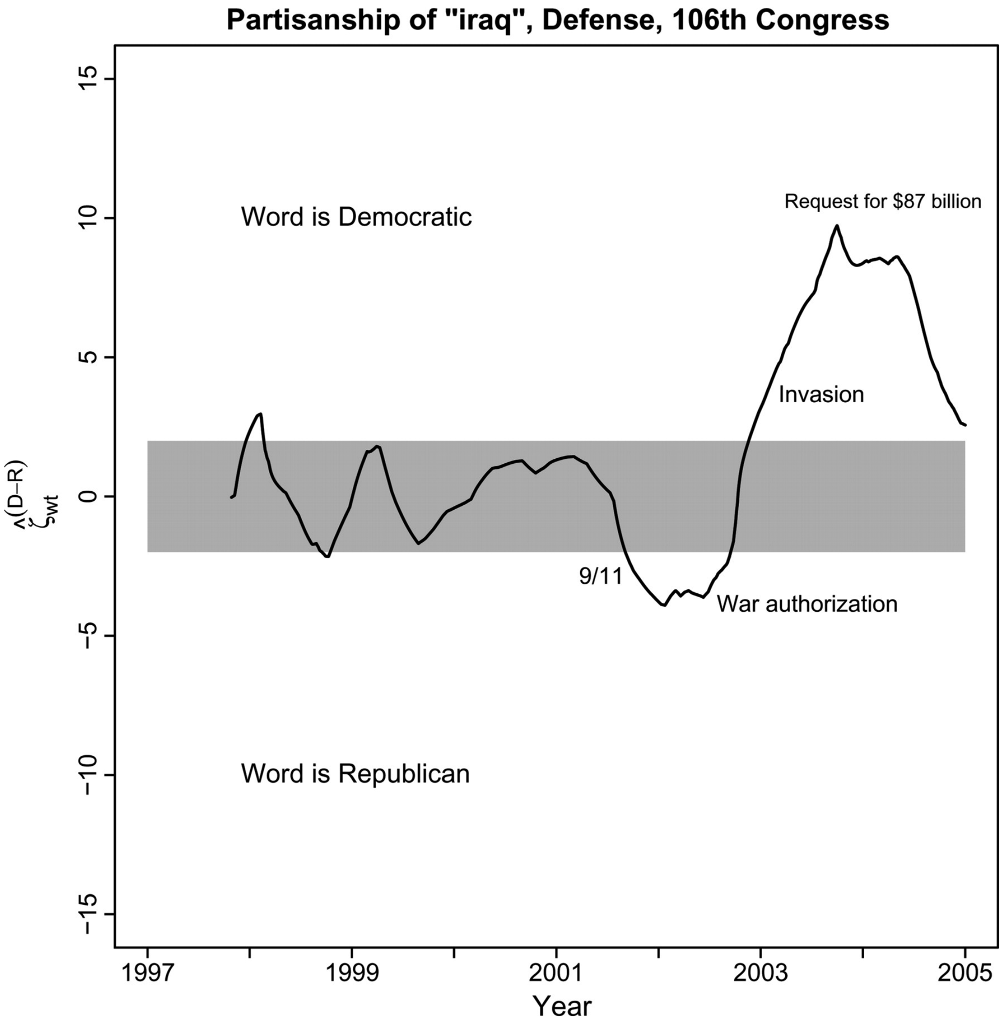
Difference of proportions: could result in lack of overall semantic validity due to the overemphasis on high-frequency words, unclear which words matter (Monroe et al. 2009).



The Laplace Model skrinks most word parameters to zero.

An additional prior means that words whose partisanship is not clear will receive partisan contrasts that are exactly zero. Identifying important words is now easier (Monroe et al. 2009).

Lexical Instability



Evaluation case study

For large numbers of categories, evaluation – even constructing a reliable confusion matrix – can be tiresome

For automated classifiers only, a lazy method is possible (King and Lowe, 2003)

Evaluation case study: events

Russian artillery south of the Chechen capital Grozny blasted Chechen positions overnight before falling silent at dawn, witnesses said on Tuesday.

Israel said on Tuesday it sent humanitarian aid to Colombia where a massive earthquake last week killed at least 938 people and injured 400.

Event data extraction

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

Israel^{S} said on Tuesday it *sent humanitarian aid*^{073} to *Colombia*^{T} where a *massive earthquake*^{S} last week *killed*^{222} at least *938 people*^{T} and injured 400.

Event data extraction

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

Israel^{S} said on Tuesday it *sent humanitarian aid*^{073} to *Colombia*^{T} where a *massive earthquake*^{S} last week *killed*^{222} at least *938 people*^{T} and injured 400.

```
20010901 RUS CHE 223
20020804 ISR COL 073
20020804 -- COL 222
```

Dyadic event data (Serbia-Bosnia)

Week	Code	Description
1995-07-11	211	SEIZE POSSESSION
	212	ARREST PERSON
	223	MILITARY ENGAGEMENT
1995-07-12	211	SEIZE POSSESSION
	223	MILITARY ENGAGEMENT
	173	SPECIF THREAT
	191	CANCEL EVENT
	211	SEIZE POSSESSION
	095	PLEAD
	111	TURN DOWN
	212	ARREST PERSON
	081	MAKE AGREEMENT
	023	NEUTRAL COMMENT
	032	VISIT
	031	MEET

Scaled dyadic event data

Week	Code	Score [-10,10)
1995-07-11	211	-9.2
	212	-9.0
	223	-10.0
1995-07-12	211	-9.2
	223	-10.0
	173	-7.0
	191	-2.2
	211	-9.2
	095	1.2
	111	-4.0
	212	-9.0
	081	6.5
	023	-0.2
	032	1.9
	031	1.0

The human elements

Coders read newswire and extract events
(e.g. GEDS projects, Swisspeace)

Experts assign scores to event types
(e.g. Goldstein 1995, Shellman 2004)

Analysts aggregate and infer conflict dynamics
(Goldstein & Pevehouse 1997, Pevehouse & Goldstein 1999)

Evaluating an event data system

Two aspects of event type evaluation
(King and Lowe 2003):

If machine says it's a use of force, is it really?
(Precision / specificity)

If it's a use of force, will the machine say it is?
(Recall / sensitivity)

$$P(T \mid M = \text{"223"}) \quad \text{vs.} \quad P(M \mid T = \text{"223"})$$

Evaluating precision

To estimate $P(T \mid M)$

1. Run the machine over all news leads
2. Select an equal number of examples from each machine assigned category M
3. Identify their true event type T

Boring (code 711 leads from $>$150 event types sampled from $P(M)$) but straightforward

Evaluating recall

We need a gold standard T, but. . .

“of the 45,000 events coded coded by the VRA Reader from news leads on the former Yugoslavia, it found 10,605 neutral comments but only 4 apologies and 35 threats of military attack.”

Coders might have to wade through ∼2500 comments to find an apology and ∼ 300 comments to find a threat of force.

and we’re lazy. . .

Solution

Stratified sampling

Choose events according to what category the machine put them in

– this is biased!

Correct for the bias

Recall from precision plus computation

$$P(M \mid T) = \frac{P(T \mid M)P(M)}{P(T)}$$

Recall from precision plus computation

$$P(M \mid T) = \frac{P(T \mid M)P(M)}{P(T)}$$

P(M) is just the event type *tabulation*

. . . run the machine on all 45,000 leads

Recall from precision plus computation

$$P(M \mid T) = \frac{P(T \mid M)P(M)}{P(T)}$$

P(M) is just the event type *tabulation*

. . . run the machine on all 45,000 leads

P(T) is a normalizing constant.

Undergrads vs. the machine

	All Codes				WEIS Codes			
	M	$U^{(1)}$	$U^{(2)}$	$U^{(3)}$	M	$U^{(1)}$	$U^{(2)}$	$U^{(3)}$
$w = 1$								
detailed	.26	.32	.23	.26	**.25**	**.44**	**.25**	**.37**
aggregate	.55	.55	.39	.48	**.62**	**.62**	**.48**	**.62**
$w = P(t)$								
detailed	.52	.48	.35	.42	.55	.64	.35	.68
aggregate	.65	.70	.53	.64	.70	.72	.56	.65
$w = 1/\sqrt{P(t)}$								
detailed	.36	.44	.33	.41	.37	.62	.34	.67
aggregate	.59	.66	.49	.62	.64	.68	.53	.63

Summary

Two points about evaluating machine

With a machine classifier we can be lazier than usual when constructing the confusion matrix

Stratification allows you choose an evaluation score that reflects the cost of mistakes in the task

The ‘inversion’ methods discussed yesterday can be applied here too. . .

Classification: Good Old Logit

It is tempting to go with methods we know (disciminative style)

$$\begin{aligned}\widehat{\theta}_{k|w} &= P(Z = k \mid W_1 \ldots W_V) \\ &= \text{logit}^{-1}(\alpha + \beta_1 W_1 + \beta_2 W_2 + \ldots)\end{aligned}$$

This is a bad idea

Why?

Bad Old Logit

The number of word types V is almost much larger than the number of documents D

Many more 'cases' than 'variables'

and no constraint on possible solutions

We need *serious* regularisation. . .

Wrapping up

Two approaches

Discriminative: Model $P(Z \mid \{W\}, \beta) = \theta_{z|w}$ directly

Generative: Model $P(\{W\} \mid Z, \beta)$ and $P(Z) = \theta_z$, then get $P(Z \mid \{W\}, \beta) = \theta_{z|w}$ via Bayes theorem

For pure blackbox efficiency, go for the first!