

---

# Text Classification

---

Michal Rosen-Zvi  
University of California, Irvine

---

# Outline

- The need for dimensionality reduction
- Classification methods
- Naïve Bayes
- The LDA model
- Topics model and semantic representation
- The Author Topic Model
  - Model assumptions
  - Inference by Gibbs sampling
  - Results: applying the model to massive datasets

# The need for dimensionality reduction

- Content-Based Ranking:
  - Ranking matching documents in a search engine according to their relevance to the user
  - Presenting documents as vectors in the words space - 'bag of words' representation
  - It is a sparse representation,  $V \gg |D|$
- A need to define conceptual closeness

# Feature Vector representation

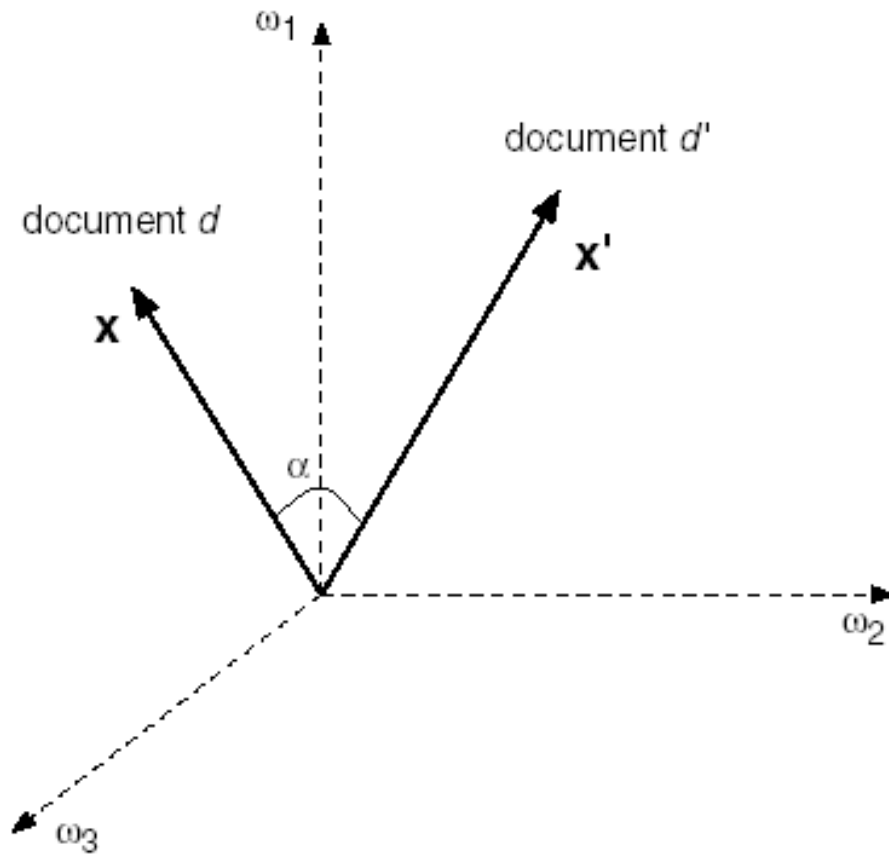


Figure 4.2 Cosine measure of document similarity.

From: *Modeling the Internet and the Web: Probabilistic methods and Algorithms*, Pierre Baldi, Paolo Frasconi, Padhraic Smyth

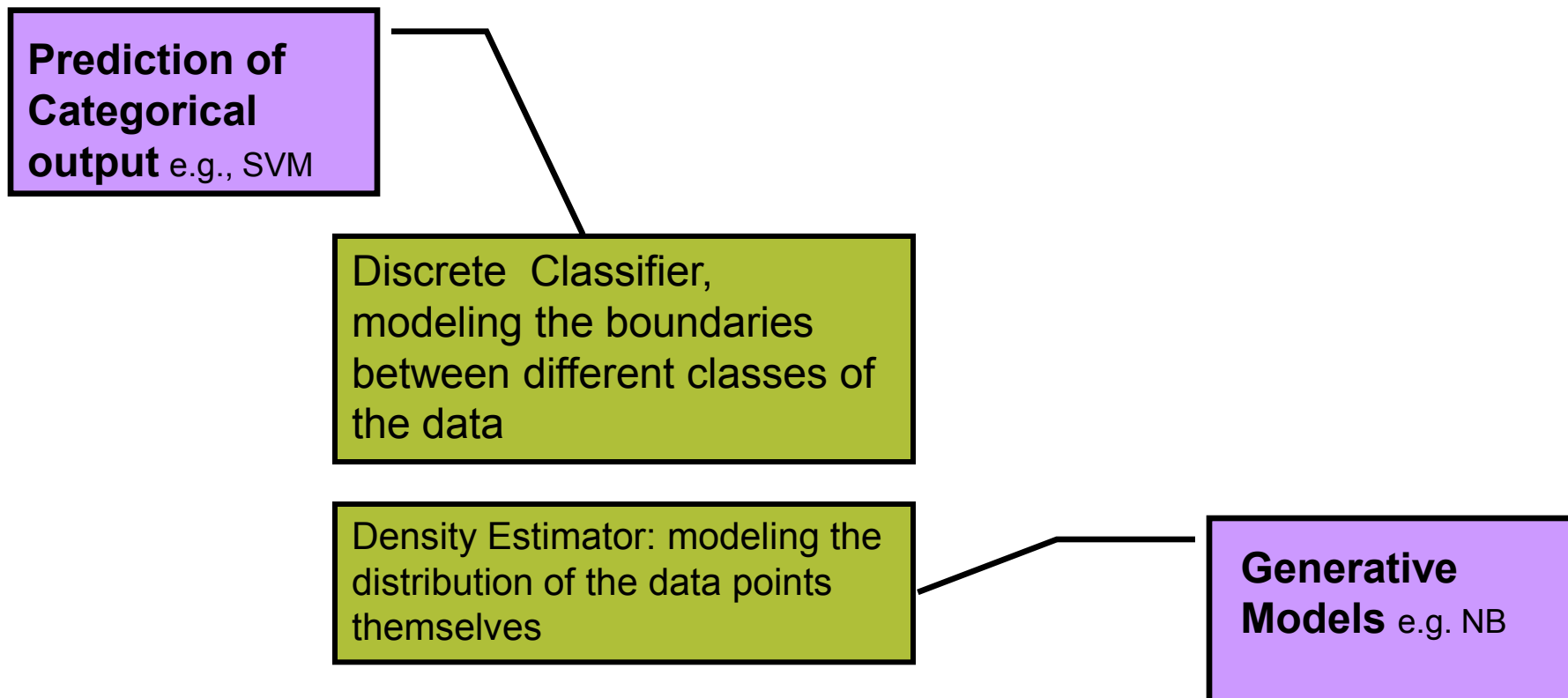
---

# What is so special about text?

- No obvious relation between features
- High dimensionality, (often larger vocabulary,  $V$ , than the number of features!)
- Importance of speed

# Classification: assigning words to topics

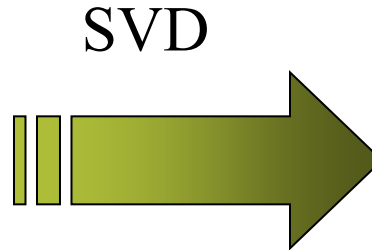
Different models for data:



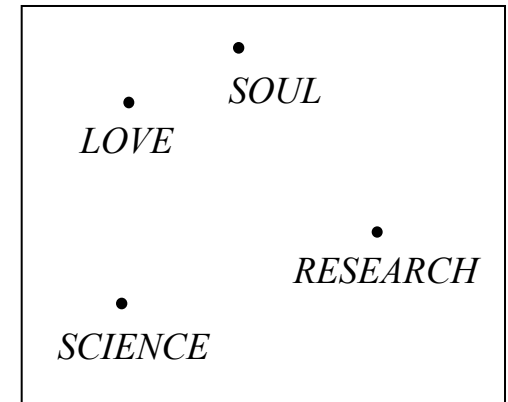
# A Spatial Representation: Latent Semantic Analysis (Landauer & Dumais, 1997)

Document/Term count matrix

	Doc1	Doc2	Doc3 ...
<i>LOVE</i>	34	0	3
<i>SOUL</i>	12	0	2
<i>RESEARCH</i>	0	19	6
<i>SCIENCE</i>	0	16	1
...	...	...	...



High dimensional space,  
not as high as  $|V|$



EACH WORD IS A *SINGLE* POINT IN A SEMANTIC SPACE

# Where are we?

- The need for dimensionality reduction
- Classification methods
- **Naïve Bayes**
- The LDA model
- Topics model and semantic representation
- The Author Topic Model
  - Model assumptions
  - Inference by Gibbs sampling
  - Results: applying the model to massive datasets



---

# The Naïve Bayes classifier

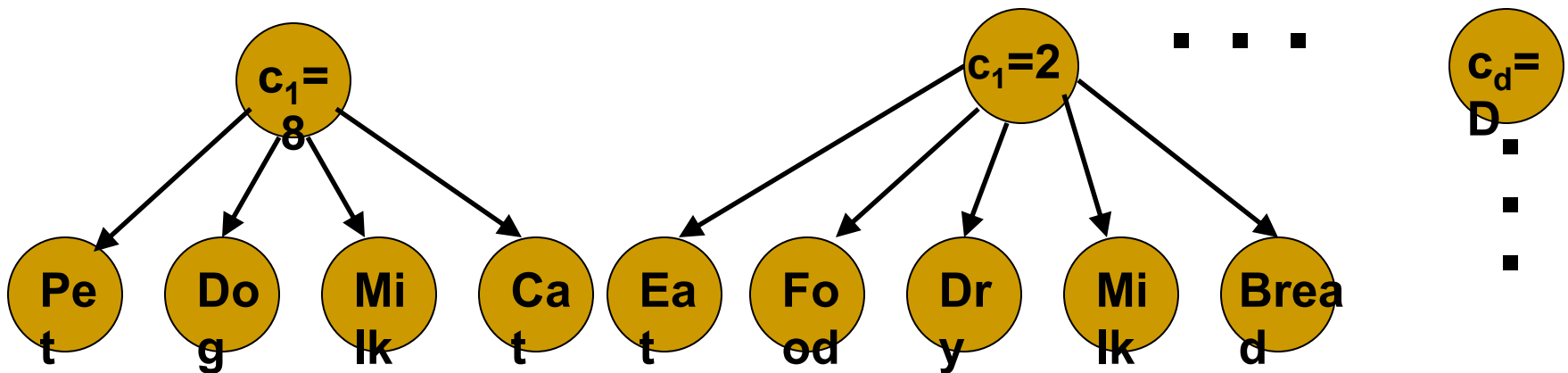
- Assumes that each of the data points is distributed independently:
- Results in a trivial learning algorithm
- Usually does not suffer from overfitting

# Naïve Bayes classifier: words and topics

A set of labeled documents is given:

$$\{C_d, \mathbf{w}_d: d=1, \dots, D\}$$

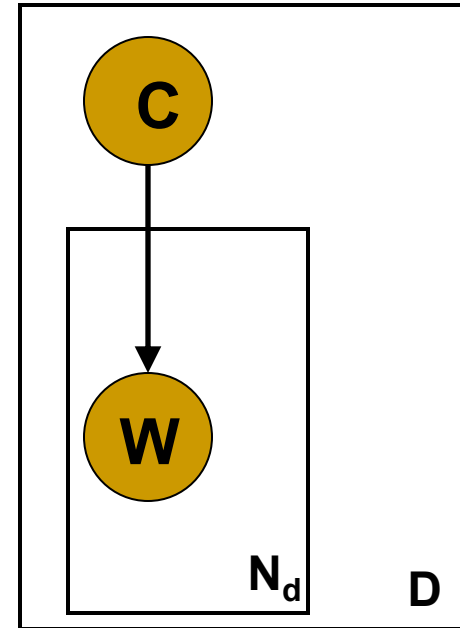
Note: classes are mutually exclusive



# Simple model for topics

Given the topic  
words are  
independent

The probability for  
a word,  $w$ , given  
a topic,  $z$ , is  $\theta_{wz}$



$$P(\{\mathbf{w}, C\} | \theta) = \prod_d P(C_d) \prod_{n_d} P(w_{n_d} | C_d, \theta)$$

# Learning model parameters

Estimating  $\theta$  from the probability:  $P(\{\mathbf{w}, c\} | \theta) = \prod_{i=1 \dots N_D} P(w_i | \theta, c_d = j) \prod_{d=1 \dots D} P(c_d = j) = \theta_{jw}^{n_j^{(w)}}$

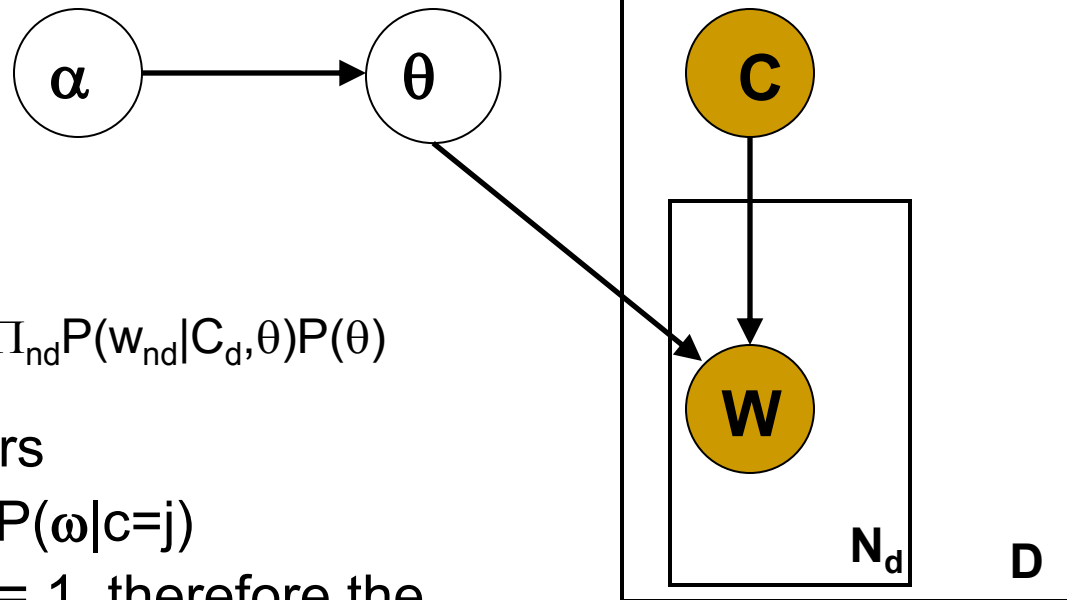
Here  $\theta_{jw}$  is the probability for word  $w$  given topic  $j$  and  $n_j^{(w)}$  is the number of times the word  $w$  is assigned to topic  $j$

Under the normalization constraint, one finds  $\hat{\theta}_{j,w} = \frac{n_{j,w}}{\sum_w n_{j,w}}$

Example of making use of the results: predicting the topic of a new document

$$P(c | \mathbf{w}, \theta) = \frac{P(\mathbf{w} | c, \theta)P(c)}{P(\mathbf{w} | \theta)} \propto P(\mathbf{w} | c, \theta)P(c)$$

# Naïve Bayes, multinomial:



$$P(\{\mathbf{w}, C\}) = \int d\theta \prod_d P(C_d) \prod_{nd} P(w_{nd} | C_d, \theta) P(\theta)$$

Generative parameters

$$\theta_{wj} = P(\omega | c=j)$$

- Must satisfy  $\sum_w \theta_{wj} = 1$ , therefore the integration is over the simplex, (space of vectors with non-negative elements that sum up to 1)
- Might have Dirichlet prior,  $\alpha$

# Inferring model parameters

One can find the distribution of  $\theta$  by sampling

$$P(\theta | c, \mathbf{w}, \alpha) = \frac{P(\mathbf{w} | c, \theta, \alpha)P(c)}{\int d\theta P(\mathbf{w} | c, \theta, \alpha)P(c)}$$

Making use of the MAP:

$$P(\mathbf{w}, c | \theta, \alpha) = \frac{P(\mathbf{w} | c, \theta, \alpha)P(c)}{P(\mathbf{w} | \theta, \alpha)} \propto P(\mathbf{w} | c, \theta, \alpha)P(c)$$

$$\hat{\theta}_{w,j} = \frac{\alpha + n_j^{(w)}}{\alpha V + \sum_{l=1}^{|V|} n_j^{(l)}}$$

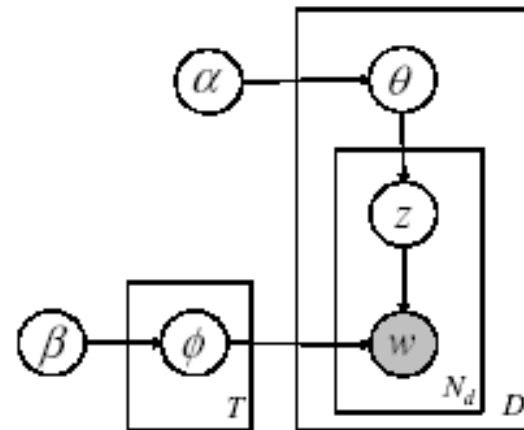
This is a point estimation of the PDF, provides the mean of the posterior PDF under some conditions provides the full PDF

# Where are we?

- The need for dimensionality reduction
- Classification methods
- Naïve Bayes
- **The LDA model**
- Topics model and semantic representation
- The Author Topic Model
  - Model assumptions
  - Inference by Gibbs sampling
  - Results: applying the model to massive datasets

# LDA: A generative model for topics

- A model that assigns Dirichlet priors to multinomial distributions: **L**atent **D**irichlet **A**llocation
- Assumes that a document is a mixture of topics





---

# LDA: Inference

Fixing the parameters  $\alpha$ ,  $\beta$  (assuming uniformity) and inferring the distribution of the latent variables:

- Variational inference (Blei et al)
- Gibbs sampling (Griffiths & Steyvers)
- Expectation propagation (Minka)

# Sampling in the LDA model

**The update rule for fixed  $\alpha, \beta$  and integrating out  $\theta$**

$$P(z_i = j | w_i = m, \mathbf{z}_{-i}, \mathbf{W}_{-i}) \propto \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \frac{C_{dj}^{DT} + \alpha}{\sum_{j'} C_{dj'}^{DT} + T\alpha}$$

**Provides point estimates to  $\theta$  and distributions of the latent variables,  $\mathbf{z}$ .**

# Making use of the topics model in cognitive science...

- The need for dimensionality reduction
- Classification methods
- Naïve Bayes
- The LDA model
- **Topics model and semantic representation**
- The Author Topic Model
  - Model assumptions
  - Inference by Gibbs sampling
  - Results: applying the model to massive datasets

# The author-topic model

- Automatically extract topical content of documents
- Learn association of topics to authors of documents
- Propose new efficient probabilistic topic model: the author-topic model
- Some queries that model should be able to answer:
  - What topics does author  $X$  work on?
  - Which authors work on topic  $X$ ?
  - What are interesting temporal patterns in topics?

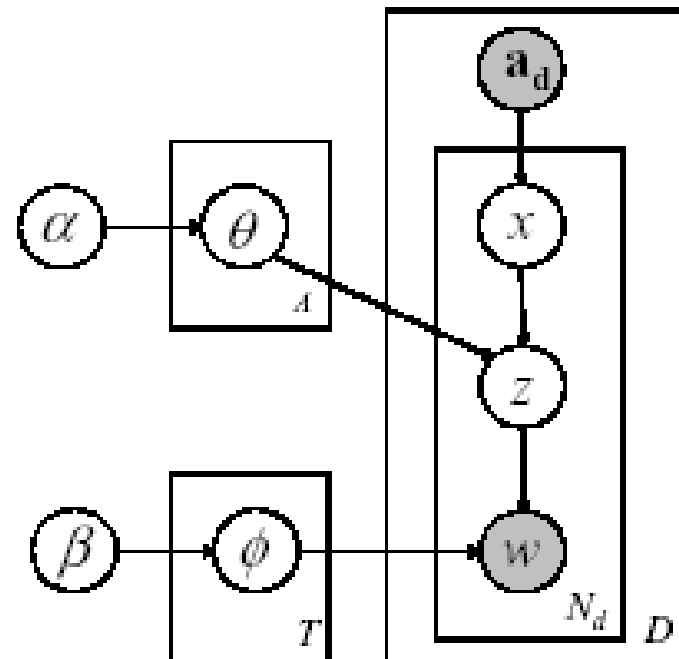
---

# The model assumptions

- Each author is associated with a topics mixture
- Each document is a mixture of topics
- With multiple authors, the document will be a mixture of the topics mixtures of the coauthors
- Each word in a text is generated from *one* topic and *one* author  
(potentially different for each word)

# The generative process

- Let's assume authors  $A_1$  and  $A_2$  collaborate and produce a paper
  - $A_1$  has multinomial topic distribution  $\theta_1$
  - $A_2$  has multinomial topic distribution  $\theta_2$
- For each word in the paper:
  1. Sample an author  $x$  (uniformly) from  $A_1, A_2$
  2. Sample a topic  $z$  from a  $\theta_x$
  3. Sample a word  $w$  from a multinomial topic distribution

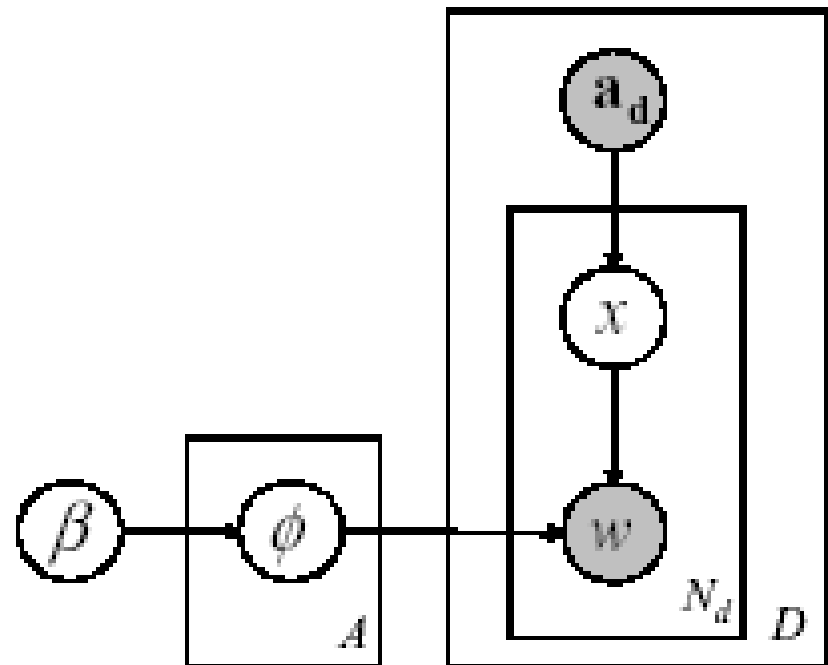


# Inference in the author topic model

- Estimate  $\mathbf{x}$  and  $\mathbf{z}$  by Gibbs sampling  
(assignments of each word to an author and topic)
- Estimation is efficient: linear in data size
- Infer from each sample using point estimations:
  - Author-Topic distributions ( $\Theta$ )
  - Topic-Word distributions ( $\Phi$ )
- View results at the [author-topic model website](#) [off-line]

# Naïve Bayes: author model

- Observed variables: authors and words on the document
- Latent variables: concrete authors that generated each word
- The probability for a word given an author is multinomial with Dirichlet prior





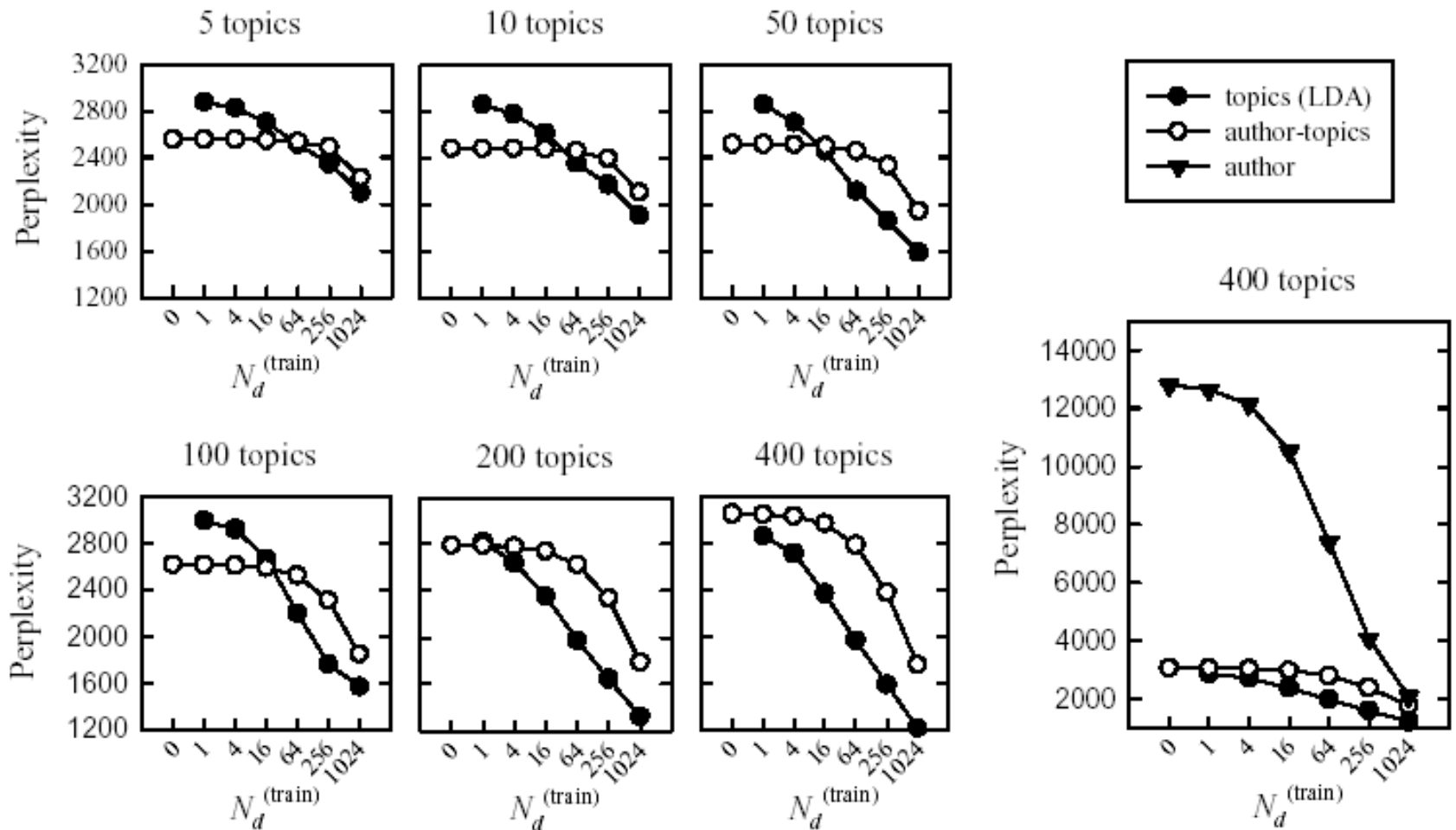
# Results: Perplexity

Lower perplexity indicates a better generalization performance

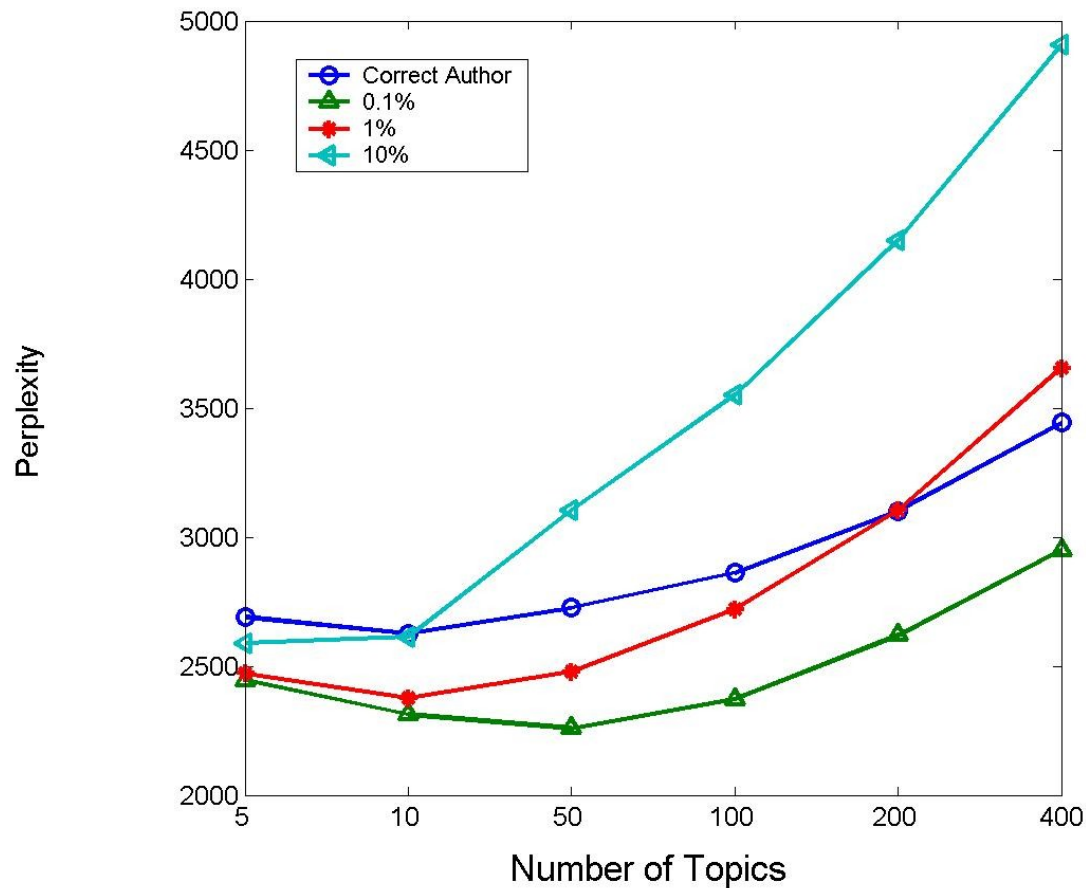
$$\text{perplexity}(\mathbf{w}_d | \mathbf{a}_d) = \exp \left[ -\frac{\ln p(\mathbf{w}_d | \mathbf{a}_d)}{N_d} \right]$$

$$p(\mathbf{w}_d | \mathbf{a}_d) = \int d\theta \int d\phi p(\theta | \mathcal{D}^{\text{train}}) p(\phi | \mathcal{D}^{\text{train}}) \cdot \prod_{m=1}^{N_d} \left[ \frac{1}{A_d} \sum_{i \in \mathbf{a}_d, j} \theta_{ij} \phi_{w_m j} \right] \cdot$$

# Results: Perplexity (cont.)



# Perplexity and Ranking results



# Perplexity and Ranking results (cont)

Can the model  
predict the  
correct  
author?

