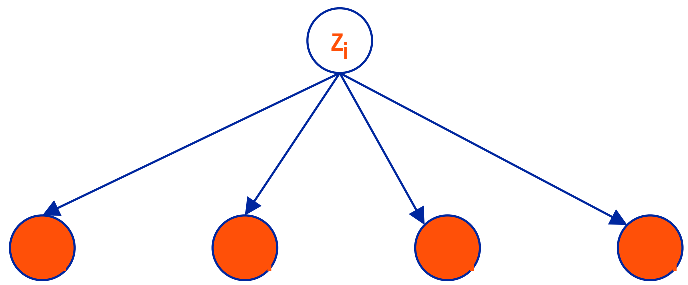
Latent Dirichlet Allocation (LDA)

Following slides borrowed (with heavily modification) from:

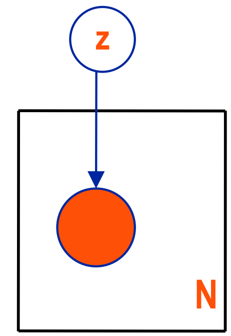
Jonathan Huang (jch1@cs.cmu.edu)

D. Blei, A. Ng, and M. Jordan. *Journal of Machine Learning Research*, 3:993-1022, 2003.

Naïve Bayes: each doc is on a topic



Mixture of Unigrams = Naïve Bayes



Equivalent Plate Model

Model: For each document:

- Choose a topic z_d with $p(topic_i) = \theta$
- Choose N words w_n by drawing each one independently from a multinomial conditioned on z_d with $p(w_n=word_j|topic_i=z) = \beta_z$
  - *Multinomial*: take a (non-uniform prior) dice with a word on each side; roll the dice N times and count how often each word comes up

In NB, we have exactly one topic per document

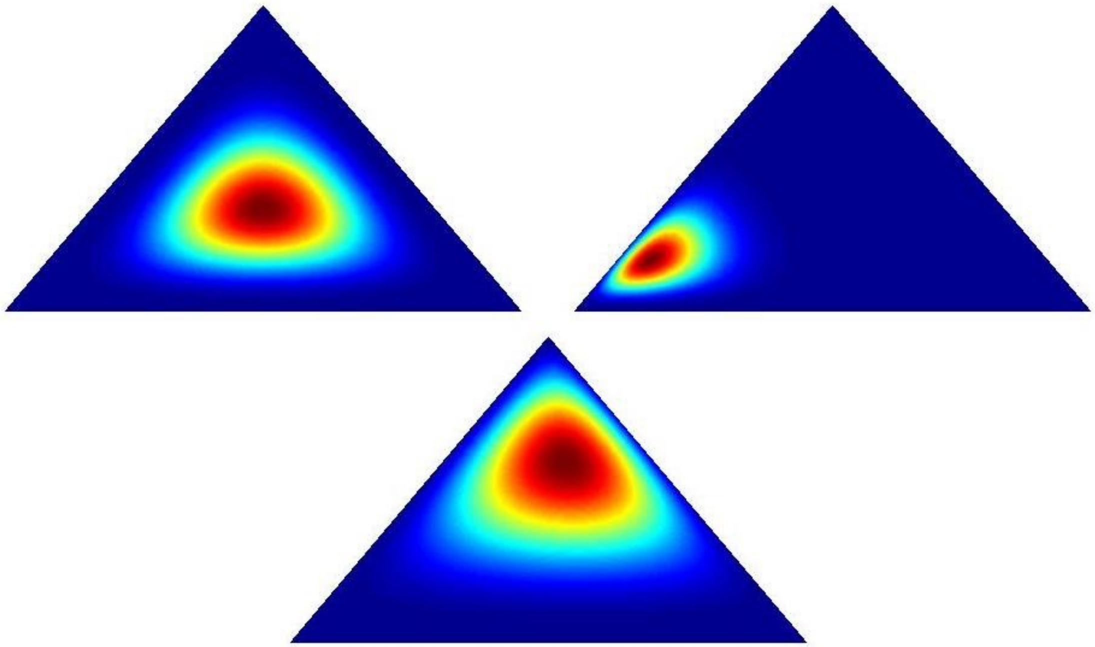
LDA: Each doc is a mixture of topics

- **LDA: each document is a (different) mixture of topics**
  - Naïve Bayes assumes each *document* is on a single topic
  - LDA lets each *word* be on a different topic
  - For each document, d:
    - Choose a *multinomial distribution* θ_d over topics for that document
    - For each of the *N* words w_n in the document
      - —Choose a topic z_n with $p(topic) = \theta_d$
      - —Choose a word w_n from a multinomial conditioned on z_n with $p(w=w_n|topic=z_n)$
      - —Note: each topic has a different probability of generating each word

Dirichlet Distributions

- **In the LDA model, we want the *topic mixture proportions* for each document to be drawn from some *distribution*.**
  - *distribution* = "probability distribution", so it sums to one
- **So, we want to put a prior distribution on multinomials. That is, k-tuples of non-negative numbers that sum to one.**
  - We want probabilities of probabilities
  - These multinomials lie in a (k-1)-*simplex*
    - *Simplex* = generalization of a triangle to (k-1) dimensions.
- **Our prior:**
  - Defined for a (k-1)-simplex.
  - Conjugate to the multinomial

3 Dirichlet Examples (over 3 topics)



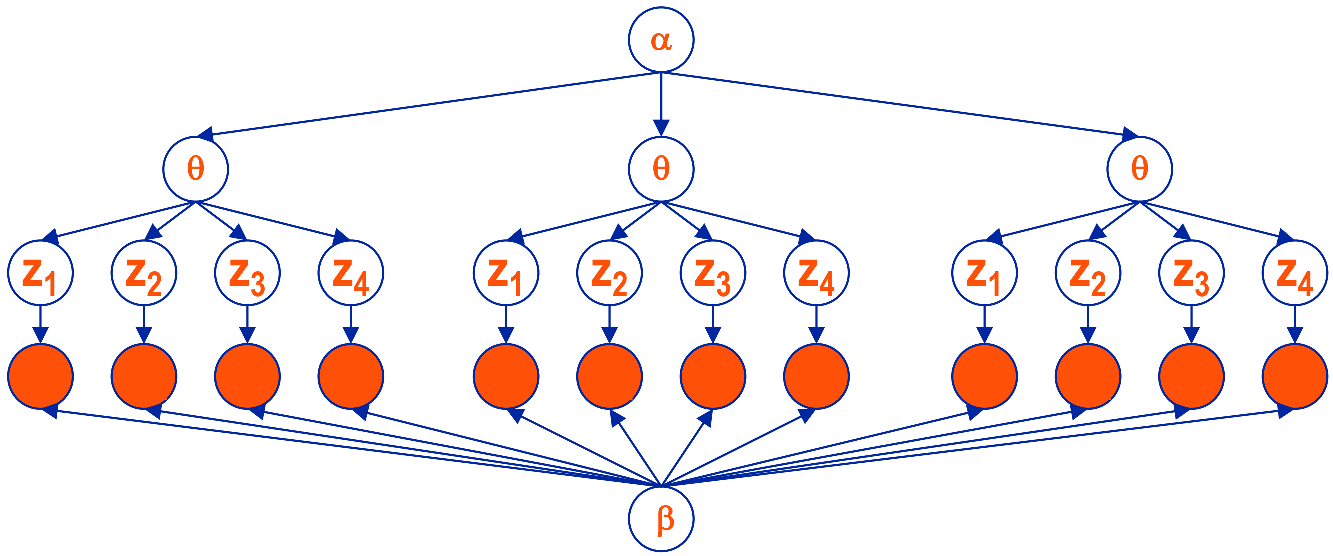
Corners: only one topic
Center: uniform mixture of topics
Colors indicate probability of seeing the topic distribution

Dirichlet Distribution

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

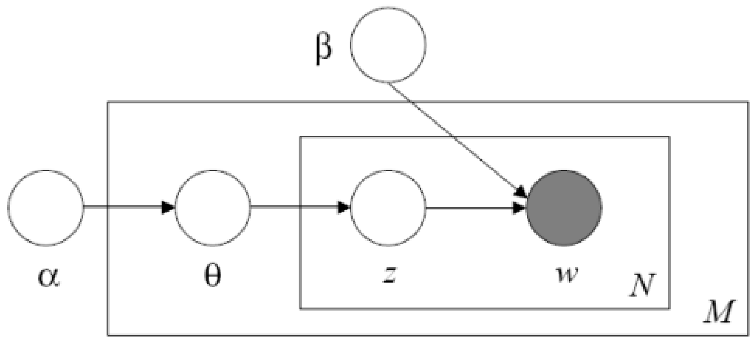
- **Dirichlet distribution**
  - is defined over a (k-1)-simplex. I.e., it takes k non-negative arguments which sum to one.
  - is the conjugate prior to the multinomial distribution.
    - I.e. if our likelihood is multinomial with a Dirichlet prior, then the posterior is also Dirichlet
  - The Dirichlet parameter α_i can be thought of as the prior count of the i^{th} class.

The LDA Model



- **For each document,**
  - Choose the topic distribution $\theta \sim \text{Dirichlet}(\alpha)$
  - For each of the N words w_n:
    - Choose a topic $z \sim \text{Multinomial}(\theta)$
    - Choose a word $w_n \sim \text{Multinomial}(\beta_z)$
      - Where each topic has a different parameter vector $\boldsymbol{\beta}$ for the words

The LDA Model: “Plate representation”



- **For each of M documents,**
  - Choose the topic distribution θ ~ Dirichlet(α)
  - For each of the N words w_n:
    - Choose a topic z ~ Multinomial(θ)
    - Choose a word w_n ~ Multinomial(β_z)

Parameter Estimation

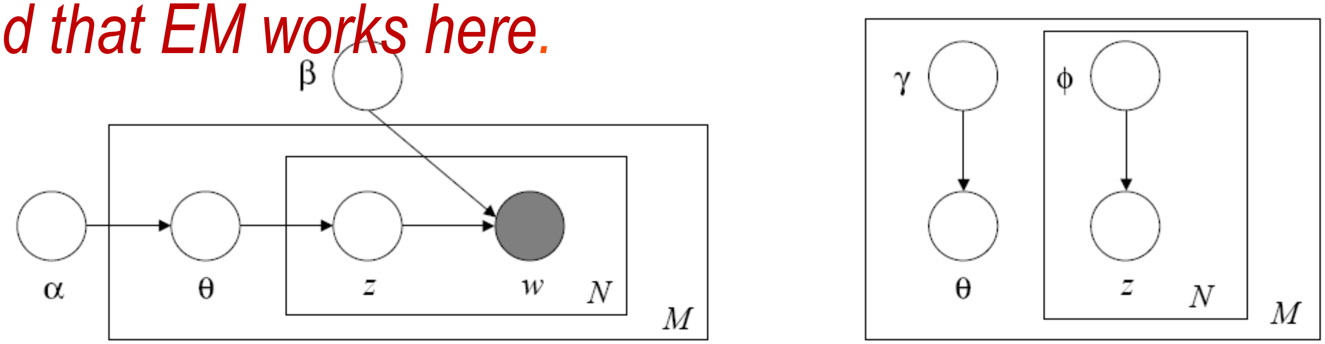
- Given a corpus of documents, find the parameters α and β which maximize the likelihood of the observed data (words in documents), marginalizing over the hidden variables θ, z

θ: topic distribution for the document
z: topic for each word in the document

- **E-step:**
  - Compute $p(\boldsymbol{\theta},\mathbf{z}|\mathbf{w},\alpha,\beta)$, the posterior of the hidden variables $(\theta,\mathbf{z})$ given each document **w**, and parameters α and β.
- **M-step**
  - Estimate parameters α and β given the current hidden variable distribution estimates
- **Unfortunately, the E-step cannot be solved in a closed form**
  - So people use a “variational” approximation

Variational Inference

You don't need to know the details; only what is hidden and what observed; and that EM works here.



- In variational inference, we consider a simplified graphical model with variational parameters γ, ϕ and minimize the KL Divergence between the variational and posterior distributions.
    - *q* approximates *p*

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma,\phi)} KL(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta))$$

Parameter Estimation: *Variational EM*

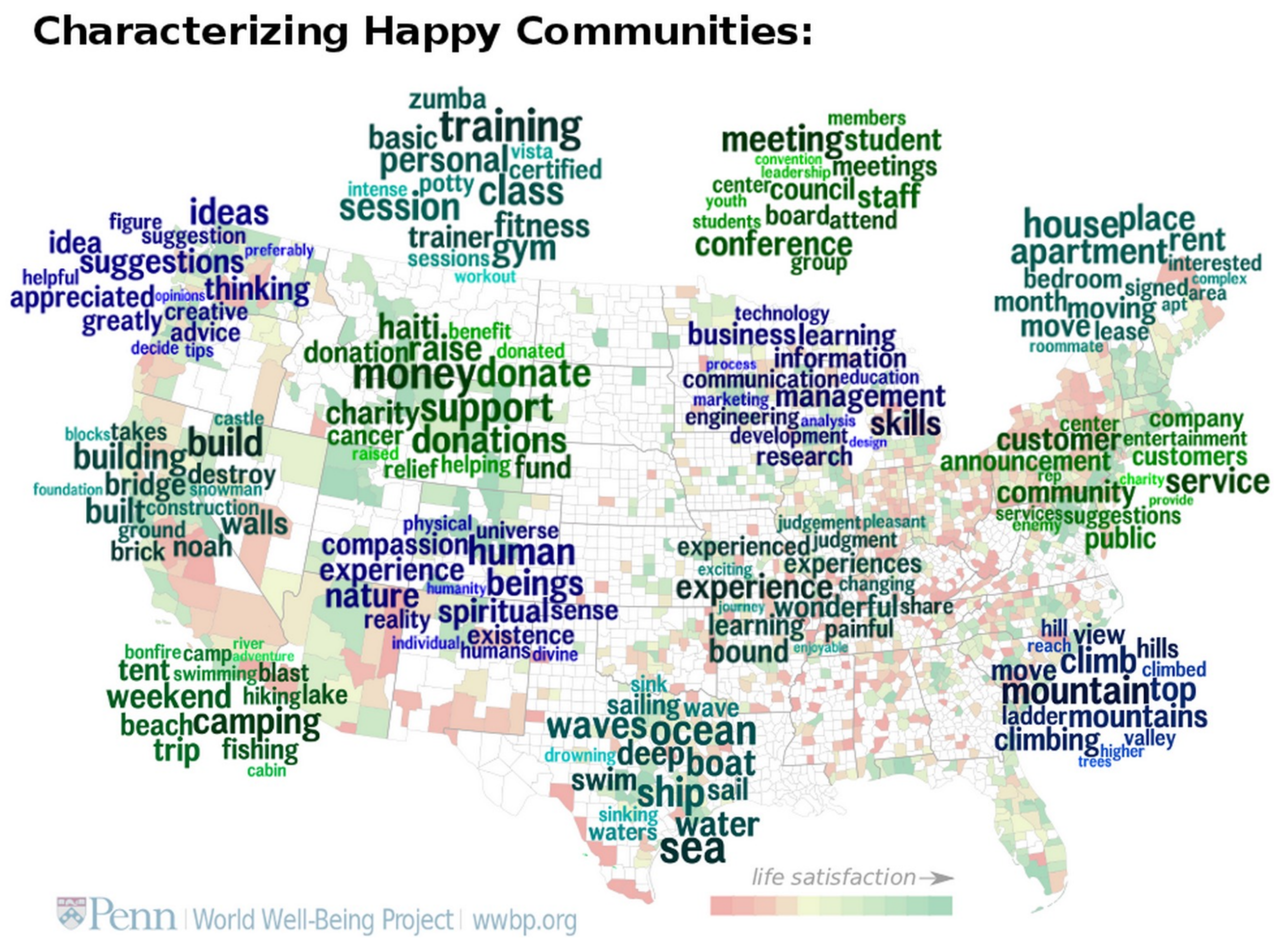
- Given a corpus of documents, find the parameters α and β which maximize the likelihood of the observed data.
- **E-step:**
  - Estimate the variational parameters γ and ϕ in $q(\gamma,\phi;\alpha,\beta)$ by minimizing the KL-divergence to p (with α and β fixed)
- **M-step**
  - Maximize (over α and β) the lower bound on the log likelihood obtained using q in place of p (with γ and ϕ fixed)

You don't need to know the details; only what is hidden and what observed; and that EM works here.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

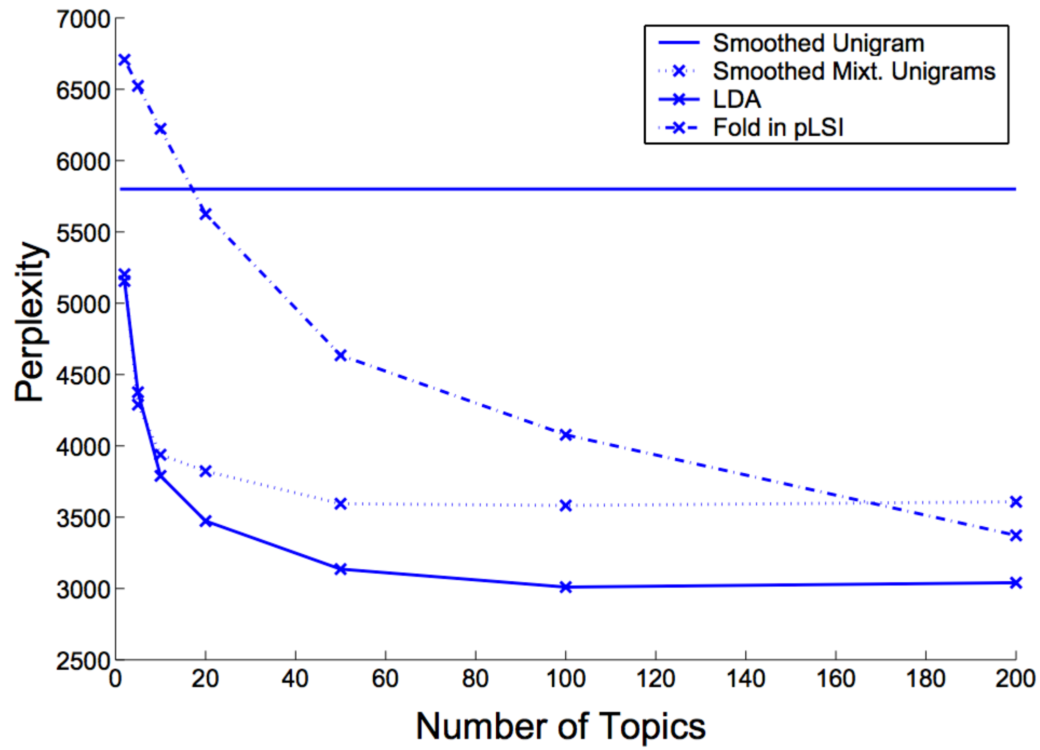
LDA topics can be used for semi-supervised learning



LDA requires fewer topics than NB

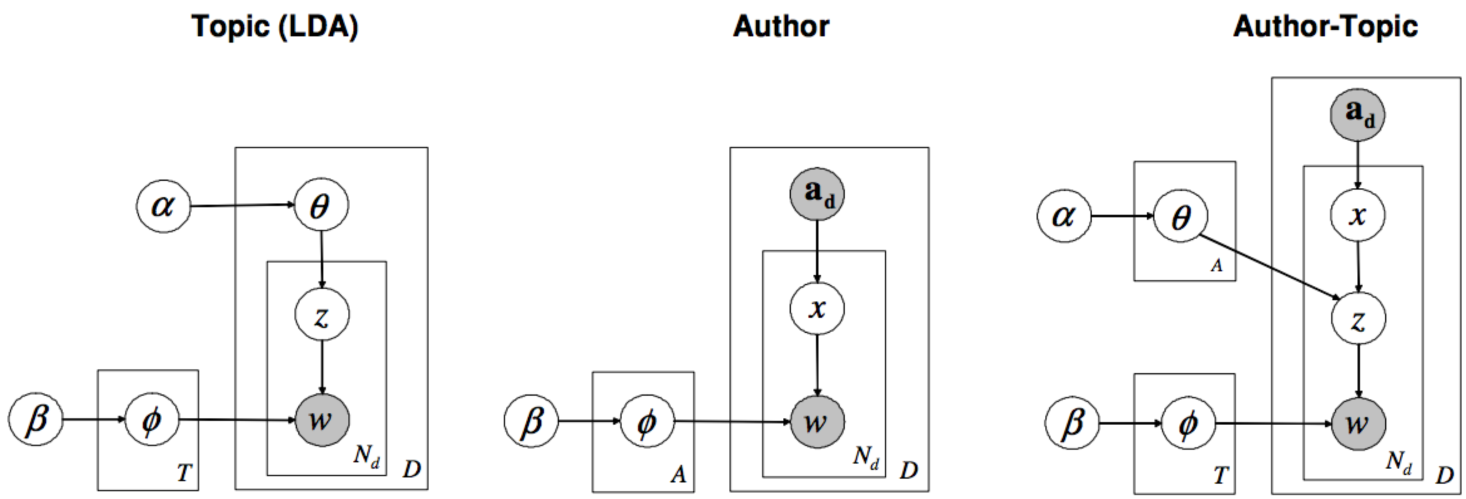
perplexity = $2^{H(p)}$
per word

I.e.,
$\log_2$(perplexity) =
entropy, H



There are *many* LDA extensions

The author-topic model

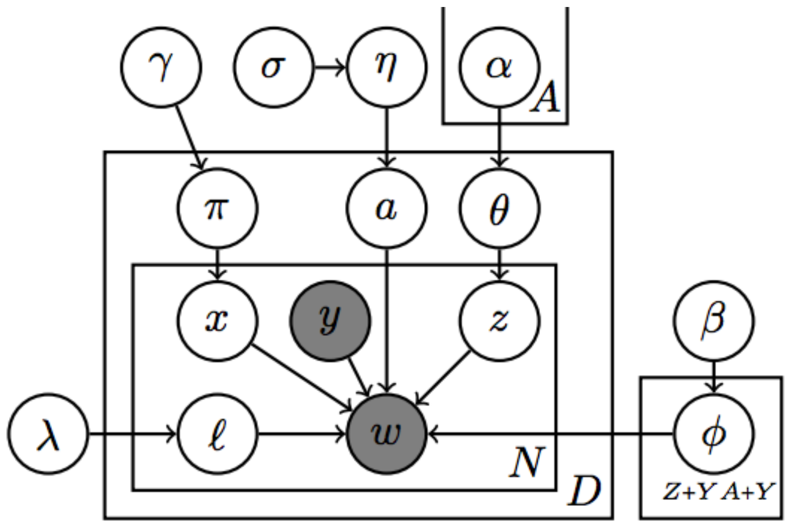


z = topic

x = author

You don't need to know the details; only that there are different hidden and observed features; and that EM works here.

Ailment Topic Aspect Model



Observed

word w

aspect y = symptom, treatment or other

Hidden

topic type: background? (l), non-ailment (x)

topic distribution

- Set the background switching binomial λ
- Draw an ailment distribution $\eta \sim \mathrm{Dir}(\sigma)$
- Draw word multinomials $\phi \sim \mathrm{Dir}(\beta)$ for the topic, ailment, and background distributions
- For each message $1 \leq m \leq D$:
  - Draw a switching distribution $\pi \sim \mathrm{Beta}(\gamma_0, \gamma_1)$
  - Draw an ailment $a \sim \mathrm{Mult}(\eta)$
  - Draw a topic distribution $\theta \sim \mathrm{Dir}(\alpha_a)$
  - For each word $w_i \in N_m$
    - Draw aspect $y_i \in \{0, 1, 2\}$ (observed)
    - Draw background switcher $\ell \in \{0, 1\} \sim \mathrm{Bi}(\lambda)$
    - If $\ell == 0$:
      - Draw $w_i \sim \mathrm{Mult}(\phi_{B,y})$ (a background)
    - Else:
      - Draw $x_i \in \{0, 1\} \sim \mathrm{Bi}(\pi)$
      - If $x_i == 0$: (draw word from topic z)
        - Draw topic $z_i \sim \mathrm{Mult}(\theta)$
        - Draw $w_i \sim \mathrm{Mult}(\phi_z)$
      - Else: (draw word from ailment a aspect y)
        - Draw $w_i \sim \mathrm{Mult}(\phi_{a,y})$

Paul & Dredze

What you should know about LDA

- **Each document is a mixture over topics**
- **Each topic looks like a Naïve Bayes model**
  - It produces words with some probability
- **Estimation of LDA is messy**
  - Requires variational EM or Gibbs sampling
- **In a plate model, each "plate" represents repeated nodes in a network**
  - The plate model shows conditional independence, but not the form of the statistical distribution (e.g. Gaussian, Poisson, Dirichlet, ….)

LDA generation - example

- **Topics** = {sports, politics}
- **Words** = {*football, baseball, TV, win, president*}

$\alpha = (0.8, 0.2)$

$\beta =$

	sports	**politics**
football	0.30	0.01
baseball	0.25	0.01
TV	0.10	0.15
win	0.30	0.25
president	0.01	0.20
OOV	0.04	0.38

LDA generation - example

- **For each document**, d
  - Pick a topic distribution, θ_d using α
  - For each word in the document
    - pick a topic, z
    - given that topic, pick a word using β

α = (0.8,0.2)

β =

	sports	politics
football	0.30	0.01
baseball	0.25	0.01
TV	0.10	0.15
win	0.30	0.25
president	0.01	0.20
OOV	0.04	0.38