# Topic Models

## Advanced Machine Learning for NLP

Jordan Boyd-Graber

OVERVIEW

## Low-Dimensional Space for Documents

- Last time: embedding space for words
- This time: embedding space for documents
- Generative story
- New inference techniques

## Why topic models?

- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes

## Why topic models?

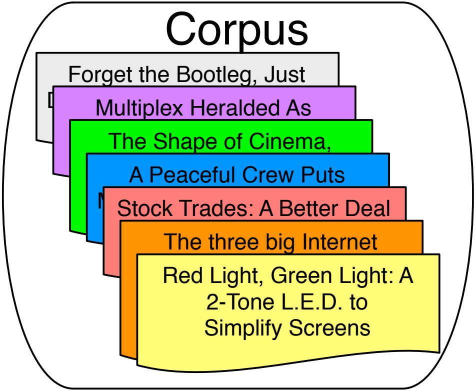
- Suppose you have a huge number of documents
- Want to know what's going on
- Can't read them all (e.g. every New York Times article from the 90's)
- Topic models offer a way to get a corpus-level view of major themes
- Unsupervised

## Roadmap

- What are topic models
- How to know if you have good topic model
- How to go from raw data to topics

## Embedding Space

From an **input corpus** and number of topics $K \rightarrow$ words to topics



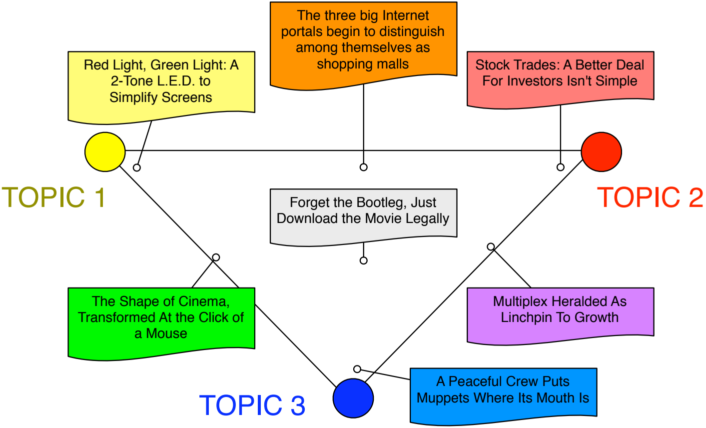
## Embedding Space

From an input corpus and number of topics $K \longrightarrow$ **words to topics**

| TOPIC 1 | TOPIC 2 | TOPIC 3 |
|---|---|---|
| computer, technology, system, service, site, phone, internet, machine | sell, sale, store, product, business, advertising, market, consumer | play, film, movie, theater, production, star, director, stage |

## Conceptual Approach

- For each document, what topics are expressed by that document?



## Topics from *Science*

| | | | |
|---|---|---|---|
| human | evolution | disease | computer |
| genome | evolutionary | host | models |
| dna | species | bacteria | information |
| genetic | organisms | diseases | data |
| genes | life | resistance | computers |
| sequence | origin | bacterial | system |
| gene | biology | new | network |
| molecular | groups | strains | systems |
| sequencing | phylogenetic | control | model |
| map | living | infectious | parallel |
| information | diversity | malaria | methods |
| genetics | group | parasite | networks |
| mapping | new | parasites | software |
| project | two | united | new |
| sequences | common | tuberculosis | simulations |

## Why should you care?

- Neat way to explore / understand corpus collections
  - E-discovery
  - Social media
  - Scientific data
- NLP Applications
  - Word Sense Disambiguation
  - Discourse Segmentation
  - Machine Translation
- Psychology: word meaning, polysemy
- Inference is (relatively) simple

## Matrix Factorization Approach

$$\underbrace{\begin{bmatrix} M \times K \end{bmatrix}}_{\text{Topic Assignment}} \times \underbrace{\begin{bmatrix} \quad K \times V \quad \end{bmatrix}}_{\text{Topics}} \approx \underbrace{\begin{bmatrix} \quad M \times V \quad \end{bmatrix}}_{\text{Dataset}}$$

**K** Number of topics

**M** Number of documents

**V** Size of vocabulary

## Matrix Factorization Approach

$$\underset{\text{Topic Assignment}}{\begin{bmatrix} M \times K \end{bmatrix}} \times \underset{\text{Topics}}{\begin{bmatrix} K \times V \end{bmatrix}} \approx \underset{\text{Dataset}}{\begin{bmatrix} M \times V \end{bmatrix}}$$

**K** Number of topics

**M** Number of documents

**V** Size of vocabulary

- If you use singular value decomposition (SVD), this technique is called latent semantic analysis.
- Popular in information retrieval.

## Alternative: Generative Model

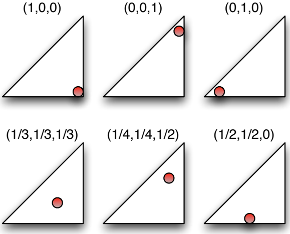
- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference

## Alternative: Generative Model

- How your data came to be
- Sequence of Probabilistic Steps
- Posterior Inference
- Blei, Ng, Jordan. Latent **Dirichlet** Allocation. JMLR, 2003.

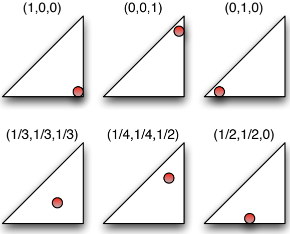
## Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation



## Multinomial Distribution

- Distribution over discrete outcomes
- Represented by non-negative vector that sums to one
- Picture representation



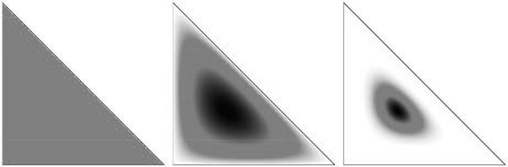

- Come from a Dirichlet distribution

## Dirichlet Distribution

$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

## Dirichlet Distribution

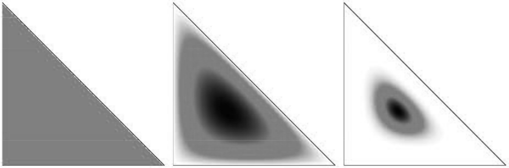
$$P(\boldsymbol{p} \mid \alpha \boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$



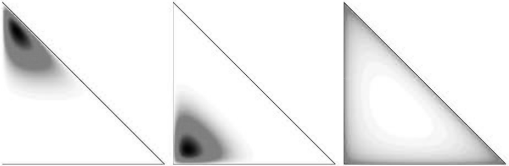
$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 6, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 30, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

## Dirichlet Distribution

$$P(\boldsymbol{p}\,|\,\alpha\boldsymbol{m}) = \frac{\Gamma(\sum_k \alpha m_k)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1}$$

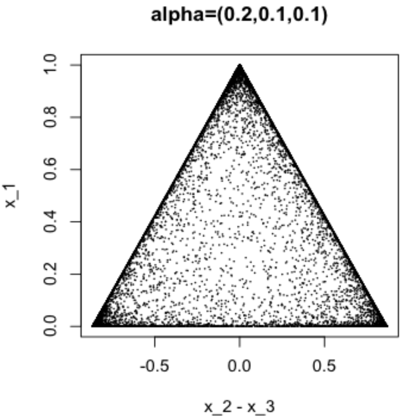


$\alpha = 3, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 6, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ $\alpha = 30, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$



$\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7})$ $\alpha = 14, \boldsymbol{m} = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7})$ $\alpha = 2.7, \boldsymbol{m} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$

## Dirichlet Distribution



- If $\boldsymbol{\phi} \sim \text{Dir}(()\alpha)$, $\boldsymbol{w} \sim \text{Mult}(()\phi)$, and $n_k = |\{w_i : w_i = k\}|$ then

$$p(\phi|\alpha, \boldsymbol{w}) \propto p(\boldsymbol{w}|\phi)p(\phi|\alpha) \tag{1}$$
$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \tag{2}$$
$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \tag{3}$$

- Conjugacy: this **posterior** has the same form as the **prior**

## Dirichlet Distribution

- If $\boldsymbol{\phi} \sim \text{Dir}(()\alpha)$, $\boldsymbol{w} \sim \text{Mult}(()\phi)$, and $n_k = |\{w_i : w_i = k\}|$ then

$$p(\phi|\alpha, \boldsymbol{w}) \propto p(\boldsymbol{w}|\phi)p(\phi|\alpha) \tag{1}$$

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1} \tag{2}$$

$$\propto \prod_k \phi^{\alpha_k + n_k - 1} \tag{3}$$

- Conjugacy: this **posterior** has the same form as the **prior**

## Generative Model

TOPIC 1

computer, technology, system, service, site, phone, internet, machine

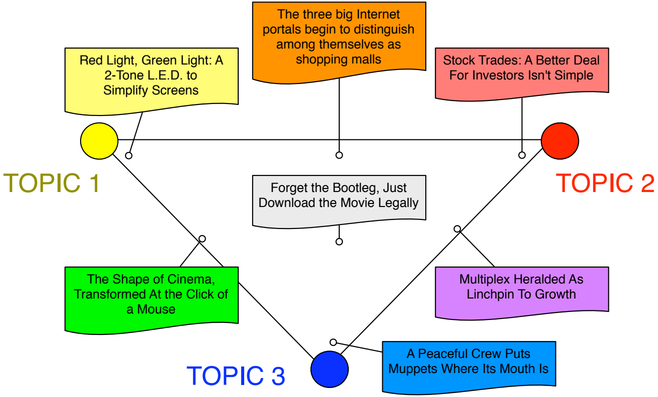
TOPIC 2

sell, sale, store, product, business, advertising, market, consumer

TOPIC 3

play, film, movie, theater, production, star, director, stage

## Generative Model



## Generative Model

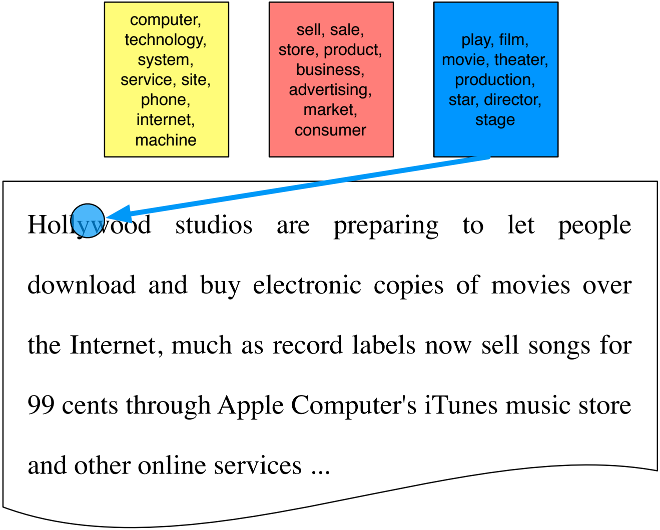
computer, technology, system, service, site, phone, internet, machine

sell, sale, store, product, business, advertising, market, consumer

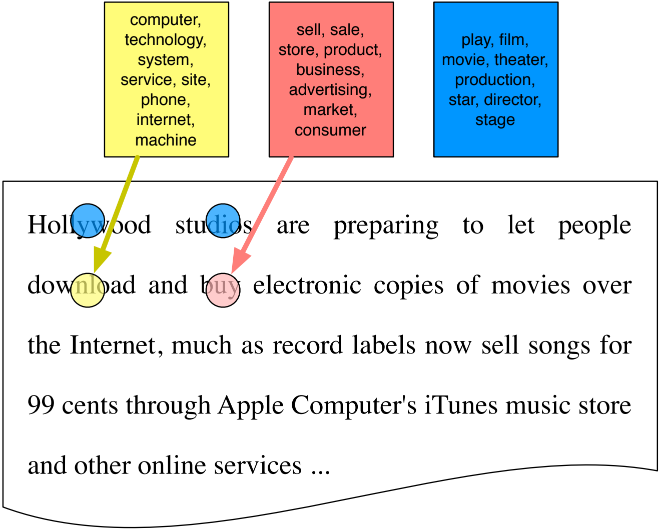
play, film, movie, theater, production, star, director, stage

Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Generative Model



## Generative Model



## Generative Model

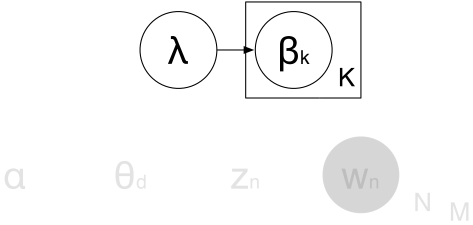
computer, technology, system, service, site, phone, internet, machine

sell, sale, store, product, business, advertising, market, consumer

play, film, movie, theater, production, star, director, stage

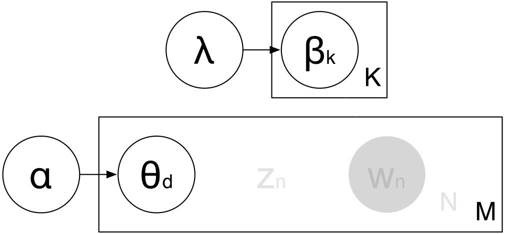
Hollywood studios are preparing to let people download and buy electronic copies of movies over the Internet, much as record labels now sell songs for 99 cents through Apple Computer's iTunes music store and other online services ...

## Generative Model Approach



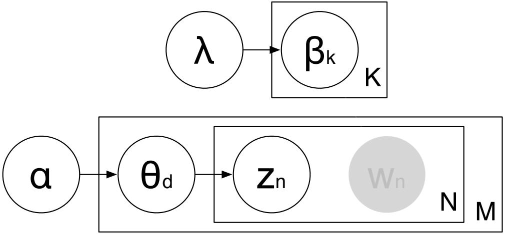

- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution $\beta_k$ from a Dirichlet distribution with parameter $\lambda$

## Generative Model Approach



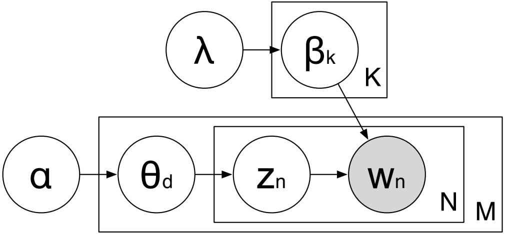

- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution $\beta_k$ from a Dirichlet distribution with parameter $\lambda$
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution $\theta_d$ from a Dirichlet distribution with parameter $\alpha$

## Generative Model Approach



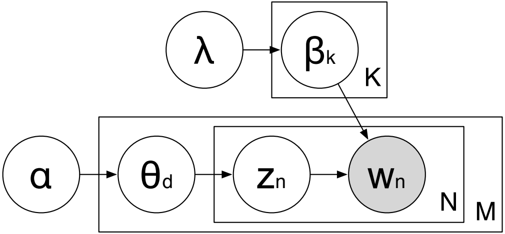

- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution $\beta_k$ from a Dirichlet distribution with parameter $\lambda$
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution $\theta_d$ from a Dirichlet distribution with parameter $\alpha$
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic $z_n$ from the multinomial distribution parameterized by $\theta$.

## Generative Model Approach




- For each topic $k \in \{1, \ldots, K\}$, draw a multinomial distribution $\beta_k$ from a Dirichlet distribution with parameter $\lambda$
- For each document $d \in \{1, \ldots, M\}$, draw a multinomial distribution $\theta_d$ from a Dirichlet distribution with parameter $\alpha$
- For each word position $n \in \{1, \ldots, N\}$, select a hidden topic $z_n$ from the multinomial distribution parameterized by $\theta$.
- Choose the observed word $w_n$ from the distribution $\beta_{z_n}$.

## Generative Model Approach




- For each topic $k \in \{1, \dots, K\}$, draw a multinomial distribution $\beta_k$ from a Dirichlet distribution with parameter $\lambda$
- For each document $d \in \{1, \dots, M\}$, draw a multinomial distribution $\theta_d$ from a Dirichlet distribution with parameter $\alpha$
- For each word position $n \in \{1, \dots, N\}$, select a hidden topic $z_n$ from the multinomial distribution parameterized by $\theta$.
- Choose the observed word $w_n$ from the distribution $\beta_{z_n}$.

## Topic Models: What's Important

- Topic models
  - Topics to word types—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA), a fully Bayesian version of pLSI, probabilistic version of LSA

## Topic Models: What's Important

- Topic models (latent variables)
  - Topics to word types—multinomial distribution
  - Documents to topics—multinomial distribution
- Focus in this talk: statistical methods
  - Model: story of how your data came to be
  - Latent variables: missing pieces of your story
  - Statistical inference: filling in those missing pieces
- We use latent Dirichlet allocation (LDA), a fully Bayesian version of pLSI, probabilistic version of LSA