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 go from raw data to topics
Embedding Space

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

Corpus

- Forget the Bootleg, Just Download the Movie Legally
- Multiplex Heralded As Linchpin To Growth The Shape of Cinema, Transformed At the Click of a Mouse
- A Peaceful Crew Puts Muppets Where Its Mouth Is
- Stock Trades: A Better Deal
- The three big Internet
- Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
Embedding Space

From an input corpus and number of topics $K \rightarrow \text{words to topics}$

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

- **TOPIC 1**
  - Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens
  - The Shape of Cinema, Transformed At the Click of a Mouse

- **TOPIC 2**
  - Stock Trades: A Better Deal For Investors Isn't Simple
  - Forget the Bootleg, Just Download the Movie Legally
  - Multiplex Heralded As Linchpin To Growth

- **TOPIC 3**
  - A Peaceful Crew Puts Muppets Where Its Mouth Is
Topics from *Science*

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

\[
\begin{bmatrix}
M 	imes K
\end{bmatrix} \times \begin{bmatrix}
K 	imes V
\end{bmatrix} \approx \begin{bmatrix}
M 	imes V
\end{bmatrix}
\]

- **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.
Matrix Factorization Approach

\[
\begin{bmatrix}
M 	imes K
\end{bmatrix}
\times
\begin{bmatrix}
K 	imes V
\end{bmatrix}
\approx
\begin{bmatrix}
M 	imes V
\end{bmatrix}
\]

- Topic Assignment
- Topics
- Dataset

**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
Multinomial Distribution

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

(1,0,0) (0,0,1) (1/2,1/2,0) (1/3,1/3,1/3) (1/4,1/4,1/2) (0,1,0)
Multinomial Distribution

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

- Come from a Dirichlet distribution
Dirichlet Distribution

\[ P(p | \alpha m) = \frac{\Gamma\left(\sum_k \alpha m_k\right)}{\prod_k \Gamma(\alpha m_k)} \prod_k p_k^{\alpha m_k - 1} \]
Dirichlet Distribution

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

\[ \alpha = 3, \ m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \alpha = 6, \ m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \alpha = 30, \ m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \]
Dirichlet Distribution

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

\[ \alpha = 3, \ m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \]
\[ \alpha = 6, \ m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \]
\[ \alpha = 30, \ m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \]

\[ \alpha = 14, \ m = (\frac{1}{7}, \frac{5}{7}, \frac{1}{7}) \]
\[ \alpha = 14, \ m = (\frac{1}{7}, \frac{1}{7}, \frac{5}{7}) \]
\[ \alpha = 2.7, \ m = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \]
Dirichlet Distribution

\( \alpha = (0.2, 0.1, 0.1) \)
• If $\phi \sim \text{Dir}(\alpha)$, $w \sim \text{Mult}(\phi)$, and $n_k = |\{w_i : w_i = k\}|$ then

$$p(\phi | \alpha, w) \propto p(w | \phi)p(\phi | \alpha)$$  \hspace{1cm} (1)

$$\propto \prod_k \phi^{n_k} \prod_k \phi^{\alpha_k - 1}$$ \hspace{1cm} (2)

$$\propto \prod_k \phi^{\alpha_k + n_k - 1}$$ \hspace{1cm} (3)

• Conjugacy: this **posterior** has the same form as the **prior**
Dirichlet Distribution

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

$$ p(\phi | \alpha, w) \propto p(w | \phi)p(\phi | \alpha) $$

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

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

- 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
Forget the Bootleg, Just Download the Movie Legally

The three big Internet portals begin to distinguish among themselves as shopping malls

Stock Trades: A Better Deal For Investors Isn’t Simple

The Shape of Cinema, Transformed At the Click of a Mouse

Red Light, Green Light: A 2-Tone L.E.D. to Simplify Screens

For the Bootleg, Just Download the Movie Legally

Multiplex Heralded As Linchpin To Growth

A Peaceful Crew Puts Muppets Where Its Mouth Is

TOPIC 1

TOPIC 2

TOPIC 3
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 ...
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 ...
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 ...
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...
• 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 \( k \in \{1, \ldots, K\} \), 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 \( \alpha \).
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 multinominal 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 $\beta_{z_n}$.
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 $\beta_{z_n}$. 
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—multinominal distribution
  - Documents to topics—multinominal 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