Topic Models
Why Topic Model

• Enormous amount of data is being generated every minute.
• Dimension reduction is necessary for effective use of the data.
User as Document

Consider that a company named M$ has collected user data from many sources. They are stored as documents (bags of words) under user ids.
Find Relevant User

Toyota wants to promote their new compact car by showing ads to users who are interested in cars.

M$ needs to extract users’ interested topics from:
• Search queries
• Tweets
• Posts
• And any sort of text records.
Today, I am #thankful for my car. I drive a red 2008 #Toyota #Scion xD. Its compact size allows it... [link]

Topics:
• Vehicle
• Compact Car

Strong interest in car.
I watched a police car crash into a public transit bus. Officers then engaged everyone onboard in a firefight. Does't show interest in car.
Topic Model

• Users' activities are diverse and numerous. Dimension reduction is necessary for effective use of the data.

• Topic model is a type of statistical models that find abstract "topics" from a collection of documents, so that each document can be represented by few topics.

• With the topic representation, we can do things like
  • Information Retrieval (find potential car buyer)
  • Classification (or prediction)
• Introduction

• **Dimension Reduction on Texts**
  • TF-IDF (1975)
  • LSI (1990)
  • Aspect Model (1999)
  • LDA (2003)
  • Deep Learning (2006)

• **Probabilistic Modeling**
  • Graphical Models
  • Multinomial and Dirichlet Distributions

• **Latent Dirichlet Allocation**
  • Generative Model
  • Inference
  • Applications
Salton, Wong and Yong proposed the famous Term Frequency - Inverse Document Frequency formula to represent each document by a fixed-length vector.

\[
\text{tfidf}(t, d, D) = \frac{\text{frequency of term } t \text{ in document } d}{\text{frequency of term } t \text{ in corpus } D}
\]

- Dimension reduction is small if vocabulary \( V \) is large.
- TF-IDF reveals little about statistical structure of the document.
Latent Semantic Indexing (1990)

SVD (Singular Value Decomposition) is the general dimension reduction technique on real-valued vectors.

(Deerwester et. al, 1990) applies SVD to identify the principle components of the tf-idf vectors and use them to represent the documents.

• Deerwester et al. argue that the derived features can capture some aspects of basic linguistic notions such as synonymy.
• LSI is not a generative model.
(Hofmann 1999) proposed pLSI (probabilistic LSI) model, aka. aspect model, as an alternative to LSI. pLSI models each word $w$ as a sample from a mixture model.

$$P(w | d) = \sum_{z \in Z} P(w | z)P(z | d)$$

The mixture components are multinomial random variables $z$ that can be viewed as "topics".
Aspect Model (1999)

The generative model is defined as follows

1. Select a document with probability $P(d), d \in D$
2. Pick a latent topic $z$ acc. to topic proportion $P(z \mid d)$
3. Generate a word $w$ acc. to word proportion $P(w \mid z)$

$$
P(w \mid d) = \sum_{z \in Z} P(w \mid z)P(z \mid d)$$

$$
P(w, d) = P(w \mid d)P(d)$$

No probabilistic model at the level of documents.

Add Dirichlet priors on aspect model, we get LDA model

$$P(\vec{w}_d, \vec{z}_d, \vec{\theta}_d, \vec{\phi} | \vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{N_d} \text{Mult}(w_{d,n} | \phi_{z_{d,n}}) \text{Mult}(z_{d,n} | \theta_d) \cdot \text{Dir}(\theta_d | \vec{\alpha}) \text{Dir}(\phi | \vec{\beta})$$

Word likelihood

$$\alpha \rightarrow \theta_d \rightarrow z_{d,n} \rightarrow w_{d,n} \rightarrow \phi_k \rightarrow \beta$$

Documents prior

Topics

Per-document topic proportions

Per-word topic assignment

Observed word

Topic hyperparameter

Dirichlet parameter
Latent Dirichlet Allocation (Blei et. al. 2003)

1. For each topic $k \in [1, K]$, sample word proportion $\bar{\phi}_k \sim \text{Dir}(\bar{\beta})$
2. For each document $d \in [1, D]$, sample topic proportion $\bar{\theta}_d \sim \text{Dir}(\bar{\alpha})$
   1. For each word $n \in [1, N_d]$, sample topic index $z_{d,n} \sim \text{Mult}(\bar{\theta}_d)$
   2. For each word $n \in [1, N_d]$, sample term $w_{d,n} \sim \text{Mult}(\bar{\phi}_{z_{d,n}})
G. E. Hinton and R. R. Salakhutdinov converts high-dimensional data to low-dimensional codes by training a multilayer neural network with a small central layer to reconstruct high-dimensional input vectors.
• Introduction

• Dimension Reduction on Texts
  • TF-IDF (1975)
  • LSI (1990)
  • Aspect Model (1999)
  • LDA (2003)
  • Deep Learning (2006)

• **Probabilistic Modeling**
  • Graphical Models
  • Multinomial and Dirichlet Distributions

• Latent Dirichlet Allocation
  • Generative Model
  • Inference
  • Applications
Probabilistic Modeling

1. Treat data as observations that arise from a generative probabilistic process that includes hidden variables.
   - For documents, the hidden variables reflect the thematic structure of a document.

2. Infer the hidden structure (topics) using posterior inference.
   - MCMC (Markov Chain Monte Carlo)

\[
P(\theta \mid X) = \frac{P(X \mid \theta)P(\theta)}{P(X)}
\]

3. Situate new data into the estimated model.
Graphical Models

- Nodes are random variables
- Edges denote possible dependence.
- Observed variables are shaded.
- Replicate structures are boxed.

\[ \tilde{\alpha}, \tilde{\theta}_d, z_{d,n}, w_{d,n}, \tilde{\varphi}_k, \tilde{\beta} \]

\[ d \in [1, D] \]

\[ n \in [1, N_d] \]

\[ k \in [1, K] \]

Dirichlet parameter
Per-document topic proportions
Per-word topic assignment
Observed word
Topics
Topic hyperparameter
Multinomial and Dirichlet Distributions

• Multinomial distribution conditioned on word proportion $\vec{p}$

$$P(\vec{w} \mid \vec{p}) = \text{Mult}(\vec{n} \mid \vec{p}) = \frac{\Gamma \left( \sum_t n_t + 1 \right)}{\prod_t \Gamma(n_t + 1)} \prod_{t=1}^V p_t^{n_t}$$

• Dirichlet distribution conditioned on pseudo-count $\vec{\alpha}$

$$P(\vec{p} \mid \vec{\alpha}) = \text{Dir}(\vec{p} \mid \vec{\alpha}) = \frac{1}{\Delta(\vec{\alpha})} \prod_{t=1}^V p_t^{\alpha_t - 1}$$

$$\Delta(\vec{\alpha}) = \frac{\prod_t \Gamma(\alpha_t)}{\Gamma \left( \sum_t \alpha_t \right)}$$
Dirichlet Pseudo-Count when $K=3$
Multinomial and Dirichlet Distributions

- Thanks to conjugacy, the Dirichlet posterior given multinomial observation $\vec{w}$ and pseudo-counts $\vec{\alpha}$ is

$$P(\vec{p} | \vec{w}, \vec{\alpha}) = \frac{\text{Mult}(\vec{w} | \vec{p}) \text{Dir}(\vec{p} | \vec{\alpha})}{P(\vec{w} | \vec{\alpha})} = \text{Dir}(\vec{p} | \vec{\alpha} + \vec{n})$$

- Likelihood($\vec{p}$) $\sim$ Mult($\vec{w} | \vec{p}$)
- Prior($\vec{p}$) $\sim$ Dir($\vec{p} | \vec{\alpha}$)

- Having new observation is equivalent with adjusting pseudo-count.
Multinomial and Dirichlet Distributions

• Integrate out $p$ to get

$$P(\vec{w} \mid \vec{\alpha}) = \int_{\vec{p}} \text{Mult}(\vec{w} \mid \vec{p}) \, \text{Dir}(\vec{p} \mid \alpha) \, d\vec{p}$$

$$= \frac{\Delta(\vec{n} + \vec{\alpha})}{\Delta(\vec{\alpha})}$$

• We’ll use this repeatedly in later derivation.
• Introduction
• Dimension Reduction on Texts
  • TF-IDF (1975)
  • LSI (1990)
  • Aspect Model (1999)
  • LDA (2003)
  • Deep Learning (2006)
• Probabilistic Modeling
  • Graphical Models
  • Multinomial and Dirichlet Distributions
• **Latent Dirichlet Allocation**
  • Generative Model
  • Inference
  • Applications
Generative Model

1. For each topic $k \in [1, K]$, sample word proportion $\vec{\phi}_k \sim \text{Dir}(\vec{\beta})$
2. For each document $d \in [1, D]$, sample topic proportion $\vec{\vartheta}_d \sim \text{Dir}(\vec{\alpha})$
   1. For each word $n \in [1, N_d]$, sample topic index $z_{d,n} \sim \text{Mult}(\vec{\vartheta}_d)$
   2. For each word $n \in [1, N_d]$, sample term $w_{d,n} \sim \text{Mult}(\vec{\phi}_{z_{d,n}})$
Generative Model

\[ P(\tilde{w}_d, \tilde{z}_d, \tilde{\theta}_d, \tilde{\phi} | \tilde{\alpha}, \tilde{\beta}) = \prod_{n=1}^{N_d} \text{Mult}(w_{d,n} | \tilde{\phi}_{z_{d,n}}) \text{Mult}(z_{d,n} | \tilde{\theta}_d) \]

\[ \cdot \text{Dir}(\tilde{\theta}_d | \tilde{\alpha}) \text{Dir}(\tilde{\phi} | \tilde{\beta}) \]

Word likelihood

Diagram:

- \( \tilde{\alpha} \)
- \( \tilde{\theta}_d \) \( d \in [1,D] \)
- \( \tilde{z}_{d,n} \) \( n \in [1,N_d] \)
- \( \tilde{w}_{d,n} \) \( n \in [1,N_d] \)
- \( \tilde{\phi}_k \) \( k \in [1,K] \)
- \( \tilde{\beta} \)

Dirichlet parameter
Per-document topic proportions
Per-word topic assignment
Observed word
Topics
Inference

• Given observations $W$, we want to infer the hidden structure (topics) from posterior probability.

• Posterior probability is computationally intractable. Approximate inference is used.
  • MCMC (Markov Chain Monte Carlo): Gibbs sampler, collapse Gibbs sampler.
  • Variational methods: replace sampling with optimization.
Inference - Strategy

• Given document $\vec{w}$, sample topic assignment $\vec{z}$ from posterior

$$P(\vec{z} | \vec{w}, \vec{\alpha}, \vec{\beta}) = \frac{P(\vec{w} | \vec{z}, \vec{\beta})P(\vec{z} | \vec{\alpha})}{P(\vec{w} | \vec{\alpha}, \vec{\beta})}$$

• Estimate topic proportion and word proportion by Dirichlet posterior

$$P(\vec{\theta}_d | \vec{z}_d, \vec{\alpha}) = \text{Dir}(\vec{\theta}_d | \vec{\alpha} + \vec{n}_d)$$
$$P(\vec{\phi}_k | \vec{z}, \vec{w}, \vec{\beta}) = \text{Dir}(\vec{\phi}_k | \vec{\beta} + \vec{n}_k)$$

$$E(\theta_{d,k}) = \frac{n_{d}^{(k)} + \alpha_k}{\sum_{k'=1}^{K} n_{d}^{(k')} + \alpha_{k'}}$$
$$E(\phi_{k,t}) = \frac{n_{k}^{(t)} + \beta_t}{\sum_{t'=1}^{V} n_{k}^{(t')} + \beta_{t'}}$$
Inference – Likelihood

• Given document $\mathbf{w}$, sample topic assignment $\mathbf{z}$ from posterior

$$P(\mathbf{z} | \mathbf{w}, \mathbf{\alpha}, \mathbf{\beta}) = \frac{P(\mathbf{w} | \mathbf{z}, \mathbf{\beta}) P(\mathbf{z} | \mathbf{\alpha})}{P(\mathbf{w} | \mathbf{\alpha}, \mathbf{\beta})}$$

• Likelihood

$$P(\mathbf{w} | \mathbf{z}, \mathbf{\beta}) = \int P(\mathbf{w} | \mathbf{z}, \mathbf{\phi}) P(\mathbf{\phi} | \mathbf{\beta}) d\mathbf{\phi}$$

$$= \int \prod_{k=1}^{K} \text{Mult}(\mathbf{n}_k | \mathbf{\phi}_k) \text{Dir}(\mathbf{\phi}_k | \mathbf{\beta}) d\mathbf{\phi}$$

$$= \prod_{k=1}^{K} \frac{\Delta(n_k + \beta)}{\Delta(\beta)}$$
Inference – Prior

• Given document $\mathbf{w}$, sample topic assignment $\mathbf{z}$ from posterior

\[
P(\mathbf{z} | \mathbf{w}, \mathbf{\alpha}, \mathbf{\beta}) = \frac{P(\mathbf{w} | \mathbf{z}, \mathbf{\beta}) P(\mathbf{z} | \mathbf{\alpha})}{P(\mathbf{w} | \mathbf{\alpha}, \mathbf{\beta})}
\]

• Prior

\[
P(\mathbf{z} | \mathbf{\alpha}) = \int P(\mathbf{z} | \mathbf{\theta}) P(\mathbf{\theta} | \mathbf{\alpha}) d\mathbf{\theta}
\]

\[
= \int \prod_{d=1}^{D} \text{Mult}(\mathbf{n}_d | \mathbf{\theta}_d) \text{Dir}(\mathbf{\theta}_d | \mathbf{\alpha}) d\mathbf{\theta}
\]

\[
= \prod_{d=1}^{D} \frac{\Delta(\mathbf{n}_d + \mathbf{\alpha})}{\Delta(\mathbf{\alpha})}
\]
Inference – Collapsed Conditional

• Given document $\vec{w}$, sample topic assignment $\vec{z}$ from posterior

$$P(\vec{z} \mid \vec{w}, \vec{α}, \vec{β}) = \frac{P(\vec{w} \mid \vec{z}, \vec{β}) P(\vec{z} \mid \vec{α})}{P(\vec{w} \mid \vec{α}, \vec{β})}$$

• For $\vec{w} = \{w_i = t, \vec{w}_{-i}\}$ and $\vec{z} = \{z_i = k, \vec{z}_{-i}\}$

$$P(z_i = k \mid \vec{z}_{-i}, \vec{w}, \vec{α}, \vec{β}) = \frac{P(w_i \mid \vec{z}, \vec{β}) P(\vec{z} \mid \vec{α})}{P(w_i \mid \vec{z}, \vec{β}) P(w_i \mid \vec{z}_{-i}, \vec{β}) P(z_i \mid \vec{α})} = \frac{P(w_i \mid \vec{z}, \vec{β})}{P(w_i \mid \vec{z}_{-i}, \vec{β})} \frac{P(z_i \mid \vec{α})}{P(z_{-i} \mid \vec{α})}$$

$$= \frac{\Delta(\vec{n}_z + \vec{β})}{\Delta(\vec{n}_{z,-i} + \vec{β})} \frac{\Delta(\vec{n}_d + \vec{α})}{\Delta(\vec{n}_{d,-i} + \vec{α})}$$

$$= \frac{n_{k,-i}^{(i)} + \beta_t}{\sum_{i' = 1}^{V} n_{k,-i}^{(i')} + \beta_t'} (n_{d,-i}^{(k)} + \alpha_k)$$
Inference – Collapsed Conditional

\[ \begin{align*}
\alpha & \xrightarrow{\text{Dirichlet parameter}} Z_{d,n} \xrightarrow{\text{Per-document topic proportions}} W_{d,n} \xrightarrow{\text{Per-word topic assignment}} \text{Observed word} \xrightarrow{\text{Topics}} \beta
\end{align*} \]

\[ d \in [1,D], \quad n \in [1,N_d], \quad k \in [1,K] \]
Inference – Gibbs Sampling

Stage 1 Initialization:
Initialize topic assignment $z$ with equal probabilities on topics.
Inference – Gibbs Sampling

\[ P(\mathbf{z}_i = k | \mathbf{z}_{-i}, \mathbf{w}, \bar{\alpha}, \bar{\beta}) \]
\[ \propto \frac{n^{(i)}_{k,-i} + \beta_t}{\sum_{t' = 1}^{V} n^{(t')}_{k,-i} + \beta'_t} (n^{(k)}_{d,-i} + \alpha_k) \]

**Stage 2 Burn-In Period:**
Randomly select topic assignment \( z \). Update the assignment acc. to the collapsed conditional.

Diagram:
- \( \tilde{\alpha} \) (Dirichlet parameter)
- \( \tilde{\beta} \) (Topic hyperparameter)
- \( d \in [1,D] \) (Per-document topic proportions)
- \( k \in [1,K] \) (Topics)
- \( \mathbf{z}_d, \mathbf{w}_d \) (Per-word topic assignment)
- \( \mathbf{z}_d, \mathbf{w}_d \) (Observed word)
Inference – Gibbs Sampling

\[
P(z_i = k | \vec{z}_{-i}, \vec{w}, \vec{\alpha}, \vec{\beta}) \\
\propto \frac{n_{k,-i}^{(i)} + \beta_t}{\sum_{t' + 1}^{V} n_{k,-i}^{(i')} + \beta_t'} (n_{d,-i}^{(k)} + \alpha_k)
\]

**Stage 2 Burn-In Period:**
Randomly select topic assignment \(z\). Update the assignment acc. to the collapsed conditional.

- \(d \in [1,D]\)
- \(k \in [1,K]\)
- Dirichlet parameter
- Per-document topic proportions
- Per-word topic assignment
- Observed word
- Topics
Inference – Gibbs Sampling

\[ P(z_i = k | z_{-i}, w, \alpha, \beta) \]

\[ \propto \frac{n_{k,-i}^{(t)} + \beta_t}{\sum_{t'=1}^{V} n_{k,-i}^{(t')} + \beta_t'} (n_{d,-i}^{(k)} + \alpha_k) \]

Stage 2 Burn-In Period:
Randomly select topic assignment \( z \). Update the assignment acc. to the collapsed conditional.
Inference – Gibbs Sampling

Stage 3 Terminate:
When stop criteria is met, read out word proportion and topic proportion from topic assignment counts.

\[
P(\hat{\theta}_d | \vec{z}_d, \vec{\alpha}) = \text{Dir}(\hat{\theta}_d | \vec{\alpha} + \vec{n}_d)
\]

\[
P(\vec{\phi}_k | \vec{z}, \vec{w}, \vec{\beta}) = \text{Dir}(\vec{\phi}_k | \vec{\beta} + \vec{n}_k)
\]

\[
E(\theta_{d,k}) = \frac{n^{(k)}_{d} + \alpha_k}{\sum_{k'=1}^{K} n^{(k')}_{d} + \alpha_k}
\]

\[
E(\phi_{k,t}) = \frac{n^{(t)}_{k} + \beta_t}{\sum_{t'=1}^{V} n^{(t')}_{k} + \beta_t}
\]
Algorithm $\{\phi, \theta, \vec{z}\} = \text{LDA}_\text{Gibbs} (\{\vec{w}\}, \tilde{\alpha}, \tilde{\beta}, K)$

//initialize counts and topic assignment $\vec{z}$
sample topic indices $\vec{z} \sim \text{Mult}(1/K)$ for every word in $\{\vec{w}\}$
set document-topic count $n^{(k)}_d = \| \{z_{d,n} = k \mid n \in [1, N_d]\} \|
set topic-term count $n^{(t)}_k = \| \{w_{d,n} = t \mid z_{d,n} = k\} \|$
Inference – Gibbs Sampler Burn-In Period

\[\text{while not finished do}\]
\[\quad \text{for all documents } d \in [1, D] \text{ do}\]
\[\quad \quad \text{//for the current assignment of topic } k \text{ to term } t \text{ for word } w_{d,n}\]
\[\quad \quad \text{decrease counts: } n_d^{(k)} \leftarrow 1, \ n_k^{(t)} \leftarrow 1\]
\[\quad \quad \text{//multinomial sampling for each component } i\]
\[\quad \quad \text{sample topic index } k' \sim P(z_i | z_{-i}, \vec{w}, \vec{\alpha}, \vec{\beta})\]
\[\quad \quad \text{//for the new assignment of } z_{d,n} \text{ to the term } t \text{ for word } w_{d,n}\]
\[\quad \quad \text{increment counts: } n_d^{(k')} \leftarrow 1, \ n_k^{(t)} \leftarrow 1\]
\[\quad \text{end for}\]

\[\text{if converged or } L \text{ sampling iterations since last read out then}\]
\[\quad \text{read out parameter estimates } \mathbf{E}(\theta_{d,k}) \text{ and } \mathbf{E}(\phi_{k,t}).\]
\[\text{end if}\]
\[\text{end while}\]
Inference - Dirichlet Parameters

• Dirichlet parameters describe our belief in the outcome.
• Dirichlet parameters less than one penalize dense topic/word proportions.

• Suggested value [Griffiths&Seyvers 2004]

\[ \alpha = \frac{50}{K} \]
\[ \beta = 0.01 \]
Dirichlet Pseudo-Count when K=3
Applications

• Given new document \( \tilde{w} \), Gibbs sampling update:

\[
P(\tilde{z}_i = k \mid \tilde{w}_i = t, \tilde{z}_{-i}, \tilde{w}_{-i}, \phi, \theta) \propto \varphi_{k,t} \times (n_{d',-i}^{(k)} + \alpha_k)
\]

\[
\mathbb{E}(\theta_{d',k}) = \frac{n_{d'}^{(k)} + \alpha_k}{\sum_{k'=1}^{K} n_{d'}^{(k')} + \alpha_{k'}}
\]

• Thus we obtain the topic proportion for documents and queries. Information retrieval and classification can be carried out on top of this representation.
References


• [http://videolectures.net/mlss09uk_blei_tm](http://videolectures.net/mlss09uk_blei_tm) by David Blei