Probabilistic Graphical Models

Latent Dirichlet Allocation (LDA) for Probabilistic Topic Modeling

Motivation for topic models, Latent Dirichlet Allocation (LDA), parameter estimation in LDA, selection of the number of topics, application of LDA, evaluation methods of topic coherence

Mingyang Xu, mxu5@ncsu.edu
PhD Student in Dr. Samatova's research lab
Department of Computer Science
North Carolina State University

NC STATE UNIVERSITY
Department of Computer Science

Topic Modeling

MOTIVATION

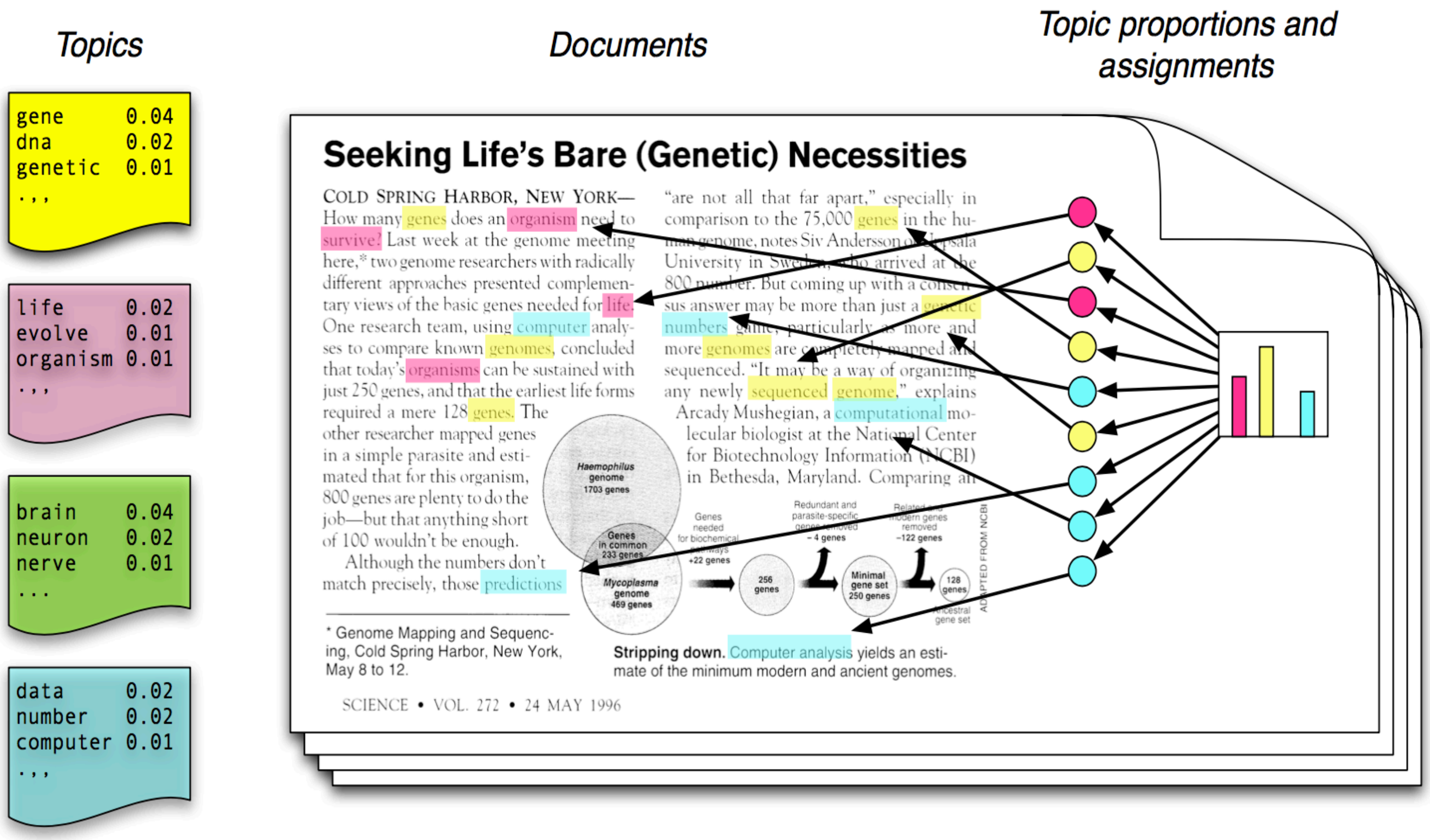
Why Topic Modeling from Unstructured Text?

- Unstructured text data is ubiquitous: online reviews, news, blogs, etc.
- It's difficult to find what we are looking for
- We need algorithms to help us organize and understand this vast amount of unstructured information

What are Topic Models Capable of?

- Automatic organization and summarization of large electronic unstructured text corpus
  - Uncover the major themes (topics) that pervade the corpus
  - Annotate the documents according to those topics
  - Use the annotations to organize and summarize the texts

Overview of a Topic Modeling



Input:

A collection of text documents

Output:

- **A set of topics; topic is the probability distribution over the unique words in the input documents**
- **Probabilistic assignment of each word to a topic**
- **Probability distribution over topics for each document**

Src: Figure from “Probabilistic Topic Models” by David Blei, April 2012 | vol. 55 | no. 4 | Communications of the ACM

Methodology

LDA: LATENT DIRICHLET ALLOCATION

What is LDA?

- A topic modelling method proposed by Prof. David Blei in JMLR 2003
- A generative model
  - Each document is assumed to be generated by a generative process
  - Presented as a probabilistic graphical model
- Unsupervised learning methodology
  - Only the number of topics is specified in advance
- In LDA, a topic is a distribution over a fixed vocabulary
  - These topics are assumed to be generated first, before the documents

Generative Model vs. Discriminative Model

Criteria	Discriminative model	Generative model
Suppose your input data: (x, y)	Learns the conditional probability $p(y\|x)$	Learns the joint probability $p(x, y)$
Suppose your observed data: (x=height, y=gender)	For a given height, what is the probability of this height to be of a male or female?	Distribution of heights for females and males
Algorithms	Logistic regression Support Vector Machines	Latent Dirichlet Allocation (LDA) Naive Bayes Classifier

Key Assumptions of LDA

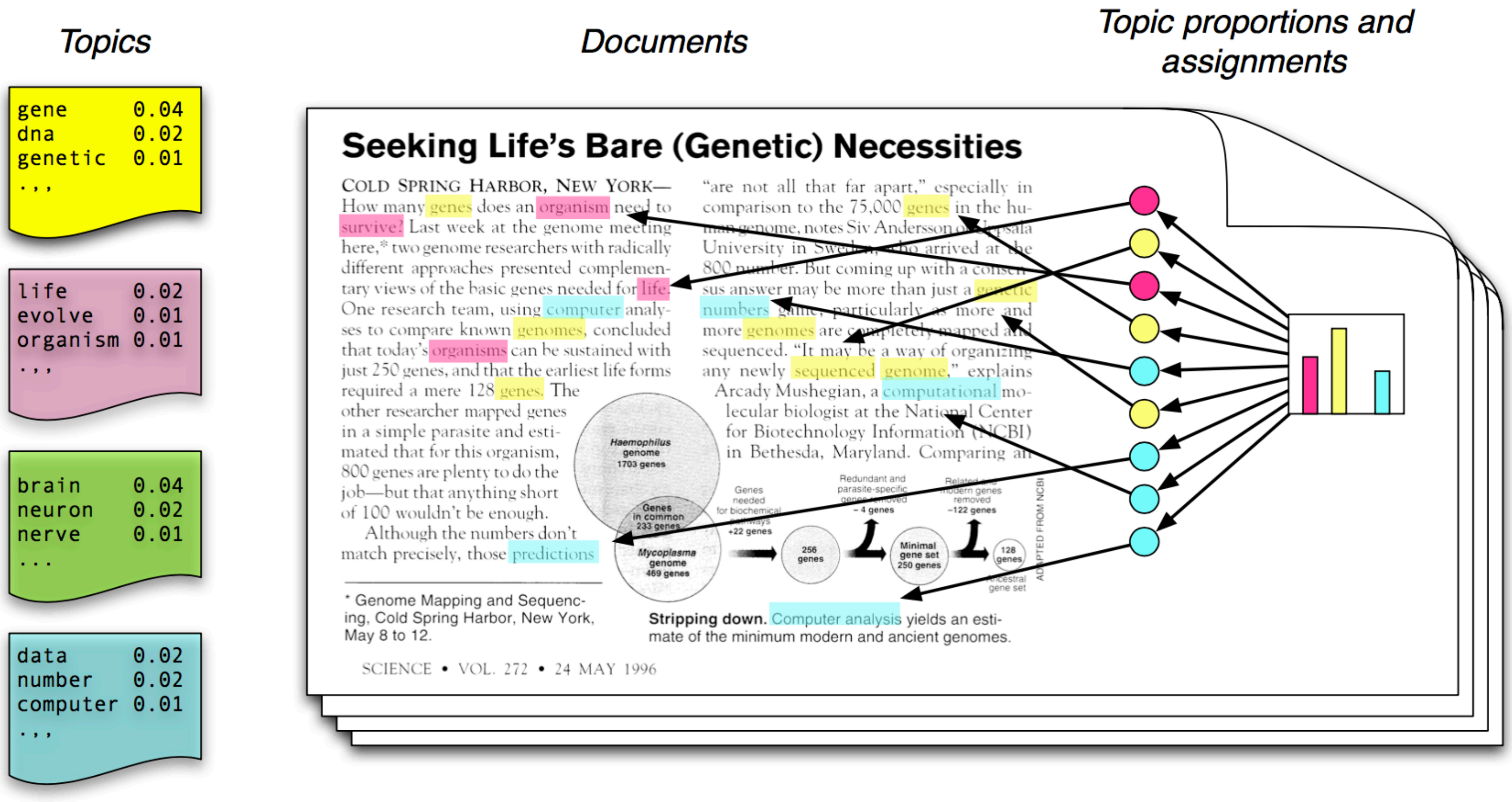
- Documents exhibit multiple topics (but not too many)
- The order of words does not matter in a document ("bag of words")
- The order of documents does not matter ("bag of documents")
- The number of topics is specified and fixed *a priori*

Latent Dirichlet Allocation

GENERATIVE PROCESS

How to Understand a Generative Process of LDA?

- **Documents are assumed to be unknown and generated by this process**
- **Topics and Topic proportions of each document are known**
- **We use these distributions to generate the documents**



Generative Process of LDA

To generate a document

1. Randomly choose a distribution over topics for each document
2. Randomly choose a distribution over words for each topic
3. For each word in each document
   a. randomly choose a topic from the distribution over topics
   b. randomly choose a word from the corresponding topic (distribution over the vocabulary)

- Step 1 and 2: Require distribution over a distribution → Dirichlet distribution
- Words are generated independently of other words (i.e., unigram of bag-of-words model)

Illustration: The Generative Process of LDA

1. Sample a topic distribution under each document and a word distribution under each topic following Dirichlet Distribution

docs	topic 0	topic 1
d_0	0.8	0.2
d_1	0.1	0.9

Per-Document Topic Distribution

2. Sample a topic, say topic 0, following multinomial distribution

topic 0		topic 1	
iphone	0.4	fast	0.5
battery	0.2	nice	0.3
......	......	......	......
brand	0.02	new	0.01

Per-Topic Word Distribution

3. Sample a word, say iphone, following multinomial distribution

Documents that LDA generates:
d_0: iphone brand happy nice new
d_1: battery iphone nice fast really low brand too

Illustration: The Generative Process of LDA

1. Sample a topic distribution under each document and a word distribution under each topic following Dirichlet Distribution

docs	topic 0	topic 1
d_0	0.8	0.2
d_1	0.1	0.9

Per-Document Topic Distribution

2. Sample a topic, say topic 0, following multinomial distribution

topic 0		topic 1	
iphone	0.4	fast	0.5
battery	0.2	nice	0.3
……	……	……	……
brand	0.02	new	0.01

Per-Topic Word Distribution

3. Sample a word, say iphone, following multinomial distribution

Documents that LDA generates:
d_0: iphone brand happy nice new
d_1: battery iphone nice fast really low brand too

Backtracking in LDA

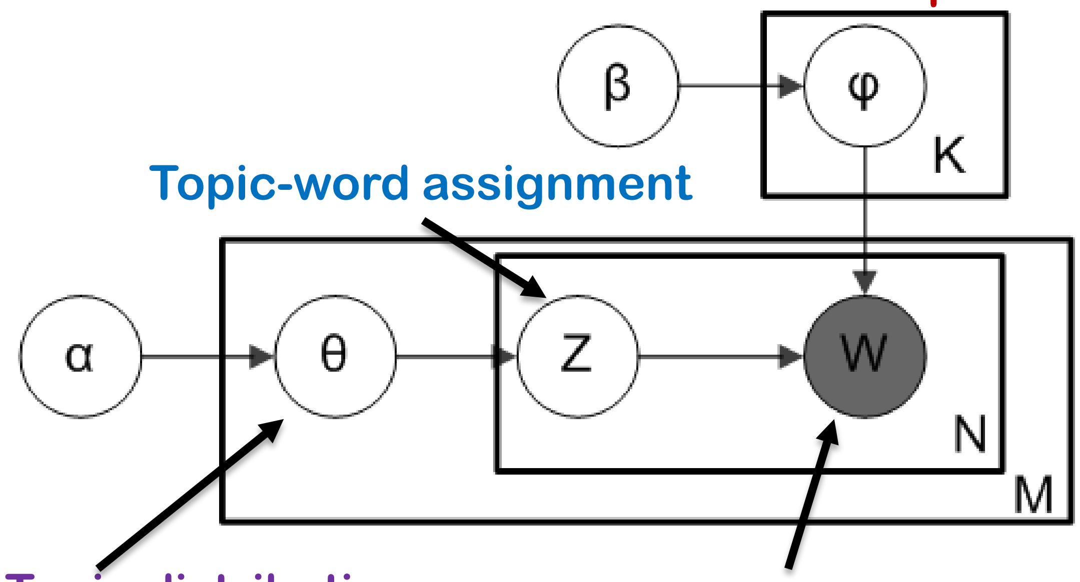
However, in reality, we only observe the documents.
The intuition of LDA is that it tries to backtrack from the input documents to estimate the hidden variables that are most likely to have generated the observed documents.

docs	topic 0	topic 1
d_0		
d_1		

topic 0		topic 1	

Documents that we observed:
d_0: **iphone happy nice brand new**
d_1: **battery low nice really fast iphone too**

Summary: Generative Process of LDA



- Each node is a variable
- Only the shaded node (words) is observed
- The other variables are latent
- Plates indicate repetition

Notations:

α, β: The parameters of Dirichlet distributions
θ: Topic distribution in each document
φ: Word distribution in each topic
Z: Topic-word assignments
W: Observed words in documents
N: The number of *words* in input corpus
M: The number of *documents* in input corpus
K: The number of *topics*

For each document d in corpus:
1. Choose θ_d ~ Dir(α)

For each topic k :
1. Choose φ ~ Dir(β)

For i_{th} word w_i in each d:
1. Choose a topic Z_i ~ Multi(θ_d)
2. Choose a word w_i~ Multi(φ_{Z_i})

What is Dirichlet Distribution?

- Denoted as Dir(α), where α is its parameter
- The prior distribution of multinomial distribution
- Distribution over distributions

Choose topic 0 for d_0 ~

docs	topic 0	topic 1
d_0	0.8	0.2
d_2	0.1	0.9

~ Dir(α)

Why Use the Dirichlet Distribution in LDA?

Dirichlet distribution is a conjugate prior of multinomial distributions and can facilitate the development of inference and parameter estimation algorithms for LDA.

Conjugacy: The form of the posterior $\mathrm{P}(\,(p_1, p_2, \ldots .\, p_k)|\alpha\,, x)$ **is *the same* as the prior** $\mathrm{P}(\,(p_1, p_2, \ldots .\, p_k)|\alpha\,)$

$$(p_1, p_2, \ldots .\, p_k) \sim Dirichlet(\alpha_1, \ldots, \alpha_k)$$

prior to collecting the data, then given observations ($x_1, x_2, \ldots x_k$**), such as the number of times each topic is assigned,**

$$(p_1, p_2, \ldots .\, p_k)|(x_1, \ldots x_k) \sim Dirichlet(\alpha_1 + x_1, \ldots, \alpha_k + x_k)$$

Latent Dirichlet Allocation

PARAMETER ESTIMATION

Parameter Estimation of LDA

Main variables of interest
φ**: distribution over vocabulary for each topic**
θ**: topic distribution for each document**

Original paper of LDA uses EM (Hoffmann 1999) algorithm

A faster algorithm is Gibbs Sampling Algorithm

- **Samples from each variable one at a time, keeping the current values of the other variables fixed**

Gibbs Sampling of LDA: Posterior Estimate

The conditional probability of assigning word w_i with topic k:

$$P(z_i = k | z^{-i}, w, \alpha, \beta) \propto \boxed{\frac{n_{d_i,k}^{-i} + \alpha}{n_{d_i}^{-i} + K\alpha}} \boxed{\frac{n_{k,w}^{-i} + \beta}{n_k^{-i} + V\beta}}$$

The proportion of assignments to topic k over all documents that come from this word w

The proportion of words in document d that are currently assigned to topic k

V : The number of unique words **K : The number of topics**

α, β: The parameters of dirichlet distributions

$n_{d_i}^{-i}$: The number of words in document d not including the current word

$n_{d_i,k}^{-i}$: The number of words in document d assigned to topic k not including current word

n_k^{-i}: The number of words assigned to topic k not including current word

$n_{k,w}^{-i}$: The number of word w assigned to topic k not including current word

Estimating Latent Variables: Posterior Estimates of φ and θ

The probability of word w in topic k is defined as:

$$\varphi_{k,w} = \frac{\boldsymbol{n_{k,w} + \beta}}{\boldsymbol{n_k + V\beta}}$$

The probability of topic k in document d is defined as:

$$\theta_{d,k} = \frac{\boldsymbol{n_{d,k} + \alpha}}{\boldsymbol{n_d + K\alpha}}$$

V : The number of unique words **K : The number of topics**

$\boldsymbol{\alpha, \beta}$: The parameters of dirichlet distributions

$\boldsymbol{n_{d_i}^{-i}}$: The number of words in document d not including the current word

$\boldsymbol{n_{d_i,k}^{-i}}$: The number of words in document d assigned to topic k not including current word

$\boldsymbol{n_{k}^{-i}}$: The number of words assigned to topic k not including current word

$\boldsymbol{n_{k,w}^{-i}}$: The number of word w assigned to topic k not including current word

Why Does LDA Work?

- LDA Trades off two goals:
  1. For each document, assigns its words to as few as topics as possible.
  2. For each topic, assigns high probability to as few terms as possible.

- However, these two goals contradict to each other:
  - Assigning each word to a single topic will make many words have equal probability in the topic.
  - Assigning a few words to each topic will make each word in each document be assigned many different topics.

- Trading off these two goals finds groups of tightly co-occurring words in the similar context, which are likely to be semantically related.

Latent Dirichlet Allocation

MODEL SELECTION

How to Choose α and β?

- **The intuition of choosing α and β:**

  α represents document-topic density - with a higher alpha, documents are made up of more topics, and with lower alpha, documents contain fewer topics.

  β represents topic-word density - with a high beta, topics are made up of most of the words in the corpus and with a low beta they consist of few words.

In practice:

There is no standard for setting α and β.

A rule of thumb given by Griffiths & Steyvers(2004) is to set:

- **α = 50/T, where T is the number of topics**
- **β = 0.1, which is a small number and can be expected to result in a fine-grained decomposition of the corpus into topics**

How to Choose Number of Topics?

- There is no best approach or standard for choosing the number of topics.
- It should be selected based on different datasets.
- The intuition: a larger number of topics can provide more detailed information, while a smaller number of topics can provide a bigger picture of your datasets.

The method proposed by Griffiths & Steyvers(2004):

- **The intuition**: Find the number of topics that can most likely generate the observed dataset

- Calculate $\log(P(w|\mathrm{T}))$ with different number of topics and select the best number of topics

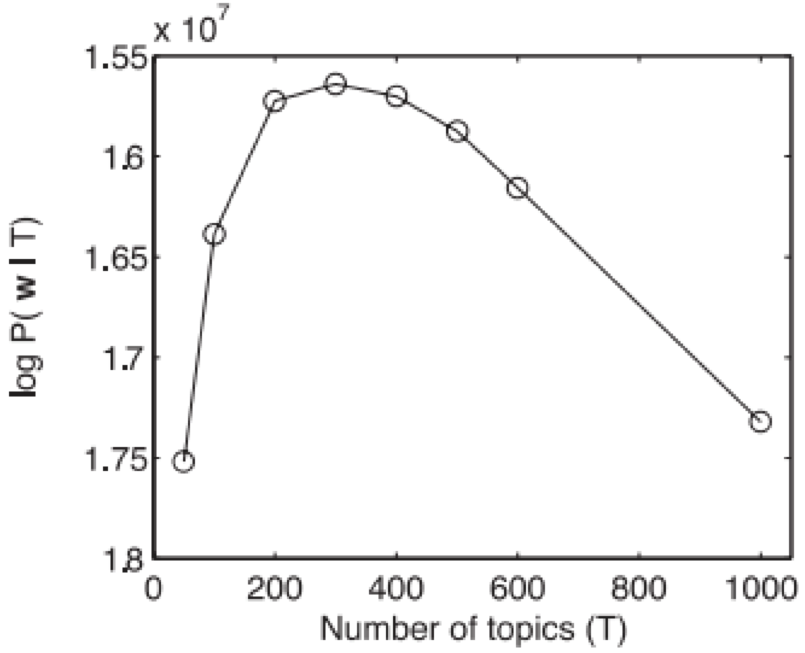


Fig. 3. Model selection results, showing the log-likelihood of the data for different settings of the number of topics, *T*. The estimated standard errors for each point were smaller than the plot symbols.

Latent Dirichlet Allocation

TOPIC MODEL PERFORMANCE EVALUATION

Topic Coherence: Model Performance Metric

Topic coherence score (maximization score: the higher, the better):

$$\mathbf{PMI}(t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_i)P(w_j)}$$ (Newman et al., 2009)

$$\mathbf{LCP}(t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{P(w_j, w_i)}{P(w_i)}$$ (Mimno et al.2011)

- N: The number of top words to keep in each topic
- $P(w_j, w_i)$: The frequency of a document containing both w_j and w_i
- $P(w_i)$: The frequency of a document containing w_i

Human Evaluation

1. Mark each topic as coherent or not.
2. Mark words as coherent to topic or not.

Evaluate the quality of topics based on:
- The percentage of coherent topics.
- P@n: The precision (percentage of coherent words) of the top n words.

Applications of LDA

- Discover the major themes of a corpus
- Keyword summarizations
- Aspects extraction
- Document clustering
- Automatic image annotation