Latent Topic Models

Davide Bacciu

Dipartimento di Informatica
Università di Pisa
bacciu@di.unipi.it

Machine Learning: Neural Networks and Advanced Models (AA2)

Today's Lecture

- Probabilistic models for document understanding
  - Use latent variables
  - Require approximate inference (most)
- Latent Dirichlet Allocation
  - Bayesian latent topic model
  - Example of variational learning
- Document understanding applications
  - Machine vision
  - Advanced topic models

Motivating Applications

- Document understanding
  - Inferring document topic
  - text $\rightarrow$ words
  - image $\rightarrow$ visual patches
- Customer profiling
  - Inferring service usage patterns
  - user $\rightarrow$ call features
  - customer $\rightarrow$ goods/services buys
- User behavior recognition
  - Inferring user habits
  - user $\rightarrow$ daily/weekly living activities

Latent Variable Models

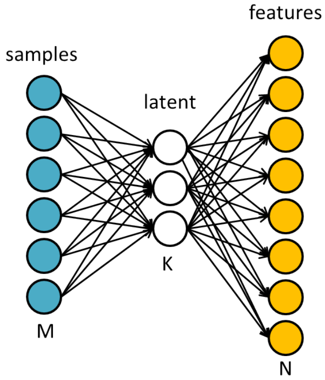
- Latent variables
  - Unobserved RV that define an hidden generative process of observed data
  - Explain complex relation between a large number of observable variables
- Latent variable models likelihood

$$P(\mathbf{X}) = \int_{\mathbf{z}} \prod_{i=1}^{N} P(X_i|\mathbf{Z} = \mathbf{z})P(\mathbf{Z} = \mathbf{z})d\mathbf{z}$$

- Something we have already seen
  - Hidden Markov models
  - Hidden states $\equiv$ Latent variables

Latent Space

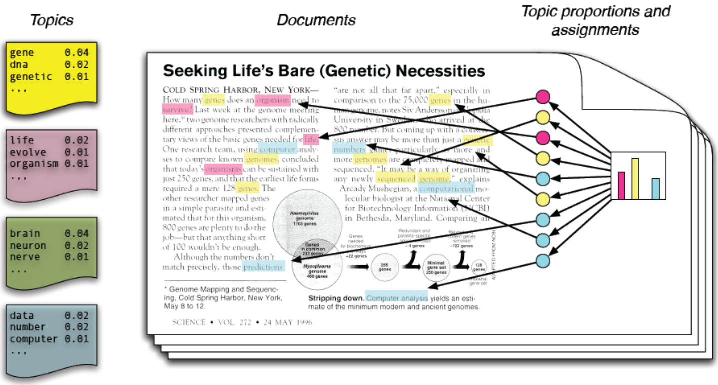
Define a latent space where high-dimensional data can be represented



Assumption

Latent variables conditional and marginal distributions are more tractable than the joint distribution $P(\mathcal{X})$ (e.g $K \ll N$)

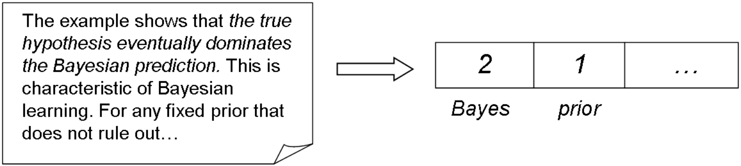
Inferring Latent Semantics of Texts



- Latent variables $\equiv$ document topics (hidden semantics of text)
- Documents are mixtures of topics (latent RV) assigned to words (observed RV)

Bag of Words (BOW) Representation

Count occurrences of dictionary words in document

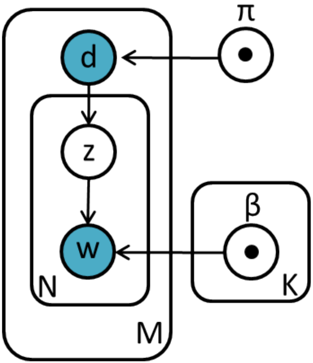


A BOW dataset (corpora) is the $N \times M$ term-document matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & \dots & x_{1i} & \dots x_{1M} \\ \dots & \dots & \dots & \dots \\ x_{j1} & \dots & x_{ji} & \dots x_{jM} \\ \dots & \dots & \dots & \dots \\ x_{N1} & \dots & x_{Ni} & \dots x_{NM} \end{bmatrix}$$

- N: number of vocabulary words w_j
- M: number of documents d_i
- $x_{ij} = n(w_j, d_i)$: number of occurrences of w_j in d_i

Probabilistic Latent Semantic Analysis



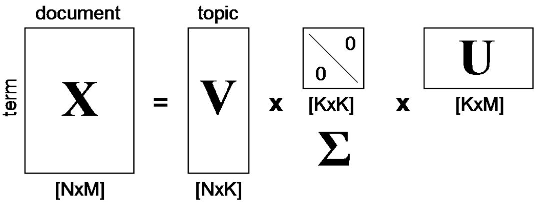
- Documents d as mixtures of topics z
  - Assigning one topic to each word w
  - A single topic for the whole document
- Generative process for the document-term matrix X
  - Select a document with probability $P(d|\pi)$
  - Pick a latent topic $z \sim P(z|d)$
  - Generate a word w with probability $P(w|z,\beta)$

$$P(w_j, d_i|\pi, \beta) = P(d_i|\pi) \sum_{k=1}^{K} P(z_k|d_i) P(w_j|z_k, \beta)$$

PLSA as Matrix Decomposition

Consider the following (equivalent) factorization

$$P(w,d) = \sum_z P(d|z)P(z)P(w|z)$$



A Non-Negative Matrix Factorization

A feature extraction approach projecting M-dimensional documents into a reduced K-dimensional topic-space

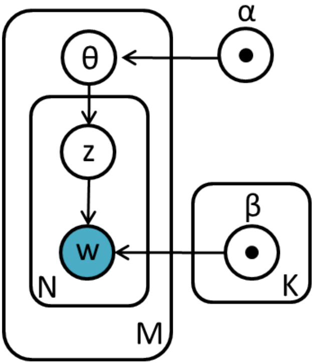
PLSA Summary

- PLSA parameters are the multinomial distributions $\pi_i = P(d_i|\pi)$, $\alpha_{ki} = P(z_k|d_i)$ and $\beta_{jk} = P(w_j|z_k, \beta)$
- Learning is by Expectation-Maximization

$$\mathcal{L}(\theta) = \sum_{i=1}^{D} \sum_{j=1}^{W} n(w_j, d_i) \log \left\{ P(d_i|\pi) \sum_{k=1}^{K} P(z_k|d_i) P(w_j|z_k, \beta) \right\}$$

  using estimated posterior $P(z_k|w_j, d_i, \pi, \beta)$
- Limitations
  - Document-topic probability $P(z_k|d_i)$ depends on training document index d_i
  - Number of parameters increases linearly with number of documents
  - Not a generative model for documents outside training set

Latent Dirichlet Allocation (LDA)



- Per-document topic proportion becomes a random variable θ
  - $P(\theta|\alpha)$ Dirichlet distribution with hyperparameter α
  - $P(z|\theta)$ multinomial topic distribution with document-specific parameter θ
  - $P(w|z, \beta)$ multinomial word-topic distribution
- Think LDA as a continuous extension of the point-wise PLSA
  - PLSA finds a set of K projection directions
  - LDA finds a set of K projection functions

Dirichlet Distribution

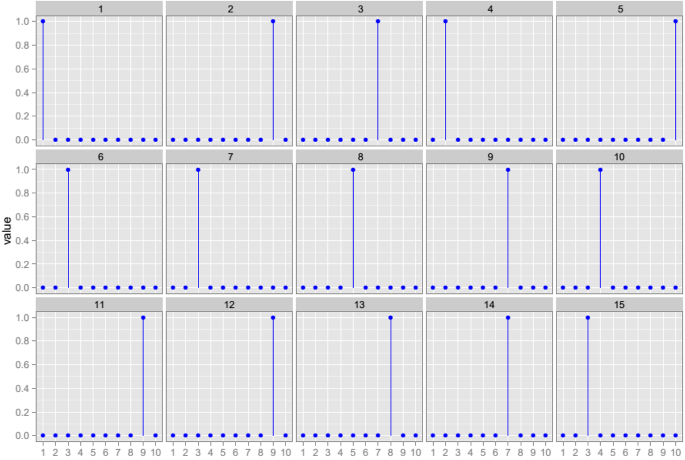
- Why a Dirichlet distribution?
  - Conjugate prior to multinomial distribution
  - If the likelihood is multinomial with a Dirichlet prior then posterior is Dirichlet
- What is a Dirichlet distribution?
  - A distribution for vectors that sum to 1 (simplex)
  - The elements of a multinomial are vector that sum to 1!
- Dirichlet distribution

$$P(\theta|\alpha) = \frac{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta_k^{\alpha_k - 1}$$

  - Dirichlet parameter α_k is a prior count of the k-th topic
  - It controls the mean shape and sparsity of multinomial parameters θ

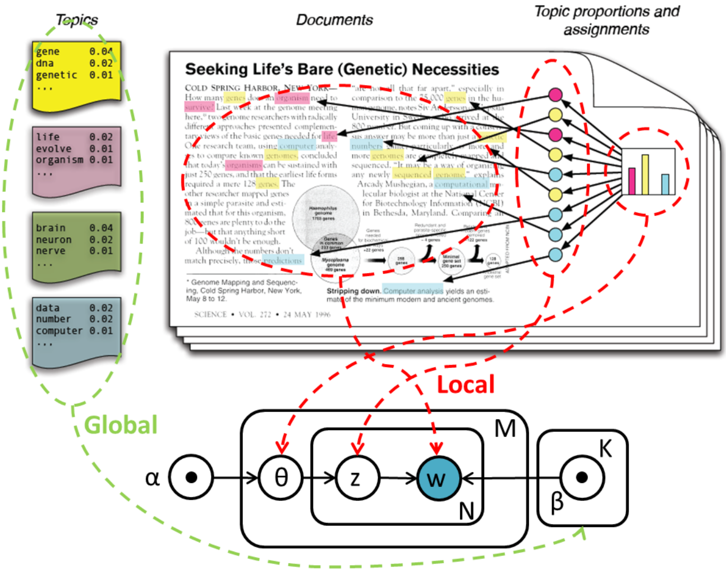
Effect of the α parameter

$\alpha = 0.001$

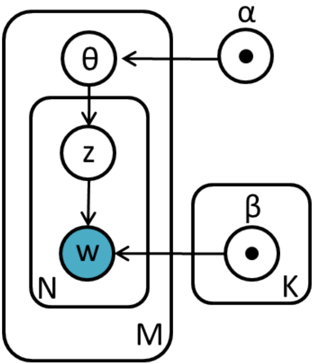


Slide Credit - Blei at KDD 2011 Tutorial

LDA and Text Analysis



LDA Generative Process



For each of the M documents

- Choose $\theta \sim \text{Dirichlet}(\alpha)$
- For each of the N words
  - Choose a topic $z \sim \text{Multinomial}(\theta)$
  - Pick a word w_j with multinomial probability $P(w_j|z,\beta)$

Multinomial topic-word parameter matrix $[\beta]_{K\times V}$

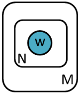
$$\beta_{kj} = P(w_j = 1|z_k = 1)$$

$$\text{or} \quad P(w_j = 1|z = k)$$

$$P(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta) = P(\theta|\alpha)\prod_{j=1}^{N} P(z_j|\theta)P(w_j|z_j, \beta)$$

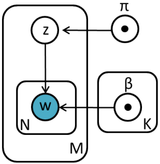
Relationship With Other Latent Variable Models

Unigram



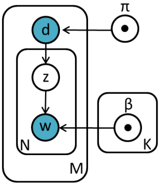
$$P(\mathbf{w}) = \prod_{j=1}^{N} w_j$$

Unigram Mixture



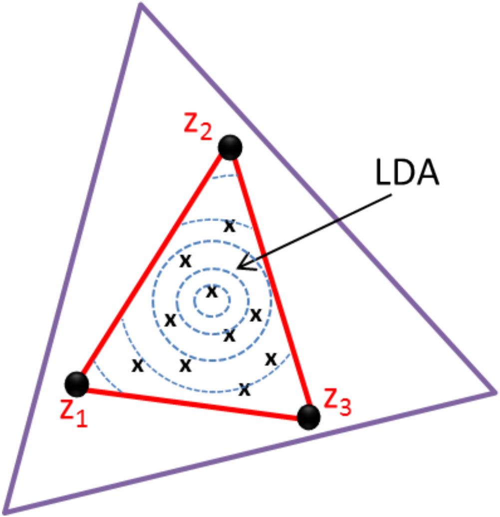
$$P(\mathbf{w}|\pi, \beta) = \sum_z P(z|\pi) \prod_{j=1}^{N} P(w_j|z, \beta)$$

PLSA



$$P(\mathbf{w}, d_i|\pi, \beta) = \prod_{j=1}^{N} P(d_i|\pi) \sum_{k=1}^{K} P(z_k|d_i) \times P(w_j|z_k, \beta)$$

Geometric Interpretation



Learning in LDA

Marginal distribution of a document **w**

$$P(\mathbf{w}|\alpha,\beta) = \int \sum_{\mathbf{z}} P(\theta,\mathbf{z},\mathbf{w}|\alpha,\beta)d\theta$$
$$= \int P(\theta|\alpha) \prod_{j=1}^{N} \sum_{z_j=1}^{k} P(z_j|\theta)P(w_j|z_j,\beta)d\theta$$

Given $\{\mathbf{w}_1,\ldots,\mathbf{w}_M\}$, find (α,β) maximizing

$$\mathcal{L}(\alpha,\beta) = \log \prod_{i=1}^{M} P(\mathbf{w}_i|\alpha,\beta)$$

Learning with hidden variables $\Rightarrow$ Expectation-Maximization

Key problem is inferring latent variables posterior

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

Posterior Inference

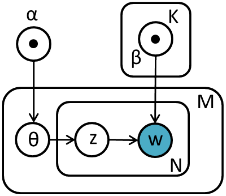
Problem comes with marginal computation

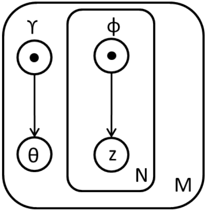
$$P(\mathbf{w}|\alpha,\beta) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \int \prod_{k=1}^{K} \theta_k^{\alpha_k - 1} \left(\prod_{j=1}^{N} \sum_{k=1}^{K} \prod_{v=1}^{V} (\theta_k \beta_{kv})^{w_j^v} \right) d\theta$$

Exact inference is intractable due to the couplings between β and θ in the summation over topics

- Gibbs Sampling (Griffiths and Steyvers, 2004)
  - Construct a Markov chain on the hidden variables whose limiting distribution is the posterior
  - Takes days to converge (but it is accurate)
- Variational Inference (Blei, Ng and Jordan, 2003)
  - Approximate the true posterior with lower bound $Q(\theta)$
  - Takes hours to converge (but it is an approximation)

Variational Posterior Inference





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

$$Q(\theta, \mathbf{z}|\gamma, \phi) = Q(\theta|\gamma) \prod_{k=1}^{K} Q(z_k|\phi)$$

Take a simplified graphical model with variational parameters γ and ϕ and find the values that make $Q(\theta, \mathbf{z}|\gamma, \phi)$ a good approximation of posterior $P(\theta, \mathbf{z}|\mathbf{w}, \alpha, \beta)$

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

Variational Expectation-Maximization

Initialize topics randomly
    repeat
        for each document **do**
            repeat
                Update topic-assignment variational parameters ϕ
                Update topic-proportions variational parameters γ
            until topic proportions change little
        end for
    Update topics distribution β from estimated variational parameters
until little likelihood improvement

Approximate posterior by finding variational parameters of $Q(\cdot)$ (E-STEP). Update model parameters using aggregated statistics from approximated posterior (M-STEP)

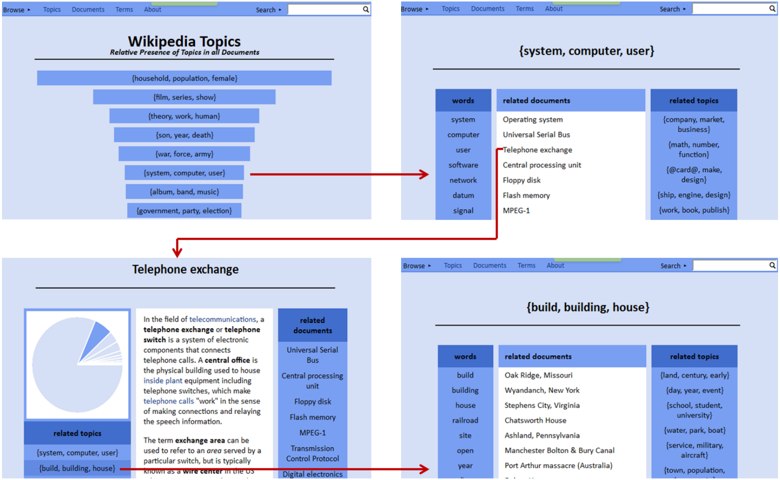
Applications

Why using LDA (and other topic models)?

- Organize large collections of documents by identifying shared topics
- Understanding the documents semantics (unsupervised)
- Documents are of different nature
  - Text
  - Images
  - Video
  - Relational data (graphs, time-series, etc..)
- In short: a model for collections of high-dimensional vectors whose attributes are multinomial distributions

Organizing and Browsing Text Collections

A Browser of 100K Wikipedia Documents

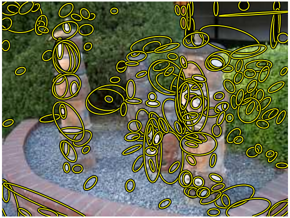


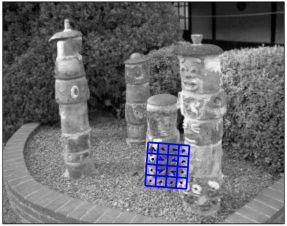
http://www.princeton.edu/~achaney/tmve/wiki100k/browse/topic-presence.html

Understanding Image Collections

Can we apply the latent topic analysis to visual documents?

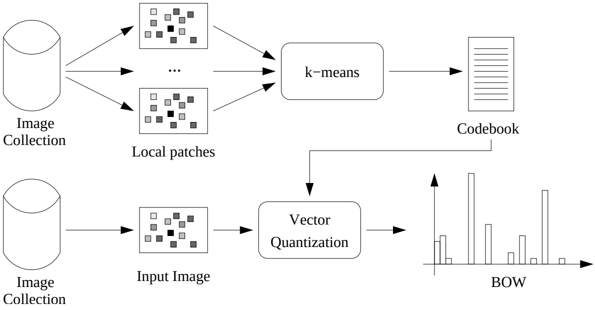
- Yes, but we need a way to represent visual content as in text
  - Text $\equiv$ collection of words
  - Image $\equiv$ collection of ?
- Visual patches
  - Local descriptors of image parts
  - How to determine what are relevant image parts
  - How to describe visual content





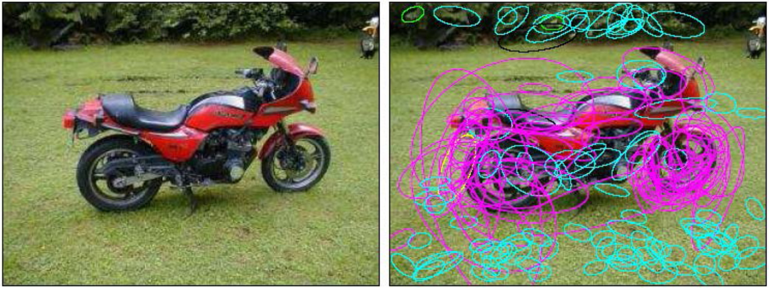
Bag of Words Image Representation

- Each image is a document and each visual patch is a word
- Count the occurrences of each dictionary visual word (visterm) in your image
- Represent the image as a vector of visterm counts (histogram)



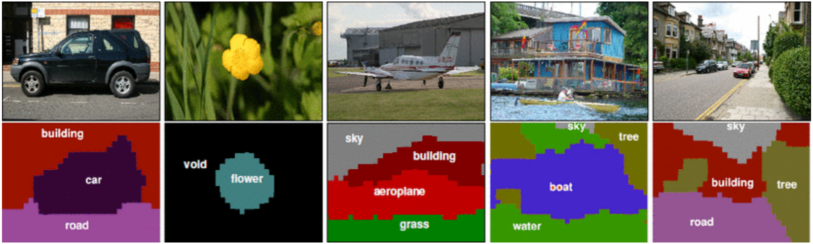
LDA Image Analysis

Assigning a topic to each visual patch



LDA Image Analysis

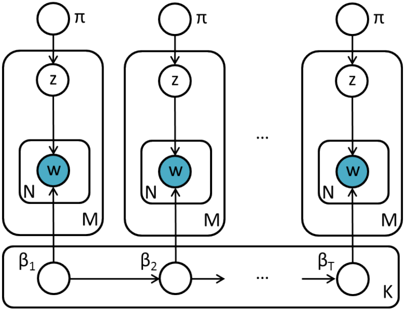
Combining topic models with Markov Random Fields



Dynamical Topic Models

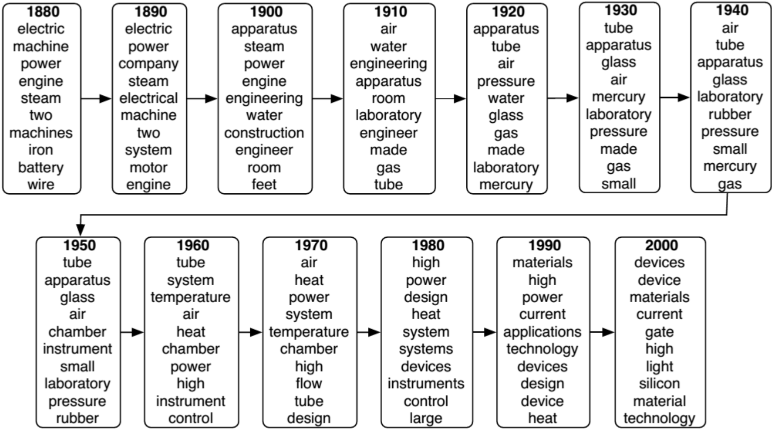
LDA assumes that the document order does not count

- What if we want to track topic evolution over time?
- Tracking how language changes over time
- Videos are image documents over time



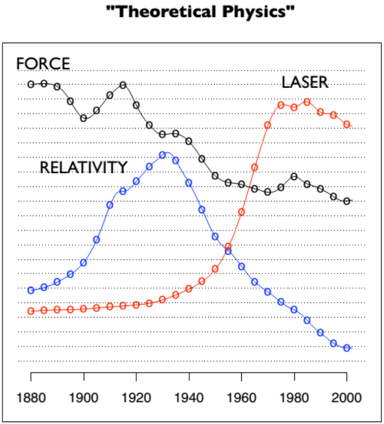
Dealing with sequential information

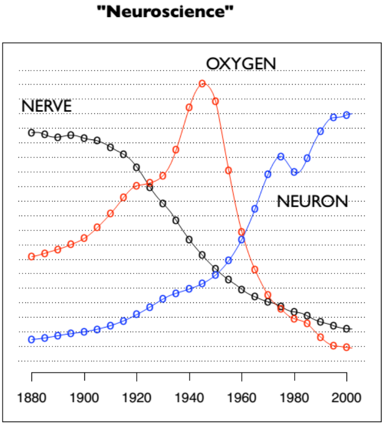
Topic Evolution over Time



http://topics.cs.princeton.edu/Science/

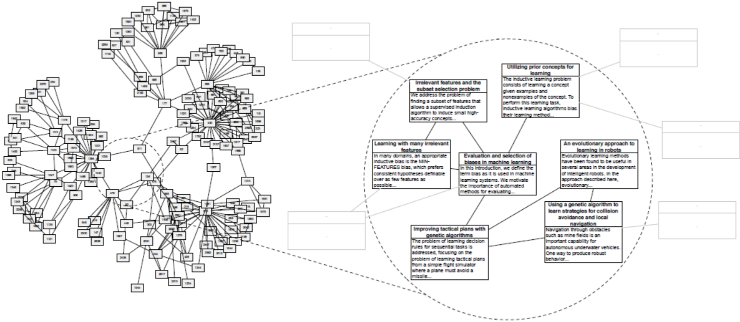
Topic Trends





http://topics.cs.princeton.edu/Science/

Relational Topic Models



- Using topic models with relational data (graphs)
- Community discovery and connectivity pattern profiles (Kemp, Griffiths, Tenenbaum, 2004)
- Joint content-connectivity analysis (Blei,Chang, 2010)

Take Home Messages

- Latent variable models
  - Introduce unobserved variables to make joint distribution tractable
  - Latent space representation of high dimensional data
- Latent Dirichlet Allocation
  - Assumes there are K topics shared by documents and each document is generated by a mixture of topics
  - Dirichlet because it is a distribution on positive sum-to-one vectors and is a conjugate of multinomial
  - Cannot perform exact inference (sampling or variational)
  - Issues? Topic number, BOW assumptions
- Loads of interesting applications to document understanding
  - Finding structure in document collections (unsupervised)
  - Applications to text, image, video, linked data