Introduction to Topic Models

Vivi Nastase

Summer semester 2012

ICL, University of Heidelberg

Course plan

Scheduling:

- Lecture: Thursdays, 14-16, here
- Office hours: Thursdays, 11-12 (Room 121)
- e-mail: nastase@cl.uni-heidelberg.de

Work:

- attend the lectures, and interact – bring pens and papers! I will rarely have slides
- a semester long project
- present and discuss an assigned paper
- oral exam

Goals

- understand the mathematical formalism behind topic models
- figure out the strengths and weaknesses of this type of approaches (the hunting joke is true!)
- look at some of the more interesting extensions of the vanilla LDA
- give you hands on experience in developing a topic model

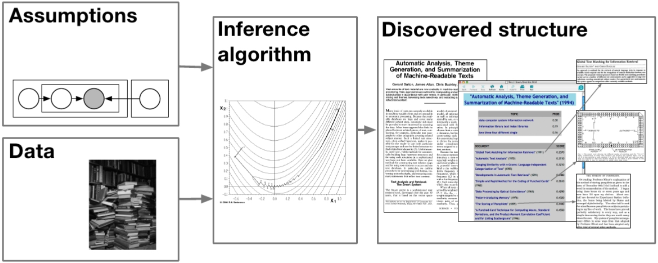
Project: LDA with your favourite extension

Homework 1, due date May 17^{th}:

- pick your favourite text collection from the ICL's resources
- implement a system that splits the input data into fragments (sentences / paragraphs/ documents) – this should be a parameter
- represent the data in a structure that matches the split
- send me an archive with your code and documentation by May 17^{th}

Why topic models?

Topic models



from David Blei, KDD-11 tutorial

- Observation: a collection of texts
- Assumption: the texts have been generated according to some model
- Output: the model that has generated the texts

Topic models

- Discover hidden topical patterns that pervade the collection through statistical regularities
- Annotate documents with these topics
- Use the topic annotations to organize, summarize, search texts ...

Topic examples

Topic 247

word	prob.
DRUGS	.069
DRUG	.060
MEDICINE	.027
EFFECTS	.026
BODY	.023
MEDICINES	.019
PAIN	.016
PERSON	.016
MARIJUANA	.014
LABEL	.012
ALCOHOL	.012
DANGEROUS	.011
ABUSE	.009
EFFECT	.009
KNOWN	.008
PILLS	.008

Topic 5

word	prob.
RED	.202
BLUE	.099
GREEN	.096
YELLOW	.073
WHITE	.048
COLOR	.048
BRIGHT	.030
COLORS	.029
ORANGE	.027
BROWN	.027
PINK	.017
LOOK	.017
BLACK	.016
PURPLE	.015
CROSS	.011
COLORED	.009

Topic 43

word	prob.
MIND	.081
THOUGHT	.066
REMEMBER	.064
MEMORY	.037
THINKING	.030
PROFESSOR	.028
FELT	.025
REMEMBERED	.022
THOUGHTS	.020
FORGOTTEN	.020
MOMENT	.020
THINK	.019
THING	.016
WONDER	.014
FORGET	.012
RECALL	.012

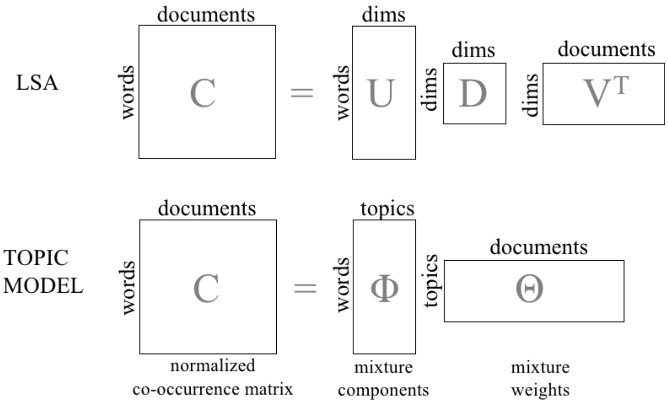
Topic 56

word	prob.
DOCTOR	.074
DR.	.063
PATIENT	.061
HOSPITAL	.049
CARE	.046
MEDICAL	.042
NURSE	.031
PATIENTS	.029
DOCTORS	.028
HEALTH	.025
MEDICINE	.017
NURSING	.017
DENTAL	.015
NURSES	.013
PHYSICIAN	.012
HOSPITALS	.011

Figure 1. An illustration of four (out of 300) topics extracted from the TASA corpus.

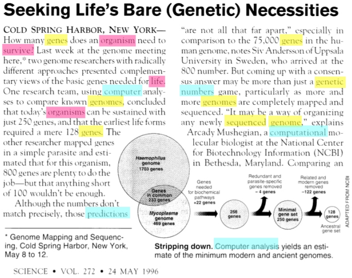
Steyvers & Griffiths, 2006

LSA and topic models



Steyvers & Griffiths, 2006

Topic models – intuition



Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those predictions "are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a genetic numbers game, particularly as more and more genomes are completely mapped and sequenced. "It may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

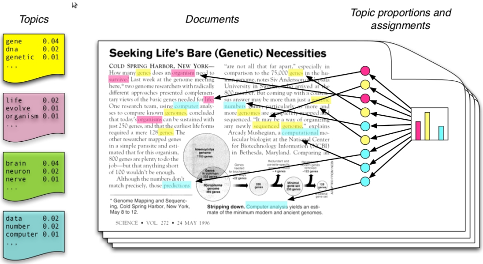
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

- Find the latent structure of "topics" or "concepts" in a text corpus, which is obscured by "word choice" noise
- Deerwester et al (1990) – LSA – co-occurrence of terms in text documents can be used to recover this latent structure, without additional knowledge.
- Latent topic representations representations of text allow modelling linguistic phenomena, like **synonymy** and **polysemy**.

Topic models



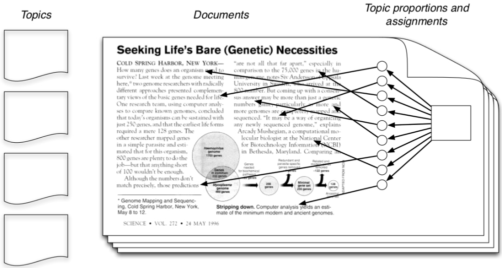
Each document is a mixture of topics:

$$\sum_k p(z_m = k) = \sum_k \theta_{m,k} = 1$$

Each word is drawn from one of its document's topics:

$$p(w_{m,n}) = \sum_k p(w_{m,n} | z_{m,n} = k) p(z_{m,n} = k) = \sum_k \varphi_k(w_{m,n}) \theta_{m,k}$$

Topic models



The **observations** are the documents: $\mathbf{w_m}, m \in 1, M$
We need to infer the **model**, i.e the underlying topic structure, i.e. the topic assignments $z_{m,n}$, the topic θ_m, $m \in 1, M$ and word distributions φ_k, $k \in 1, K$

Priors:

$\theta \sim$ distribution with hyperparameter α

$\varphi \sim$ distribution with hyperparameter β

Topic models – Latent Dirichlet Allocation

$$p(\theta|\alpha) = \frac{1}{B(\alpha)} \prod_k \theta_k^{\alpha_k - 1}$$

$$\sum_k \theta_{m,k} = 1$$

α controls the mean shape and sparsity of θ
The topic proportions (θ_m) are a K-dimensional Dirichlet
$z_{m,n}$ are multinomial distributions from θ_m

$$p(z_{m,n}|\theta_m) = \frac{N!}{\prod_{k=1}^{K} n_k!} \prod_{k=1}^{K} \theta_{m,k}^{n_k}$$

Topic models – Latent Dirichlet Allocation

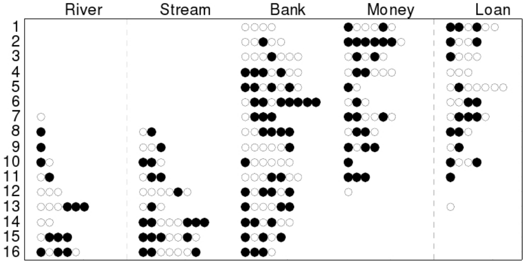
$$p(\varphi|\beta) = \frac{1}{B(\beta)} \prod_{v} \varphi_v^{\beta_v - 1}$$

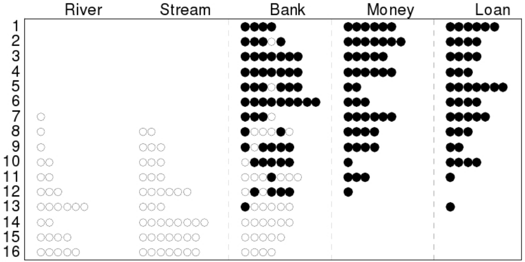
$$\sum_{v} \varphi_{k,v} = 1$$

β controls the mean shape and sparsity of φ
The topics (φ_k) are a V-dimensional Dirichlet
$w_{m,n}$ are multinomial distributions from $\varphi_{z_{m,n}}$

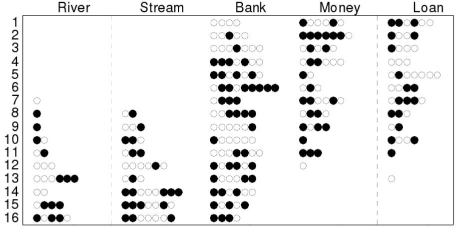
$$p(w_{m,n}|\varphi_k) = \frac{V!}{\prod_{v=1}^{V} n_v!} \prod_{v=1}^{V} \varphi_{k,v}^{n_v}$$

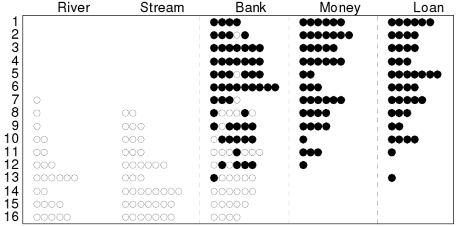
Topic models – inference via Gibbs sampling





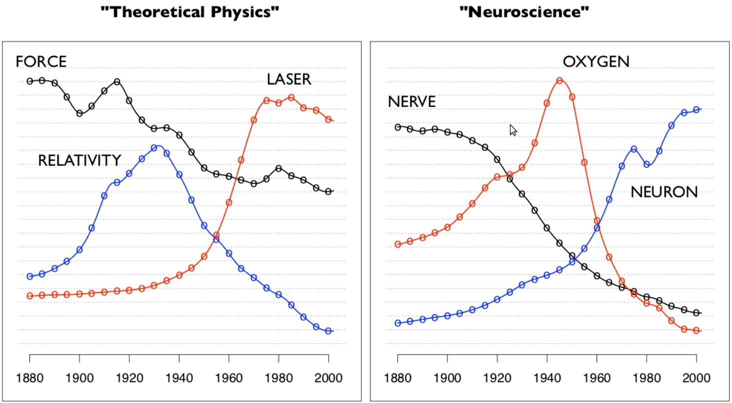
Topic models – inference via Gibbs sampling





$$p(x = 1|\mathcal{O}, \alpha_h, \alpha_t) = \frac{p(x = 1, \mathcal{O}|\alpha_h, \alpha_t)}{p(\mathcal{O}|\alpha_h, \alpha_t)} = \frac{n_h + \alpha_h}{N + \alpha_h + \alpha_t}$$

Topic examples



Topic examples

SKY WATER TREE
MOUNTAIN PEOPLE

SCOTLAND WATER
FLOWER HILLS TREE

SKY WATER BUILDING
PEOPLE WATER

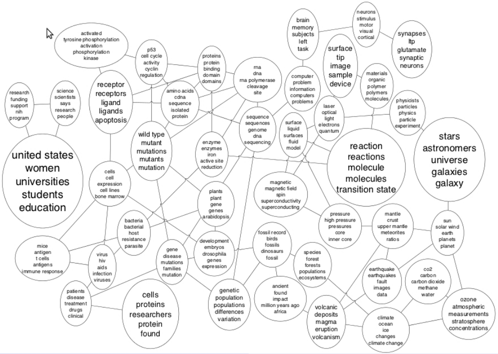
FISH WATER OCEAN
TREE CORAL

PEOPLE MARKET PATTERN
TEXTILE DISPLAY

BIRDS NEST TREE
BRANCH LEAVES

Object $\equiv$ bag of words with labels

Topic examples



Basic components:

- A set of entities (e.g. documents, images, individuals, genes)
- A set of relations (e.g. citation, coauthor, co-tag, friends, pathways)

Topic models in machine learning

- generative – assume an underlying model (probability distribution, parameters) generated the observed data
- the class is a hidden variable
- can handle a large number of classes
- difference relative to discriminative models?

Topic models in machine learning

- generative – assume an underlying model (probability distribution, parameters) generated the observed data
- the class is a hidden variable
- can handle a large number of classes
- difference relative to discriminative models?

  discriminative: $P(Y|X)$
  generative: $P(Y, X)$

References

- *Probabilistic topic models*, Mark Steyvers, Tom Griffiths
- *Parameter estimation for text analysis*, Gregor Heinrich
- *Topic Models*, David Blei (tutorial, videolectures.net)
- Any of the many tutorials you can find on-line

Probabilities refresher

probability/probable

late 14c., from O.Fr. probable (14c.), from L. probabilis "provable," from probare "to try, to test"

Wahrsheinlichkeit/wahrsheinlich

seems to be true

Probabilities refresher

An experiment whose outcome depends on chance

random variable $\mathbf{X}$ captures the outcome of the experiment

sample space S the set of all possible outcomes

event $E \subseteq S$

$\mathbf{X}$ can be

discrete if S is finite or countably infinite

continuous

Examples?

Distributions and probabilities

The distribution function:

$$p : S \rightarrow [0, 1]$$

$$p(x) \geq 0, \forall x \in S$$

$$\sum_{x \in S} p(x) = 1$$

Distributions and probabilities

The distribution function:

$$p : S \to [0, 1]$$

$$p(x) \geq 0, \forall x \in S$$

$$\sum_{x \in S} p(x) = 1$$

Probability of an event:

$$P(E) = \sum_{x \in E} p(x)$$

$$P(\{x\}) = p(x)$$

A bit of practice

1. dice rolling
2. tossing two coins

Properties of probabilities

$$P(E) \geq 0, \forall E \subseteq S$$

Properties of probabilities

$$P(E) \geq 0, \forall E \subseteq S$$

$$P(S) = 1$$

Properties of probabilities

$$P(E) \geq 0, \forall E \subseteq S$$

$$P(S) = 1$$

$$E \subset F \subset S \rightarrow P(E) \leq P(F)$$

Properties of probabilities

$$P(E) \geq 0, \forall E \subseteq S$$

$$P(S) = 1$$

$$E \subset F \subset S \rightarrow P(E) \leq P(F)$$

$$E \cap F = \emptyset \rightarrow P(E \cup F) = P(E) + P(F)$$

Properties of probabilities

$$P(E) \geq 0, \forall E \subseteq S$$

$$P(S) = 1$$

$$E \subset F \subset S \rightarrow P(E) \leq P(F)$$

$$E \cap F = \emptyset \rightarrow P(E \cup F) = P(E) + P(F)$$

$$P(\bar{\mathrm{E}}) = 1 - P(E)$$

Properties of probabilities

$$P(E) \geq 0, \forall E \subseteq S$$

$$P(S) = 1$$

$$E \subset F \subset S \rightarrow P(E) \leq P(F)$$

$$E \cap F = \emptyset \rightarrow P(E \cup F) = P(E) + P(F)$$

$$P(\bar{\mathrm{E}}) = 1 - P(E)$$

Proofs?

Examples of probabilities in language models

- the sample space
- the events
- distributions

Expected value

Discrete:

$$E(X) = \sum_{x \in S} xP(x)$$

Continuous:

$$E(X) = \int_a^b xp(x)dx$$

Common discrete distributions

Uniform(n) : $|S| = n$, n is finite

$$P(X = x) = \frac{1}{n}$$

Common discrete distributions

Uniform(n) : $|S| = n$, n is finite

$$P(X = x) = \frac{1}{n}$$

Bernoulli(p) : $p \in [0, 1]; X \in 0, 1$:

$$P(X = 1) = p; P(X = 0) = 1 - p$$

Common discrete distributions

Uniform(n) : $|S| = n$, n is finite

$$P(X = x) = \frac{1}{n}$$

Bernoulli(p) : $p \in [0, 1]; X \in 0, 1$:

$$P(X = 1) = p; P(X = 0) = 1 - p$$

Binomial(p,n) : $p \in [0, 1]; X \in 0, 1, ..., n; n \in \mathbb{N}$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{(n-x)}$$

Common discrete distributions

Uniform(n) : $|S| = n$, n is finite

$$P(X = x) = \frac{1}{n}$$

Bernoulli(p) : $p \in [0, 1]; X \in 0, 1$:

$$P(X = 1) = p; P(X = 0) = 1 - p$$

Binomial(p,n) : $p \in [0, 1]; X \in 0, 1, ..., n; n \in \mathbb{N}$

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{(n-x)}$$

Multinomial($p_1, ..., p_k; x_1, ..., x_k; n$) : $\sum_i x_i = n$

$$P(X_1 = x_1, ..., X_k = x_k) = \frac{n!}{x_1!...x_k!} p_1^{x_1} ... p_k^{x_k}$$

...

Common continuous distributions

$$P(X \leq x) = \int_{-\infty}^{x} p(y)dy$$

Common continuous distributions

$$P(X \leq x) = \int_{-\infty}^{x} p(y)dy$$

Uniform(a,b) : $a, b \in \mathbb{R}, a < b, X \in [a, b]$

$$p(x) = \frac{1}{b-a}$$

Common continuous distributions

$$P(X \leq x) = \int_{-\infty}^{x} p(y)dy$$

Uniform(a,b) : $a, b \in \mathbb{R}, a < b, X \in [a, b]$

$$p(x) = \frac{1}{b - a}$$

Beta(α, β) : $\alpha, \beta \in \mathbb{R}_{++}, X \in [0, 1]$

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

Common continuous distributions

$$P(X \leq x) = \int_{-\infty}^{x} p(y)dy$$

Uniform(a,b) : $a, b \in \mathbb{R}, a < b, X \in [a, b]$

$$p(x) = \frac{1}{b - a}$$

Beta(α, β) : $\alpha, \beta \in \mathbb{R}_{++}, X \in [0, 1]$

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1}$$

Dirichlet(α) : generalization of $Beta(\alpha, \beta)$

Common continuous distributions

$$P(X \leq x) = \int_{-\infty}^{x} p(y)dy$$

Uniform(a,b) : $a, b \in \mathbb{R}, a < b, X \in [a, b]$

$$p(x) = \frac{1}{b-a}$$

Beta(α, β) : $\alpha, \beta \in \mathbb{R}_{++}, X \in [0, 1]$

$$p(x; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

Dirichlet(α) : generalization of $Beta(\alpha, \beta)$

Normal(μ, σ^2) : $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_{++}, X \in \mathbb{R}$

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Test

Test

Two random variables
thought they were discrete
but I heard them continuously.

Next week sneak preview

Next week sneak preview

Bayes' law and conjugate distributions