INSTITUTE FOR HEALTH RESEARCH AND POLICY

UIC

MRC Chalk Talk

Getting Started with Probabilistic Topic Models

Yoonsang Kim, PhD
Methodology Research Core
Institute for Health Research and Policy
November 10, 2015

Topic Models

- Analyze the words of texts to discover the topics and how each document exhibits the topics
- Unsupervised learning
- Probabilistic models for uncovering the underlying semantic/thematic structure of a document collection based on a hierarchical Bayesian analysis

Tobacco smoking on Twitter

Topic	Most Likely Topic Components (n-grams)	%
1	smoking weed, smoking gun, smoking crack, stop smoking, cigarette burns, external cell phones, hooka bar, youre smoking, smoke cigars, smoking kush, hand smoke, im taking, smoking barrels, hookah house, hes smoking, ryder cup, dont understand, talking bout, im ready, twenty years people	0.16
2	quit smoking, stop smoking, cigar guy, smoking cigarettes, hookah bar, usa protect, quitting smoking, started smoking, electronic cigarette, cigars link, smoking addiction, cigar shop, quit smoking cigarettes, chronical green smoke, link quit smoking naturally, smoking pot, youtube video, link quit smoking, link holistic remedies, chronical protect	0.27
3	cigarette smoke, dont smoke, quit smoking, stop smoking, smoking pot, im gonna, hookah tonight, smoking ban, drink specials, free food, ladies night, electronic cigarettes, good times, smoking session, cigarette break, secondhand smoke, everythings real, effective steps, smoking cigs, smoking tonight	0.22
4	smoking weed, cont link, ladies free, piedmont cir, start smoking, hate smoking, hookahs great food, cigarette butts, thingswomenshouldstopdoing smoking, lol rt, sunday spot, cigarettes today, fletcher knebel smoking, pot smoking, film stars, external cell, fetishize holding, smoking room, halloween party, million people	0.25
5	smoke cigarettes, smoking hot, im smoking, smoking section, stopped smoking, chewing tobacco, smoking kills, chain smoking, smoking area, ban smoking, people die, ring ring hookah ring ring, love lafayette, link rt, damn cigarette, healthiest smoking products, theyre smoking, hate cigarettes, world series, hideout apartment	0.06

Twitter Search API, 14 days in Oct 2010 (Prier et al. 2011)

Outline

- Latent Dirichlet Allocation
- Notation and terminology
- Generative process
- How to summarize a corpus
- Parameter Estimation
- Limitations and extensions
- Applications

Latent Dirichlet Allocation

- LDA models documents as arising from multiple topics – generative process
- What is a ***topic***?

  A distribution over a fixed vocab of terms
- Hidden structure (latent topical structure) in the observed data (words of each document) are learned using posterior probabilistic inference

Notation & Terminology

- **Topic**

  Distribution over words

- **Word**

  Basic unit of data

  A word is an item from a vocab {1, …, V}

- **Document**

  A sequence of words {1,…,N}

- **Corpus**

  A collection of documents {1,…,D}

Multinomial Distribution

- The outcome of each trial falls into one of K classes
- For a K-dim random variable y

$$p(y|\theta) = \frac{(\sum_i y_i)!}{y_1!\, y_2!\, \ldots y_K!}\, \theta_1^{y_1} \ldots \theta_K^{y_K}$$

- y_i=num of trials for which the outcome falls into class i
- θ =class probabilities
- e.g. Classify people into 5 income brackets; Throw a dice and observe which number shows up; assign a topic to a word

Dirichlet Distribution

- Distribution over non-negative numbers that sum to 1.
- For a K-dim random variable θ

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \cdots \theta_K^{\alpha_K - 1}$$

- $\theta_i \geq 0, \sum_1^K \theta_i = 1$
- α =parameter vector with $\alpha_i > 0$
- Used as a distribution over discrete distributions (more on this in next slide)

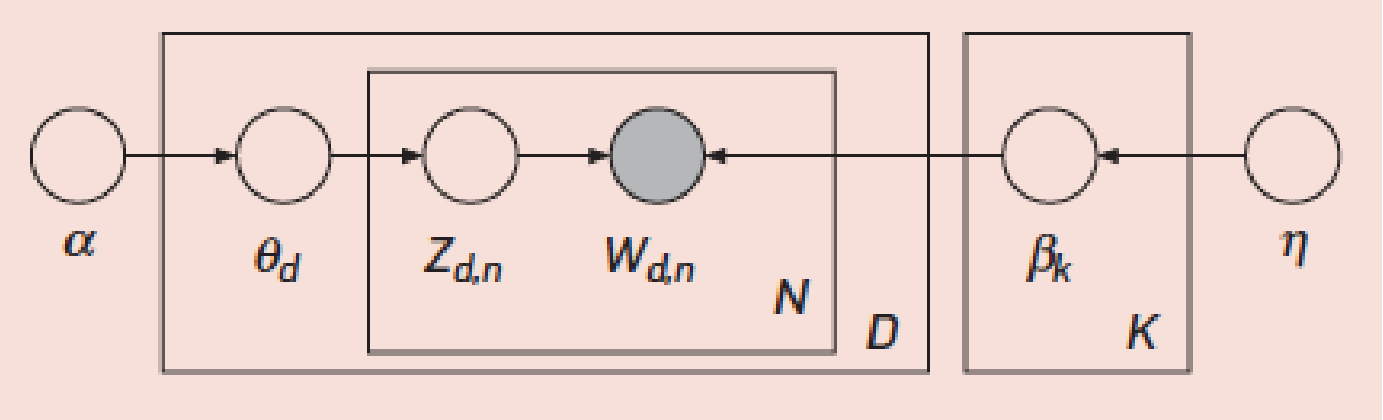
LDA Generative Process

K = num of topics; V = size of vocab
N = num of words in each document d, d $\in \{1, \dots, D\}$, D = size of corpus
K = num of topics (fixed)

1. **For each topic k**
   a. Draw a distribution over words $\beta_k \sim Dir_V(\eta)$
      $\beta = K \times V$ matrix of per-topic word probabilities; represents the relationship between words and topics
2. **For each document d**
   a. Draw a vector of topic proportions $\theta_d \sim Dir(\alpha)$
      θ_d and α are K-dim vectors
   b. For each word
      i. Draw a topic assignment $Z_{d,n} \sim Multinom(\theta_d)$, $Z_{d,n} \in \{1, \dots, K\}$
      ii. Draw a word $W_{d,n} \sim Multinom\left(\beta_{Z_{d,n}}\right)$, $W_{d,n} \in \{1, \dots, V\}$

LDA Generative Process

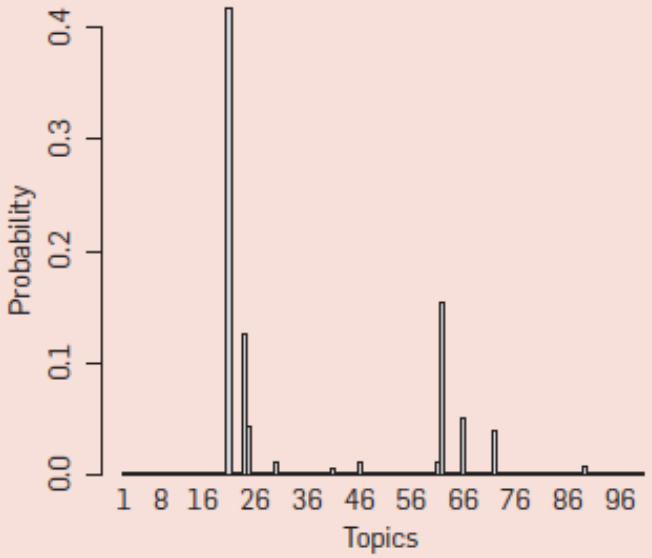
- Graphical model (plate notation)



- Each node is a random variable
- The latent nodes (topic proportions, topic assignments, topics) are unshaded
- The observed nodes (words) are shared
- The “plates” indicate replications

Seeking Life's Bare (Genetic) Necessities

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



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

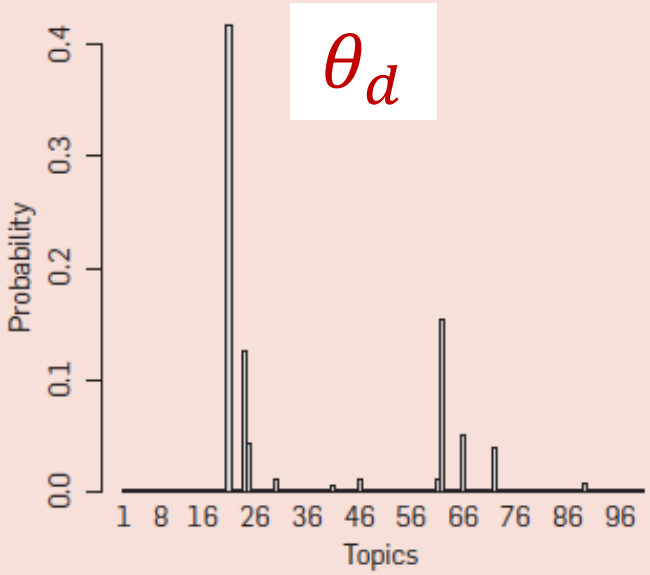
Probabilistic topic models. Blei 2012

Exploring a corpus

- **Visualize a topic**
  - Per-topic word probabilities $\hat{\beta}$
  - Term-score for k^{th} topic, v^{th} word: function of $\hat{\beta}$, down-weights words that have high prob under all topics
- **Visualize a document**
  - Which topics the document is about: plot posterior topic proportions θ
  - Most likely topic assigned to each word: posterior topic assignment z
- **Find similar documents**
  - Document similarity: topic-based similarity btwn documents
  - Hellinger distance (d,d') = $\sum_1^K \left(\sqrt{\hat{\theta}_d} - \sqrt{\hat{\theta}_{d'}}\right)^2$

Seeking Life's Bare (Genetic) Necessities

Figure 2. Real inference with LDA. We fit a 100-topic LDA model to 17,000 articles from the journal *Science*. At left are the inferred topic proportions for the example article in Figure 1. At right are the top 15 most frequent words from the most frequent topics found in this article.



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

Ranked by term-score

Probabilistic topic models. Blei 2012

Parameter Estimation

- **Posterior inference**
  - $p(\theta, z, \beta|w)$= conditional distribution of the topic structure given the documents
  - Intractable ⇒ need approximation
  - Compute posterior means of θ, z, β
- **Variational method**
  - Mean field variation inference (Blei et al 2003, Hoffman et al 2010)
  - Minimize Kullback-Leibler divergence btwn the variational distribution and the true posteriors
- **Sampling-based method**
  - Gibbs sampling (Steyvers and Griffiths 2006)
  - Approximate the posterior with samples

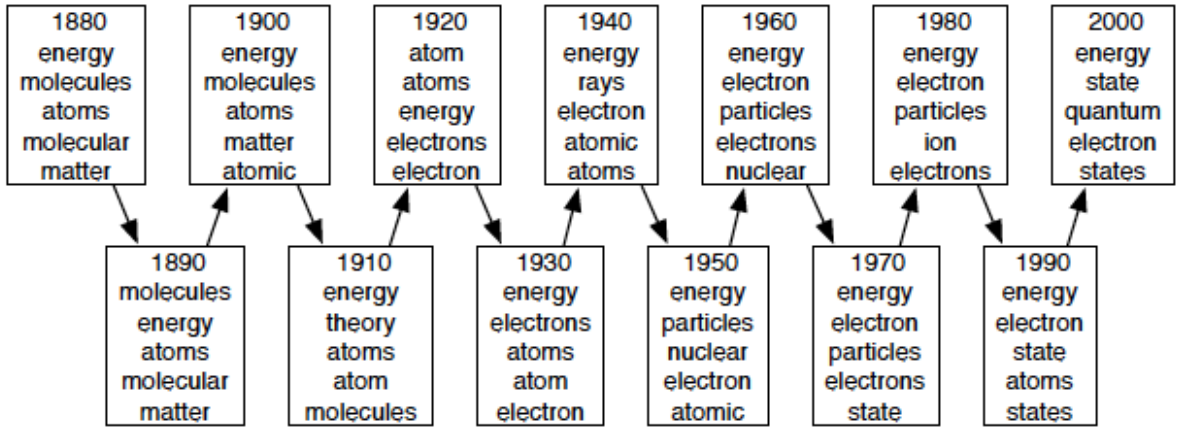
Limitations & Extensions

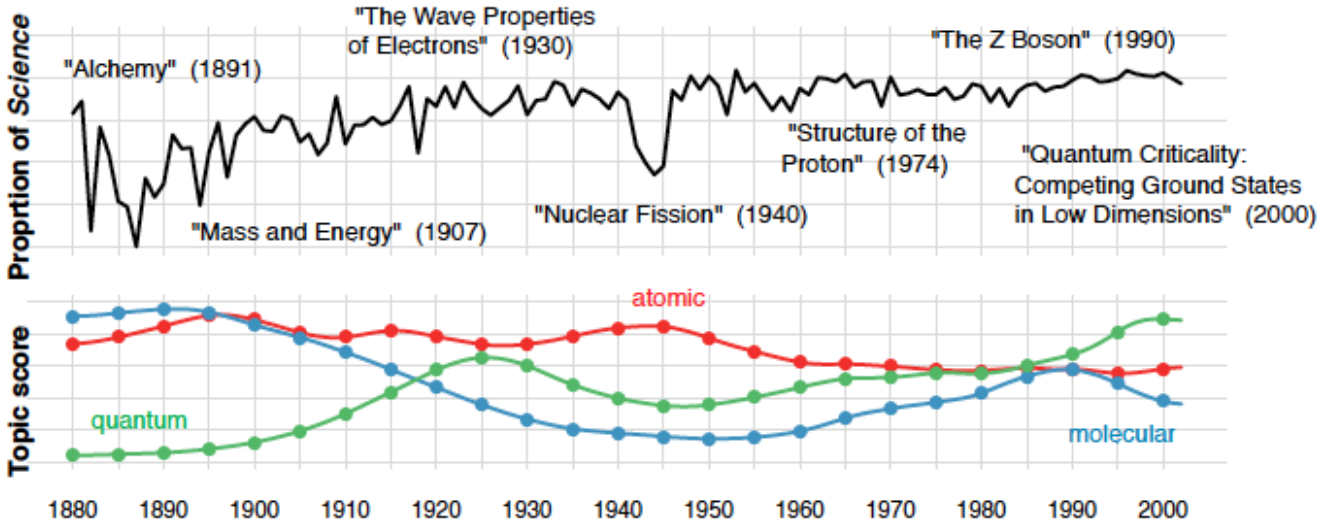
- **Bag of words**: the words are exchangeable in the document
  - e.g. The words in "The William Randolph Hearst Foundation" assigned to multiple topics (Blei, Ng, Jordan 2003)
- **The documents are exchangeable** in the corpus
  - Unrealistic when the topics change over time
  - Dynamic topic model
- **The topics are not correlated**
  - Correlated topic model (Blei and Lafferty 2007)
- **The num of topics is known** and fixed
  - Nonparametric topic model using hierarchical Dirichlet processes (Teh, Jordan, Beal, Blei 2012)
- The **authors** are ignored
  - Author-topic model

Dynamic Topic Model

- Sequentially organize corpus of documents
- Divide the data by time slice, e.g. by year
- Model the documents of each slice
- Topics associated with slice t depend on the topics associated with slice t-1
- Logistic-normal distribution for topic proportions
- Blei and Lafferty (2009) analyzed the entire archive of Science from 1880-2002. The corpus had 140K documents.

Dynamic Topic Model



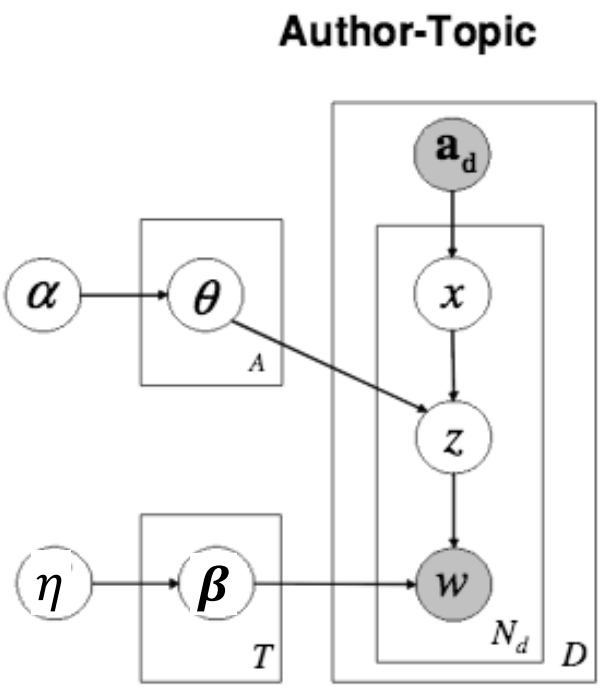


Author-Topic Model

- Explore authors and topics
- Which topics an author writes about? Which authors produce similar work?
- Extension of LDA
- Rosen-Zvi et al. 2004

x = author of a given word, chosen uniformly from the set of authors $\mathbf{a}_d$

Each author is associated with a distribution over topics θ, which is used to select a topic **z**



Author-Topic Model

TOPIC 10	
WORD	PROB.
SPEECH	0.1134
RECOGNITION	0.0349
WORD	0.0295
SPEAKER	0.0227
ACOUSTIC	0.0205
RATE	0.0134
SPOKEN	0.0132
SOUND	0.0127
TRAINING	0.0104
MUSIC	0.0102
AUTHOR	PROB.
Waibel_A	0.0156
Gauvain_J	0.0133
Lamel_L	0.0128
Woodland_P	0.0124
Ney_H	0.0080
Hansen_J	0.0078
Renals_S	0.0072
Noth_E	0.0071
Boves_L	0.0070
Young_S	0.0069

TOPIC 209	
WORD	PROB.
PROBABILISTIC	0.0778
BAYESIAN	0.0671
PROBABILITY	0.0532
CARLO	0.0309
MONTE	0.0308
DISTRIBUTION	0.0257
INFERENCE	0.0253
PROBABILITIES	0.0253
CONDITIONAL	0.0229
PRIOR	0.0219
AUTHOR	PROB.
Friedman_N	0.0094
Heckerman_D	0.0067
Ghahramani_Z	0.0062
Koller_D	0.0062
Jordan_M	0.0059
Neal_R	0.0055
Raftery_A	0.0054
Lukasiewicz_T	0.0053
Halpern_J	0.0052
Muller_P	0.0048

TOPIC 87	
WORD	PROB.
USER	0.2541
INTERFACE	0.1080
USERS	0.0788
INTERFACES	0.0433
GRAPHICAL	0.0392
INTERACTIVE	0.0354
INTERACTION	0.0261
VISUAL	0.0203
DISPLAY	0.0128
MANIPULATION	0.0099
AUTHOR	PROB.
Shneiderman_B	0.0060
Rauterberg_M	0.0031
Lavana_H	0.0024
Pentland_A	0.0021
Myers_B	0.0021
Minas_M	0.0021
Burnett_M	0.0021
Winiwarter_W	0.0020
Chang_S	0.0019
Korvemaker_B	0.0019

TOPIC 20	
WORD	PROB.
STARS	0.0164
OBSERVATIONS	0.0150
SOLAR	0.0150
MAGNETIC	0.0145
RAY	0.0144
EMISSION	0.0134
GALAXIES	0.0124
OBSERVED	0.0108
SUBJECT	0.0101
STAR	0.0087
AUTHOR	PROB.
Linsky_J	0.0143
Falcke_H	0.0131
Mursula_K	0.0089
Butler_R	0.0083
Bjorkman_K	0.0078
Knapp_G	0.0067
Kundu_M	0.0063
Christensen-J	0.0059
Cranmer_S	0.0055
Nagar_N	0.0050

Analyzed 162K abstracts from CiteSeer, 85K authors, >11M word tokens

Dimension Reduction

- LDA as a dimensionality reduction
- LDA reduces any document to a set of real-valued features $\gamma^*(\boldsymbol{w})$ in low-dim
- But, how much information is lost?
- Blei, Ng, and Jordan (2009)

  Conducted two binary classification experiments using the Reuters dataset: 8,000 documents and 15,818 words

  i. Estimated $\gamma^*(\boldsymbol{w})$ using an 50-topic LDA and fit SVM on those LDA-based features

  ii. Fit SVM on all the word features

Dimension Reduction

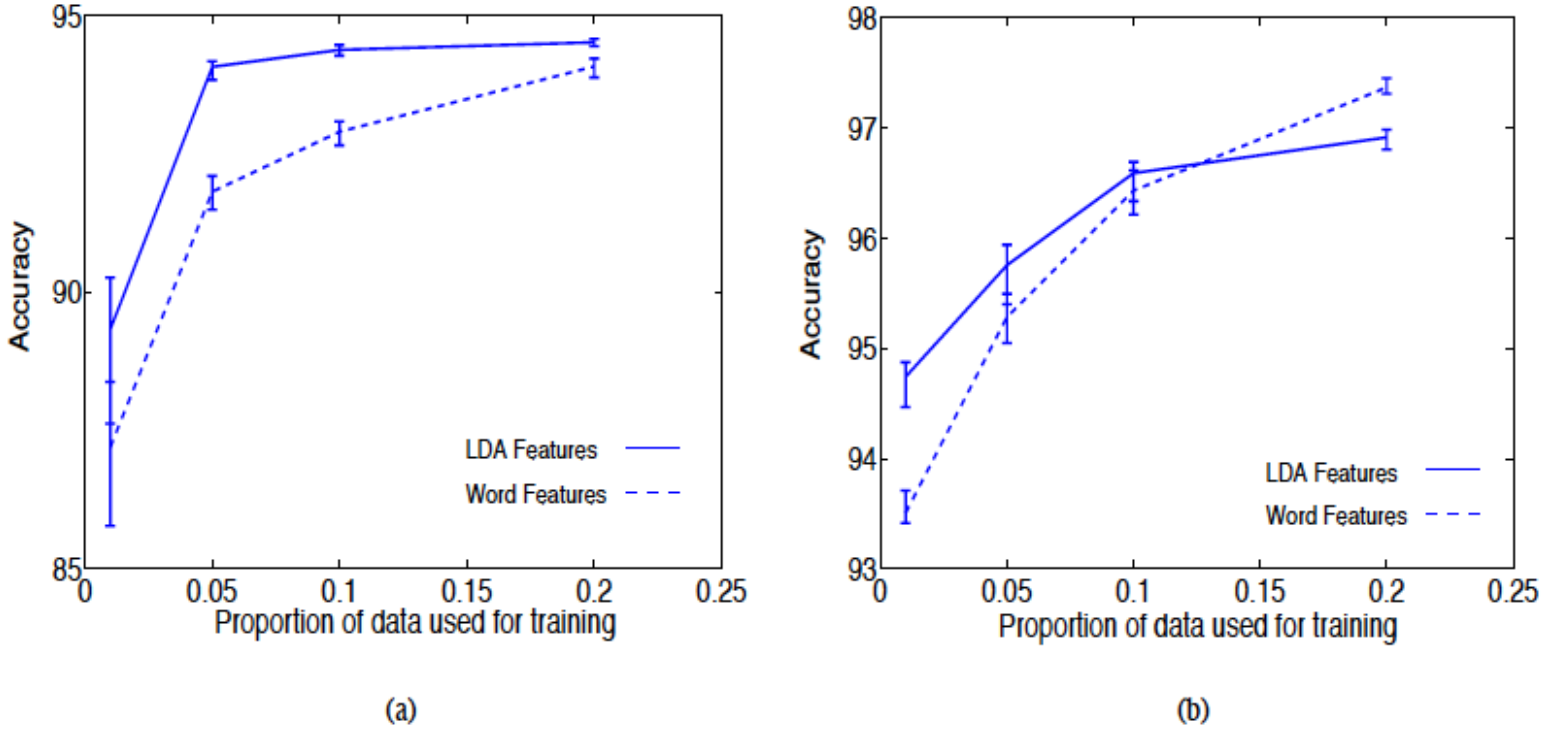


Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

Tools

- https://www.cs.princeton.edu/~blei/topicmodeling.html
- MALLET toolkit http://mallet.cs.umass.edu/
- R package mallet – operate MALLET within R

Other Applications

- **Genetic data**: to find ancestral populations; to characterize the genetic patterns and identify how each individual expresses them
- **Image analysis**: each image exhibit a combination of visual patterns, which recur throughout a collection of images
- **Social networks**: author-recipient topic model; to characterize topic distributions based on the direction-sensitive messages sent between people (McCallum et al. 2007)

Comments

- LDA is a useful **exploratory tool.** Topical structures found with topic models are **not definitive**.
- New key terms to retrieve social media data can be discovered using topic models.
- The “LDA” can mean another method!

References

- Blei, David M. “Probabilistic Topic Models.” *Communications of the ACM* 55, no. 4 (April 1, 2012): 77. doi:10.1145/2133806.2133826.
- Blei, David M., and J. Lafferty. “Topic Models.” *Text Mining: Classification, Clustering, and Applications* 10 (2009): 71.
- Blei, David M., Andrew Y. Ng, and Michael I. Jordan. “Latent Dirichlet Allocation.” *The Journal of Machine Learning Research* 3 (2003): 993–1022.
- Grün, Bettina, and Kurt Hornik. “Topicmodels: An R Package for Fitting Topic Models.” *Journal of Statistical Software* 40, no. 13 (2011): 1–30.
- McCallum, Andrew, Andres Corrada-Emmanuel, and Xuerui Wang. “Topic and Role Discovery in Social Networks.” *Computer Science Department Faculty Publication Series*, 2005, 3.
- Prier, Kyle W., Matthew S. Smith, Christophe Giraud-Carrier, and Carl L. Hanson. “Identifying Health-Related Topics on Twitter.” In *Social Computing, Behavioral-Cultural Modeling and Prediction*, 18–25. Springer, 2011. http://link.springer.com/chapter/10.1007/978-3-642-19656-0_4.
- Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. “The Author-Topic Model for Authors and Documents.” In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, 487–94. AUAI Press, 2004. http://dl.acm.org/citation.cfm?id=1036902.