An Introduction to Latent Dirichlet Allocation(LDA)

Tutorial 4

YE Rong

CONTENTS

Prerequisite

Part I

- Conjugacy distribution
- Dirichlet distribution
- Statistical Language Model
- Topic models

Bayesian inference and conjugacy distribution

- Bayes' theorem

$$p(A|B)=\frac{p(B|A)\bullet p(A)}{p(B)}\propto p(B|A)\bullet p(A)$$

$$posterior\propto prior\times likelihood$$

- Conjugacy distribution:
    - Posterior is in the same probability distribution family as the prior.
    - Conjugacy prior (given likelihood…)

Example - Binomial & Beta

- Given likelihood of a binomial distribution

$$p(X = k \mid \theta, n) = \binom{n}{k} \theta^k (1-\theta)^{n-k} \propto \theta^k (1-\theta)^{n-k}$$

- Consider a *beta* distribution
  - Form: $p(\theta \mid \alpha, \beta) = const \times \theta^{\alpha-1}(1-\theta)^{\beta-1}$

$$const = \mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

- Update θ -- posterior

$$p(\theta \mid X = k, n) \propto p(X = k \mid \theta, n) \times p(\theta)$$

$$= \theta^{k+\alpha-1}(1-\theta)^{n-k+\beta-1}$$

$$= Beta(k+\alpha, n-k+\beta)$$

Multinomial case

- Multinomial likelihood

$$p(X_1 = x_1, X_2 = x_2, \ldots X_k = x_k \mid \vec{\theta}, n) \propto \prod_{i=1..k} \theta_i^{x_i}$$

- What about prior?
  - $p(\vec{\theta})$ should in form $\prod_{i=1}^{k} \theta_i^{???}$
  - Dirichlet Distribution!

$$p(\vec{\theta} \mid \vec{\alpha}) = \frac{1}{\Delta(\vec{\alpha})} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \qquad \Delta(\vec{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} = \text{const.}$$

Statistical Language Model

- Text(sequence of words) representation

$$p(W) = p(w_1, w_2, ..., w_n) = p(w_1)p(w_2 | w_1)p(w_3 | w_2, w_1)\dots p(w_n | w_{n-1}, w_{n-2}...w_1)$$

$$p(\text{I am a student}) = p(\text{I})p(\text{am} | \text{I})p(\text{a} | \text{am, I})p(\text{student} | \text{a, am, I})$$

- Markov property
  - Bi-gram: $p(W) = p(w_1)p(w_2 | w_1)p(w_3 | w_2)\dots p(w_n | w_{n-1})$

    $$p(\text{I am a student}) = p(\text{I})p(\text{am} | \text{I})p(\text{a} | \text{am})p(\text{student} | \text{a})$$

  - Tri-gram, N-gram
- Independence - **Unigram** model
  - $p(W) = p(w_1)p(w_2)\dots p(w_N)$
  - Bag-of-words: $\vec{n} = (n_1, n_2 ... n_V)$

    $$p(W) = p(\vec{n}) = multi(\vec{n} | \vec{p}, N)$$

Topic models - intro

- Care about content: Topic model

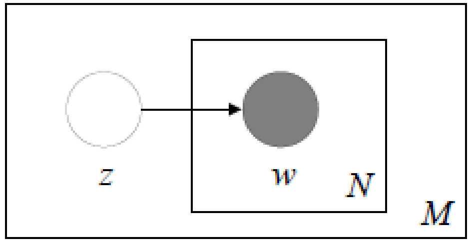
- Variable about topic: $\boldsymbol{z}$

Topic model1 – Mixture of Unigram

- Care about content: Topic models
- Mixture of unigram
  - Each document is generated by first choosing a topic z and then generating N words independently.

  - Mixture of several topics:

$$p(W) = \sum_{z} p(z) \bullet p(W \mid z) = \sum_{z} p(z) \bullet \prod_{i} p(w_i \mid z)$$

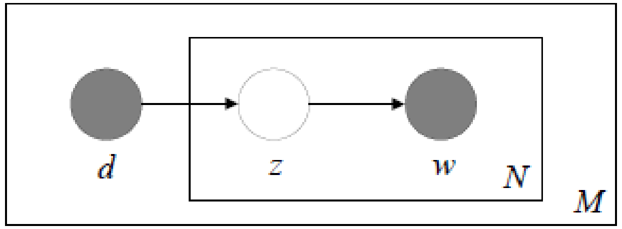


(b) mixture of unigrams

Topic model2 - pLSA

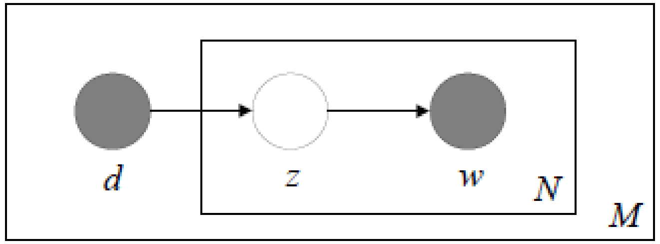
- Probabilistic Latent Sematic Analysis(pLSA)
  - Also, Probabilistic Latent Sematic Indexing(pLSI)
  - K topics, V words, M documents
  - Given document d, $p(w|d) = \prod_i p(w_i|d)$

$$= \prod_i \sum_{z=1}^{K} p(w_i|z)p(z|d)$$



(c) pLSI/aspect model

Topic model2 - pLSA



(c) pLSI/aspect model

- Observable variables: *d*,*w*

$$p(w,d) = \sum_{d_j} p(d_j) \prod_i p(w_i \mid d_j)$$

$$= \sum_{d_j} p(d_j) \prod_i \sum_{z=1}^{K} p(w_i \mid z) p(z \mid d_j)$$

- Drawbacks
  - *d*: a *r.v.*, probability?
  - *(KV+KM)* Parameters – overfit

Latent Dirichlet Allocation (LDA)

Part II

- LDA – intro
- Priors
- Generation procedure
- Bayes' net
- Likelihoods

LDA - a Bayesian version of PLSA

- LDA vs. pLSA:
  - Recall pLSA:

$$p(w|d) = \prod_i \sum_{z=1}^{K} p(w_i|z)\,p(z|d)$$

pLSA: only decided by the topic of a document
LDA: decided by the topic of both document(z) and word itself, ϕ.

pLSA: d is observable
LDA: About Topic – a mixture proportion for topics, θ.

Assumptions of LDA model

- LDA vs. pLSA:
  - LDA :

$$p(w \mid \phi, \theta) = \prod_{i} \sum_{z=1}^{K} p(w_i \mid z, \phi)\, p(z \mid \theta)$$

pLSA: only decided by the topic of a document
LDA: decided by the topic of both document(z) and word itself, ϕ.

pLSA: d is observable
LDA: About Topic – a mixture proportion for topics, θ.

- θ:
  - 文档A = 0.5 娱乐 + 0.3 明星 + 0.2 音乐， θ_A = (0.5, 0.3, 0.2)
  - 文档B = 0.3 娱乐 + 0.1 明星 + 0.5 音乐， θ_B = (0.3, 0.1, 0.5)
- ϕ:
  - 娱乐 = 0.3 “王菊” + 0.2 “周杰伦” + 0.4 “创造101”， $\phi_{娱乐}$ = (0.3, 0.2, 0.4)
  - 明星 = 0.6 “王菊” + 0.4 “周杰伦” + 0.0 “创造101”， $\phi_{明星}$ = (0.6, 0.4, 0.0)
  - 音乐 = 0.1 “王菊” + 0.8 “周杰伦” + 0.1 “创造101”， $\phi_{音乐}$ = (0.1, 0.8, 0.1)

Assumptions of LDA model

- LDA vs. pLSA:
  - LDA :

$$p(w|\phi,\theta)=\prod_{i}\sum_{z=1}^{K}p(w_i|z,\phi)\,p(z|\theta)$$

pLSA: only decided by the topic of a document
LDA: decided by the topic of both document(z) and word itself, ϕ.

A multinomial distribution for word w_i, when given z,

pLSA: d is observable
LDA: About Topic – a mixture proportion for topics, θ.

A multinomial distribution

- Use Dirichlet distribution for the priors of θ, ϕ

Priors of LDA

- Use Dirichlet distribution for the priors of θ, ϕ
  - Θ : a topic mixture proportion for document, M×K matrix

$$\theta_m = (\theta_{m1}, \theta_{m2} \ldots \theta_{mK}) \qquad \theta_1, \theta_2 \ldots \theta_M \overset{iid}{\sim} Dir(\vec{\alpha}), \alpha \in \mathbb{R}^K$$

  - Φ : a word mixture proportion for topic, K×V matrix

$$\phi_k = (\phi_{k1}, \phi_{k2} \ldots \phi_{kN}) \qquad \phi_1, \phi_2 \ldots \phi_K \overset{iid}{\sim} Dir(\vec{\beta}), \beta \in \mathbb{R}^V$$

LDA – a generative model

- Generation procedure
  - Topic level ⟶ word level ⟵ document level

LDA generation procedure

***Topic level*:**

Sample topic $\phi_k \sim Dir(\beta), \beta \in \mathbb{R}^V, k = 1,2\ldots K$

***Document level*:**

for each document m=1,2…M:

Sample mixture proportion $\theta_m \sim Dir(\alpha), \alpha \in \mathbb{R}^K$

Sample document length $N_m \sim poisson(\xi)$

***Word level*:**

For word $n = 1\ldots N_m$:

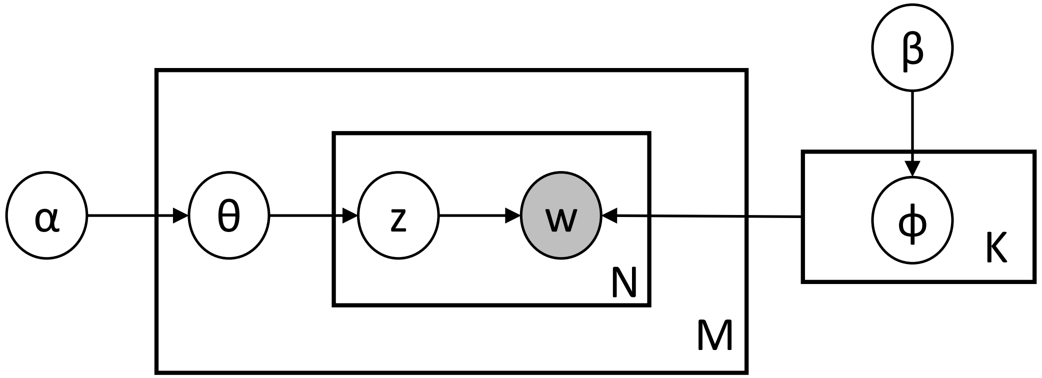
Sample topic index $Z_{m,n} \sim Multi(\theta_m)$

Sample term for word $W_{m,n} \sim Multi(\phi_{Z_{m,n}})$

Related to both ϕ and *Zm,n*

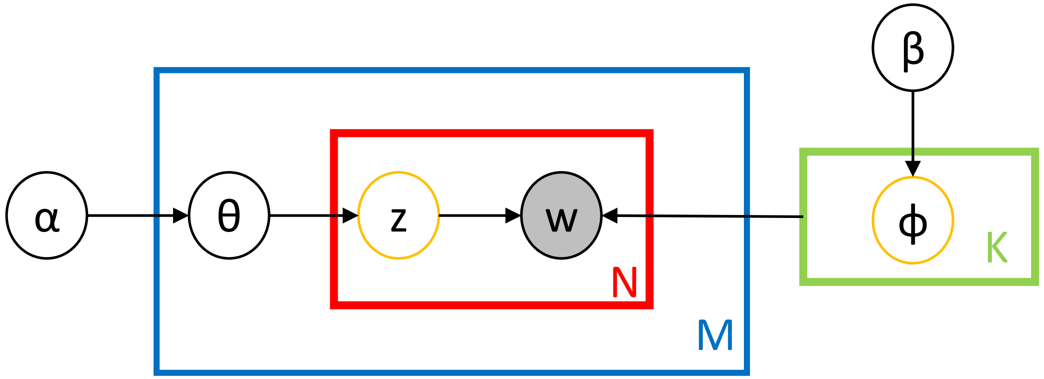
LDA – a generative model

- Bayes' net representation



LDA – Likelihood

- For document m, joint distribution:

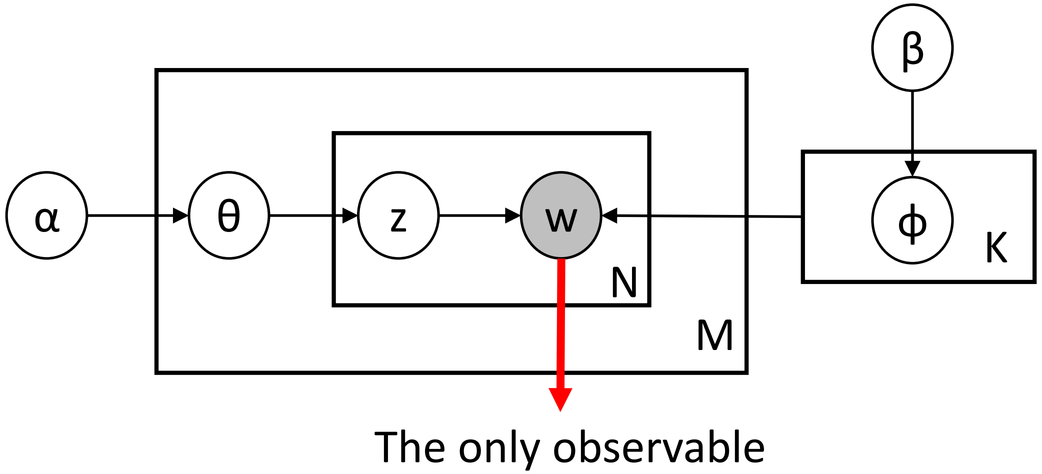


Related to both ϕ and word-topic *Zm,n*

$$p(\vec{w}_m, \vec{Z}_m, \vec{\theta}_m, \Phi \mid \alpha, \beta) = \underbrace{\underbrace{\prod_{n=1}^{N_m} p(w_{m,n} \mid \phi_{Z_{m,n}})\, p(Z_{m,n} \mid \vec{\theta}_m)}_{\text{Word level}} \times p(\vec{\theta}_m \mid \alpha)}_{\text{Document level}} \times \underbrace{p(\Phi \mid \beta)}_{\text{Topic level}}$$

LDA – Likelihood

- Likelihood for documents



$$p(\overrightarrow{w_m} \mid \alpha, \beta) = \iint p(\overrightarrow{\theta_m} \mid \alpha) \times p(\Phi \mid \beta) \bullet \prod_{n=1}^{N_m} \sum_{z_{m,n}=1}^{K} p(w_{m,n} \mid \phi_{z_{m,n}}) p(z_{m,n} \mid \overrightarrow{\theta_m}) \, \mathrm{d}\overrightarrow{\theta_m} \, \mathrm{d}\Phi$$

$$p(W \mid \alpha, \beta) = \prod_{m=1}^{M} p(\overrightarrow{w_m} \mid \alpha, \beta)$$

Parameter inference

Part III

- Parameters –intro
- Method 1:
  Variational inference
- Method 2:
  Via Gibbs sampling

Parameters inference - intro

- Parameters we want to know:
  - θ: a topic mixture proportion for document
  - ϕ: a word mixture proportion for topic

- Maximize a posterior (MAP):

$$p(\underbrace{Z,\theta,\Phi}_{\text{All the hiddens}} \mid \alpha,\beta,W) = \frac{p(W,Z,\theta,\Phi \mid \alpha,\beta)}{p(W \mid \alpha,\beta)}$$

  - Intractable ☹

Method 1: Variational inference

- Idea: find parameters Λ of a family of distributions *q(.| Λ)* on the latent variables (usually exponential family)
  - $$\Lambda^* = \arg\min_{\Lambda} D_{KL}(q(\theta, z, \phi \mid \Lambda) \,||\, p(\theta, z, \phi \mid w, \alpha, \beta))$$
  - It can be proved that

$$\Lambda^* = \arg\max_{\Lambda} E_q[\log p(\theta, z, \phi, W \mid \alpha, \beta) - \log q(\theta, z, \phi \mid \Lambda)]$$

  - More detailed: *Blei et.al.(2003)*

Method 2: Via Gibbs sampling

- Idea: Consider $p(\vec{z} \mid \vec{w})$, and adjust word-count/topic counts etc. according to $\vec{z}$ follows $p(\vec{z} \mid \vec{w})$.

- Steps (general):
  - Initialization
  - Gibbs sampling and adjust word-counts/topic-counts etc.
  - Compute $\hat{\theta}$ and $\hat{\phi}$ according to hyperparameters and word-counts/topic-counts

Gibbs sampling algorithm for LDA

Initialize

⊡ initialisation
zero all count variables, $n_m^{(k)}, n_m, n_k^{(t)}, n_k$
for all documents $m \in [1, M]$ **do**
  for all words $n \in [1, N_m]$ in document m **do**
    sample topic index $z_{m,n}=k \sim \text{Mult}(1/K)$
    increment document–topic count: $n_m^{(k)} + 1$
    increment document–topic sum: $n_m + 1$
    increment topic–term count: $n_k^{(t)} + 1$
    increment topic–term sum: $n_k + 1$
  end for
end for

- $n_m^{(k)}$ - word count for topic k in document m
- $n_k^{(t)}$ - word count for word t in topic k

Sample and adjust

⊡ Gibbs sampling over burn-in period and sampling period
while not finished **do**
  for all documents $m \in [1, M]$ **do**
    for all words $n \in [1, N_m]$ in document m **do**
      ⊡ for the current assignment of k to a term t for word $w_{m,n}$:
      decrement counts and sums: $n_m^{(k)} - 1; n_m - 1; n_k^{(t)} - 1; n_k - 1$
      ⊡ multinomial sampling acc. to Eq. 79 (decrements from previous step):
      sample topic index $\tilde{k} \sim p(z_i | \vec{z}_{\neg i}, \vec{w})$
      ⊡ use the new assignment of $z_{m,n}$ to the term t for word $w_{m,n}$ to:
      increment counts and sums: $n_m^{(\tilde{k})} + 1; n_m + 1; n_{\tilde{k}}^{(t)} + 1; n_{\tilde{k}} + 1$
    end for
  end for

Adjust $n_m^{(k)}$, $n_k^{(t)}$ etc. according to $\tilde{k}$

Estimate parameters

  ⊡ check convergence and read out parameters
  if converged and L sampling iterations since last read out **then**
    ⊡ the different parameters read outs are averaged.
    read out parameter set $\underline{\Phi}$ according to Eq. 82
    read out parameter set $\underline{\Theta}$ according to Eq. 83
  end if
end while

Algo: *Gregor(2006)*

Gibbs sampling algorithm for LDA

- Joint $$p(\vec{w},\vec{z}\,|\,\vec{\alpha},\vec{\beta})=\prod_{k=1}^{K}\frac{\Delta(\vec{n_k}+\vec{\beta})}{\Delta(\vec{\beta})}\bullet\prod_{m=1}^{M}\frac{\Delta(\vec{n_m}+\vec{\alpha})}{\Delta(\vec{\alpha})}$$

- Conditional (eq.79):

$$P(z_i=k\,|\,\vec{z_{-i}},\vec{w})\propto\frac{n_{k,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,-i}^{(t)}+\beta_t}\bullet\frac{n_{m,-i}^{(k)}+\alpha_k}{[\sum_{k=1}^{K}n_{m,-i}^{(k)}+\alpha_k]-1}$$

- Estimation (eq.82,83):

$$\varphi_{kt}=\frac{n_{k,-i}^{(t)}+\beta_t}{\sum_{t=1}^{V}n_{k,-i}^{(t)}+\beta_t}$$

$$\theta_{mk}=\frac{n_{m,-i}^{(k)}+\alpha_k}{\sum_{k=1}^{K}n_{m,-i}^{(k)}+\alpha_k}$$

- Detailed derivation: *Gregor(2006)*

Model evaluation

Part IV

- Perplexity
- Example: Perplexity for N-gram model
- Perplexity for LDA

Perplexity – how good is our model

- Definition:
  - In NLP, perplexity is a measure of the ability of a model to generalize to unseen data (test set).
    - For a sentence:

$$Perplexity(W) = p(w_1 w_2 \ldots w_N)^{-\frac{1}{N}}$$

    - For a corpus:

$$Perplexity(D_{test}) = \exp\left(- \frac{\sum_{d=1}^{|D_{test}|} \log P(\overrightarrow{w_d})}{\sum_{d=1}^{|D_{test}|} N_d} \right)$$

Example: Perplexity in N-gram model

- Perplexity: The lower the better.
  - Training 38 million words, test 1.5 million words, Wall Street Journal *(Stanford cs124)*

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

Unigram

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

Bigram

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Trigram

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Perplexity for LDA

- LDA (for given topic number = K)

$$P(\overrightarrow{w_d}) = \prod_{n=1}^{N_d}\left(\sum_{k=1}^{K} P(w_n \mid z_n = k) \bullet P(z_n = k \mid d)\right)$$

$$= \prod_{t=1}^{V}\left(\sum_{k=1}^{K} \theta_{d,k} \bullet \varphi_{k,t}\right)^{n_d^{(t)}}$$

$$\log P(\overrightarrow{w_d}) = \sum_{t=1}^{V} n_d^{(t)} \log\left(\sum_{k=1}^{K} \theta_{d,k} \bullet \varphi_{k,t}\right)$$

Via Gibbs

Not related to test corpus

- Topic number chosen & perplexity
    - Usually choose K = argmin perplexity(D_{test})

Applications & extensions

Part V

- Document modeling
- Document classification
- Collaborative filtering
- Extensions

Document modeling

- (Blei, 2003) on AP corpus: 16,333 newswire articles

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Document modeling

- (Blei, 2003) on AP corpus: 16,333 newswire articles

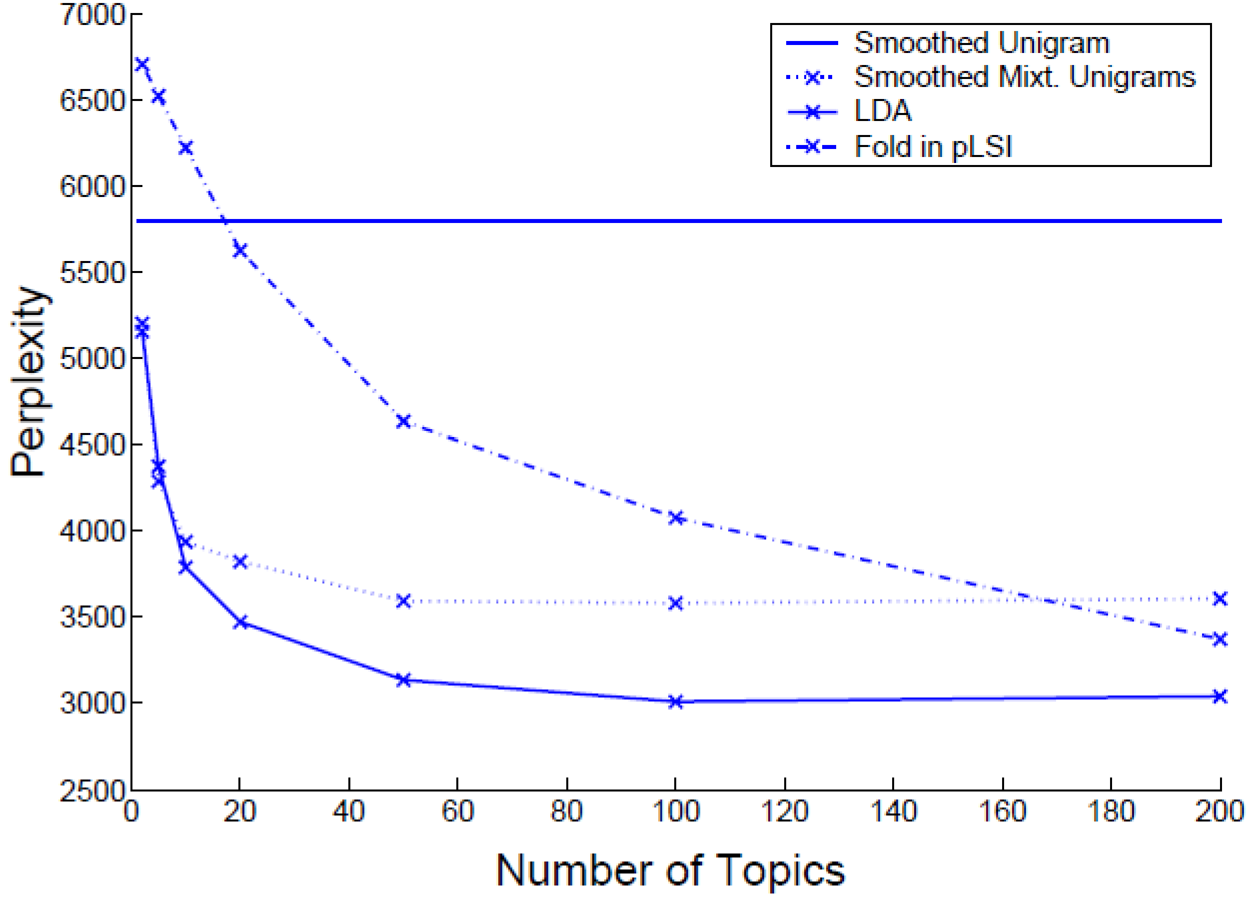
“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Document modeling

- Perplexity vs topic number (Blei, 2003)



Document classification

- binary classification, Reuters-21578 dataset (Blei, 2003)
  - 8000 documents
  - Low-dimensional representations from LDA(50 topic) for document vs. word features
  - When proportion of training data is low…

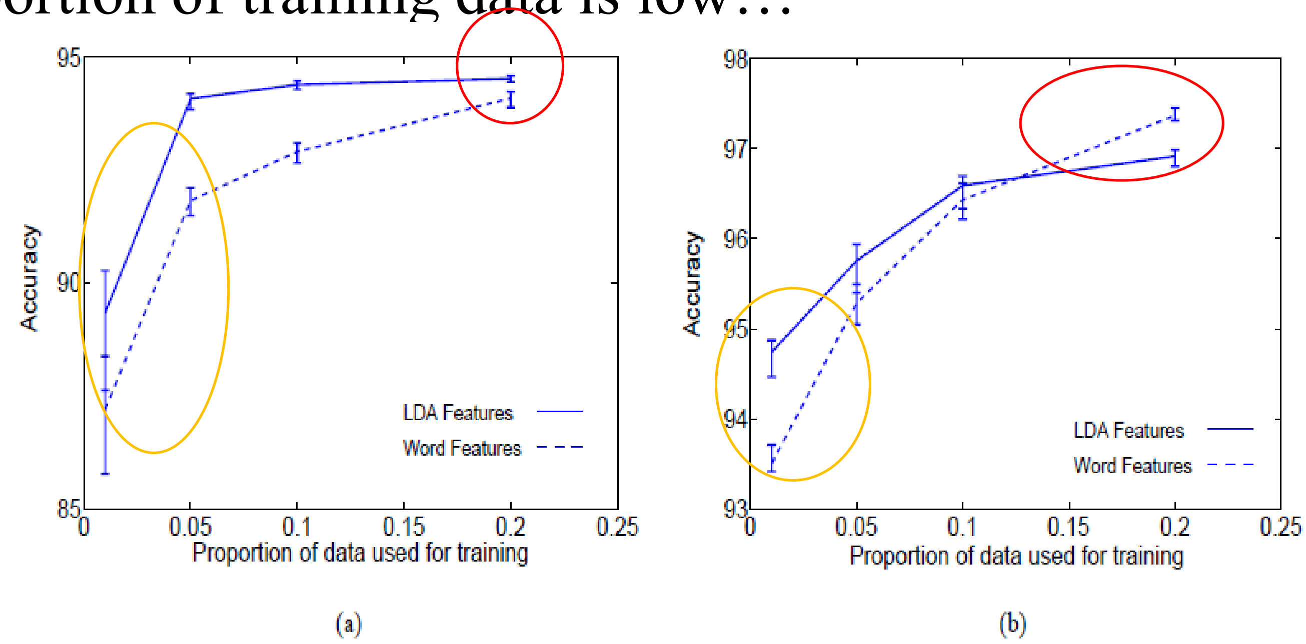
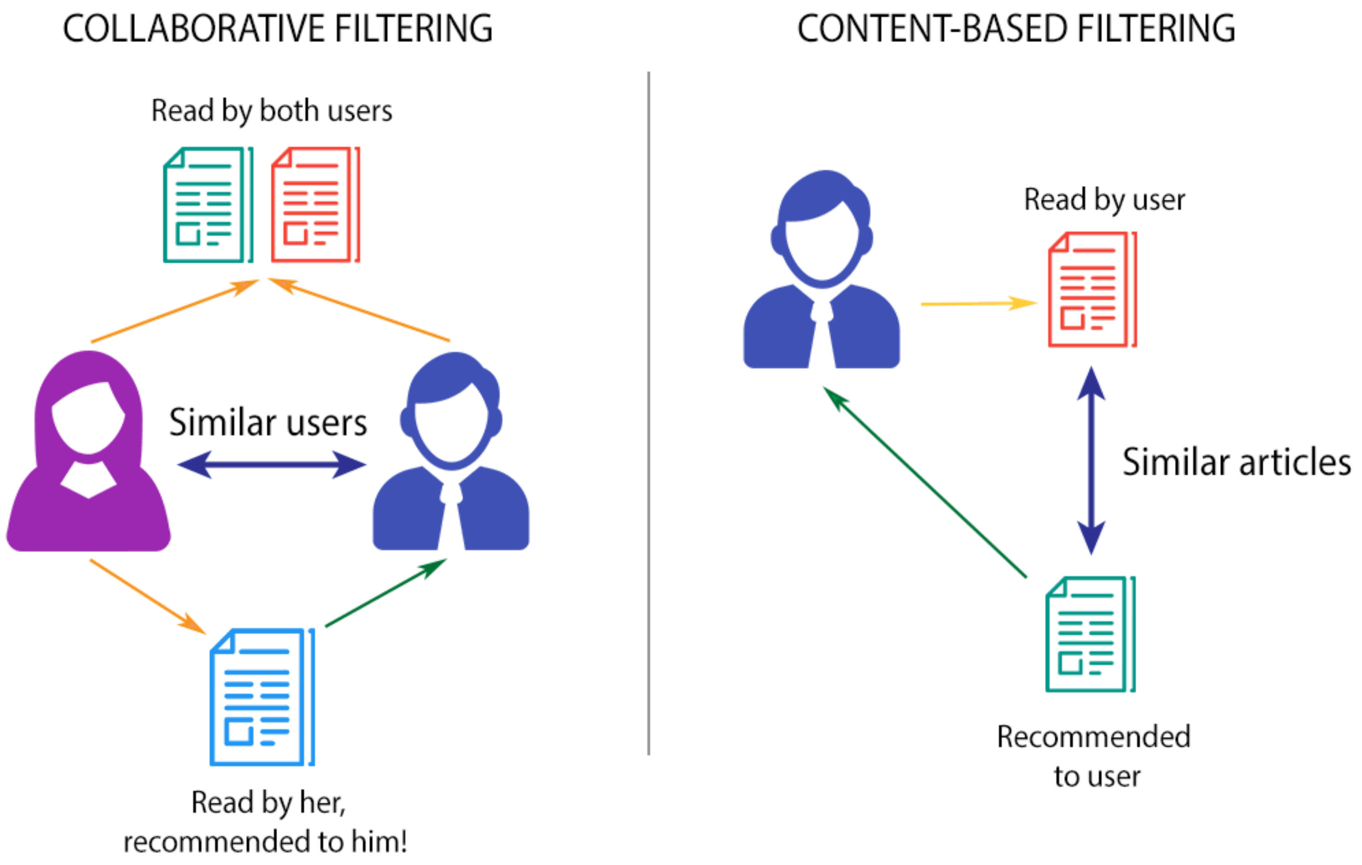


Figure 10: Classification results on two binary classification problems from the Reuters-21578 dataset for different proportions of training data. Graph (a) is EARN vs. NOT EARN. Graph (b) is GRAIN vs. NOT GRAIN.

Collaborative filtering

- Similarity → preference



Collaborative filtering

- User similarity → preference
- Dataset: EachMovie (user: preferred movie choices)
- Analogy:
  - User – document
  - Movie chosen – words in document
  - 1600 movies – vocabulary
  - Movie genre – topic
    - User – genre θ:

      Eg. Mary preference = 0.5 Romance + 0.4 comedy + 0.1 sci-fi
    - Movie–genre ϕ :

      Eg. Sci-fi = 0.1 *Ready Player One* + 0.2 *Marvel's The Avengers* + 0.3 *Matrix* + ...

Collaborative filtering

- 3300 training users; 390 testing users
- Predictive-Perplexity:

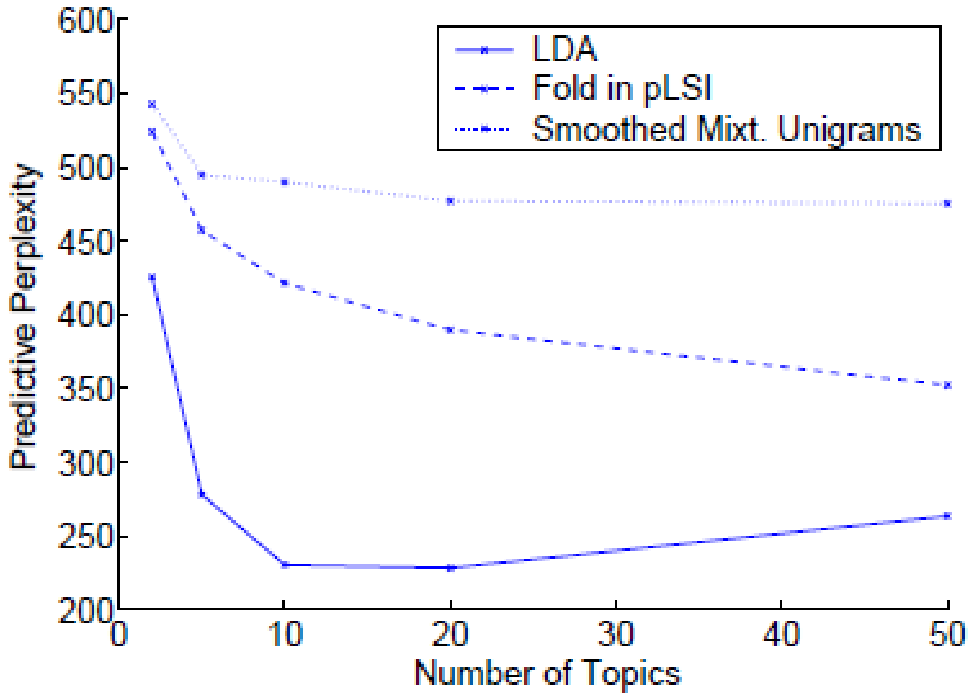


Figure 11: Results for collaborative filtering on the EachMovie data.

Extensions

- Drawbacks of LDA model
  - Short text topics?
  - Noisy data?
  - …

- Some extension models:
  - TwitterLDA (Zhao et.al, 2011):
    - Assumption: one topic for one tweets.
  - Labeled-LDA(Ramage et.al 2009, ACL):
    - Used in document classification
  - Hierarchically LDA(Teh et.al.2005):
    - Find K automatically
  - …

Reference

- Blei D M, Ng A Y, Jordan M I. Latent dirichlet allocation[J]. Journal of machine Learning research, 2003, 3(Jan): 993-1022.
- Heinrich G. Parameter Estimation for Text Analysis[J]. Technical Report, 2008.
- Ramage D, Hall D, Nallapati R, et al. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora[C]//Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1. Association for Computational Linguistics, 2009: 248-256.
- Teh Y W, Jordan M I, Beal M J, et al. Sharing clusters among related groups: Hierarchical Dirichlet processes[C]//Advances in neural information processing systems. 2005: 1385-1392.
- Zhao W X, Jiang J, Weng J, et al. Comparing twitter and traditional media using topic models[C]//European Conference on Information Retrieval. Springer, Berlin, Heidelberg, 2011: 338-349.