

# Regularized Matrix Factorization for Topic Modeling of Text Collections

Konstantin Vorontsov

Oleksandr Frei • Murat Apishev  
Andrey Shadrikov • Alexander Plavin

(MIPT, MSU, CC RAS, Yandex • Moscow, Russia)

4th International Conference on Matrix Methods in  
Mathematics and Applications MMMA-2015

Moscow • 24–28 August 2015

- 1 Probabilistic Topic Modeling**
  - Approximate stochastic matrix factorization
  - Basic topic models PLSA and LDA
  - The paradigm of Exploratory Search
- 2 ARTM: Additive Regularization of Topic Models**
  - Additive regularization and modalities
  - Regularization examples
  - BigARTM open source project
- 3 Experiments**
  - Time and memory performance
  - Initialization
  - Choosing the number of topics

## Sparse stochastic matrix factorization under KL-loss

Given a matrix  $Z = \|z_{ij}\|_{n \times m}$ ,  $(i, j) \in \Omega \subseteq \{1..n\} \times \{1..m\}$

Find matrices  $X = \|x_{it}\|_{n \times k}$  and  $Y = \|y_{tj}\|_{k \times m}$  such that

$$\|Z - XY\|_{\Omega, d} = \sum_{(i,j) \in \Omega} d\left(z_{ij}, \sum_t x_{it} y_{tj}\right) \rightarrow \min_{X, Y}$$

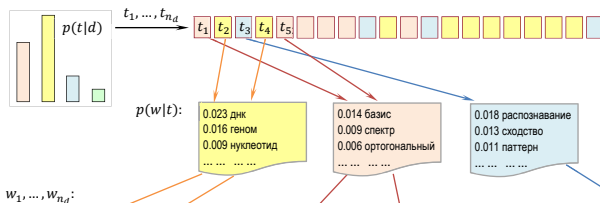
Variants of the problem:

- quadratic loss:  $d(z, \hat{z}) = (z - \hat{z})^2$
- Kullback–Leibler loss:  $d(z, \hat{z}) = z \ln(z/\hat{z}) - z + \hat{z}$
- nonnegative matrix factorization:  $x_{it} \geq 0, y_{tj} \geq 0$
- stochastic matrix factorization:  $x_{it} \geq 0, y_{tj} \geq 0, \sum_i x_{it} = 1, \sum_t y_{tj} = 1$
- sparse input data:  $|\Omega| \ll nm$
- sparse output factorization  $X, Y$

# Probabilistic Topic Model (PTM) generating a text collection

Topic model explains terms  $w$  in documents  $d$  by topics  $t$ :

$$p(w|d) = \sum_t p(w|t)p(t|d)$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в геномных последовательностях. Метод основан на разномасштабном оценивании сходства нуклеотидных последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим ортогональным базисам. Найдены условия оптимальной аппроксимации, обеспечивающие автоматическое распознавание повторов различных видов (прямых и инвертированных, а также тандемных) на спектральной матрице сходства. Метод одинаково хорошо работает на разных масштабах данных. Он позволяет выявлять следы сегментных дубликаций и мегасателлитные участки в геноме, районы синтении при сравнении пары геномов. Его можно использовать для детального изучения фрагментов хромосом (поиска размытых участков с умеренной длиной повторяющегося паттерна).

Inverse problem: text collection  $\rightarrow$  PTM

**Given:**  $D$  is a set (collection) of documents

$W$  is a set (vocabulary) of terms

$n_{dw}$  = how many times term  $w$  appears in document  $d$

**Find:** parameters  $\phi_{wt} = p(w|t)$ ,  $\theta_{td} = p(t|d)$  of the topic model

$$p(w|d) = \sum_t \phi_{wt} \theta_{td}.$$

**The problem** of log-likelihood maximization under constraints:

$$\mathcal{L}(\Phi, \Theta) = \sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

$$\phi_{wt} \geq 0, \quad \sum_{w \in W} \phi_{wt} = 1; \quad \theta_{td} \geq 0, \quad \sum_{t \in T} \theta_{td} = 1.$$

Hofmann T. Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.

## EM-algorithm for likelihood maximization [Hofmann, 1999]

From KKT conditions for the constrained maximization problem

### Theorem

Maximum of  $\mathcal{L}(\Phi, \Theta)$  satisfies the system of equations with model parameters  $\phi_{wt}$ ,  $\theta_{td}$  and auxiliary variables  $p_{tdw}$ ,  $n_{wt}$ ,  $n_{td}$ :

$$\begin{cases} \text{E-step:} & p_{tdw} = \frac{\phi_{wt}\theta_{td}}{\sum_{t'} \phi_{wt'}\theta_{t'd}}; \\ \text{M-step:} & \begin{cases} \phi_{wt} = \frac{n_{wt}}{\sum_{w'} n_{w't}}; & n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}; \\ \theta_{td} = \frac{n_{td}}{\sum_{t'} n_{t'd}}; & n_{td} = \sum_{w \in D} n_{dw} p_{tdw}; \end{cases} \end{cases}$$

EM-algorithm alternates E-step and M-step until convergence.  
EM-algorithm is equivalent to a simple iteration method.

## LDA — Latent Dirichlet Allocation [Blei, 2003]

**Assumption.** Column vectors  $\phi_t = (\phi_{wt})_{w \in W}$  and  $\theta_d = (\theta_{td})_{t \in T}$  are generated from Dirichlet distributions,  $\alpha \in \mathbb{R}^{|T|}$ ,  $\beta \in \mathbb{R}^{|W|}$ :

$$\text{Dir}(\phi_t | \beta) = \frac{\Gamma(\beta_0)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{wt}^{\beta_w - 1}, \quad \beta_0 = \sum_w \beta_w, \quad \beta_t \geq 0;$$

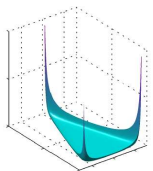
$$\text{Dir}(\theta_d | \alpha) = \frac{\Gamma(\alpha_0)}{\prod_t \Gamma(\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}, \quad \alpha_0 = \sum_t \alpha_t, \quad \alpha_t \geq 0;$$

**Example:**

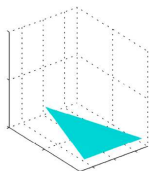
$$\text{Dir}(\theta | \alpha)$$

$$|T| = 3$$

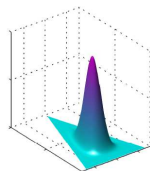
$$\theta, \alpha \in \mathbb{R}^3$$



$$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$$



$$\alpha_1 = \alpha_2 = \alpha_3 = 1$$



$$\alpha_1 = \alpha_2 = \alpha_3 = 10$$

## The main difference between LDA and PLSA

The estimates of conditionals  $\phi_{wt} \equiv p(w|t)$ ,  $\theta_{td} \equiv p(t|d)$ :

- in PLSA — unbiased maximum likelihood estimates:

$$\phi_{wt} = \frac{n_{wt}}{n_t}, \quad \theta_{td} = \frac{n_{td}}{n_d}$$

- in LDA — smoothed Bayesian estimates:

$$\phi_{wt} = \frac{n_{wt} + \beta_w}{n_t + \beta_0}, \quad \theta_{td} = \frac{n_{td} + \alpha_t}{n_d + \alpha_0}.$$

The difference is significant for small  $n_{wt}$ ,  $n_{td}$  only.

Robust LDA and robust PLSA produce almost identical models.

*Asuncion A., Welling M., Smyth P., Teh Y. W.* On smoothing and inference for topic models. Int'l Conf. on Uncertainty in Artificial Intelligence, 2009.

*Potapenko A. A., Vorontsov K. V.* Robust PLSA Performs Better Than LDA. ECIR-2013, Moscow, Russia, 24-27 March 2013. LNCS, Springer. Pp. 784–787.

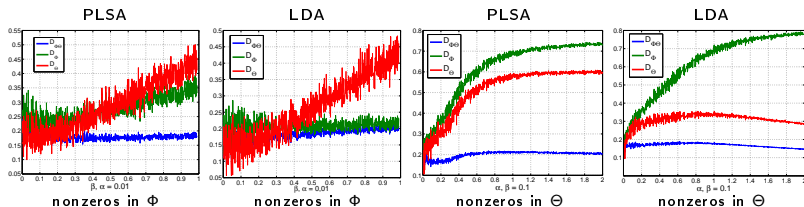


## Topic Modeling as an ill-posed inverse problem

The *nonuniqueness* and *instability* of matrix factorization:  
 $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$  for all  $S$  such that  $\Phi', \Theta'$  are stochastic.

**Experiment:** recovering known  $\Phi, \Theta$  on synthetic dataset,  
 $|D| = 500$ ,  $|W| = 1000$ ,  $|T| = 30$ .

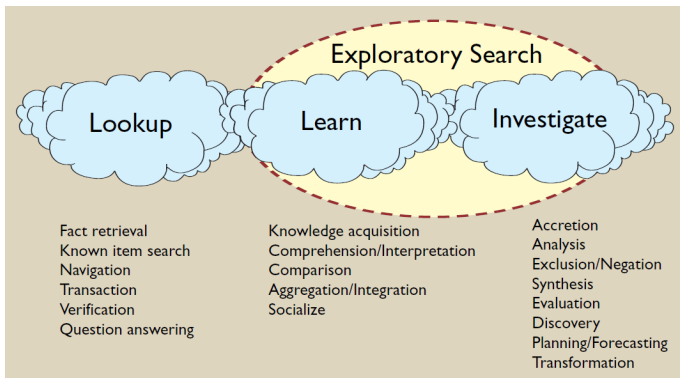
**Result:** product  $\Phi\Theta$  is always recovered well, however  
 matrix  $\Phi$  and matrix  $\Theta$  are recovered if being highly sparse only:



**Conclusions:** Dirichlet prior is too weak as a regularizer;  
 stronger regularization is needed to ensure a stable solution.

# Exploratory Search for learning, knowledge acquisition and discovery

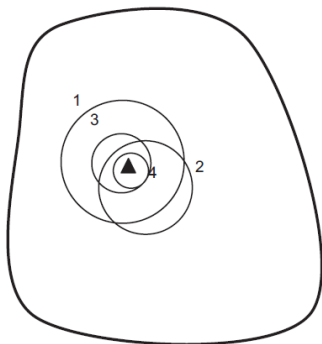
- what if the user doesn't know which keywords to use?
- what if the user isn't looking for a single answer?



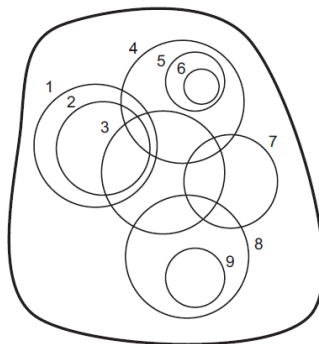
*Gary Marchionini*. Exploratory Search: from finding to understanding. Communications of the ACM. 2006, 49(4), p. 41–46.

# Iterative “query-browse-refine” search vs Exploratory Search

## Iterative Search



## Exploratory Search



▲ Search target      ◻ Information space

○ Result sets (larger = more results, intersection = overlap, # = iteration)

*R.W.White, R.A.Roth. Exploratory Search: beyond the Query-Response paradigm. San Rafael, CA: Morgan and Claypool, 2009.*

## Exploratory search scenario

### Search query:

- a document of any length or even a set of documents

### Search intents:

- what topics does it contain?
- what else is known on these topics?
- what is the structure of this domain area?
- what is most important, useful, popular, recent here?

### Search scenario:

- 1 given a text (of any length) at hand (in any application)
- 2 identify topics and sub-topics it contains
- 3 show textual and graphical representations of these topics

## Exploratory search: the prototype of graphical user interface

Color topic bar is a starting GUI element for exploratory search

The screenshot shows the BigARTM web interface in a browser window. The browser address bar shows 'www.machinelearning.us'. The page title is 'BigARTM'. The main content area contains text about the library and its features, including a 'Теоретическое введение' (Theoretical Introduction) section. On the right side of the page, there is a vertical color bar with a gradient from red at the top to blue at the bottom, representing different topics. The interface also includes a navigation menu on the left and a search bar at the top right.

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.  
ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выделенных тематических документов. Тематическая модель включает в себя два дискретных распределения на множестве термов: каждый документ — дискретное распределение на множестве термов, тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантизации текстов.

Тематическая модель — это представление наблюдаемого условного распределения  $p(\mathbf{w}|\mathcal{D})$  термов (слов или словосочетаний)  $\mathbf{w}$  в документе  $\mathcal{D}$ :

$$p(\mathbf{w}|\mathcal{D}) = \sum_{\mathcal{T} \in \mathcal{T}} p(\mathbf{w}|\mathcal{T})p(\mathcal{T}|\mathcal{D}),$$

где  $\mathcal{T}$  — множество тем.

$\phi_{\mathcal{w}|\mathcal{T}} = p(\mathbf{w}|\mathcal{T})$  — известное распределение термов в теме  $\mathcal{T}$ ;  
 $\theta_{\mathcal{T}|\mathcal{D}} = p(\mathcal{T}|\mathcal{D})$  — известное распределение тем в документе  $\mathcal{D}$ .

Параметры тематической модели — матрицы  $\Phi = (\phi_{\mathcal{w}|\mathcal{T}})$  и  $\Theta = (\theta_{\mathcal{T}|\mathcal{D}})$  — являются ключевым решением задачи максимизации правдоподобия

$$\sum_{\mathcal{D} \in \mathcal{D}} \sum_{\mathbf{w} \in \mathcal{W}} n_{\mathcal{D},\mathbf{w}} \log \sum_{\mathcal{T} \in \mathcal{T}} \phi_{\mathcal{w}|\mathcal{T}} \theta_{\mathcal{T}|\mathcal{D}} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормированности

## Exploratory search: the prototype of graphical user interface

Click on the **color topic bar** is a topic query

The screenshot shows the BigARTM web interface. The browser address bar displays "www.machinelearning.ru". The page title is "BigARTM". The main content area contains the following text:

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределённая реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выделения тематик коллекций документов. Тематическая модель включает в себя два дискретных распределения на множестве термов, каждый документ — дискретное распределение на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантизации текстов.

Тематическая модель — это представление наблюдаемого условного распределения  $p(w|d)$  термов (слов или словосочетаний)  $w$  в документе  $d$  коллекции  $D$ :

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d),$$

где  $T$  — множество тем;

$\phi_{wt} = p(w|t)$  — известное распределение термов в теме  $t$ ;

$\theta_{td} = p(t|d)$  — известное распределение тем в документе  $d$ .

Параметры тематической модели — матрицы  $\Phi = (\phi_{wt})$  и  $\Theta = (\theta_{td})$  — находят путь решения задачи максимизации правдоподобия

$$\sum_{d \in D} \sum_{w \in V} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях нормировки и неотрицательности

On the right side of the page, there is a vertical "color topic bar" with a rainbow gradient. Below it is a vertical toolbar with several icons, including a document, a list, a pie chart, a map, a bar chart, and a summation symbol.



# Exploratory search: the prototype of graphical user interface

## Topics of the query document

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная статья: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для вычлечения тематик коллекций документов. Тематическая модель включает в себя функцию распределения на множестве термиче, каждый документ — дисперсионное распределение на множестве термиче, используется для информационного поиска, классификации, категоризации, аннотирования, семантики текстов.

Тематическая модель — это представление наблюдаемого условного распределения  $p(w|d)$  термиче (слов или словосочетаний)  $w$  в документе  $d$  коллекции  $D$ :

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d),$$

где  $T$  — множество термиче,  
 $\phi_{wt} = p(w|t)$  — известное распределение термиче в теме  $t$ ;  
 $\theta_{dt} = p(t|d)$  — известное распределение тем в документе  $d$ .

Параметры тематической модели — матрица  $\Phi = (\phi_{wt})$  и  $\Theta = (\theta_{dt})$  — находят путем решения задачи максимизации правдоподобия

$$\sum_{d \in D} \sum_{w \in V} n_{dw} \log \sum_{t \in T} \phi_{wt} \theta_{dt} \rightarrow \max_{\Phi, \Theta},$$

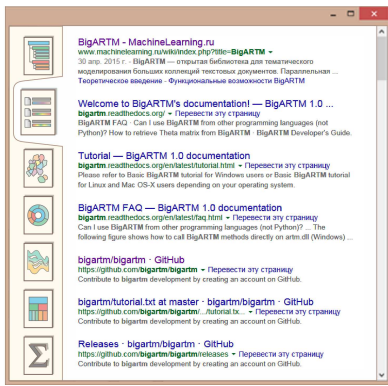
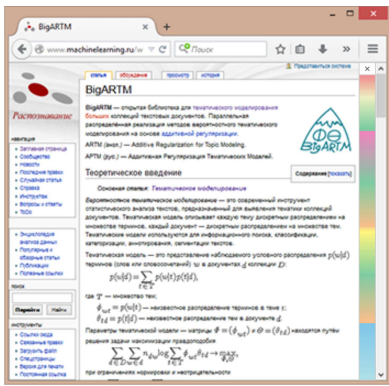
при ограничениях неотрицательности и нормированности

Topics in «BigARTM» [English] [Russian]

- Natural language processing
  - Statistical text analysis
    - Probabilistic topic modeling
- Probability theory
  - Likelihood maximization
- Mathematical programming
  - Nonconvex optimization
    - Constrained nonconvex optimization
- Machine Learning
  - Topic Modeling
    - Probabilistic Topic Modeling
- Matrix Factorization
  - Nonnegative Matrix Factorization
    - Probabilistic Topic Modeling
- Parallel computing
- Big Data

# Exploratory search: the prototype of graphical user interface

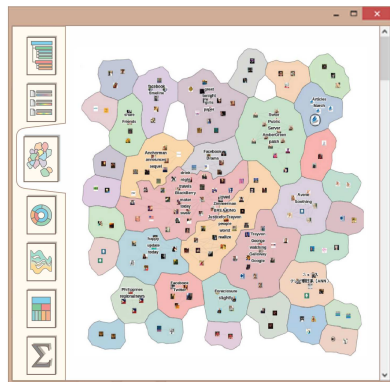
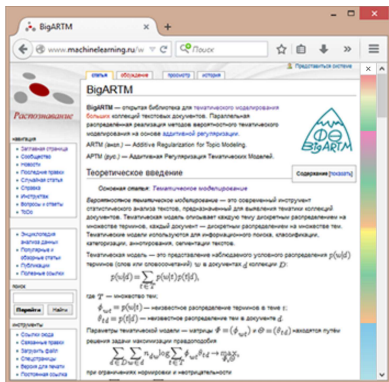
## Similar documents and objects ranked by relevance





# Exploratory search: the prototype of graphical user interface

## Topic roadmap: clustering of relevant documents



## Exploratory search: the prototype of graphical user interface

## Topic hierarchy: topical structure of the domain area

BigARTM

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная идея: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель включает в себя группу дискретных распределений на множестве термов, каждый документ — дискретное распределение на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантизации текстов.

Тематическая модель — это представление наблюдения условного распределения  $p(w|d)$  термов (слов или словосочетаний)  $w$  в документе  $d$  коллекции  $D$ :

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d),$$

где  $T$  — множество тем;

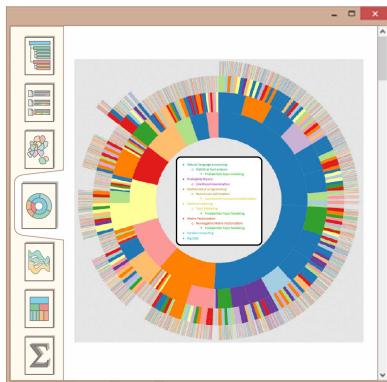
$\phi_{wt} = p(w|t)$  — известное распределение термов в теме  $t$ ;

$\theta_{dt} = p(t|d)$  — известное распределение тем в документе  $d$ .

Параметры тематической модели — матрицы  $\Phi = (\phi_{wt})$  и  $\Theta = (\theta_{dt})$  — находят путем решения задачи максимизации правдоподобия

$$\sum_{d \in D} \sum_{w \in V} p_w \log \sum_{t \in T} \phi_{wt} \theta_{dt} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях неотрицательности и нормированности



## Exploratory search: the prototype of graphical user interface

## Topic river: evolution of the domain area

BigARTM

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределённая реализация метода вероятностного тематического моделирования на основе аддитивной регуляризации.

ARTM (англ.) — Additive Regularization for Topic Modeling.

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

Теоретическое введение

Основная идея: Тематическое моделирование

Вероятностное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для вычисления тематик коллекций документов. Тематическая модель включает в себя два дискретных распределения на множестве терминов, каждый документ — дискретный распределение на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, семантики текстов.

Тематическая модель — это представление наблюдаемого условного распределения  $p(w|d)$  термов (или их эмбедингов)  $w$  в документе  $d$  коллекции  $D$ :

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d),$$

где  $T$  — множество тем;

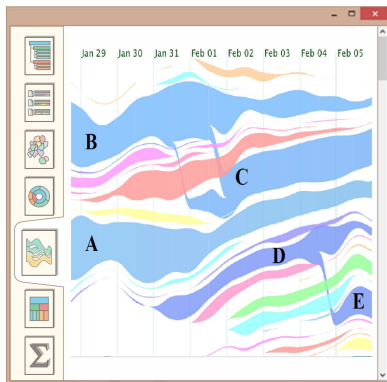
$\phi_{wt} = p(w|t)$  — известное распределение термов в теме  $t$ ;

$\theta_{td} = p(t|d)$  — известное распределение тем в документе  $d$ .

Параметры тематической модели — матрицы  $\Phi = (\phi_{wt})$  и  $\Theta = (\theta_{td})$  — находят путем решения задачи максимизации правдоподобия

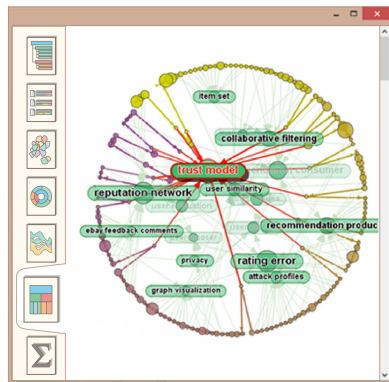
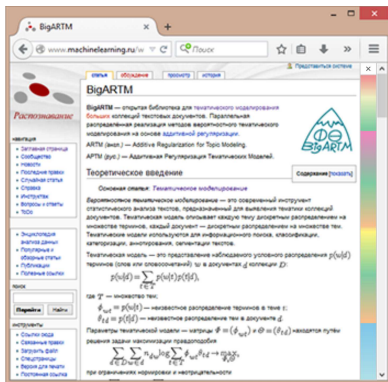
$$\sum_{d \in D} \sum_{w \in V} p_w \log \sum_{t \in T} \phi_{wt} \theta_{td} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях нормировки и неотрицательности



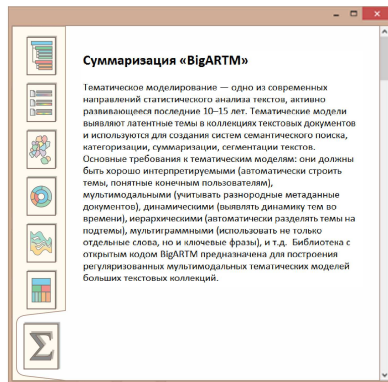
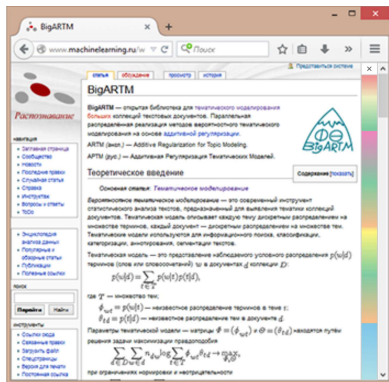
# Exploratory search: the prototype of graphical user interface

## Topic bar: segmentation of the query document



# Exploratory search: the prototype of graphical user interface

## Summarization of the query document



<http://textvis.lnu.se>

## A visual survey of 170 text visualization techniques



## The elements of Exploratory Search technology

- ① Web crawling ..... ready-made solutions
- ② Content filtering ..... ready-made solutions
- ③ **Topic modeling** ..... **ongoing research**
- ④ Building the inverted index ..... ready-made solutions
- ⑤ Ranking ..... ready-made solutions
- ⑥ Visualization ..... ready-made solutions

## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable:** each topic should be well interpretable by humans and labeled automatically



## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable**: each topic should be well interpretable by humans and labeled automatically
- 2 **Multigram**: keyphrases should be extracted automatically

## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable:** each topic should be well interpretable by humans and labeled automatically
- 2 **Multigram:** keyphrases should be extracted automatically
- 3 **Multilingual:** cross-language and multi-language search should be supported

## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable**: each topic should be well interpretable by humans and labeled automatically
- 2 **Multigram**: keyphrases should be extracted automatically
- 3 **Multilingual**: cross-language and multi-language search should be supported
- 4 **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model

## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable**: each topic should be well interpretable by humans and labeled automatically
- 2 **Multigram**: keyphrases should be extracted automatically
- 3 **Multilingual**: cross-language and multi-language search should be supported
- 4 **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
- 5 **Temporal**: topic dynamics over time should be identified

## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable:** each topic should be well interpretable by humans and labeled automatically
- 2 **Multigram:** keyphrases should be extracted automatically
- 3 **Multilingual:** cross-language and multi-language search should be supported
- 4 **Multimodal:** authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
- 5 **Temporal:** topic dynamics over time should be identified
- 6 **Hierarchical:** granularity of topics should be user-adjustable

## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable**: each topic should be well interpretable by humans and labeled automatically
- 2 **Multigram**: keyphrases should be extracted automatically
- 3 **Multilingual**: cross-language and multi-language search should be supported
- 4 **Multimodal**: authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
- 5 **Temporal**: topic dynamics over time should be identified
- 6 **Hierarchical**: granularity of topics should be user-adjustable
- 7 **Segmented**: the topical text segmentation should be supported beyond the bag-of-words (BoW) model

## Exploratory Search requires that the Topic Model was...

- 1 **Interpretable:** each topic should be well interpretable by humans and labeled automatically
- 2 **Multigram:** keyphrases should be extracted automatically
- 3 **Multilingual:** cross-language and multi-language search should be supported
- 4 **Multimodal:** authors, categories, sources, links, tags, named entities, users, etc. should be involved in the model
- 5 **Temporal:** topic dynamics over time should be identified
- 6 **Hierarchical:** granularity of topics should be user-adjustable
- 7 **Segmented:** the topical text segmentation should be supported beyond the bag-of-words (BoW) model
- 8 **Semi-supervised:** the corrections from experts should be used to improve the model

## Additive Regularization for Topic Modeling (ARTM)

Additional *regularization* criteria  $R_i(\Phi, \Theta) \rightarrow \max, i = 1, \dots, n$ .

The problem of **regularized** log-likelihood maximization under non-negativeness and normalization constraints:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathcal{L}(\Phi, \Theta)} + \underbrace{\sum_{i=1}^n \tau_i R_i(\Phi, \Theta)}_{R(\Phi, \Theta)} \rightarrow \max_{\Phi, \Theta}$$

$$\phi_{wt} \geq 0; \quad \sum_{w \in W} \phi_{wt} = 1; \quad \theta_{td} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1$$

where  $\tau_i > 0$  are *regularization coefficients*.

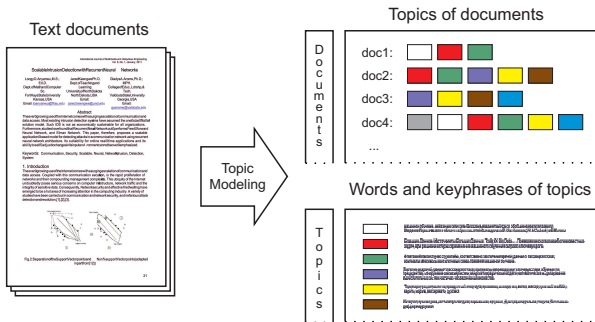
PLSA:  $R(\Phi, \Theta) = 0$

LDA:  $R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$



# Multimodal Probabilistic Topic Modeling

Given a text document collection *Probabilistic Topic Model* finds:  
 $p(t|d)$  — topic distribution for each document  $d$ ,  
 $p(w|t)$  — term distribution for each topic  $t$ .





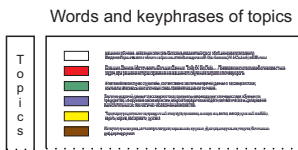
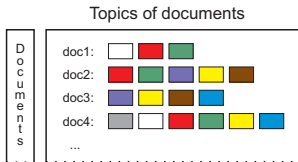
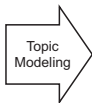
# Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms  $p(w|t)$ , authors  $p(a|t)$ , time  $p(y|t)$ , **objects on images  $p(o|t)$** ,

**Metadata:**  
 Authors  
 Data Time  
 Conference  
 Organization  
 URL  
 etc.

Text documents

Images

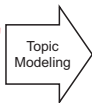


# Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms  $p(w|t)$ , authors  $p(a|t)$ , time  $p(y|t)$ , objects on images  $p(o|t)$ , linked documents  $p(d'|t)$ ,

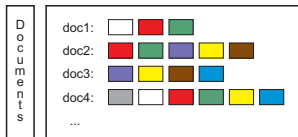
**Metadata:**  
 Authors  
 Data Time  
 Conference  
 Organization  
 URL  
 etc.

**Text documents**

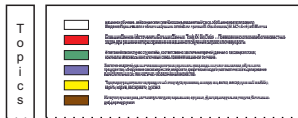


Images Links

**Topics of documents**



**Words and keyphrases of topics**

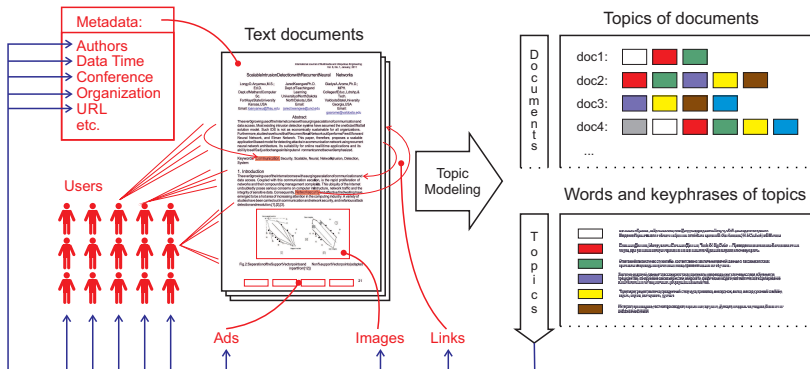






# Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms  $p(w|t)$ , authors  $p(a|t)$ , time  $p(y|t)$ , objects on images  $p(o|t)$ , linked documents  $p(d'|t)$ , advertising banners  $p(b|t)$ , users  $p(u|t)$ , and binds all these modalities into a single topic model.



## Multimodal ARTM: combining multimodality and regularization

$M$  is the set of modalities

$W^m$  is a vocabulary of tokens of  $m$ -th modality,  $m \in M$

$W = W^1 \sqcup \dots \sqcup W^M$  is a joint vocabulary of all modalities

The problem of **multimodal regularized** log-likelihood maximization under non-negativeness and normalization constraints:

$$\sum_{m \in M} \lambda_m \underbrace{\sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}}_{\text{modality log-likelihood } \mathcal{L}_m(\Phi, \Theta)} + \underbrace{\sum_{i=1}^n \tau_i R_i(\Phi, \Theta)}_{R(\Phi, \Theta)} \rightarrow \max_{\Phi, \Theta}$$

$$\phi_{wt} \geq 0, \quad \sum_{w \in W^m} \phi_{wt} = 1, \quad m \in M; \quad \theta_{td} \geq 0, \quad \sum_{t \in T} \theta_{td} = 1.$$

where  $\lambda_m > 0$ ,  $\tau_i > 0$  are *regularization coefficients*.



## EM-algorithm for multimodal ARTM

EM-algorithm is a simple-iteration method for a system of equations

**Theorem.** The local maximum  $(\Phi, \Theta)$  satisfies the following system of equations with auxiliary variables  $p_{tdw} = p(t|d, w)$ :

$$p_{tdw} = \mathop{\text{norm}}_{t \in T} (\phi_{wt} \theta_{td});$$

$$\phi_{wt} = \mathop{\text{norm}}_{w \in W^m} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right); \quad n_{wt} = \sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw};$$

$$\theta_{td} = \mathop{\text{norm}}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right); \quad n_{td} = \sum_{w \in D} \lambda_{m(w)} n_{dw} p_{tdw};$$

where  $\mathop{\text{norm}}_{t \in T} x_t = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$  is nonnegative normalization;

$m(w)$  is the modality of the term  $w$ , so that  $w \in W^{m(w)}$ .

# ARTM vs. Graphical Models and Bayesian Inference

In Bayesian approach, a lot of calculus to be done *for each model* to go from the problem statement to the solution algorithm:

$$\begin{aligned}
 p(Z, W | \alpha, \beta) &= p(W | Z, \beta) p(Z | \alpha) \\
 p(W | Z, \beta) &= \int p(W | Z, \Phi) p(\Phi | \beta) d\Phi \\
 p(\Phi | \beta) &= \prod_{k=1}^K p(\phi_k | \beta) = \prod_{k=1}^K \frac{1}{\text{B}(\beta)} \prod_{v=1}^V \phi_{k,v}^{\beta_{k,v}-1} \\
 p(W | Z, \Phi) &= \prod_{i=1}^N \theta_{i,\alpha_i} = \prod_{k=1}^K \prod_{v=1}^V \Phi_{k,v}^{k(i,v)} \\
 \Phi(k, v) &= \sum_{z=1}^K \mathbb{I}\{w_i = v \wedge z_i = k\} \\
 p(W | Z, \beta) &= \int \prod_{i=1}^N \frac{1}{\text{B}(\beta)} \prod_{k=1}^K \prod_{v=1}^V \Phi_{k,v}^{k(i,v)+\beta_{k,v}-1} d\phi_k \\
 &= \int \prod_{k=1}^K f_k(\phi_k) d\phi_1 \dots d\phi_K = \prod_{k=1}^K \int f_k(\phi_k) d\phi_k \\
 p(W | Z, \beta) &= \prod_{k=1}^K \left( \int \frac{1}{\text{B}(\beta)} \prod_{v=1}^V \Phi_{k,v}^{k(i,v)+\beta_{k,v}-1} d\phi_k \right) \\
 &= \prod_{k=1}^K \left( \frac{1}{\text{B}(\beta)} \prod_{v=1}^V \int \Phi_{k,v}^{k(i,v)+\beta_{k,v}-1} d\phi_k \right) \\
 p(W | Z, \beta) &= \prod_{k=1}^K \frac{\text{B}(\Psi_k + \beta)}{\text{B}(\beta)}
 \end{aligned}$$

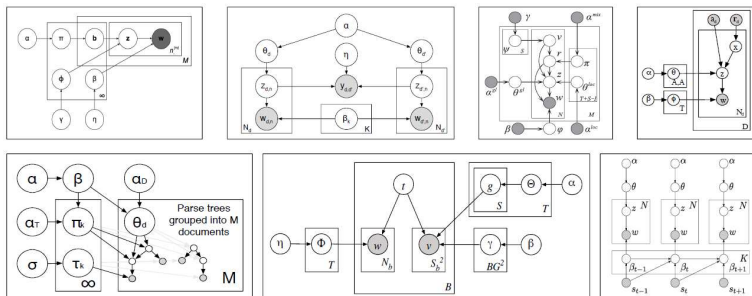
$$\begin{aligned}
 p(\Theta | \alpha) &= \prod_{d=1}^D p(\theta_{d,\alpha}) = \prod_{d=1}^D \frac{1}{\text{B}(\alpha)} \prod_{k=1}^K \theta_{d,k}^{\alpha_{d,k}-1} \\
 p(Z | \Theta) &= \prod_{d=1}^D \theta_{d,\alpha_d} = \prod_{d=1}^D \prod_{k=1}^K \theta_{d,k}^{d(d,k)} \\
 p(Z | \alpha) &= \int p(Z | \Theta) p(\Theta | \alpha) d\Theta \\
 &= \prod_{d=1}^D \left( \int \frac{1}{\text{B}(\alpha)} \prod_{k=1}^K \theta_{d,k}^{d(d,k)+\alpha_{d,k}-1} d\theta_k \right) \\
 &= \prod_{d=1}^D \frac{\text{B}(\Omega_d + \alpha)}{\text{B}(\alpha)} \\
 \Omega(d, k) &= \sum_{i=1}^N \mathbb{I}\{d_i = m \wedge z_i = k\} \\
 p(Z, W | \alpha, \beta) &= p(W | Z, \beta) p(Z | \alpha) \\
 &= \prod_{k=1}^K \frac{\text{B}(\Psi_k + \beta)}{\text{B}(\beta)} \cdot \prod_{d=1}^D \frac{\text{B}(\Omega_d + \alpha)}{\text{B}(\alpha)} \\
 p(z_i = k | Z_{-i}, W, \alpha, \beta) &= \frac{p(z_i = k, Z_{-i}, W | \alpha, \beta)}{p(Z_{-i}, W | \alpha, \beta)} \\
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{p(Z, W | \alpha, \beta)}{p(Z_{-i}, W | \alpha, \beta)} \\
 p(Z, W | \alpha, \beta) &= p(W | Z, \beta) p(Z | \alpha) \\
 &= \prod_{k=1}^K \frac{\text{B}(\Psi_k^+ + \beta)}{\text{B}(\beta)} \cdot \prod_{d=1}^D \frac{\text{B}(\Omega_d^+ + \alpha)}{\text{B}(\alpha)}
 \end{aligned}$$

$$\begin{aligned}
 \Psi^{-1}(k, v) &= \sum_{\substack{1 \leq i \leq N \\ j \neq i}} \mathbb{I}\{w_j = v \wedge z_j = k\} \\
 \Omega^{-1}(d, k) &= \sum_{\substack{1 \leq i \leq N \\ j \neq i}} \mathbb{I}\{d_j = d \wedge z_j = k\} \\
 \Phi(k, v) &= \begin{cases} \Psi^{-1}(k, v) + 1 & \text{if } v = w_i \text{ and } k = z_i; \\ \Psi^{-1}(k, v) & \text{all other cases.} \end{cases} \\
 \Omega(d, k) &= \begin{cases} \Omega^{-1}(d, k) + 1 & \text{if } d = d_i \text{ and } k = z_i; \\ \Omega^{-1}(d, k) & \text{all other cases.} \end{cases} \\
 \sum_{i=1}^V n(i; z_i) &= 1 + \sum_{i=1}^V n_{-i}(z_i) \\
 \text{B}(x) &= \frac{\Gamma(x) \Gamma(x)}{\Gamma(\sum_{i=1}^m x_i)} \\
 \sum_{k=1}^K n(z_i; d_k) &= 1 + \sum_{k=1}^K n_{-i}(z_i; d_k) \\
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{\text{B}(n(z_i) + \beta)}{\text{B}(n_{-i}(z_i) + \beta)} \cdot \frac{\text{B}(n(z_i; m) + \alpha)}{\text{B}(n_{-i}(z_i; m) + \alpha)} \\
 &= \frac{\prod_{k=1}^K \frac{\Gamma(n(z_i; k) + \beta_k)}{\Gamma(\sum_{k=1}^K n(z_i; k) + \beta)} \cdot \frac{\Gamma(n(z_i; m) + \alpha_m)}{\Gamma(\sum_{k=1}^K n(z_i; k) + \alpha_k)}}{\prod_{k=1}^K \frac{\Gamma(n_{-i}(z_i; k) + \beta_k)}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; k) + \beta)} \cdot \frac{\Gamma(n_{-i}(z_i; m) + \alpha_m)}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; k) + \alpha_k)}}
 \end{aligned}$$

$$\begin{aligned}
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{n(z_i; z_i) + \beta_{z_i} - 1}{\left[ \sum_{k=1}^K n(z_i; z_k) + \beta_k \right] - 1} \cdot \frac{n(z_i; d_i) + \alpha_{z_i} - 1}{\left[ \sum_{k=1}^K n(z_i; d_k) + \alpha_{z_k} \right] - 1} \\
 p(z_i | Z_{-i}, W, \alpha, \beta) &= \frac{n(z_i; z_i) + \beta_{z_i} - 1}{\left[ \sum_{k=1}^K n(z_i; z_k) + \beta_k \right] - 1} \cdot [n(z_i; d_i) + \alpha_{z_i} - 1] \\
 \phi_{k,i} &= p(w = i; z = k, W, Z, \beta) \\
 \theta_{k,i} &= p(z = k | Z, \alpha) \\
 \phi_{k,i} \cdot \theta_{k,i} &= p(w = i | z = k, W, Z, \beta) \cdot p(z = k | Z, \alpha) \\
 &= p(w = i, z = k | W, Z, \alpha, \beta) \\
 &= \frac{p(W, Z | \alpha, \beta)}{p(W, Z | \alpha, \beta)} \\
 \phi_{k,i} \cdot \theta_{k,i} &= \frac{p(w = i, z = k | W, Z, \alpha, \beta)}{p(W, Z | \alpha, \beta)} \\
 &= \frac{\frac{\Gamma(n(z_i; k) + 1 + \beta_k)}{\Gamma(\sum_{k=1}^K n(z_i; k) + 1 + \beta_k)} \cdot \frac{\Gamma(n(z_i; d_i) + 1 + \alpha_{z_i})}{\Gamma(\sum_{k=1}^K n(z_i; d_k) + 1 + \alpha_{z_k})}}{\frac{\Gamma(n_{-i}(z_i; k) + \beta_k)}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; k) + \beta_k)} \cdot \frac{\Gamma(n_{-i}(z_i; d_i) + \alpha_{z_i})}{\Gamma(\sum_{k=1}^K n_{-i}(z_i; d_k) + \alpha_{z_k})}} \\
 \phi_{k,i} \cdot \theta_{k,i} &= \frac{n(z_i; k) + \beta_k}{\left( \sum_{k=1}^K n(z_i; k) + \beta_k \right)} \cdot \frac{n(z_i; d_i) + \alpha_{z_i}}{\left( \sum_{k=1}^K n(z_i; d_k) + \alpha_{z_k} \right)} \\
 \phi_{k,i} &= \frac{n(z_i; k) + \beta_k}{\left( \sum_{k=1}^K n(z_i; k) + \beta_k \right)} \\
 \theta_{k,i} &= \frac{n(z_i; d_i) + \alpha_{z_i}}{\left( \sum_{k=1}^K n(z_i; d_k) + \alpha_{z_k} \right)}
 \end{aligned}$$

# ARTM vs. Graphical Models and Bayesian Inference

In Bayesian approach, Graphical Models are used to make model representation clear:



## ARTM vs. Graphical Models and Bayesian Inference

In ARTM, a general system of equations holds *for all the models*, each model represented by its own regularizer  $R(\Phi, \Theta)$ :

$$\left\{ \begin{array}{l} p_{tdw} = \text{norm}_t(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \text{norm}_w\left(\sum_d n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \text{norm}_t\left(\sum_w n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{array} \right.$$

## ARTM vs. Graphical Models and Bayesian Inference

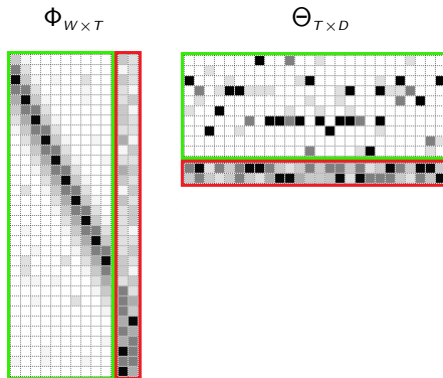
In ARTM, we can combine models, simply making a weighted sum of their regularizers  $\tau_1 R_1 + \dots + \tau_k R_k$ :

$$\left\{ \begin{array}{l} p_{tdw} = \text{norm}_t(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \text{norm}_w\left(\sum_d n_{dw} p_{tdw} + \phi_{wt} \sum_i \tau_i \frac{\partial R_i}{\partial \phi_{wt}}\right) \\ \theta_{td} = \text{norm}_t\left(\sum_w n_{dw} p_{tdw} + \theta_{td} \sum_i \tau_i \frac{\partial R_i}{\partial \theta_{td}}\right) \end{array} \right.$$

## Assumptions: what topics would be well-interpretable?

Topics  $S \subset T$  contain domain-specific terms  
 $p(w|t)$ ,  $t \in S$  are sparse and different (weakly correlated)

Topics  $B \subset T$  contain background terms  
 $p(w|t)$ ,  $t \in B$  are dense and contain common lexis words



## Smoothing regularization (rethinking LDA)

**The non-sparsity assumption** for background topics  $t \in B$ :

$\phi_{wt}$  are similar to a given distribution  $\beta_w$ ;

$\theta_{td}$  are similar to a given distribution  $\alpha_t$ .

$$\sum_{t \in B} \text{KL}_w(\beta_w \parallel \phi_{wt}) \rightarrow \min_{\Phi}; \quad \sum_{d \in D} \text{KL}_t(\alpha_t \parallel \theta_{td}) \rightarrow \min_{\Theta}.$$

We minimize the sum of these KL-divergences to get a regularizer:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in B} \alpha_t \ln \theta_{td} \rightarrow \max.$$

The regularized M-step applied for all  $t \in B$  coincides with LDA:

$$\phi_{wt} \propto n_{wt} + \beta_0 \beta_w, \quad \theta_{td} \propto n_{td} + \alpha_0 \alpha_t,$$

which is new non-Bayesian interpretation of LDA [Blei 2003].

## Sparsing regularizer (further rethinking LDA)

The **sparsity assumption** for domain-specific topics  $t \in S$ :  
distributions  $\phi_{wt}$ ,  $\theta_{td}$  contain many zero probabilities.

We maximize the sum of KL-divergences  $\text{KL}(\beta \parallel \phi_t)$  and  $\text{KL}(\alpha \parallel \theta_d)$ :

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in S} \alpha_t \ln \theta_{td} \rightarrow \max.$$

The regularized M-step gives “anti-LDA”, for all  $t \in S$ :

$$\phi_{wt} \propto (n_{wt} - \beta_0 \beta_w)_+, \quad \theta_{td} \propto (n_{td} - \alpha_0 \alpha_t)_+.$$

---

*Varadarajan J., Emonet R., Odohez J.-M.* A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.



## Regularization for topics decorrelation

**The dissimilarity assumption** for domain-specific topics  $t \in S$ :  
if topics are interpretable then they must differ significantly.

We maximize covariances between column vectors  $\phi_t$ :

$$R(\Phi) = -\frac{\tau}{2} \sum_{t \in S} \sum_{s \in S \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \rightarrow \max.$$

The regularized M-step makes columns of  $\Phi$  more distant:

$$\phi_{wt} \propto \left( n_{wt} - \tau \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \right)_+.$$

---

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp. 224–228.

## ARTM: available regularizers

- topic smoothing ( $\Leftrightarrow$  Latent Dirichlet Allocation)
- topic sparsing
- topic decorrelation
- topic selection via entropy sparsing
- topic coherence maximization
- supervised learning for classification and regression
- semi-supervised learning
- using documents citation and links
- modeling temporal topic dynamics
- using vocabularies in multilingual topic models
- etc.

---

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models. Machine Learning Journal. Springer, 2014.

## BigARTM project

### BigARTM features:

- Parallel + Online + Multimodal + Regularized Topic Modeling
- Out-of-core processing of Big Data
- Built-in library of regularizers and quality measures

### BigARTM community:

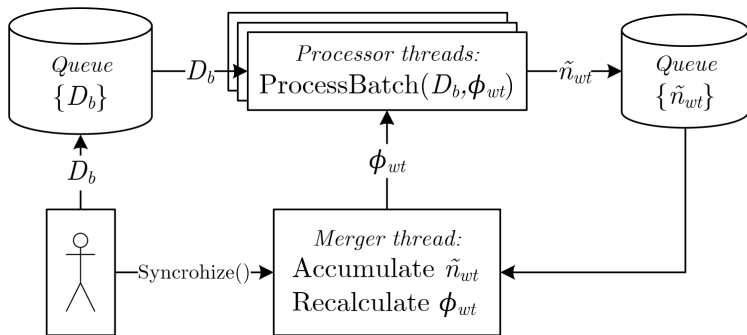
- Open-source <https://github.com/bigartm>  
(discussion group, issue tracker, pull requests)
- Documentation <http://bigartm.org>



### BigARTM license and programming environment:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

## The BigARTM project: parallel architecture



- Concurrent processing of batches  $D = D_1 \sqcup \dots \sqcup D_B$
- Simple single-threaded code for *ProcessBatch*
- User controls when to update the model in online algorithm
- Deterministic (reproducible) results from run to run

## Fast online EM-algorithm for regularized multimodal PTMs

**Input:** collection  $D$  split into batches  $D_b$ ,  $b = 1, \dots, B$ ;

**Output:** matrix  $\Phi$ ;

- 1 initialize  $\phi_{wt}$  for all  $w \in W$ ,  $t \in T$ ;
- 2  $n_{wt} := 0$ ,  $\tilde{n}_{wt} := 0$  for all  $w \in W$ ,  $t \in T$ ;
- 3 **for all** batches  $D_b$ ,  $b = 1, \dots, B$
- 4     iterate each document  $d \in D_b$  at a constant matrix  $\Phi$ :  
    $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \mathbf{ProcessBatch}(D_b, \Phi)$ ;
- 5     **if** (synchronize) **then**
- 6          $n_{wt} := n_{wt} + \tilde{n}_{dw}$  for all  $w \in W$ ,  $t \in T$ ;
- 7          $\phi_{wt} := \mathop{\text{norm}}_{w \in W^m} \left( n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$  for all  $w \in W^m$ ,  $m \in M$ ,  $t \in T$ ;
- 8          $\tilde{n}_{wt} := 0$  for all  $w \in W$ ,  $t \in T$ ;

## Fast online EM-algorithm for Multi-ARTM

ProcessBatch iterates documents  $d \in D_b$  at a constant matrix  $\Phi$ .

matrix  $(\tilde{n}_{wt}) := \text{ProcessBatch}$  (set of documents  $D_b$ , matrix  $\Phi$ )

- 1  $\tilde{n}_{wt} := 0$  for all  $w \in W, t \in T$ ;
- 2 **for all**  $d \in D_b$
- 3     initialize  $\theta_{td} := \frac{1}{|T|}$  for all  $t \in T$ ;
- 4     **repeat**
- 5          $p_{tdw} := \text{norm}_{t \in T}(\phi_{wt}\theta_{td})$  for all  $w \in d, t \in T$ ;
- 6          $n_{td} := \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw}$  for all  $t \in T$ ;
- 7          $\theta_{td} := \text{norm}_{t \in T}(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}})$  for all  $t \in T$ ;
- 8     **until**  $\theta_d$  converges;
- 9      $\tilde{n}_{wt} := \tilde{n}_{wt} + \lambda_{m(w)} n_{dw} p_{tdw}$  for all  $w \in d, t \in T$ ;

## Summary of ARTM approach

### EM-algorithm is computationally effective:

- It has linear time complexity  $O(n \cdot |T| \cdot n_{\text{iter}})$
- Its online version makes only one pass through big collection
- Parallelism is possible for both multi-core CPUs and clusters

### ARTM reduces barriers to entry into PTM research field:

- General EM-algorithm for many models and their combinations
- PLSA, LDA, and 100s of PTMs are covered by ARTM
- Combining multiple modalities and regularizers is easy
- No complicated Bayesian inference and graphical models

### Open problem / Under development:

- Adaptive optimization of regularization coefficients  $\tau_i, \lambda_m$

## BigARTM vs Gensim vs Vowpal Wabbit

- 3.7M articles from Wikipedia, 100K unique words

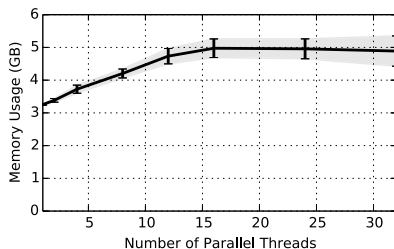
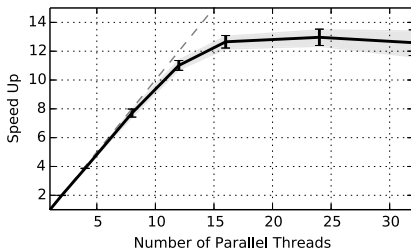
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

- *procs* = number of parallel threads
- *inference* = time to infer  $\theta_d$  for 100K held-out documents
- *perplexity* is calculated on held-out documents.



## Running BigARTM in parallel

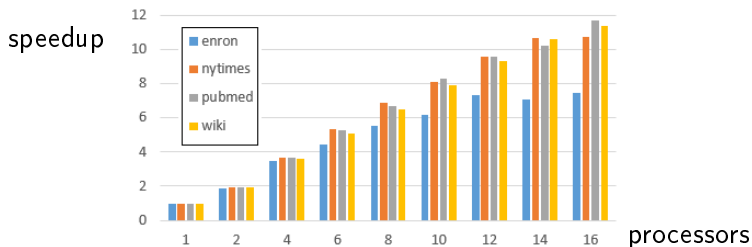
- 3.7M articles from Wikipedia, 100K unique words



- Amazon EC2 c3.8xlarge (16 physical cores + hyperthreading)
- No extra memory cost for adding more threads

## Running BigARTM on large collections

collection	$ W , 10^3$	$ D , 10^6$	$n, 10^6$	size, GB
enron	28	0.04	6.4	0.07
nytimes	103	0.3	100	0.13
pubmed	141	8.2	738	1.0
wiki	100	3.7	1009	1.2

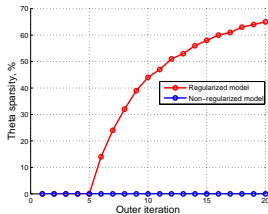
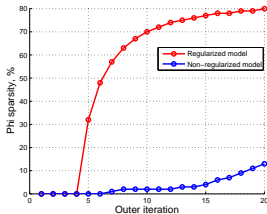
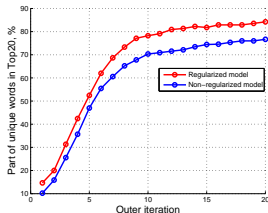
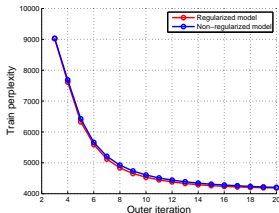


Amazon EC2 cc2.8xlarge instance:

16 cores + hyperthreading, Intel<sup>®</sup> Xeon<sup>®</sup> CPU E5-2670 2.6GHz.

## Running BigARTM with multiple regularizers

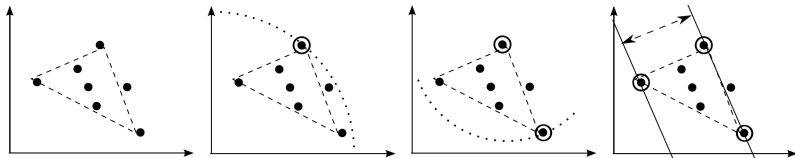
ARTM combines regularizers to improve sparsity and number of topical words without a loss of the perplexity.



## Arora's algorithm based on anchor words recovery

**Def.** The word  $w$  is an *anchor word* of the topic  $t$  if  $p(w|t) = 1$ .

**Arora's algorithm** finds  $\Phi$  with the identity submatrix of anchors



- ⊕ The fastest algorithm for Topic Modeling
- ⊕ Theoretical guarantees for polynomial time and global optimum
- ⊖ The hypothesis that  $\forall t$  the anchor word exists is restrictive
- ⊖ The algorithm is not so fast for big vocabularies  $|W|$

---

Sanjeev Arora et al. A Practical Algorithm for Topic Modeling with Provable Guarantees. ICML 2013.

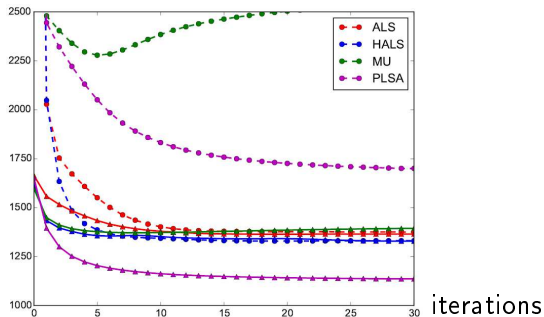
## Arora's algorithm for EM-algorithms initialization

ALS, HALS, MU — methods for nonnegative matrix factorization

NIPS collection:  $|D| = 1500$ ,  $|W| = 12419$ ,  $|T| = 25$ .

$$\text{Perplexity} = \exp\left(-\frac{1}{n}\mathcal{L}(\Phi, \Theta)\right)$$

perplexity



iterations

solid lines — initialization by Arora's algorithm

dotted lines — random initialization

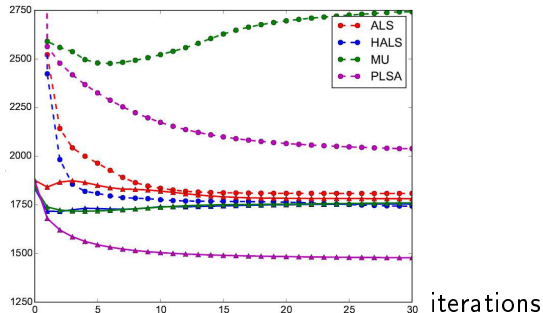
## Arora's algorithm for EM-algorithms initialization

ALS, HALS, MU — methods for nonnegative matrix factorization

Daily Kos collection:  $|D| = 3430$ ,  $|W| = 6906$ ,  $|T| = 25$ .

Perplexity =  $\exp\left(-\frac{1}{n}\mathcal{L}(\Phi, \Theta)\right)$

perplexity



solid lines — initialization by Arora's algorithm

dotted lines — random initialization

## Conclusions about Arora's initialization

- Arora's initialization greatly improves PLSA (PLSA — Probabilistic Latent Semantic Analysis is equivalent to ARTM without regularization)
- Arora's initialization does not improve quadratic loss minimizers
- PLSA is capable of improving the Arora's initialization, perhaps, because of restrictive assumptions of Arora's algorithm do not hold in the real data

## Regularization for topic selection

Let us maximize KL-divergence:  $\text{KL}\left(\frac{1}{|T|} \parallel p(t)\right) \rightarrow \max$   
to make distribution over topics  $p(t)$  sparse:

$$R(\Theta) = -\tau n \sum_{t \in S} \frac{1}{|T|} \ln \underbrace{\sum_{d \in D} p(d) \theta_{td}}_{p(t)} \rightarrow \max.$$

The regularized M-step formula results in  $\Theta$  row sparsing:

$$\theta_{td} = \text{norm}_{t \in T} \left( n_{td} \left( 1 - \tau \frac{n}{n_t |T|} \right) \right).$$

**The row sparsing effect:**

if  $n_t < \tau \frac{n}{|T|}$  then all values in the  $t$ -th row turn into zeros.



## The experiments with topic selection

**Real dataset:** NIPS (Neural Information Processing System)

- $|D| = 1566$  preprocessed papers from NIPS conference;
- vocabulary:  $|W| \approx 1.3 \cdot 10^4$ ; hold-out set:  $|D'| = 174$ .

**Synthetic dataset:**

- 500 EM iterations for PLSA with  $|T_0| = 50$  topics on NIPS
- generate synthetic dataset ( $n_{dw}^0$ ) using obtained  $\Phi$  and  $\Theta$ :

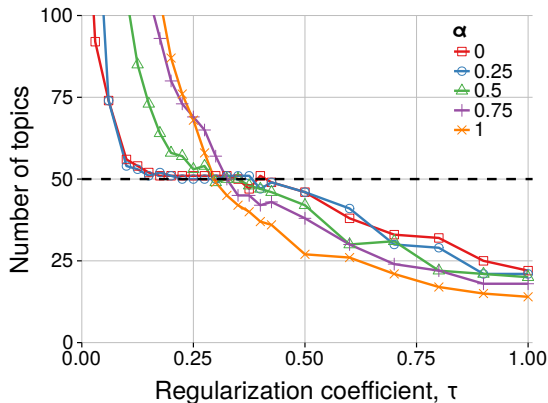
$$n_{dw}^0 = n_d \sum_{t \in T} \phi_{wt} \theta_{td}$$

**Parametric family of semi-real datasets:**

- $(n_{dw}^\alpha)$  is a mixture of synthetic ( $n_{dw}^0$ ) and real ( $n_{dw}$ ) datasets:

$$n_{dw}^\alpha = \alpha n_{dw} + (1 - \alpha) n_{dw}^0$$

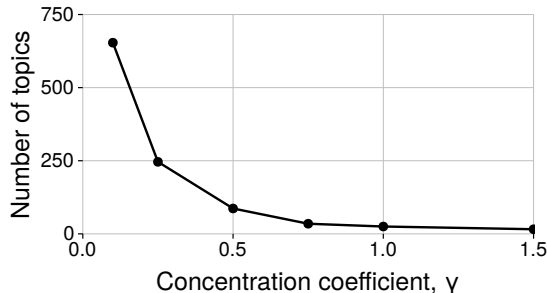
## Number of topics determination



- For synthetic dataset ARTM reliably finds the truth:  $|T| = 50$ .
- The range of  $\tau$  values leading to the correct number is wide.
- For real data the number of topics is not clear.

## Comparison to HDP topic model

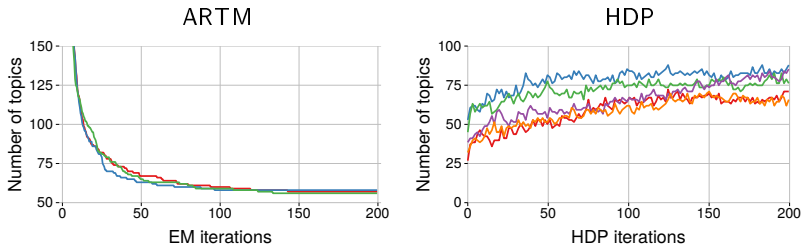
HDP (Hierarchical Dirichlet Process, Teh et. al, 2006)  
is the state-of-art approach for a number of topics optimization.



- The choice of the concentration coefficient  $\gamma$  of Dirichlet process may lead to nearly any number of topics.

## Stability of ARTM vs HDP

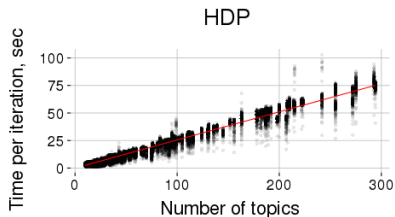
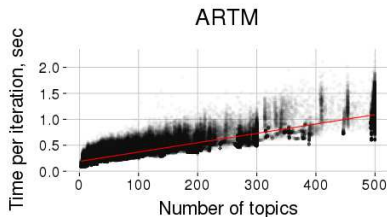
Starting ARTM and HDP many times from different initializations:



- 1 HDP is less stable in two ways:
  - 1 The number of topics fluctuates from iteration to iteration
  - 2 The results for several random starts significantly differ
- 2 The “recommended” parameters  $\gamma$  for HDP and  $\tau$  for ARTM give the similar number of topics  $\approx 60$

## Running time of ARTM vs HDP

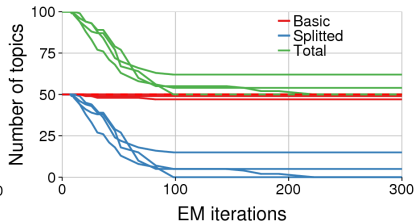
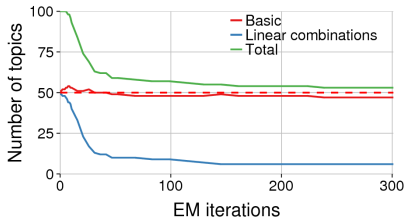
Comparing the running time per iteration (sec) of ARTM vs HDP (one iteration = one pass through the collection)



- Our method is 100 times faster!

## Elimination of linearly dependent topics and nested subtopics

- Add 50 linear combinations of topics in synthetic dataset.
- Add 50 nested subtopics in synthetic dataset.



- Our regularizer effectively eliminates both linearly dependent topics and nested subtopics from the model
- More diverse topics of the original model remain.

## Conclusions about number of topics

- It seems that the “true number of topics” does not exist in real text collections.
- ARTM has a special regularizer for topic selection, which eliminates small, linearly dependent, and nested topics.
- It is faster and more stable than state-of-the-art HDP.

- Topic Modeling is an applied area of optimization and matrix factorization in text analysis
- ARTM (Additive Regularization) is a semi-probabilistic non-Bayesian multicriteria view on Topic Modeling
- BigARTM is open source project for parallel online multimodal regularized Topic Modeling of large collections

### Contacts:

Konstantin Vorontsov: voron@forecsys.ru

Wiki [www.MachineLearning.ru](http://www.MachineLearning.ru) User:Vokov (in Russian)