

# Additive Regularization of Topic Models: Towards Exploratory Search and Other Multi-Criteria Applications

Konstantin Vorontsov

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

Machine Learning: Prospects and Applications

Berlin • October 5–8, 2015

## 1 Exploratory Search

- The paradigm of exploratory search
- The prototype GUI for exploratory search
- The keystone of exploratory search

## 2 Topic Modeling

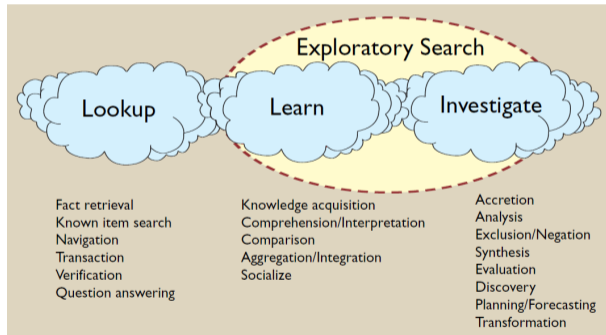
- ARTM: additive regularization for topic modeling
- BigARTM open source project
- Experiments

## 3 Symbolic Dynamics for Medical Diagnostics

- Electrocardiography
- Informational analysis of ECG signals
- Multi-disease ECG diagnostics

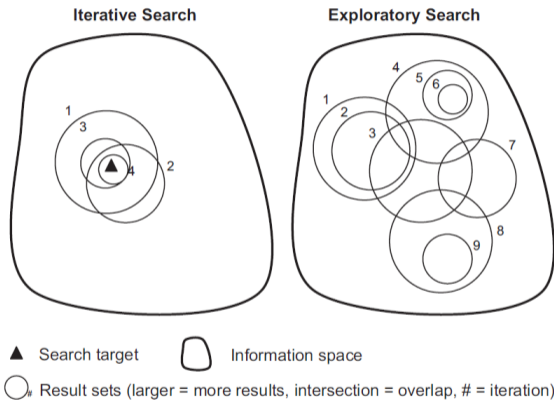
## Exploratory Search for learning, knowledge acquisition and discovery

- 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



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

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

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

ARTM (рус.) — Аддитивная Регуляризация Тематических Моделей.

**Теоретическое введение**

Основная идея: Тематическое моделирование

Беремословное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель описывает каждую тему дискретным распределением на неизвестные термине, каждый документ — дискретным распределением на известные тек. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, ориентации текстов.

Тематическая модель — это представление наблюдаемого условного распределения  $p(u|d)$  термине (слово или словосочетание)  $u$  в документе  $d$  коллекции  $D$ :

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

где  $T$  — множество тем;

$$\theta_{ut} = p(u|t) \text{ — неизвестное распределение термине в теме } t;$$

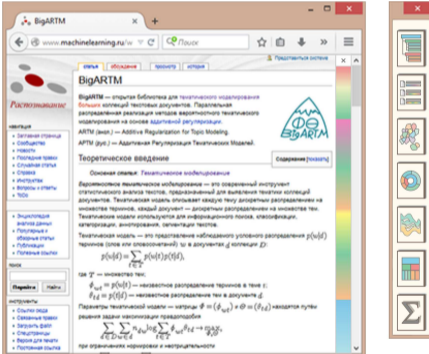
$$\phi_{td} = p(t|d) \text{ — неизвестное распределение тем в документе } d;$$

Параметры тематической модели — матрицы  $\Phi = (\phi_{td})$  и  $\Theta = (\theta_{ut})$  называются матрицами решетки задачи максимизации правдоподобия

$$\sum_{d \in D} \sum_{u \in U} n_{ud} \ln \sum_{t \in T} \theta_{ut} \phi_{td} \rightarrow \max_{\Phi, \Theta},$$

при ограничении 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. On the right side of the main content area, there is a vertical color bar representing a topic query. The interface includes a navigation menu on the left, a main content area with text and a logo, and a sidebar on the right with various icons.

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

Березинское вероятностное моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель описывает каждую тему дискретным распределением на неизвестные термины, каждый документ — дискретным распределением на известные тек. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, оптимизации текстов.

Тематическая модель — это представление наблюдаемого условного распределения  $p(w|d)$  термине (слове или словосочетании)  $w$  в документе  $d$  коллекции  $D$ :

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

где  $T$  — множество тем;

$\theta_{wt} = p(w|t)$  — неизвестное распределение термине в теме  $t$ ;

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

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

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

при ограничениях нормировки и неотрицательности

## Exploratory search: the prototype of graphical user interface

## Topics of the query document

BigARTM — специализированная библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределённая реализация методов 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

Березинское вероятностное моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель описывает каждую тему дискретным распределением на неизвестные термины, каждый документ — дискретным распределением на известные тек. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

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

$$p(\mathbf{d}|\mathbf{z}) = \sum_{t \in \mathcal{T}} p_t(\mathbf{d})z_t(\mathbf{z}),$$

где  $\mathcal{T}$  — множество тем;

$$\delta_{\mathbf{d}|\mathbf{z}} = p(\mathbf{d}|\mathbf{z}) - \text{неизвестное распределение терминов в теме } t;$$

$$\theta_{t\mathbf{d}} = p(\mathbf{d}|\mathbf{z}) - \text{неизвестное распределение тем в документе } \mathbf{d}.$$

Параметры тематической модели — матрицы  $\Phi = (\delta_{\mathbf{d}|\mathbf{z}_t})$  и  $\Theta = (\theta_{t\mathbf{d}})$  находят путь решения задачи максимизации правдоподобия

$$\sum_{\mathbf{d} \in \mathcal{D}} \sum_{t \in \mathcal{T}} n_{\mathbf{d}t} \ln \sum_{t' \in \mathcal{T}} \delta_{\mathbf{d}|\mathbf{z}_{t'}} \rightarrow \max_{\Phi, \Theta},$$

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

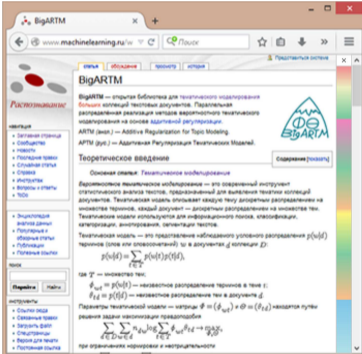
## Similar documents and objects ranked by relevance

The screenshot shows the BigARTM website interface. At the top, there are navigation tabs: "главная", "облачные", "применения", and "статьи". The main content area features the BigARTM logo and a description of the system as a library for topic modeling. A sidebar on the left contains a "навигация" menu with links to "Главная страница", "Облачные", "Новости", "Последние темы", "Слайдовая презентация", "История", "Видео и статьи", and "Тесты". Below the main text, there is a section titled "Теоретическое введение" with a sub-section "Основная идея: Тематическое моделирование". It includes a paragraph about the system and a mathematical formula for the joint distribution of topics and documents:  $p(\theta, d) = \prod_{t \in T} p(\theta(t)) p(d(t))$ . Below this, it defines  $\theta$  as a matrix of topic probabilities and  $d$  as a document, with the joint distribution  $p(\theta, d) = \prod_{t \in T} p(\theta(t)) p(d(t))$ . It also mentions the task of maximizing the log-likelihood of the observed data.

The screenshot shows a search results page for "BigARTM". The results are ranked by relevance. The top result is "BigARTM" from machinelearning.ru, with a snippet: "BigARTM реализует мультиязычные модели, позволяющие обрабатывать метаданные любого числа типов одновременно." Below this are several other results, including "Welcome to BigARTM's documentation! — BigARTM..." from docs.bigartm.org, "bigartm/bigartm" on GitHub, and "BigARTM — NLPub" from nlpub.ru. The results are presented in a clean, organized layout with icons for each item.

## Exploratory search: the prototype of graphical user interface

## Topic roadmap: clustering of relevant documents



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

Березинское тематическое моделирование — это современный инструмент статистического анализа текстов, разработанный для выделенных тематик коллекций документов. Тематическая модель описывает каждую тему дискретным распределением на неизвестные термины, каждый документ — дискретным распределением на известные темы. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

Тематическая модель — это предельное наблюдаемое условное распределение  $p(v|d)$  терминов (слов или словосочетаний)  $v$  в документе  $d$ :

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

где  $T$  — множество тем;

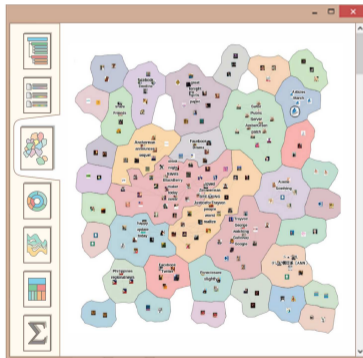
$$\delta_{v|t} = p(v|t) — \text{известное распределение терминов в теме } t;$$

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

Параметры тематической модели — матрицы  $\Phi = (\delta_{v|t})$  и  $\Theta = (\theta_{t|d})$  канонический путь решения задачи максимизации правдоподобия

$$\sum_{d \in D} \sum_{v \in V} n_{v,d} \log \sum_{t \in T} \delta_{v|t} \theta_{t|d} \rightarrow \max_{\Phi, \Theta},$$

при ограничениях: нормировка и неотрицательность

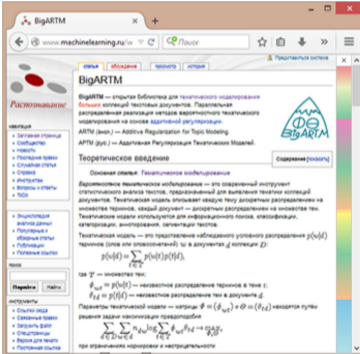


E.R.Gansner, Y.Hu, S.North. Visualizing Streaming Text Data with Dynamic Maps. 2012.



## Exploratory search: the prototype of graphical user interface

## Topic river: evolution of the domain area



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

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

ARTM (рус.) — Аддитивная Регуляризация Тематических Модели.

**Теоретическое введение**

Основная идея: Тематическое моделирование

Березинское вероятностное моделирование — это современный инструмент статистического анализа текстов, предназначенный для выявления тематик коллекций документов. Тематическая модель описывает каждую тему дискретным распределением на множестве термине, каждый документ — дискретным распределением на множестве тем. Тематические модели используются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов.

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

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

где  $T$  — множество тем;

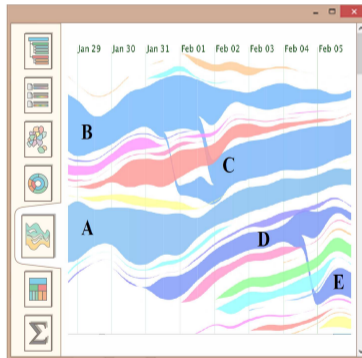
$\delta_{w|t} = p(\mathbf{w}|\mathbf{t})$  — неизвестное распределение термине в теме  $t$ ;

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

Параметры тематической модели — матрицы  $\Phi = (\delta_{w|t})$  и  $\Theta = (\theta_{t|\mathbf{d}})$  являются решением задачи максимизации правдоподобия

$$\sum_{\mathbf{d} \in D} \sum_{w \in V} n_{w\mathbf{d}} \ln \sum_{t \in T} \delta_{w|t} \theta_{t|\mathbf{d}} \rightarrow \max_{\Phi, \Theta},$$

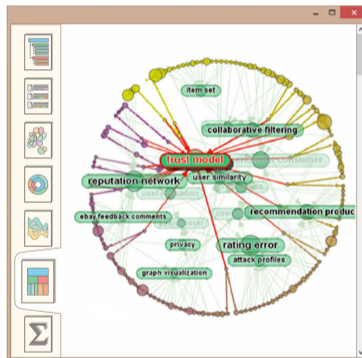
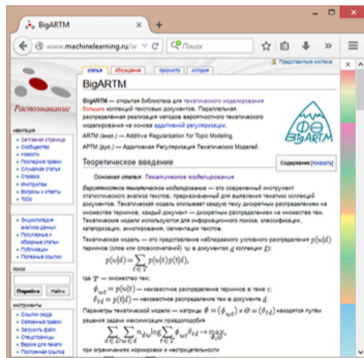
при ограничениях нормировки и неотрицательности



Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. IEEE Trans. Vis. Comput. Graph. 2014.

## Exploratory search: the prototype of graphical user interface

### Topic bar: segmentation of the query document



Gretarsson B., O'Donovan J., Bostandjiev S. et al. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

## Exploratory search: the prototype of graphical user interface

## Summarization of the query document

**BigARTM**

BigARTM — открытая библиотека для тематического моделирования больших коллекций текстовых документов. Параллельная распределенная реализация метода вариационного 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

Естественное тематическое моделирование — это современный инструмент статистического анализа текстов, предназначенный для выделения тематик коллекций документов. Тематическая модель описывает каждую тему дисперсным распределением на множестве термов, каждый документ — дисперсным распределением на множестве тем. Тематические модели используются для информативного поиска, классификации, категоризации, аннотирования, сегментации текстов.

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

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

где  $T$  — множество тем;

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

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

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

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

при ограничениях неотрицательности.

**Суммаризация «BigARTM»**

Тематическое моделирование — одно из современных направлений статистического анализа текстов, активно развивающееся последние 10–15 лет. Тематические модели выявляют латентные темы в коллекциях текстовых документов и используются для создания системы семантического поиска, категоризации, суммаризации, сегментации текстов. Основные требования к тематическим моделям: они должны быть хорошо интерпретируемыми (автоматически строить темы, понятные конечным пользователям), мультимодальными (учитывать разнородные метаданные документов), динамическими (выявлять динамику тем во времени), иерархическими (автоматически разделять темы на подтемы), мультиграммными (использовать не только отдельные слова, но и ключевые фразы), и т.д. Библиотека с открытым кодом BigARTM предназначена для построения регуляризованных мультимодальных тематических моделей больших текстовых коллекций.

<http://textvis.lnu.se>

## A visual survey of 220 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



## Topic Model used for Exploratory Search must be...

- 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 model
- 8 **Semi-supervised**: labeling should be used to improve the model
- 9 **Online, parallel, distributed**: big data should be processed

## What is “topic”?

- *Topic* is a specific terminology of a particular domain area.
- *Topic* is a set of coherent terms (words or phrases) that often co-occur in documents.

More formally,

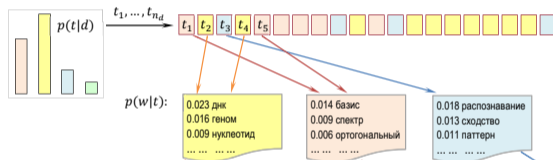
- *topic* is a probability distribution over terms:  
 $p(w|t)$  is (unknown) frequency of word  $w$  in topic  $t$ .
- *document profile* is a probability distribution over *topics*:  
 $p(t|d)$  is (unknown) frequency of topic  $t$  in document  $d$ .

When writing term  $w$  in document  $d$  author thought of topic  $t$ .

*Topic model* tries to uncover latent topics from observable terms in a text collection.

# Probabilistic Topic Model (PTM) generating a text collection

Topic model  $p(w|d) = \sum_t p(w|t)p(t|d)$  explains terms  $w$  in documents  $d$  by topics  $t$ :



$w_1, \dots, w_{n_d}$ :

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

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

under nonnegativity and normalization constraints

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

**The ill-posed problem** of matrix factorization has infinitely many solutions:

$$\Phi \Theta = (\Phi S)(S^{-1} \Theta) = \Phi' \Theta'$$

for any  $T \times T$ -matrix  $S$  such that  $\Phi', \Theta'$  are stochastic.

## PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Constrained maximization of the log-likelihood:

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

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

$$\begin{cases} \text{E-step:} & p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} \right) \\ \theta_{td} = \text{norm}_{t \in T} \left( \sum_{w \in W} n_{dw} p_{tdw} \right) \end{cases} \end{cases}$$

where  $\text{norm}_{t \in T} x_t = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$  is vector normalization,  $p_{tdw} = p(t|d, w)$ .

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

Maximum a posteriori (MAP) with Dirichlet prior:

$$\underbrace{\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td}}_{\text{log-likelihood } \mathcal{L}(\Phi, \Theta)} + \underbrace{\sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}}_{\text{regularization criterion } R(\Phi, \Theta)} \rightarrow \max_{\Phi, \Theta}$$

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

$$\begin{cases} \text{E-step:} & p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} + \beta_w \right) \\ \theta_{td} = \text{norm}_{t \in T} \left( \sum_{w \in W} n_{dw} p_{tdw} + \alpha_t \right) \end{cases} \end{cases}$$

## ARTM — Additive Regularization of Topic Model [Vorontsov, 2014]

Maximum log-likelihood with additive regularization criterion  $R$ :

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

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

$$\begin{array}{l} \text{E-step:} \\ \text{M-step:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \operatorname{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \phi_{wt} = \operatorname{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \operatorname{norm}_{t \in T} \left( \sum_{w \in W} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right.$$

## Many Bayesian PTMs can be reinterpreted as regularizers in ARTM

- smoothing for background and stop-words topics (LDA)
- **sparsing for domain-specific topics (anti-LDA)**
- topic decorrelation
- topic coherence maximization
- supervised learning for classification and regression
- semi-supervised learning
- using document citations and links
- **determining number of topics via entropy sparsing**
- modeling topical hierarchies
- modeling temporal topic dynamics
- using vocabularies in multilingual topic models
- etc.

---

Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Volume 101, Issue 1 (2015), Pp. 303-323.

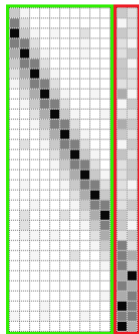


## Examples of regularization. Joint use of sparse and smoothed topics

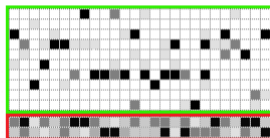
$S \subset T$ : topics of *domain-specific terminology*, with sparse and distinct  $p(w|t)$

$B \subset T$ : *background topics* of common lexis words, with dense  $p(w|t)$

$\phi_{wt}$  terms  $\times$  topics



$\theta_{td}$  topics  $\times$  documents



## Example 1. Regularizer for topic smoothing (rethinking LDA)

The **high-density assumption** for background topics  $t \in B$ :  
distributions  $\phi_{wt}$ ,  $\theta_{td}$  are similar to given distributions  $\beta_w$ ,  $\alpha_t$ .

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

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

The regularized M-step gives a non-Bayesian interpretation of LDA:

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

## Example 2. Regularizer for topic sparsing (further rethinking LDA)

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

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., Odobez 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.

### Example 3. Regularizer for topics decorrelation

**The dissimilarity assumption:**

domain-specific topics  $t \in S$  must be as distant as possible.

Maximize covariances between all pairs of column vectors  $\phi_t, \phi_s$ :

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

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.

## Example 4. Regularizer for topic selection

**Assumption:** insignificant topics are not well-interpretable.

Maximize  $\text{KL}\left(\frac{1}{|T|} \parallel p(t)\right)$  to make distribution over topics  $p(t) = \sum_d p(d)\theta_{td}$  sparse:

$$R(\Theta) = -\tau \sum_{t \in S} \ln \sum_{d \in D} p(d)\theta_{td} \rightarrow \max.$$

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

$$\theta_{td} \propto \left( n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.$$

**Effect:** if  $n_t$  is small then all values in the  $t$ -th row may turn into zeros.

---

Vorontsov K. V., Potapenko A. A., Plavin A. V. Additive regularization of topic models for topic selection and sparse factorization // SLDS 2015, Royal Holloway, University of London, UK. pp.193–202.

## Combining topic models by adding their regularizers

Maximum log-likelihood **with additive combination of regularizers**:

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{i=1}^n \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

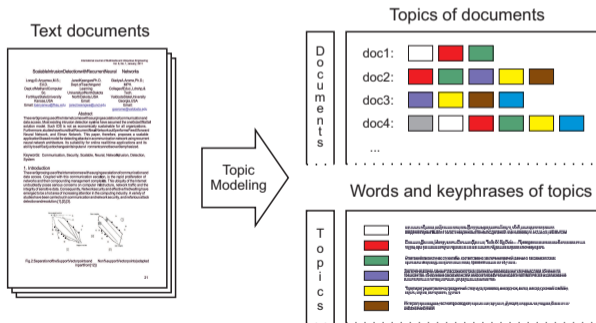
where  $\tau_i$  are regularization coefficients.

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

$$\begin{cases} \text{E-step:} & p_{tdw} = \mathop{\text{norm}}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \mathop{\text{norm}}_{w \in W} \left( \sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \sum_{i=1}^n \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left( \sum_{w \in D} n_{dw} p_{tdw} + \theta_{td} \sum_{i=1}^n \tau_i \frac{\partial R_i}{\partial \theta_{td}} \right) \end{cases} \end{cases}$$

# Multimodal Probabilistic Topic Modeling

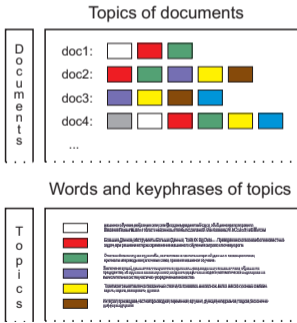
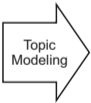
Given a text document collection *Probabilistic Topic Model* finds:  
 $p(t|d)$  — topic distribution for each document  $d$ ,  
 $p(w|t)$  — term distribution for each topic  $t$ .



# Multimodal Probabilistic Topic Modeling

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

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





# 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.

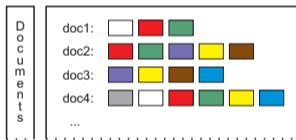
**Text documents**

Содержание  
 Abstract  
 1. Introduction

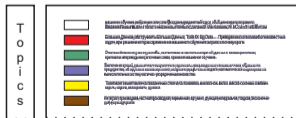
Images



Topics of documents

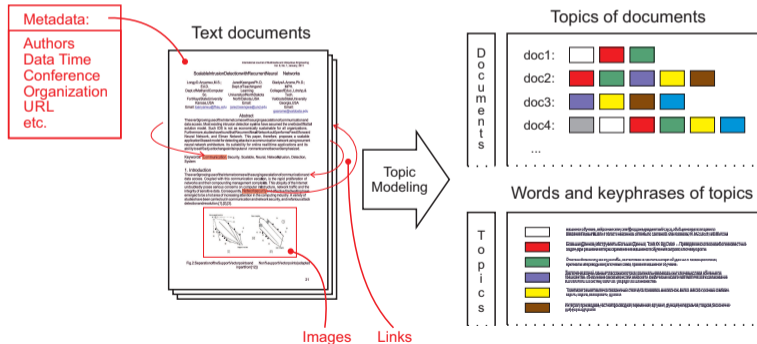


Words and keyphrases of topics



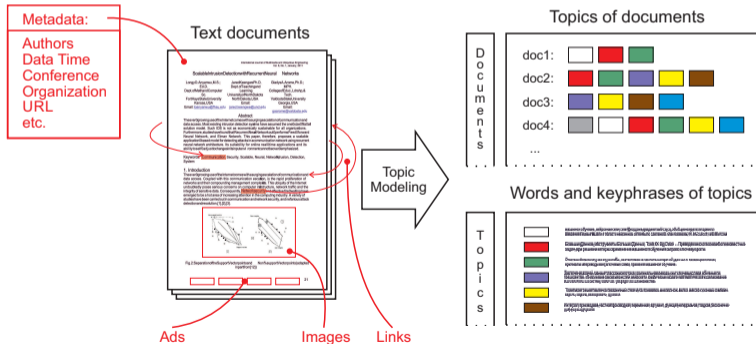
# Multimodal Probabilistic Topic Modeling

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**  $p(d'|t)$ ,



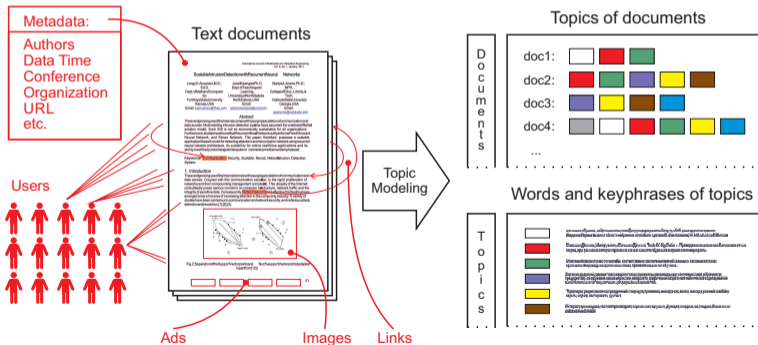
# Multimodal Probabilistic Topic Modeling

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  $p(d'|t)$ , **banner ads**  $p(b|t)$ ,



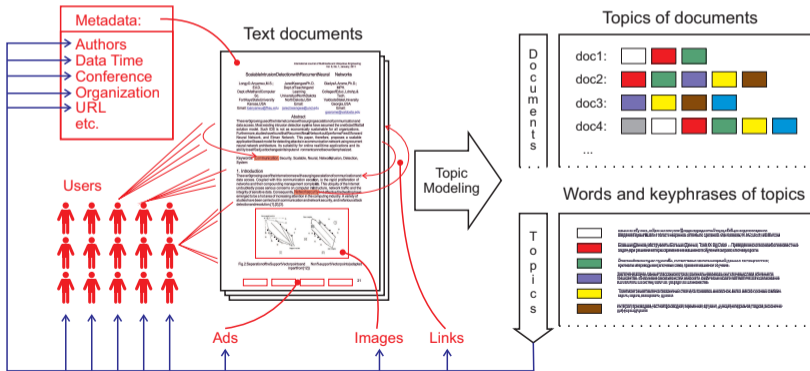
# Multimodal Probabilistic Topic Modeling

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



# 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  $p(o|t)$ , linked documents  $p(d'|t)$ , banner ads  $p(b|t)$ , users  $p(u|t)$ , and binds all these modalities into a single topic model.



## Multimodal extension of ARTM [Vorontsov, 2015]

$W^m$  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

Maximum **multimodal** log-likelihood with regularization:

$$\sum_{m \in M} \lambda_m \sum_{d \in D} \sum_{w \in W^m} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

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

$$\begin{array}{l} \text{E-step:} \\ \text{M-step:} \end{array} \left\{ \begin{array}{l} p_{tdw} = \mathop{\text{norm}}_{t \in T} (\phi_{wt} \theta_{td}) \\ \phi_{wt} = \mathop{\text{norm}}_{w \in W^m} \left( \sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \mathop{\text{norm}}_{t \in T} \left( \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right.$$

## Out-of-bag-of-words extension of ARTM [Potapenko, 2015]

Each document  $d$  is a sequence of  $n_d$  tokens:  $w_1, w_2, \dots, w_{n_d}$

Maximum log-likelihood with positional regularizers  $R_{di}$ :

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt} \theta_{td} + \sum_{d \in D} \sum_{i=1}^{n_d} R_{di}(p_{1dw_i}, \dots, p_{Tdw_i}) + R(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

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

$$\begin{cases} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}); \quad \tilde{p}_{tdw} = \frac{p_{tdw}}{n_{dw}} \sum_{i: w_i=w} \left( 1 + \frac{\partial R_{di}}{\partial p_{tdw}} - \sum_{s \in T} p_{sdw} \frac{\partial R_{di}}{\partial p_{sdw}} \right); \\ \text{M-step:} & \left\{ \begin{array}{l} \phi_{wt} = \text{norm}_{w \in W} \left( \sum_{d \in D} n_{dw} \tilde{p}_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right) \\ \theta_{td} = \text{norm}_{t \in T} \left( \sum_{w \in W} n_{dw} \tilde{p}_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right) \end{array} \right. \end{array} \right. \end{cases}$$

## BigARTM project

### BigARTM features:

- **Parallel + Online** + Multimodal + Regularized Topic Modeling
- Out-of-core one-pass 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>

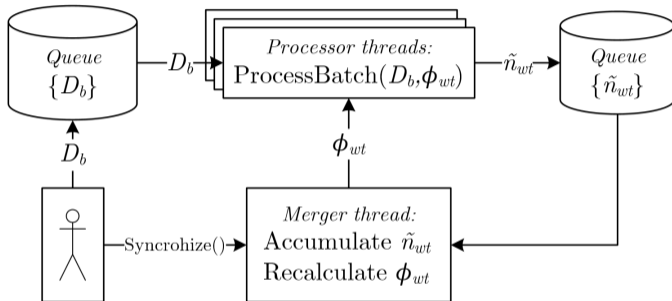


### BigARTM license and programming environment:

- 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



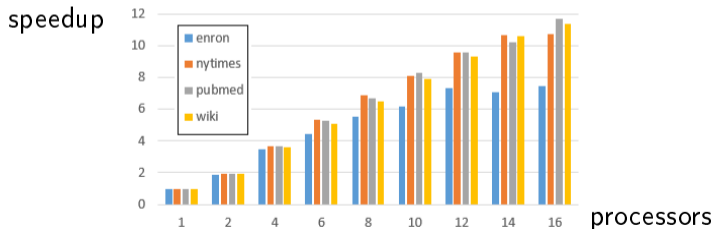
## The BigARTM project: parallel architecture



- 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

## Experiment 1: 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.

## Experiment 2: BigARTM vs Gensim vs Vowpal Wabbit

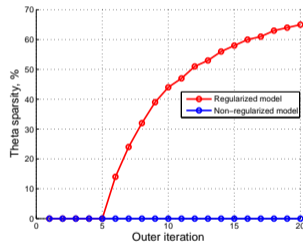
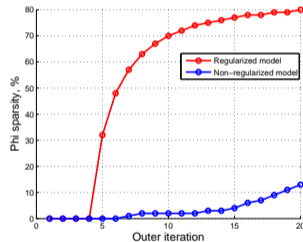
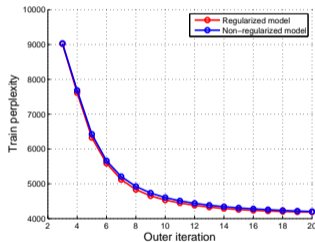
- 3.7M articles from Wikipedia, 100K unique words

	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

## Experiment 3: Running BigARTM with multiple regularizers

ARTM combines regularizers to improve sparsity without a loss of the perplexity



## Experiment 4: Hierarchical topic model for MPMR-IIP conferences

### Collection:

$|D| = 865$ ,

$|W| = 42\,000$ ,

in Russian,

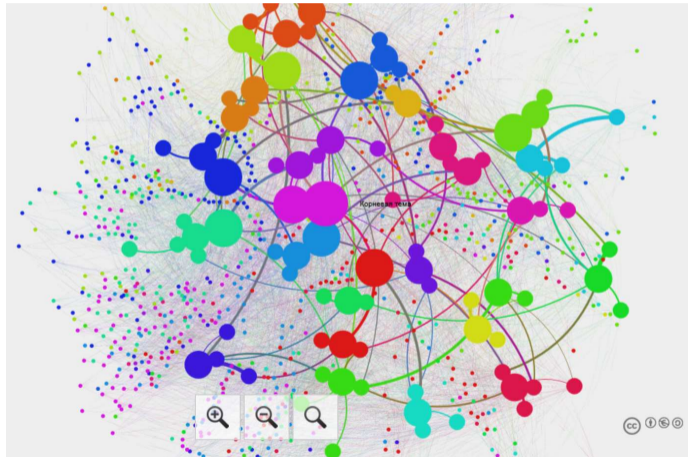
$n$ -grams

### BigARTM:

7 regularizers,

3-level hierarchy

<http://explore-mmro.ru>



## Experiment 5: The interpretability of $n$ -gram models

MMPR-IIP collection,  $|D| = 865$ , in Russian. Two modalities: unigrams & bigrams

pattern recognition in bioinformatics		optimization and computational complexity	
unigrams	bigrams	unigrams	bigrams
объект	задача распознавания	задача	разделять множества
задача	множество мотивов	множество	конечное множество
множество	система масок	подмножество	условие задачи
мотив	вторичная структура	условие	задача о покрытии
разрешимость	структура белка	класс	покрытие множества
выборка	распознавание вторичной	решение	сильный смысл
маска	состояние объекта	конечный	разделяющий комитет
распознавание	обучающая выборка	число	минимальный аффинный
информативность	оценка информативности	аффинный	аффинный комитет
состояние	множество объектов	случай	аффинный разделяющий
закономерность	разрешимость задачи	покрытие	общее положение
система	критерий разрешимости	общий	множество точек
структура	информативность мотива	пространство	случай задачи
значение	первичная структура	схема	общий случай
регулярность	тупиковое множество	комитет	задача MASC

## Experiment 6. Temporal topic model

1. Sparsing  $p(t|y) = \sum_{d \in D_y} \theta_{td} p(d)$  distributions for each time instance  $y \in Y$ :

$$R_1(\Theta) = -\tau_1 \sum_{y \in Y} \sum_{t \in T} \ln p(t|y) \rightarrow \max.$$

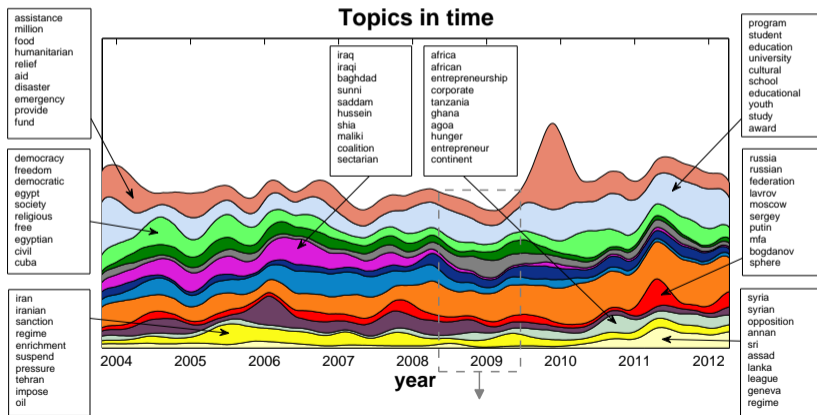
where  $D_y \subset D$  — all documents labeled by  $y$ .

2. Penalizing noisy variations of  $p(y|t)$ , a probability time series for a topic:

$$R_2(\Theta) = -\tau_2 \sum_{y \in Y} \sum_{t \in T} |p(y|t) - p(y-1|t)| \rightarrow \max.$$

## Experiment 6. Temporal topic model of political press-releases

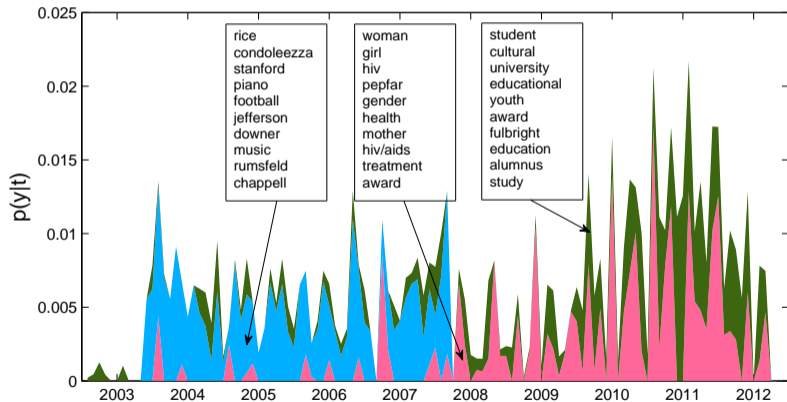
20 000 press-releases from 2003 to 2013, 180Mb. Examples of most valuable topics:





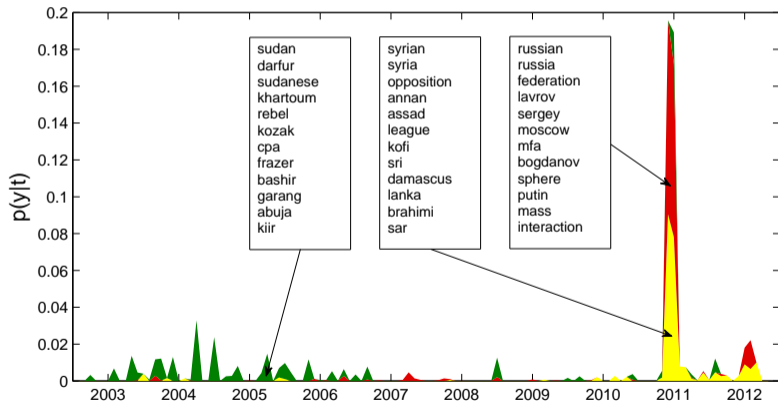
## Experiment 6. Temporal topic model of political press-releases

20 000 press-releases from 2003 to 2013, 180Mb. Examples of permanent topics:



## Experiment 6. Temporal topic model of political press-releases

20 000 press-releases from 2003 to 2013, 180Mb. Examples of **event topics**:



## Brief summary

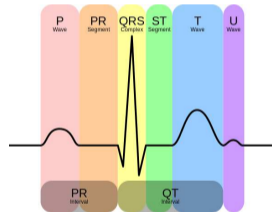
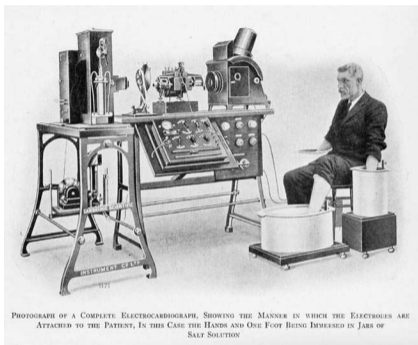
- **Exploratory Search:** a paradigm of Information Retrieval for professionals, researchers, students, and inquisitive persons
- **Multi-criteria Topic Modeling:** a way to meet multiple requirements coming from Exploratory Search
- **ARTM:** a novel non-Bayesian approach for multi-criteria optimization and combining Topic Models
- **BigARTM:** open source project for parallel online multimodal **Additively Regularized Topic Modeling** of large collections



<http://bigartm.org>

- Join BigARTM community!

# Electrocardiography



1872 — first record of the electrical activity of the heart

1911 — an early commercial ECG device (photo)

1924 — Nobel Prize in Medicine for the description of the ECG features of a number of cardiovascular disorders (Willem Einthoven)

## Theory of Information Function of the Heart (Uspenskiy, 2008)

### Assumptions:

- ECG signal carries information about the functioning of not only the heart, but all the systems of the body
- Each disease exhibits a specific modulation of the amplitudes and intervals of cardiac cycles
- This modulation can be detected at any stage of the disease including latent and preclinical stages

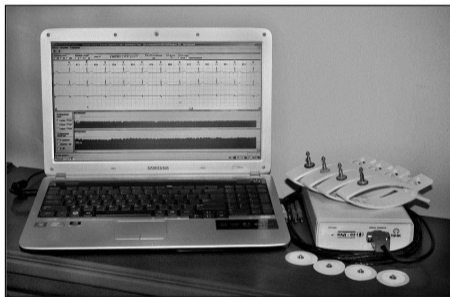
**Bold idea: early diagnosis of many diseases from one ECG**

---

V. Uspenskiy. Information Function of the Heart. *Clinical Medicine*, vol. 86, no. 5 (2008), pp. 4–13.

V. Uspenskiy. Diagnostic System Based on the Information Analysis of Electrocardiogram. *MECO 2012. Advances and Challenges in Embedded Computing* (Bar, Montenegro, June 19-21, 2012), pp. 74–76.

## Multidisease Diagnostic System «Skrinfaks»



- more than 30 years of research (from 1978)
- more than 15 years of experimental exploitation
- more than 20 000 cases (ECG record + diagnosis)
- more than 40 internal diseases can be detected

## Preprocessing step 1: Variability of R-amplitudes and RR-intervals

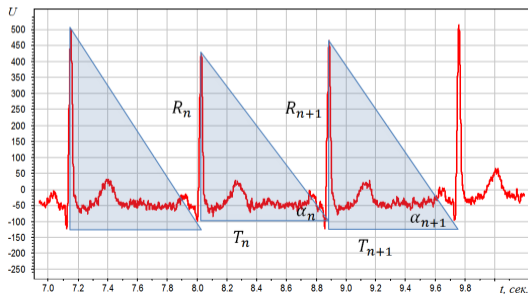
**Input:** a detailed raw ECG signal (3Mb file)

**Output:** a sequence of increment signs (225b,  $10^4$  times compression!)

amplitude  $dR_n = R_{n+1} - R_n$

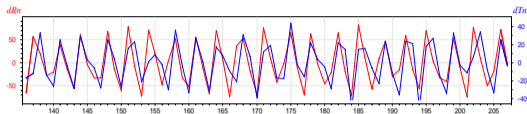
interval  $dT_n = T_{n+1} - T_n$

angle  $d\alpha_n = \alpha_{n+1} - \alpha_n$ , where  $\alpha_n = \arctg \frac{R_n}{T_n}$

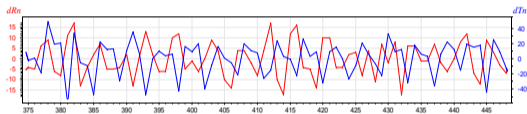


## Variability of increments $dR_n$ and $dT_n$ for ill and healthy persons

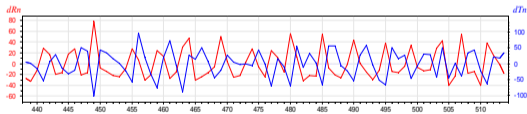
healthy:



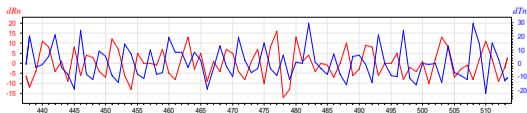
peptic ulcer:



hypertension:



cancer:





## Preprocessing step 2: Discretization and symbolic representation

**Input:** intervals and amplitudes  $(T_1, R_1), \dots, (T_N, R_N)$

**Output:** *codogram*  $x = (s_1, \dots, s_{N-1})$

is a sequence of symbols from the alphabet  $\mathcal{A} = \{A, B, C, D, E, F\}$

if  $R_n < R_{n+1}, T_n < T_{n+1}, \alpha_n < \alpha_{n+1}$  then  $s_n = A$

if  $R_n \geq R_{n+1}, T_n \geq T_{n+1}, \alpha_n < \alpha_{n+1}$  then  $s_n = B$

if  $R_n < R_{n+1}, T_n \geq T_{n+1}, \alpha_n < \alpha_{n+1}$  then  $s_n = C$

if  $R_n \geq R_{n+1}, T_n < T_{n+1}, \alpha_n \geq \alpha_{n+1}$  then  $s_n = D$

if  $R_n < R_{n+1}, T_n < T_{n+1}, \alpha_n \geq \alpha_{n+1}$  then  $s_n = E$

if  $R_n \geq R_{n+1}, T_n \geq T_{n+1}, \alpha_n \geq \alpha_{n+1}$  then  $s_n = F$

## Preprocessing step 3: Vectorization

Input: a codogram  $x = (s_1, \dots, s_{N-1})$  as a text string

```

DBEEACFDAAFBABDDAADFAFFFEACFEACFBAEFFAABFFAAFFAAFFAAFFAEAEBAEBFEAAFCAAFFAAD
FCAFFAADFCADFCDFDACFFACDFAEFFACFFEADFCABBCADFFECFFAAFFAAFFAEFFCACFCAEFFCAD
DAADBFAAFFAEBAABFACDFFAABBAADFADFAAFCECFCEDFCEECAEFBECBBBADBACFFAAFFA
CFFCECFDAAEDAEFFAFFFCEBFAAFFAEFFAEFBACFBAEDEAFFFCAFFDAAFFAEBDADBBADFDAFF
EABFCCAFDEEBDECFACFFAABFAADFBAFFACFFFAEFFACFFACFFCECFBAFFFFAAFFAAFFAADFBA
AABFACDFDAEFFAABBAEFFEAFFBCECFDECCFBAAFFAADFDACDFAFFAADFCADFAEFBAFFCADFE
AFFCECFCEFFAAFFABCFDAAFFADBFCAEFFAABFACBFABFBFAEBFAFFBAFFAAFFDADFADBFB
CAFFAECFFACFFACDFCADFDABFAREDDABBFACDDBAFAAFFCADFAADFACFFAEDFCACFCREBCE

```

Output: triplet frequency  $f_j(x)$  — how many times the triplet  $j$  appears in the codogram  $x$ ,  $j = 1, \dots, n$ ,  $n = 6^3 = 216$

1. FFA - 42	17. EFF - 10	33. CEC - 6	49. EAC - 3
2. FAA - 33	18. DAA - 10	34. ADB - 5	50. DDA - 3
3. AFF - 32	19. ECF - 9	35. FFE - 5	51. CAC - 3
4. AAF - 30	20. FFC - 9	36. EBF - 5	52. EDF - 3
5. ADF - 18	21. FEA - 9	37. CFD - 5	53. EFB - 3
6. FCA - 18	22. DFC - 8	38. AFB - 4	54. DBA - 3
7. ACF - 17	23. ABF - 8	39. AAE - 4	55. FCC - 2
8. AAD - 15	24. AAB - 8	40. CFC - 4	56. AFC - 2
9. CFF - 14	25. FCE - 8	41. CAE - 4	57. EAA - 2
10. AEF - 13	26. AEB - 7	42. DAC - 4	58. CED - 2
11. FDA - 13	27. DFD - 7	43. DBF - 4	59. CAA - 2
12. FAE - 12	28. ACD - 6	44. BFC - 4	60. BCA - 2
13. FAC - 12	29. CDF - 6	45. CFB - 4	61. BBA - 2
14. FBA - 11	30. DFA - 6	46. AED - 3	62. DFF - 2
15. BFA - 11	31. CAF - 6	47. FFF - 3	63. BDA - 2
16. BAA - 11	32. CAD - 6	48. FBC - 3	64. DAE - 2

## Machine learning step: why Topic Modeling?

Multimodal Topic Model for document classification:

- document  $\leftrightarrow$  codogram extracted from the ECG record
- modality #1: word  $\leftrightarrow$  triplet from  $\{AAA, AAB, \dots, FFF\}$
- modality #2: class  $\leftrightarrow$  disease
- topic  $\leftrightarrow$  diagnostic pattern of the class

### Healthy:

topic 1: AED, BCE, CED, DBD, DDC, EDF, EFC, FCA, FCE

topic 2: BCE, CAD, DBD, DDC, EDB, EDF, FCA, FCE

topic 3: AED, CED, DBD, DFC, EDB, EFC, FCE

### Disease (diabetes):

topic 1: AFC, CAF, AFA, FAE, AFB, BAF, BAD, EFC, EFA, CFC

topic 2: AFC, CAF, AFA, FAB, ABB, BAF, BCD, EFF

## Cross-validation experiments

Training set — for learning model parameters  $w_j$ ,  $j = 1, \dots, 216$

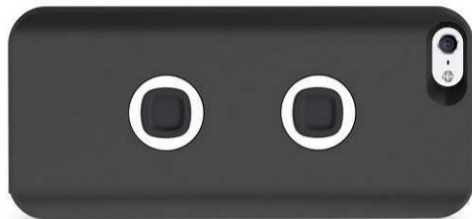
Testing set — for evaluating sensitivity, specificity and AUC

40×10-fold cross-validation to build 95% confidence intervals

disease	cases	AUC, %	spec, % (sens=95%)
femoral head necrosis	327	99.19 ± 0.10	96.6 ± 1.76
cholelithiasis	277	98.98 ± 0.23	94.4 ± 1.54
coronary heart disease	1262	97.98 ± 0.14	91.1 ± 1.86
gastritis	321	97.76 ± 0.11	88.3 ± 2.64
hypertensive disease	1891	96.76 ± 0.09	84.7 ± 1.99
diabetes	868	96.75 ± 0.19	85.3 ± 2.18
benign prostatic hyperplasia	257	96.49 ± 0.13	80.1 ± 3.19
cancer	525	96.49 ± 0.28	82.2 ± 2.38
nodular goiter thyroid	750	95.57 ± 0.16	73.5 ± 3.41
chronic cholecystitis	336	95.35 ± 0.12	74.8 ± 2.46
biliary dyskinesia	714	94.99 ± 0.16	70.3 ± 4.67
urolithiasis	649	94.99 ± 0.11	69.3 ± 2.14

CardioQVARK project

<http://cardioqvark.ru>



## A first toy experiment with CardioQVARK data

*Data from CardioQVARK database*

2611 cases of two classes:

- 26% smoke
- 74% don't smoke

*Cross-validation AUC, %:*

HRV	2T	4RT	6RTA	HRV + 6RTA
86.4	81.8	87.1	87.8	91.6

HRV: standard features from Heart Rate Variability analysis

2T: 2-symbol encoding of intervals

4RT: 4-symbol encoding of intervals and amplitudes

6RTA: 6-symbol encoding of intervals, amplitudes, and their ratios

## Brief summary

- The high-accuracy diagnostics of multiple internal diseases via a single ECG record is possible!
- A wide spread of portable devices leads to the accumulation of BigData of biomedical signals that can be used for remote health care services
- Symbolic Dynamics and Topic Modeling can be used for mining diagnostic patterns from biomedical signals

### Contacts:

Konstantin Vorontsov: [voron@yandex-team.ru](mailto:voron@yandex-team.ru)

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

-  *Hofmann T.* Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.
-  *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003. No. 3, pp. 993–1022.
-  *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.
-  *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. AIST'2014, Analysis of Images, Social networks and Texts. Springer, 2014. CCIS, Vol. 436. pp. 29–46.
-  *Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A., Yanina A. O.* Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. Topic Models: Post-Processing and Applications, CIKM 2015 TM Workshop, Melbourne, Australia.