Topic modeling as a key technology for exploratory search and social media mining

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev,
Anna Potapenko, Peter Romov

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

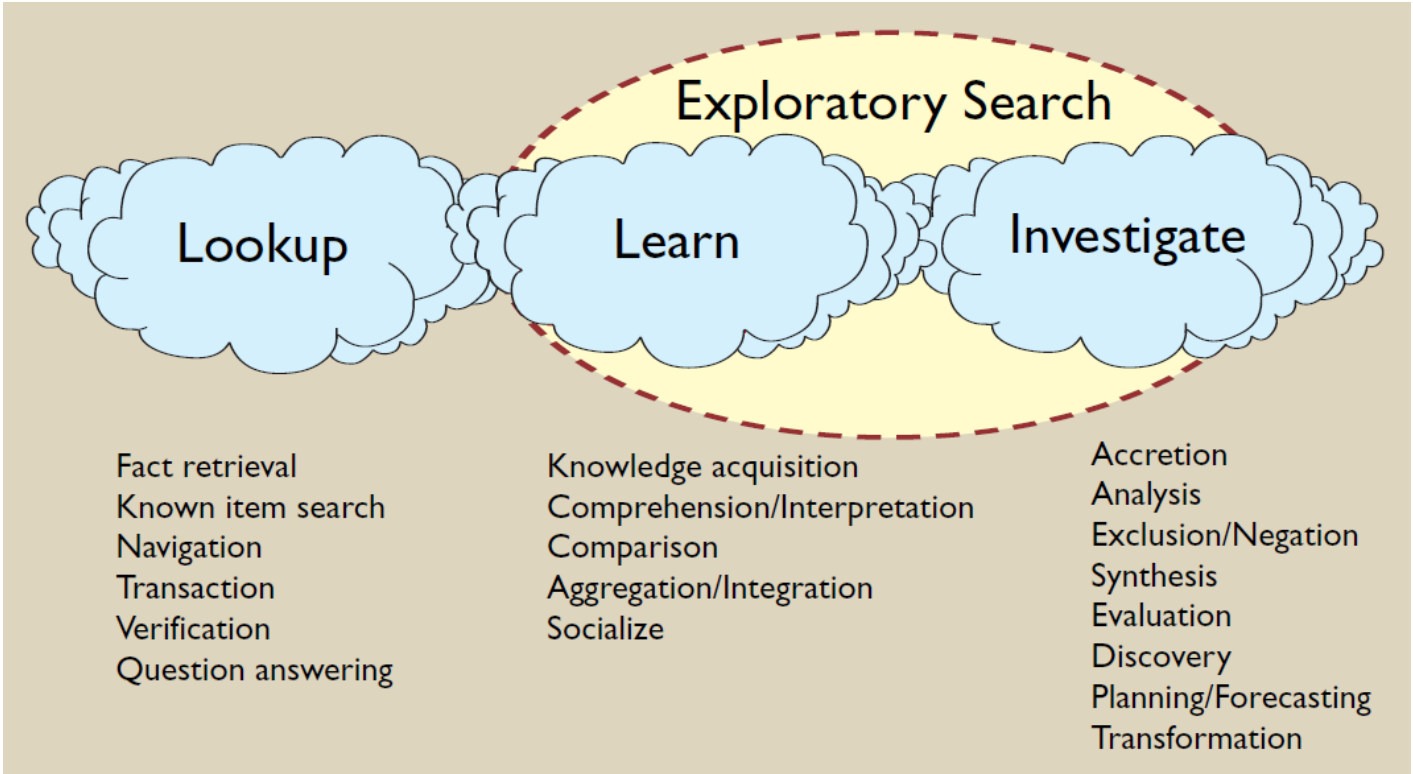
Artificial Intelligence and Natural Language &
Information Extraction, Social Media and Web Search
FRUCT Conference

St-Petersburg, Russia • 9–14 November 2015

Contents

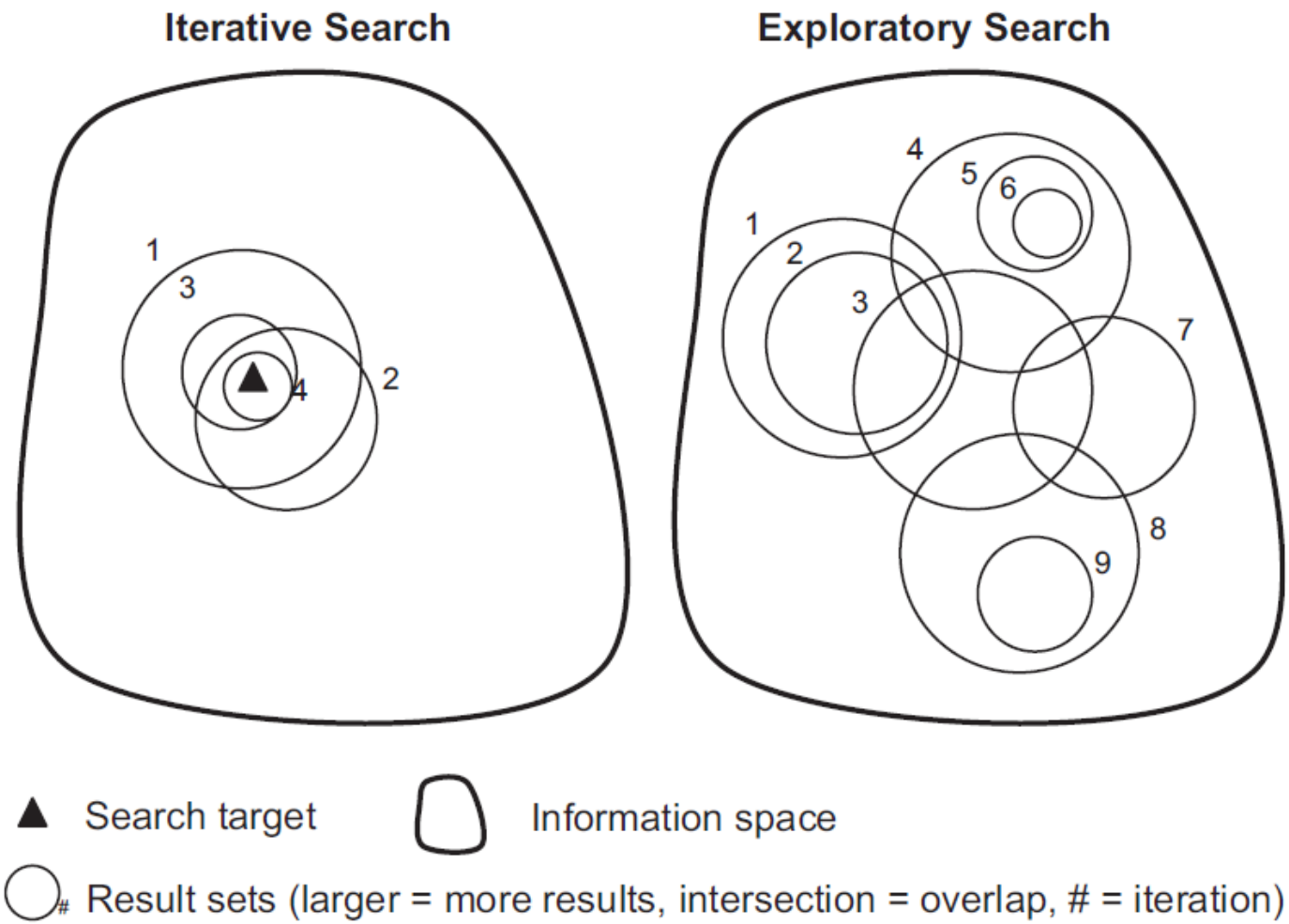
Exploratory Search for learning, knowledge acquisition and discovery

- 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:

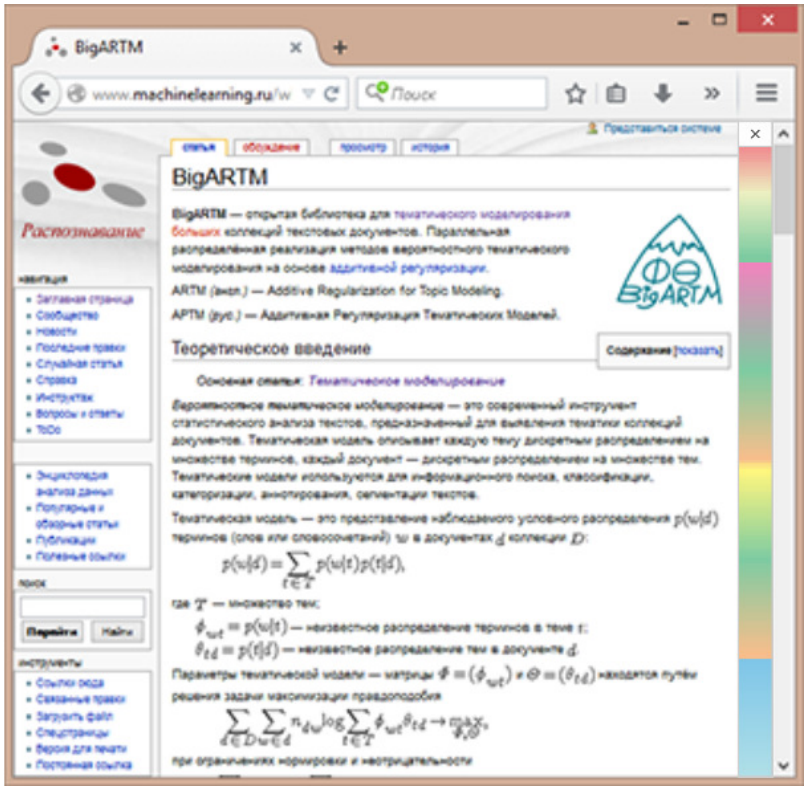
- what topics does it contain?
- what else is known on these topics?
- what is the structure of this domain area?
- what is most important, useful, popular, recent here?

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

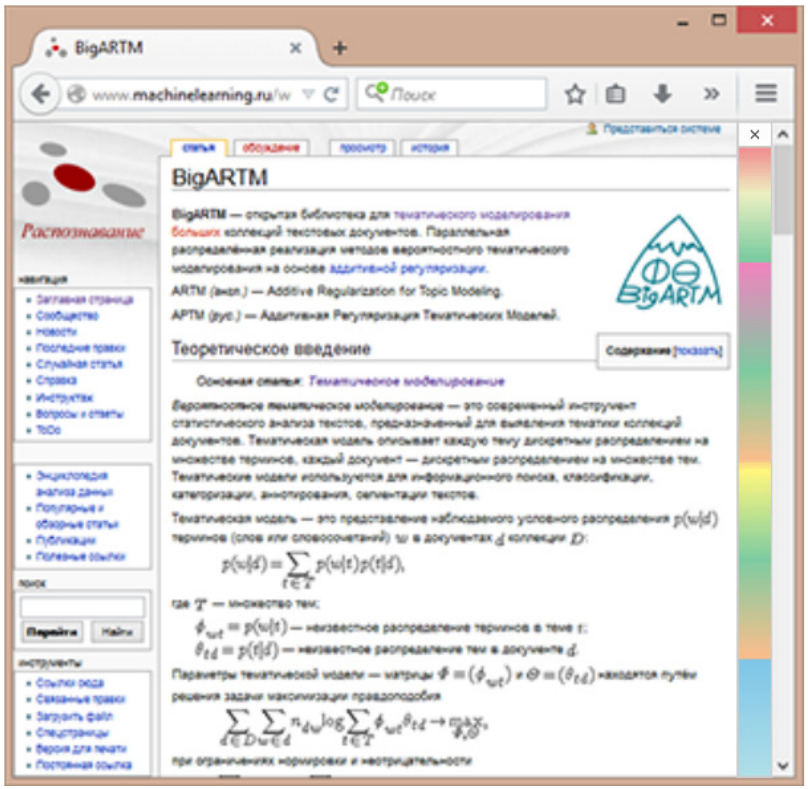
Exploratory search: the prototype of graphical user interface

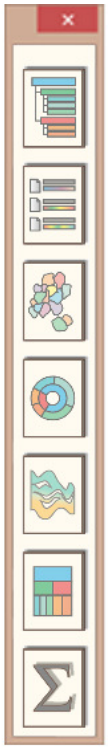
Color topic bar is a starting GUI element for exploratory search



Exploratory search: the prototype of graphical user interface

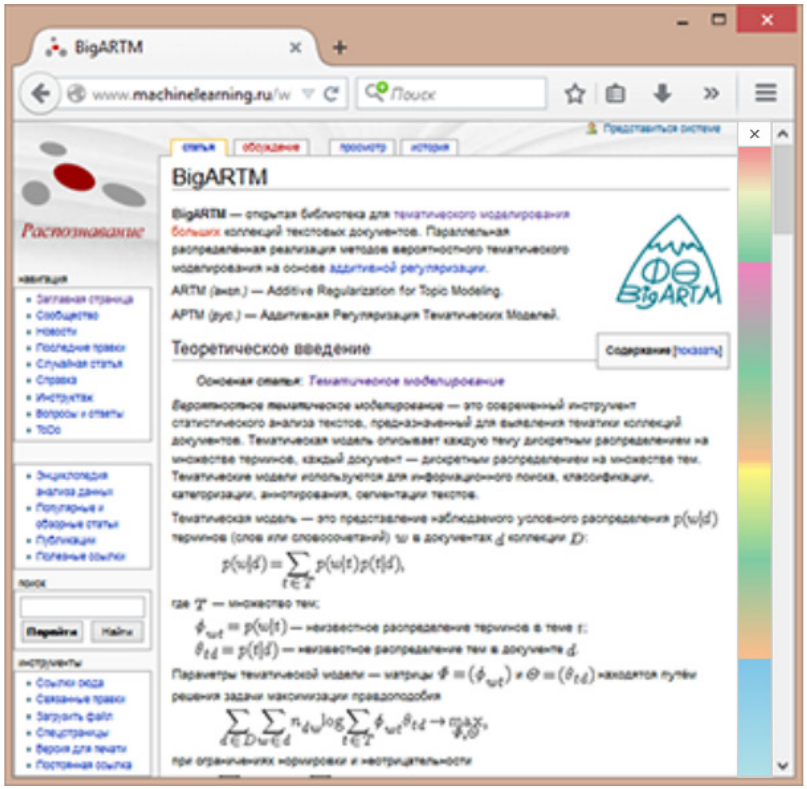
Click on the color topic bar is a topic query

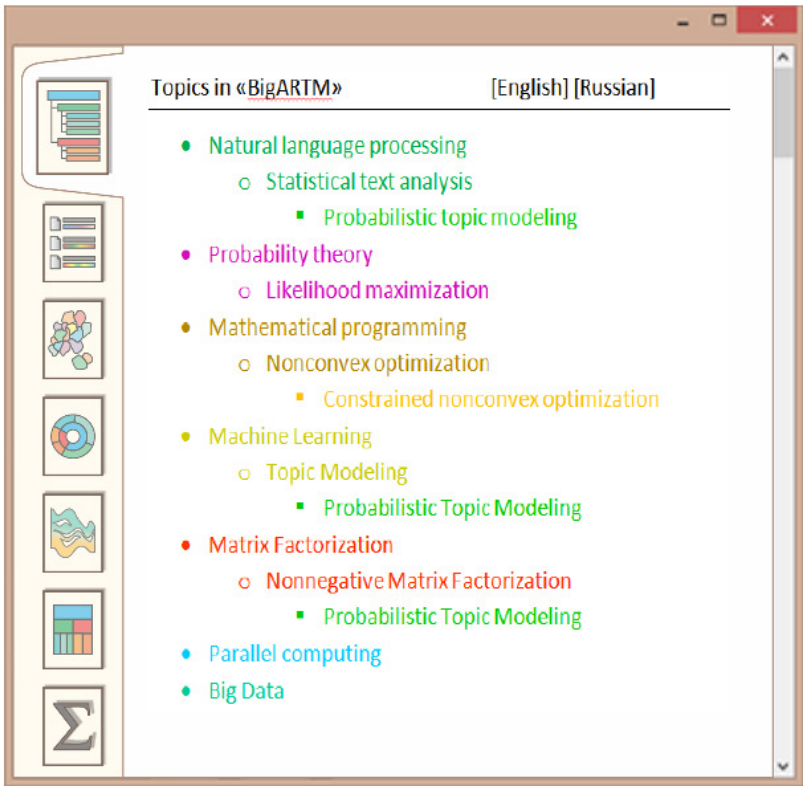




Exploratory search: the prototype of graphical user interface

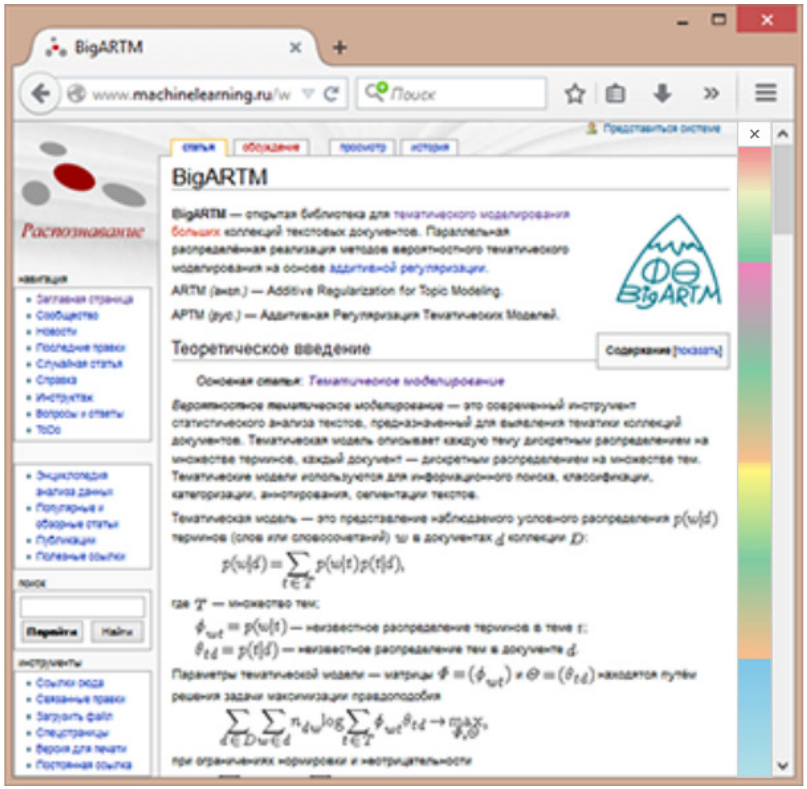
Topics of the query document

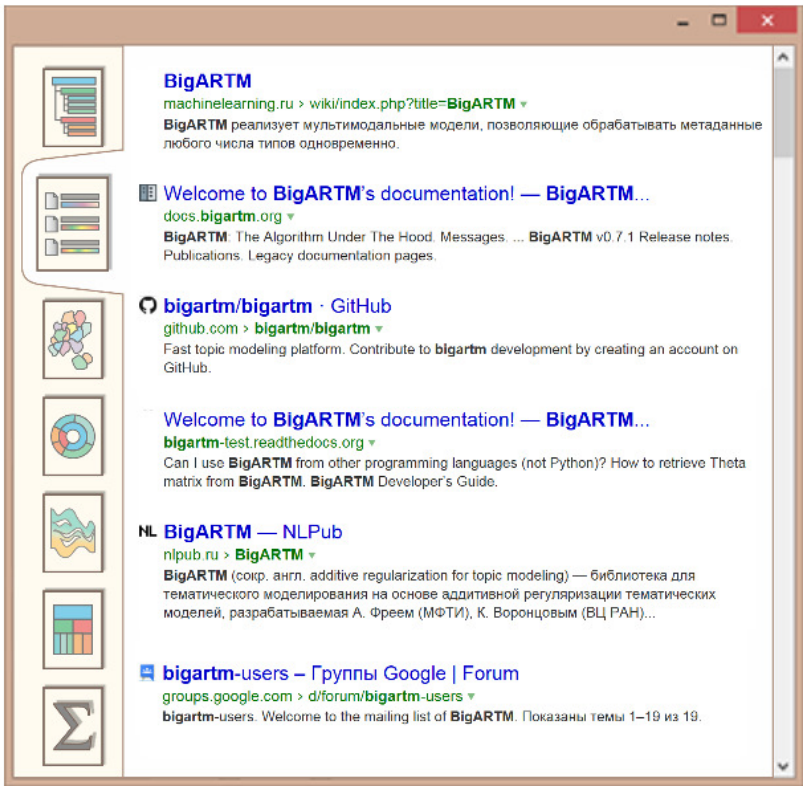




Exploratory search: the prototype of graphical user interface

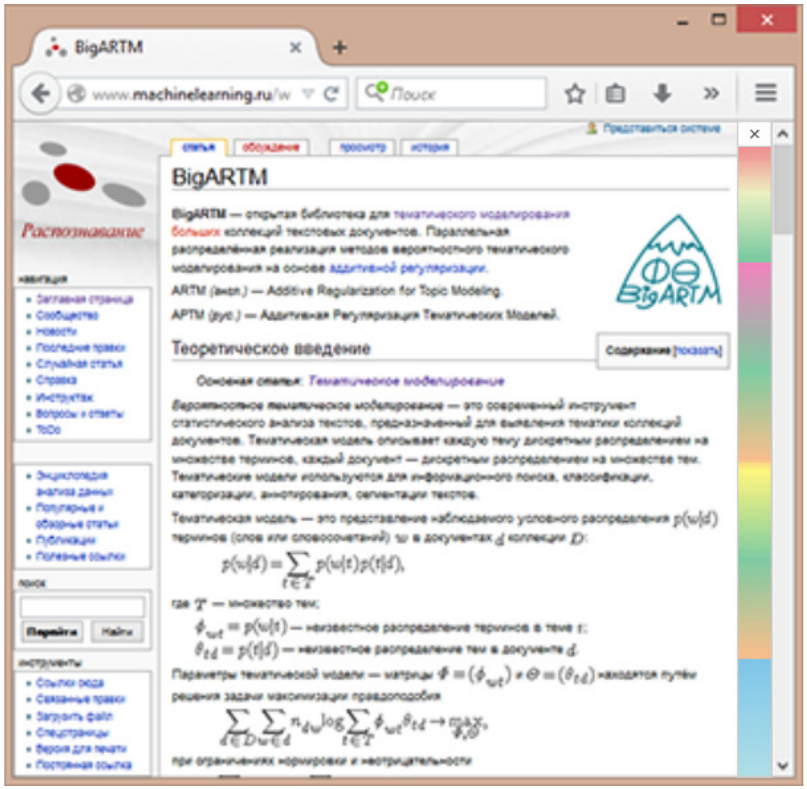
Similar documents and objects ranked by relevance

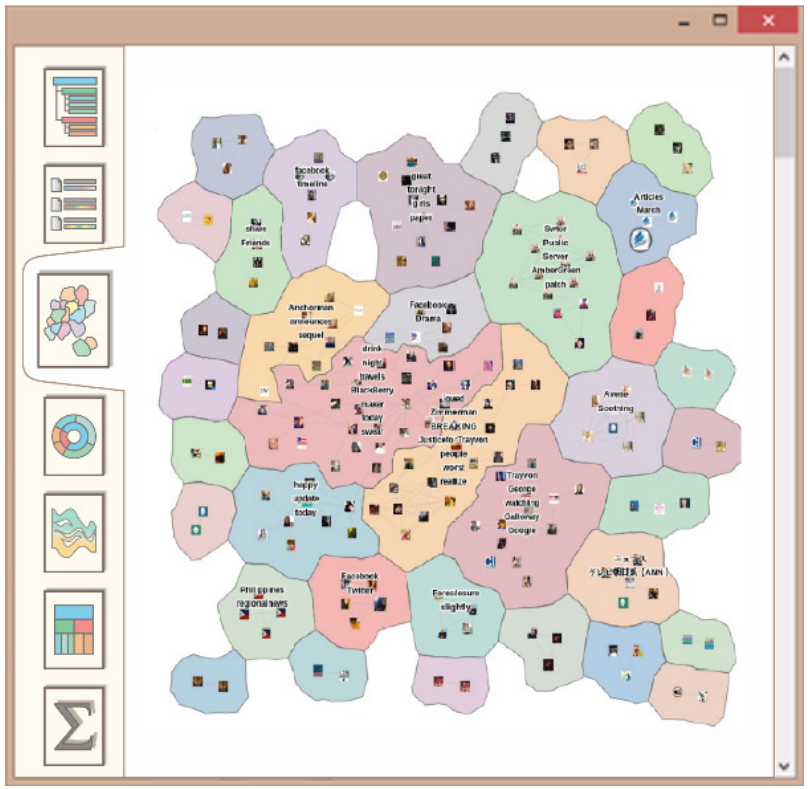




Exploratory search: the prototype of graphical user interface

Topic roadmap: clustering of relevant documents

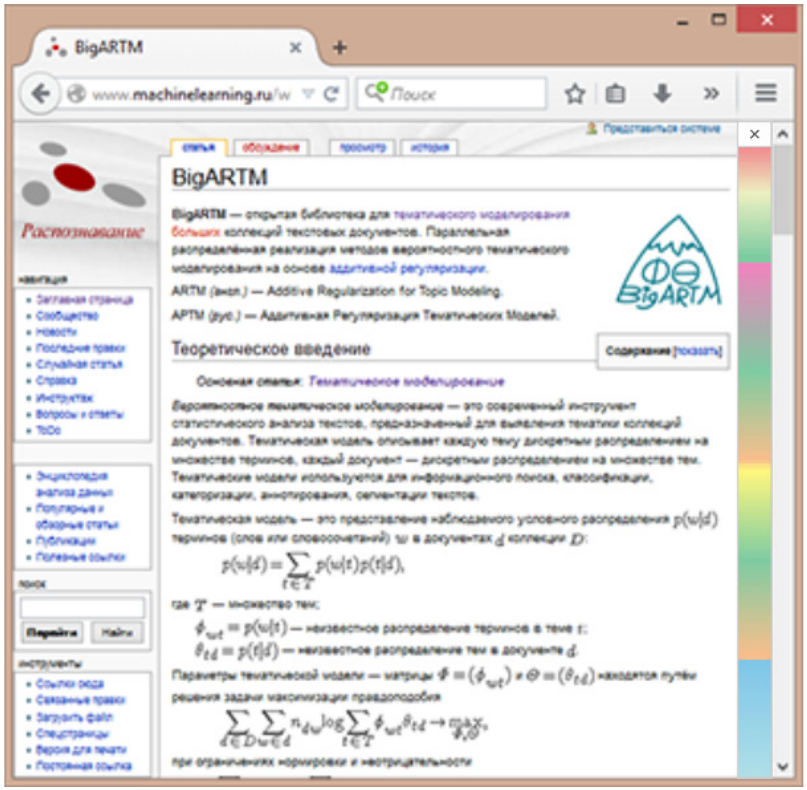


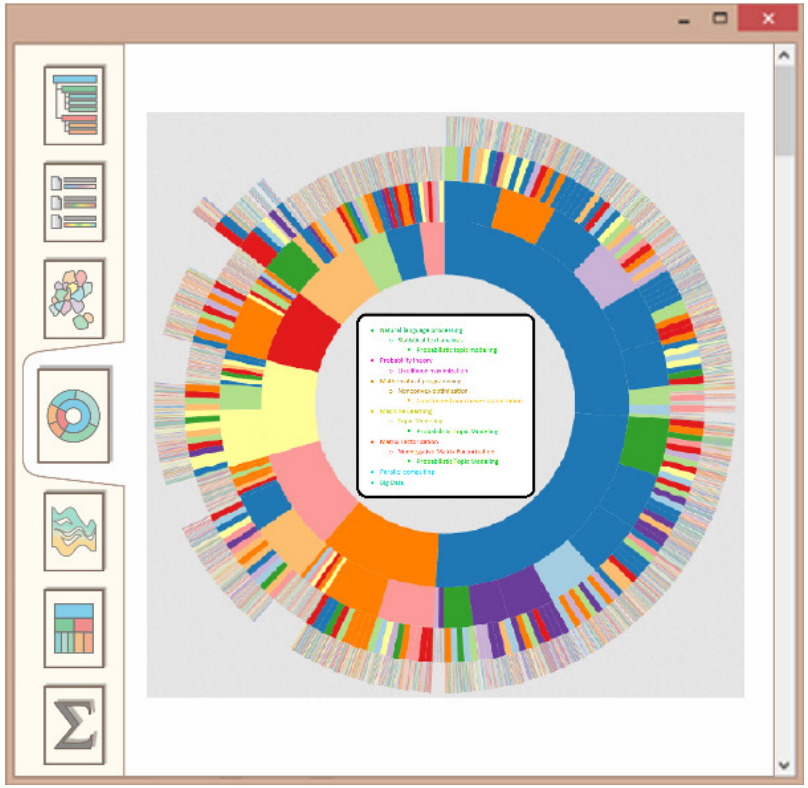


E.R.Gansner, Y.Hu, S.North. Visualizing Streaming Text Data with Dynamic Maps. 2012.

Exploratory search: the prototype of graphical user interface

Topic hierarchy: topical structure of the domain area

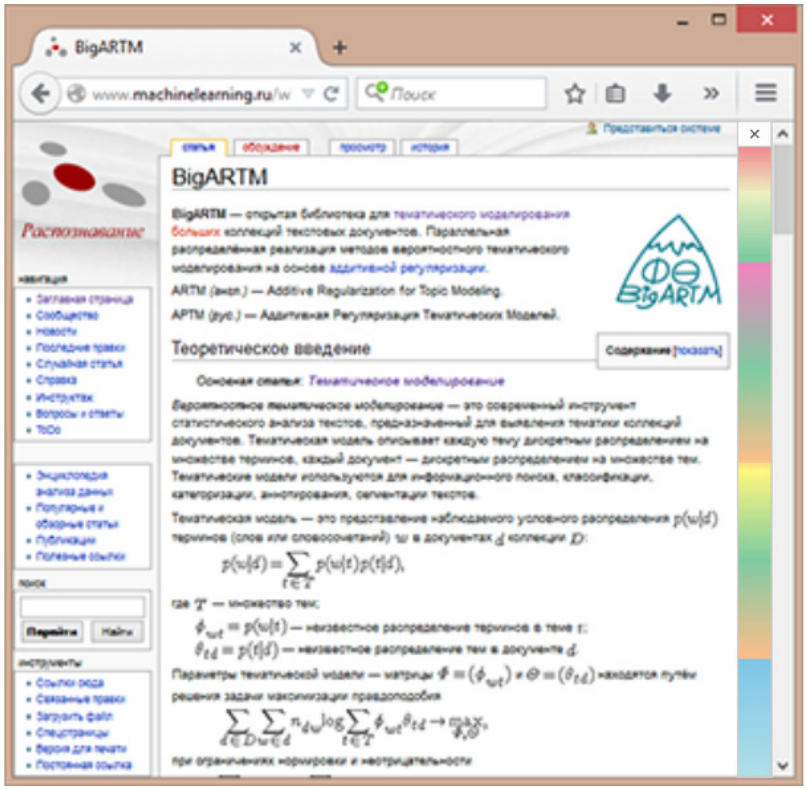


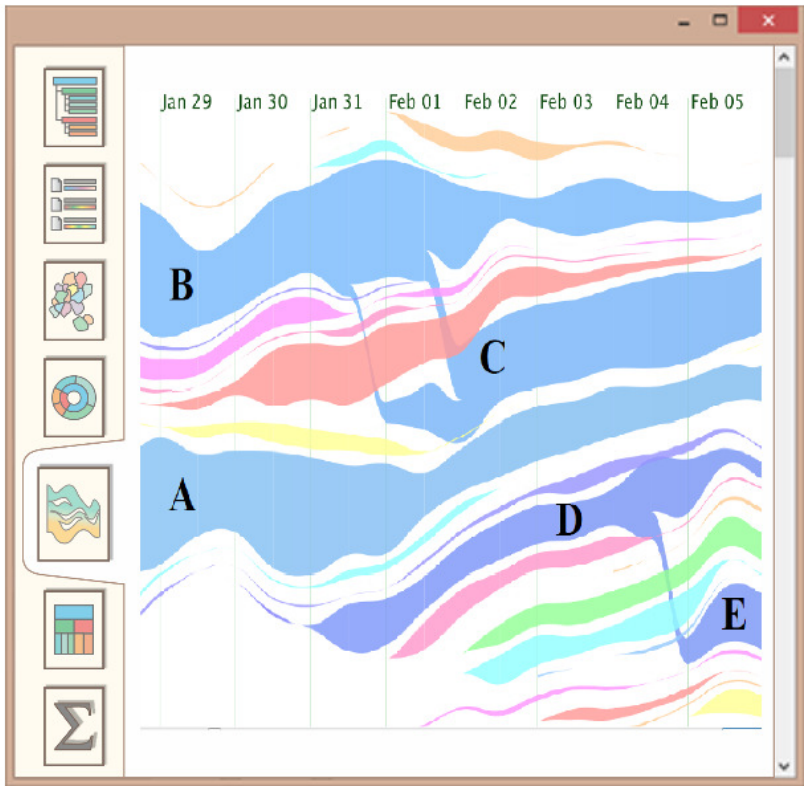


Smith A., Hawes T., Myers M.. Hiérarchie: interactive visualization for hierarchical topic models. Workshop on Interactive Language Learning, Visualization, and Interfaces, ACL, 2014.

Exploratory search: the prototype of graphical user interface

Topic river: evolution of the domain area

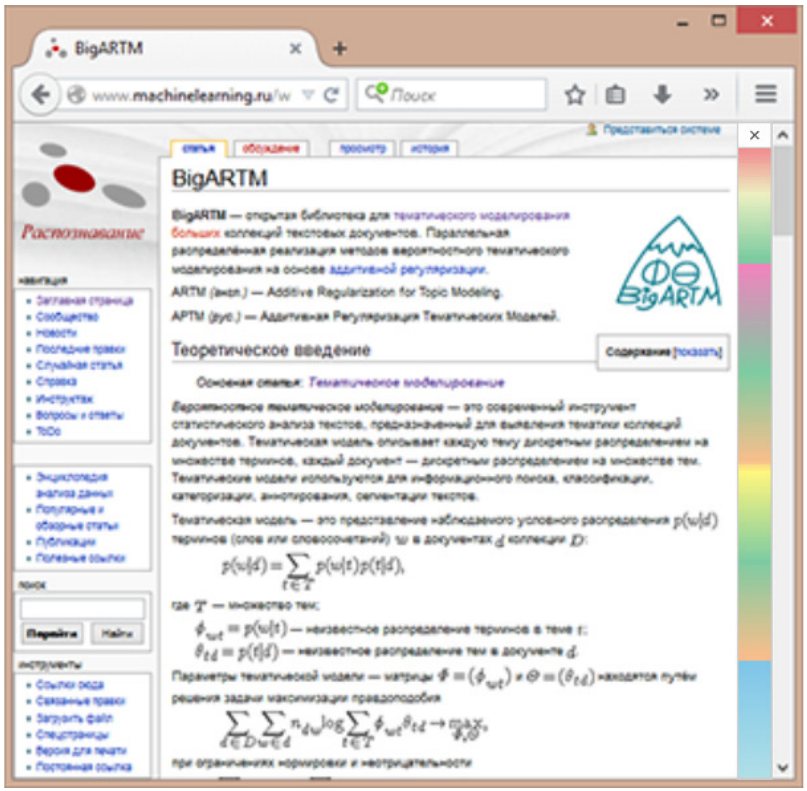


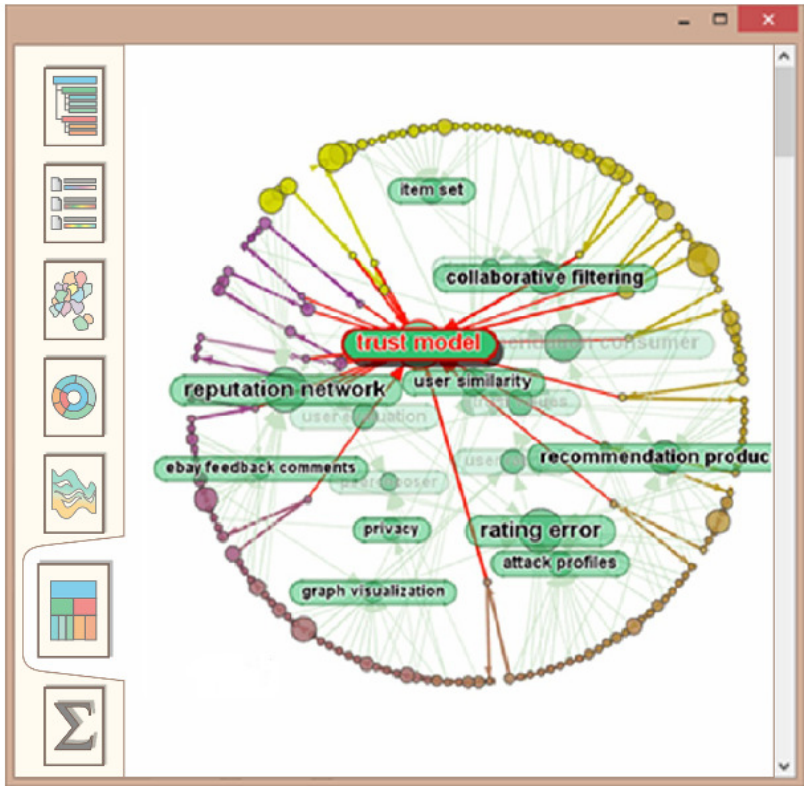


Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. IEEE Trans. Vis. Comput. Graph. 2014.

Exploratory search: the prototype of graphical user interface

Topic bar: segmentation of the query document

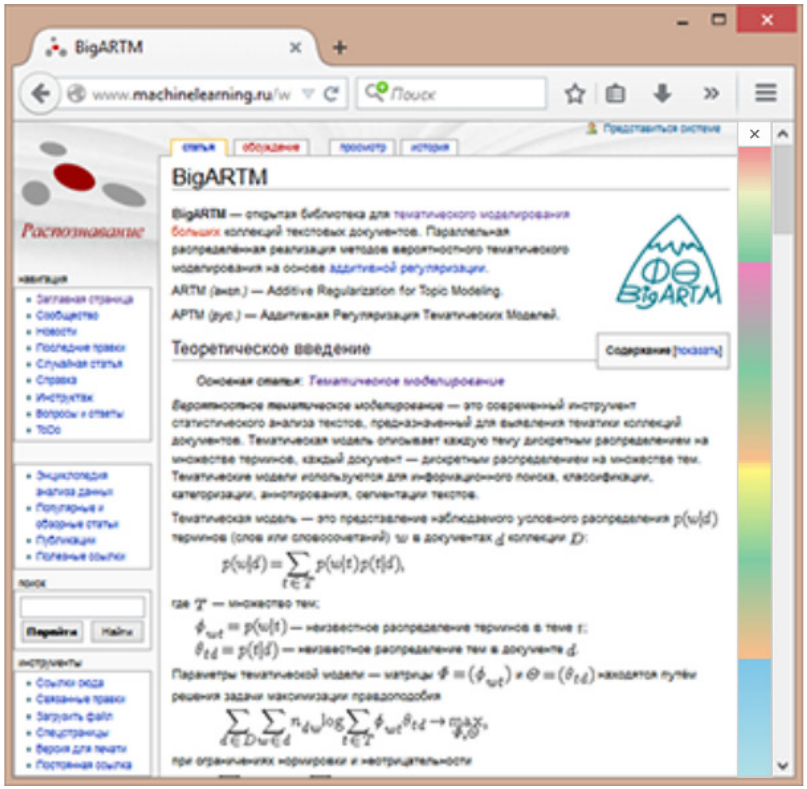


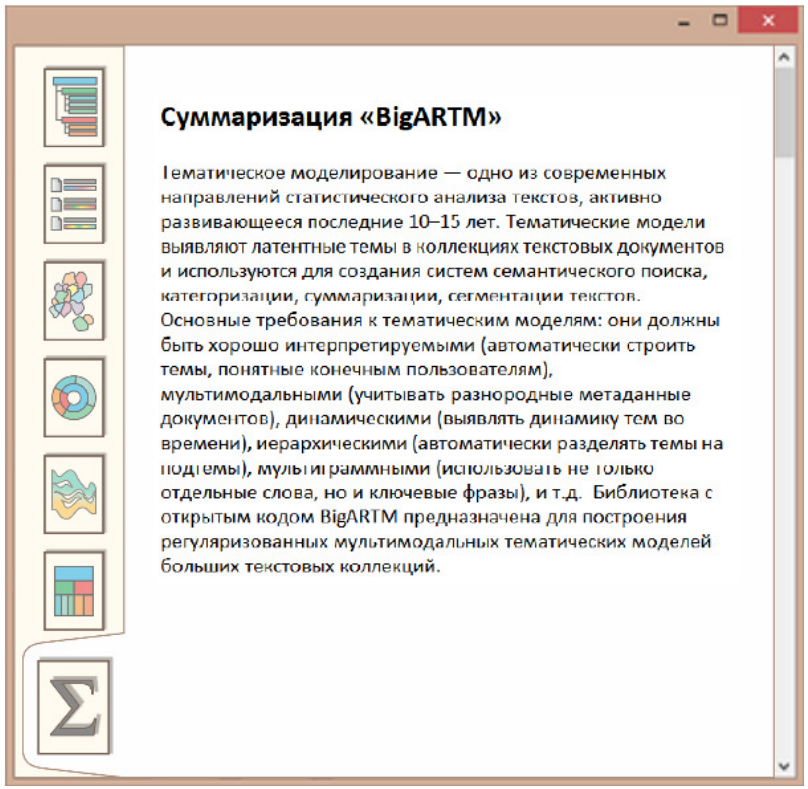


Gretarsson B., O'Donovan J., Bostandjiev S., Hollerer T., Asuncion A., Newman D., Smyth P. TopicNets: visual analysis of large text corpora with topic modeling. ACM Trans. on Intelligent Systems and Technology. 2012.

Exploratory search: the prototype of graphical user interface

Summarization of the query document





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,

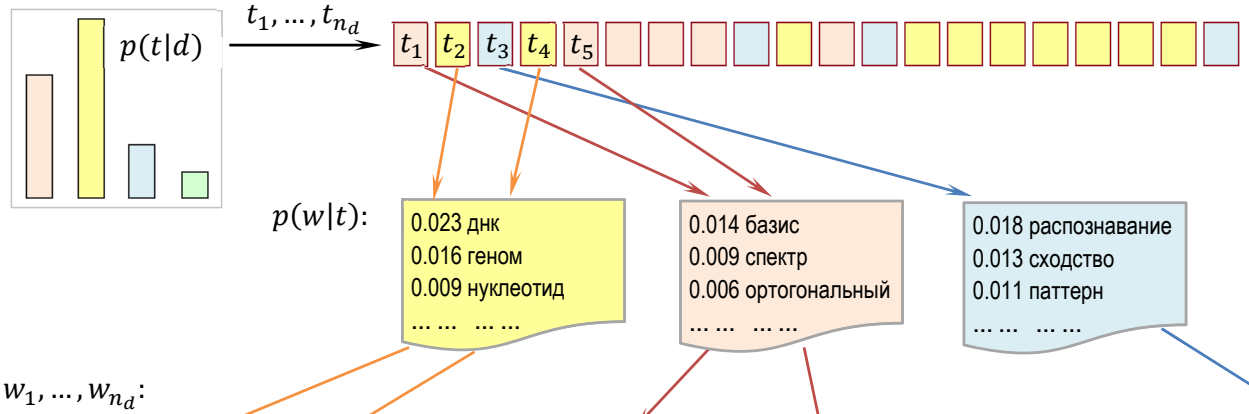
- *topic* is a probability distribution over terms: $p(w|t)$ is (unknown) frequency of word w in topic t.
- *document profile* is a probability distribution over *topics*: $p(t|d)$ is (unknown) frequency of topic t in document d.

When writing term w in document d author thought of topic t.
Topic model tries to uncover latent topics in a text collection.

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 дупликаций 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 → 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

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

The ill-posed problem of matrix factorization:

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

for all S such that Φ', Θ' are stochastic.

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Constrained maximization of the log-likelihood:

$$\mathscr{L}(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_{t} \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}$$

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

$$\begin{cases} \text{E-step:} & p_{tdw} \equiv p(t|d,w) = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right) \\ \text{M-step:} & \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw}\right) \\ & \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw}\right) \end{cases}$$

where $\underset{t\in T}{\text{norm}}\, x_t = \dfrac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ is vector normalization.

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

Maximum a posteriori (MAP) with Dirichlet prior:

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

EM-algorithm is a simple iteration method for the system

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

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

Maximum log-likelihood with additive regularization criterion R:

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

EM-algorithm is a simple iteration method for the system

$$\text{E-step:} \quad p_{tdw} = \underset{t\in T}{\text{norm}}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-step:} \quad \begin{cases} \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum\limits_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\dfrac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{t\in T}{\text{norm}}\left(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\dfrac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

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.

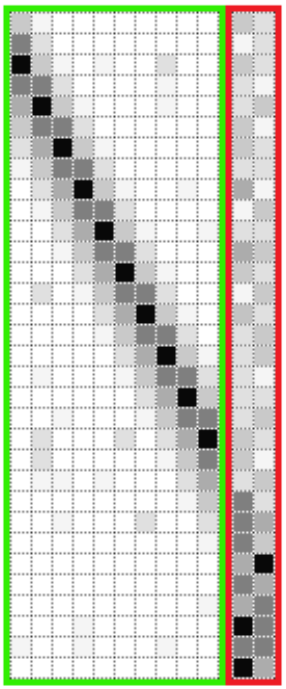
Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Volume 101, Issue 1 (2015), Pp. 303-323.

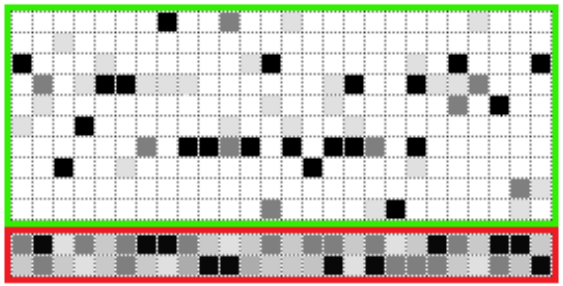
Assumptions: what topics would be well-interpretable?

Specific topics S contain domain-specific terms,
$p(w|t)$ are sparse and decorrelated, $p(t|d)$ are sparse.

Background topics B contain common lexis words,
$p(w|t)$ and $p(t|d)$ are dense.

ϕ_{wt} terms×topics θ_{td} topics×documents





Smoothing regularization (rethinking LDA)

The non-sparsity assumption for background topics $t \in B$:
ϕ_{wt} are similar to a given distribution β_w;
θ_{td} are similar to a given distribution α_t.

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

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

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

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

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

David M. Blei. Probabilistic topic models // Communications of the ACM, 2012. Vol. 55, No. 4., Pp. 77–84.

Sparsing regularizer (further rethinking LDA)

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

Maximize the sum of KL-divergences $\mathrm{KL}(\beta\|\phi_t)$ and $\mathrm{KL}(\alpha\|\theta_d)$:

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

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

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

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.

Regularization for topics decorrelation

The dissimilarity assumption:
domain-specific topics $t \in S$ must be as distant as possible.

Maximize covariances between column vectors ϕ_t:

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

The regularized M-step makes columns of Φ 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.

Regularization for topic selection

Assumption: infrequent topics are not well-interpretable.

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

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

The regularized M-step formula results in Θ rows sparsing:

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

Effect: if n_t is small then 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:

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

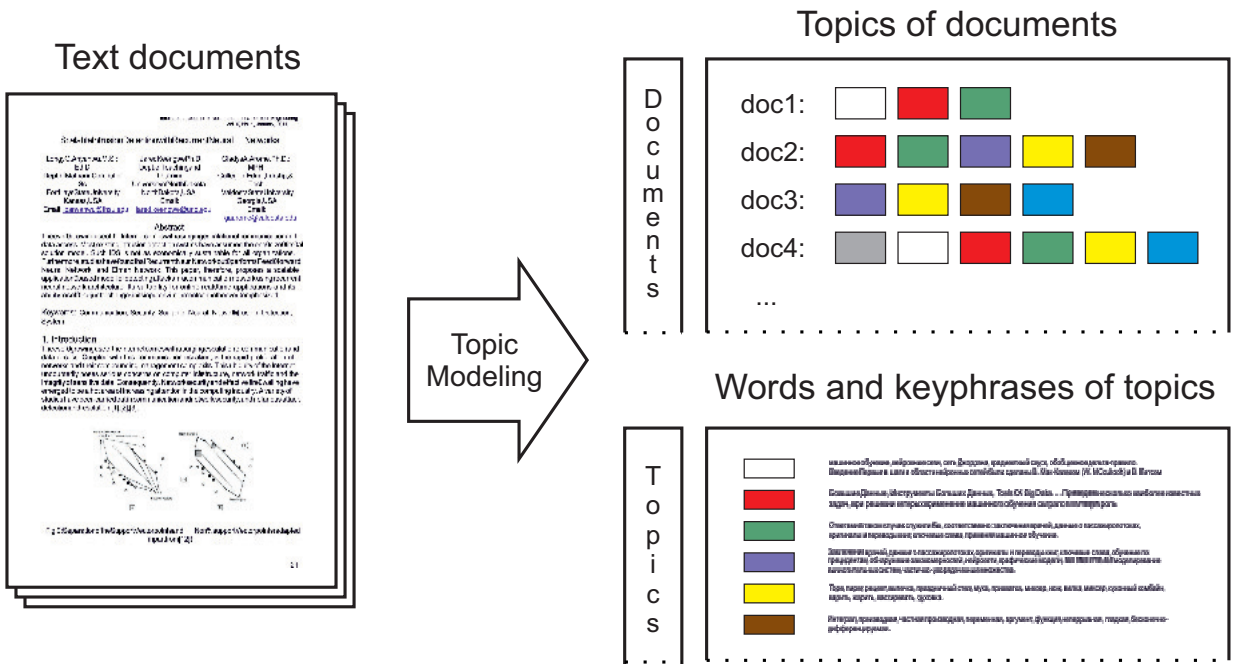
where τ_i are regularization coefficients.

EM-algorithm is a simple iteration method for the system

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

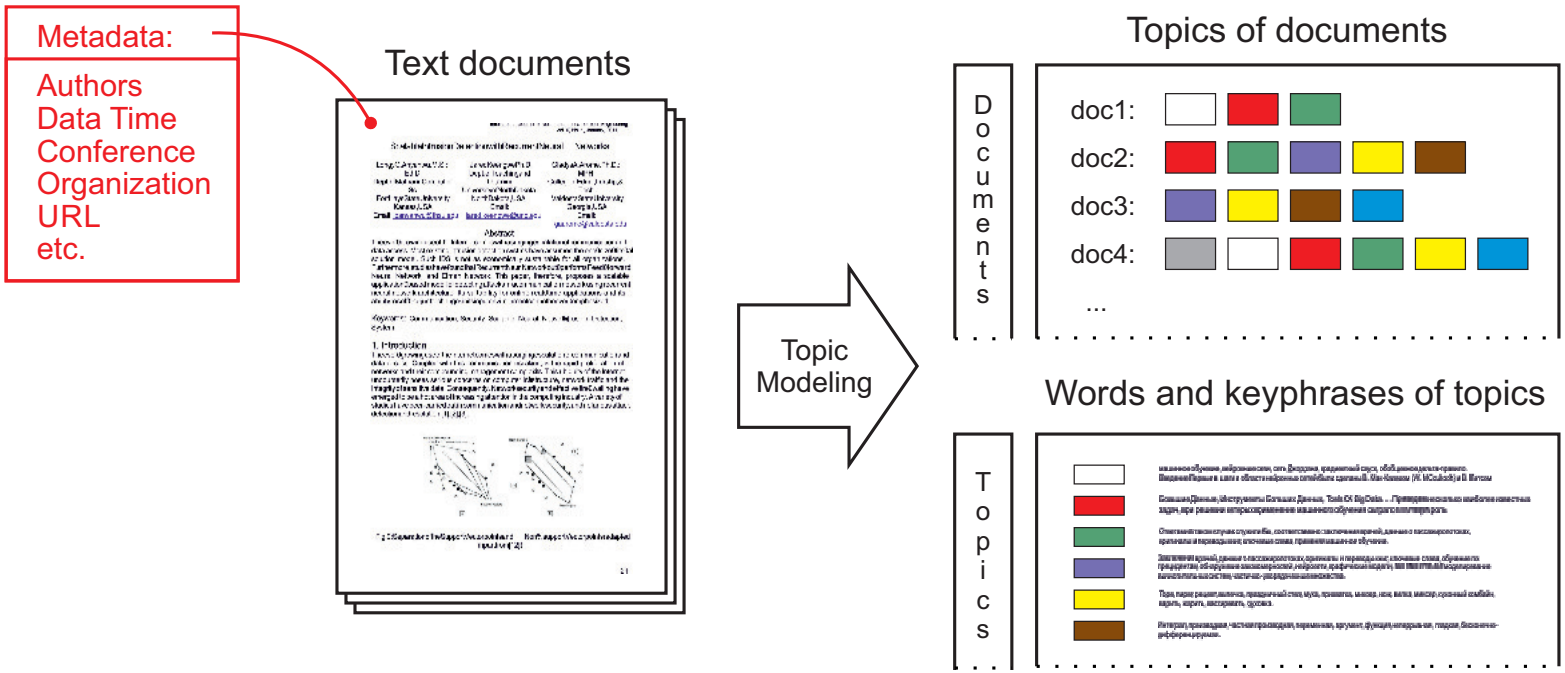
Multimodal Probabilistic Topic Modeling

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.



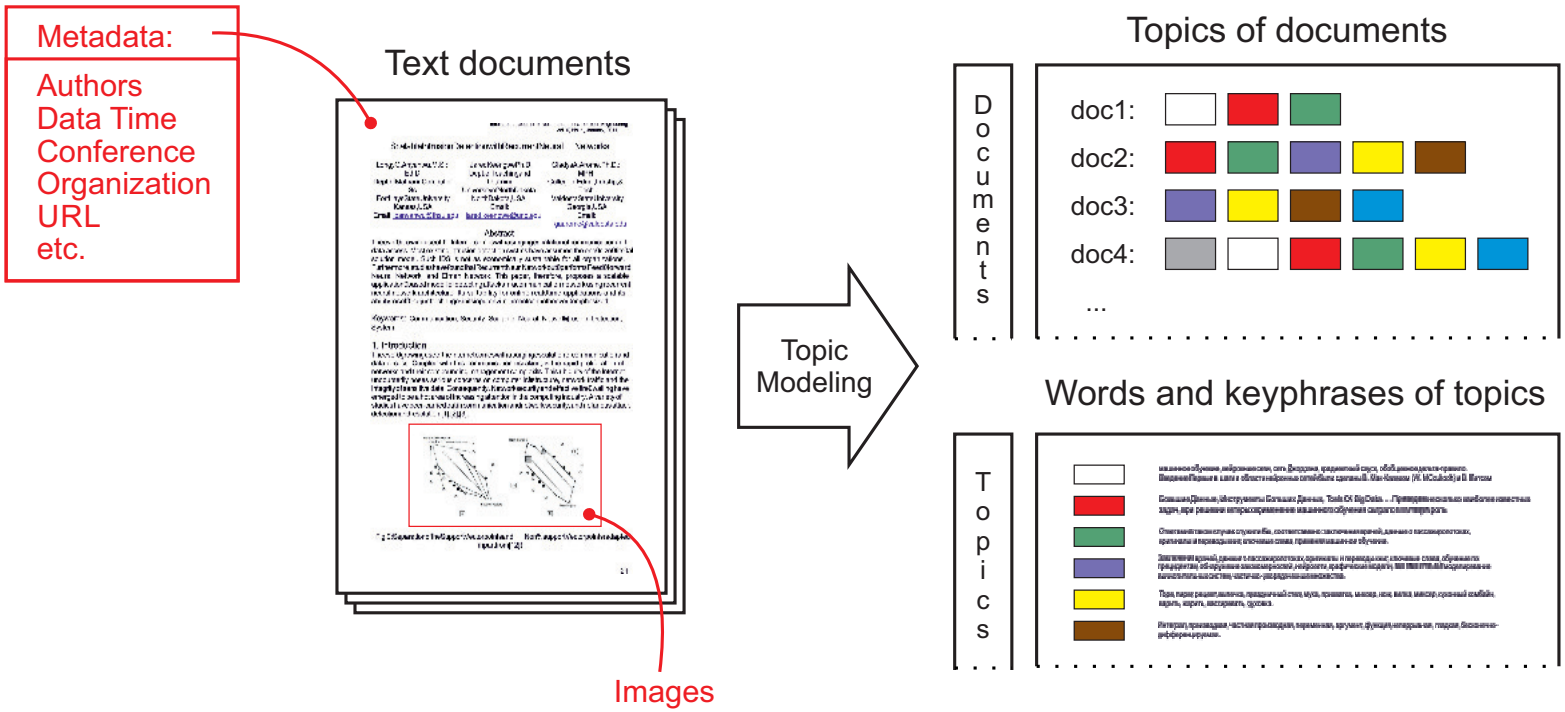
Multimodal Probabilistic Topic Modeling

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



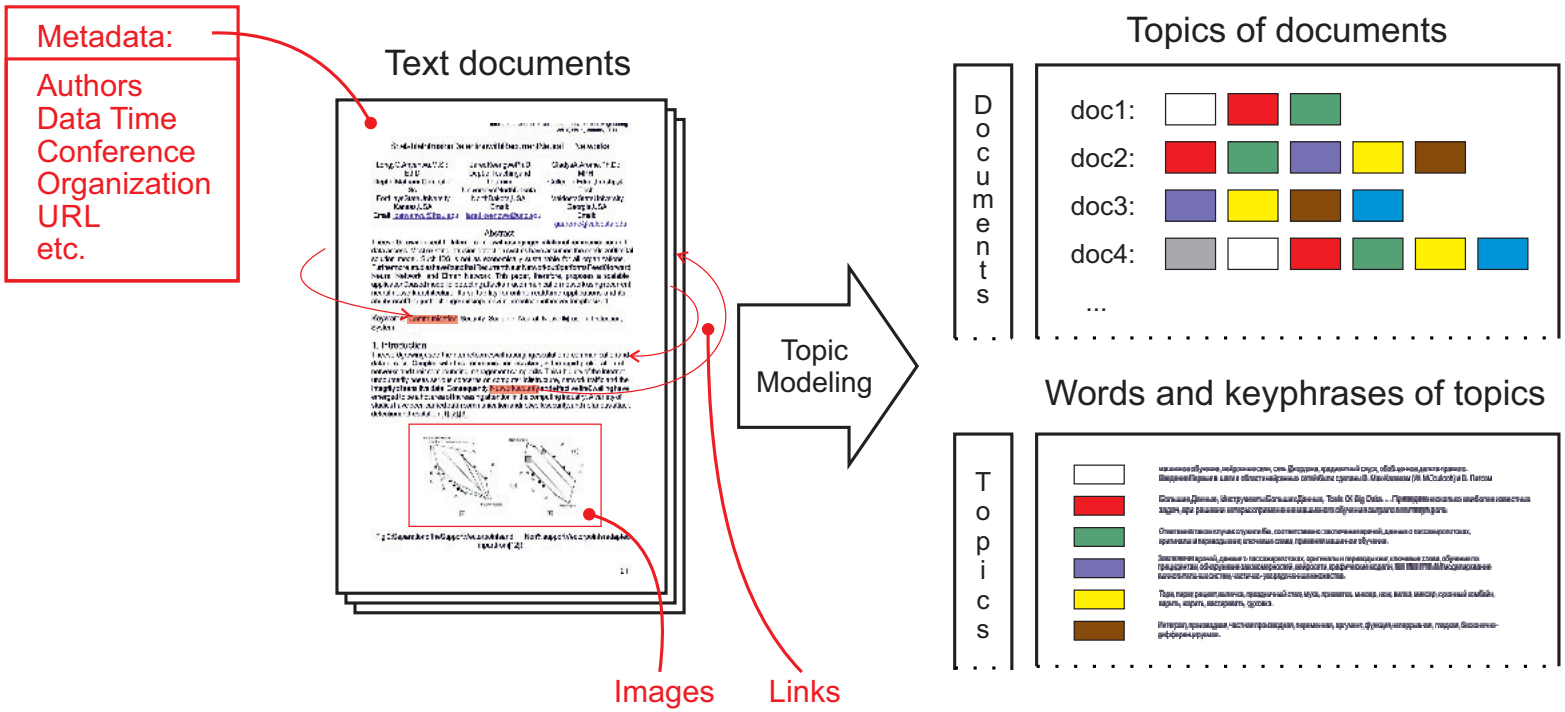
Multimodal Probabilistic Topic Modeling

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



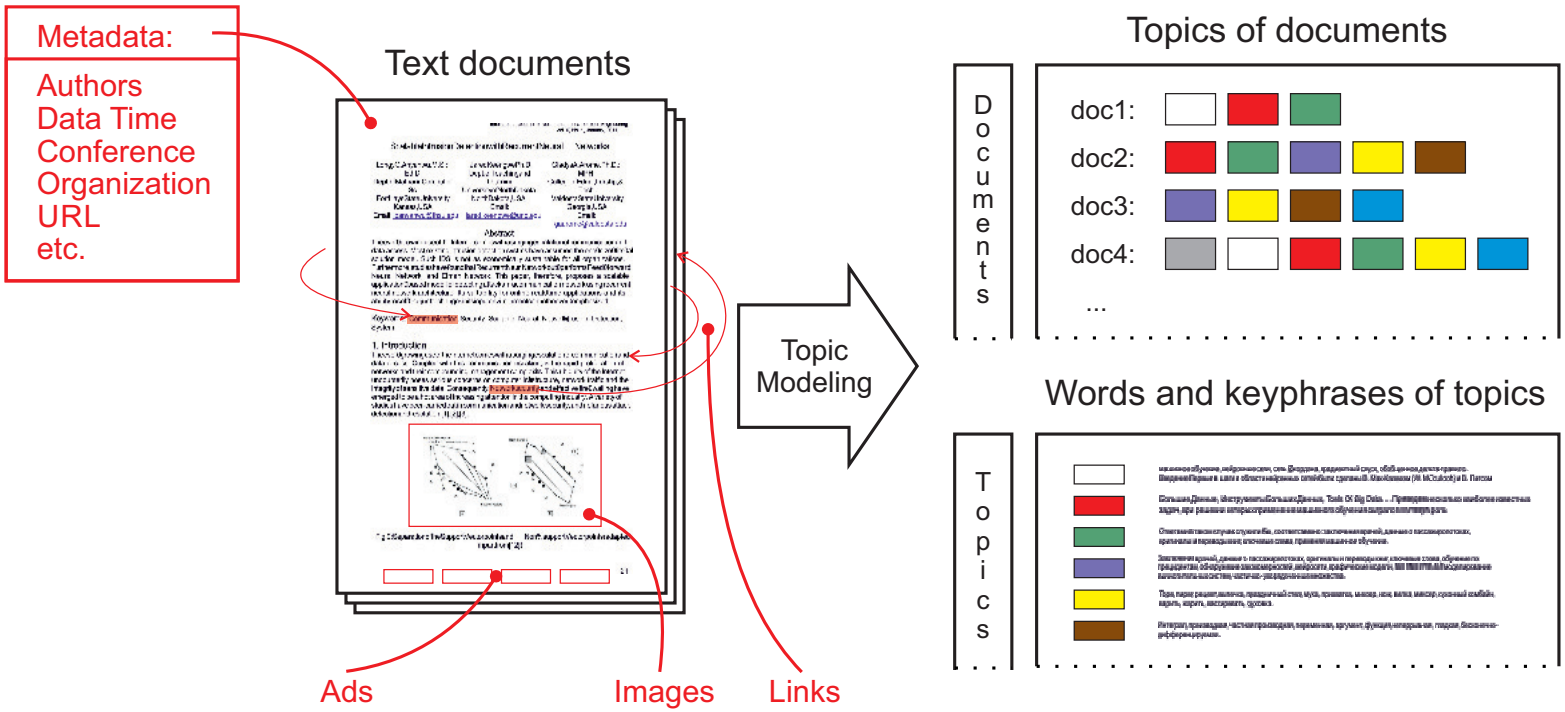
Multimodal Probabilistic Topic Modeling

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 $p(d'|t)$,



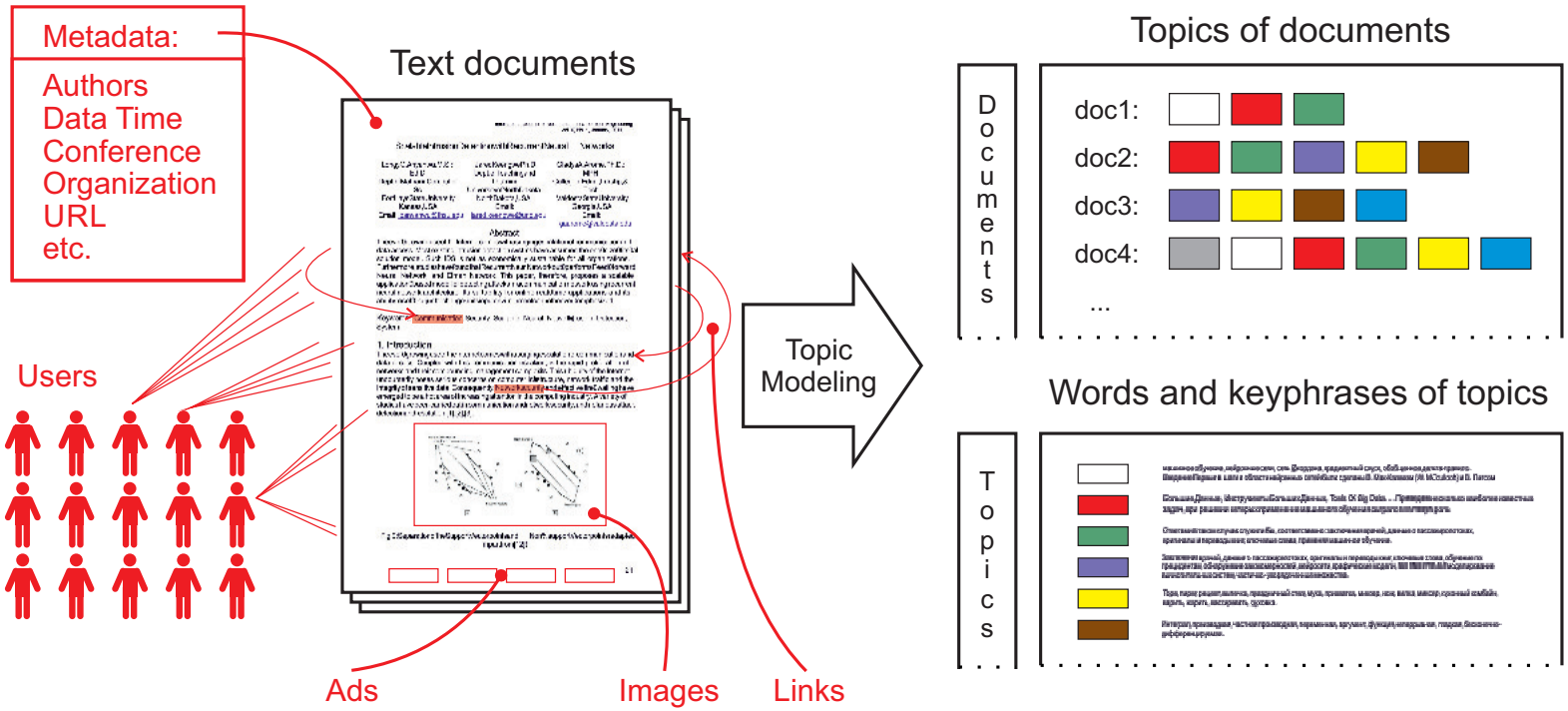
Multimodal Probabilistic Topic Modeling

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 $p(d'|t)$, advertising banners $p(b|t)$,



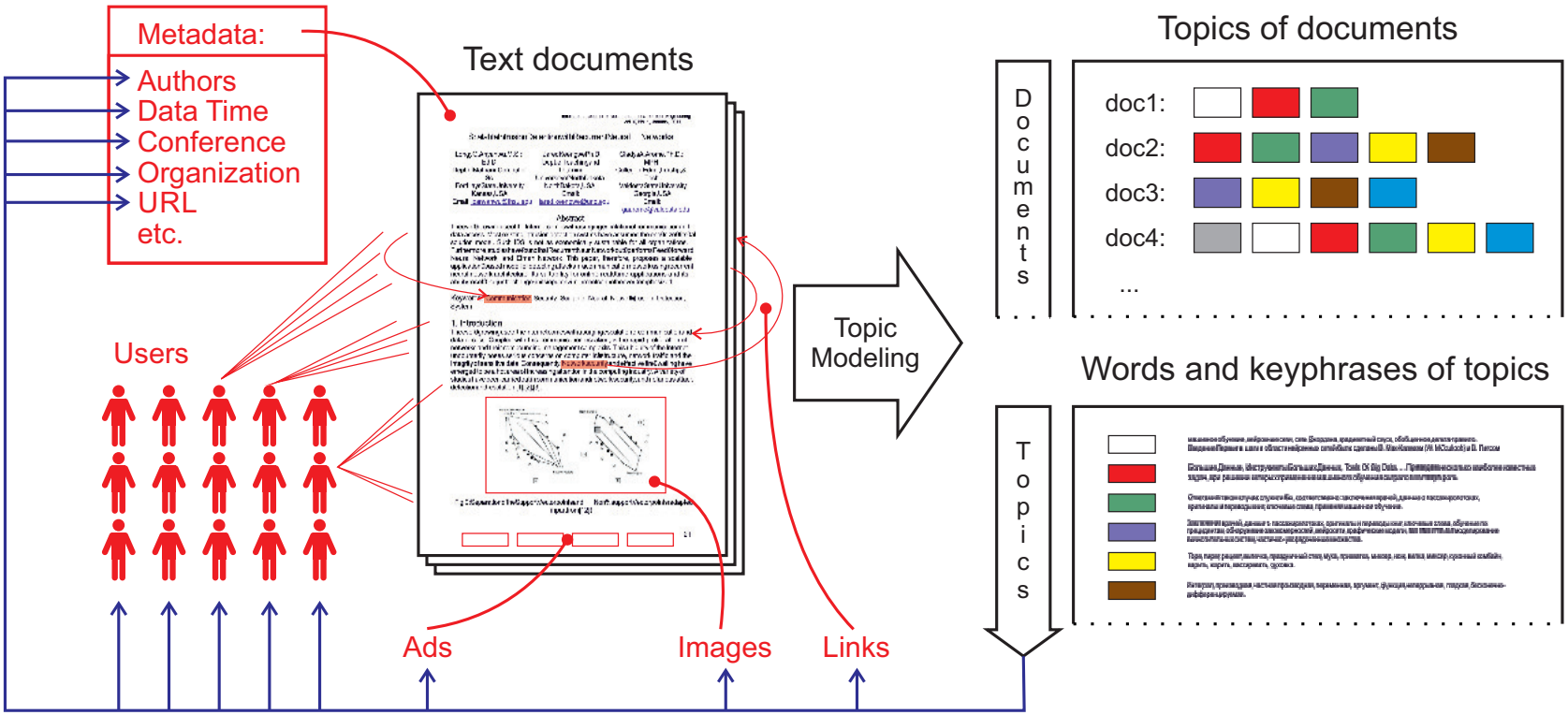
Multimodal Probabilistic Topic Modeling

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 $p(d'|t)$, advertising banners $p(b|t)$, 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 $p(d'|t)$, advertising banners $p(b|t)$, 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, $m \in M$
$W = W^1 \sqcup \cdots \sqcup W^M$ is a joint vocabulary of all modalities

Maximum multimodal log-likelihood with regularization:

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

EM-algorithm is a simple iteration method for the system

$$\text{E-step:} \quad p_{tdw} = \underset{t \in T}{\text{norm}}\left(\phi_{wt} \theta_{td}\right)$$

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

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

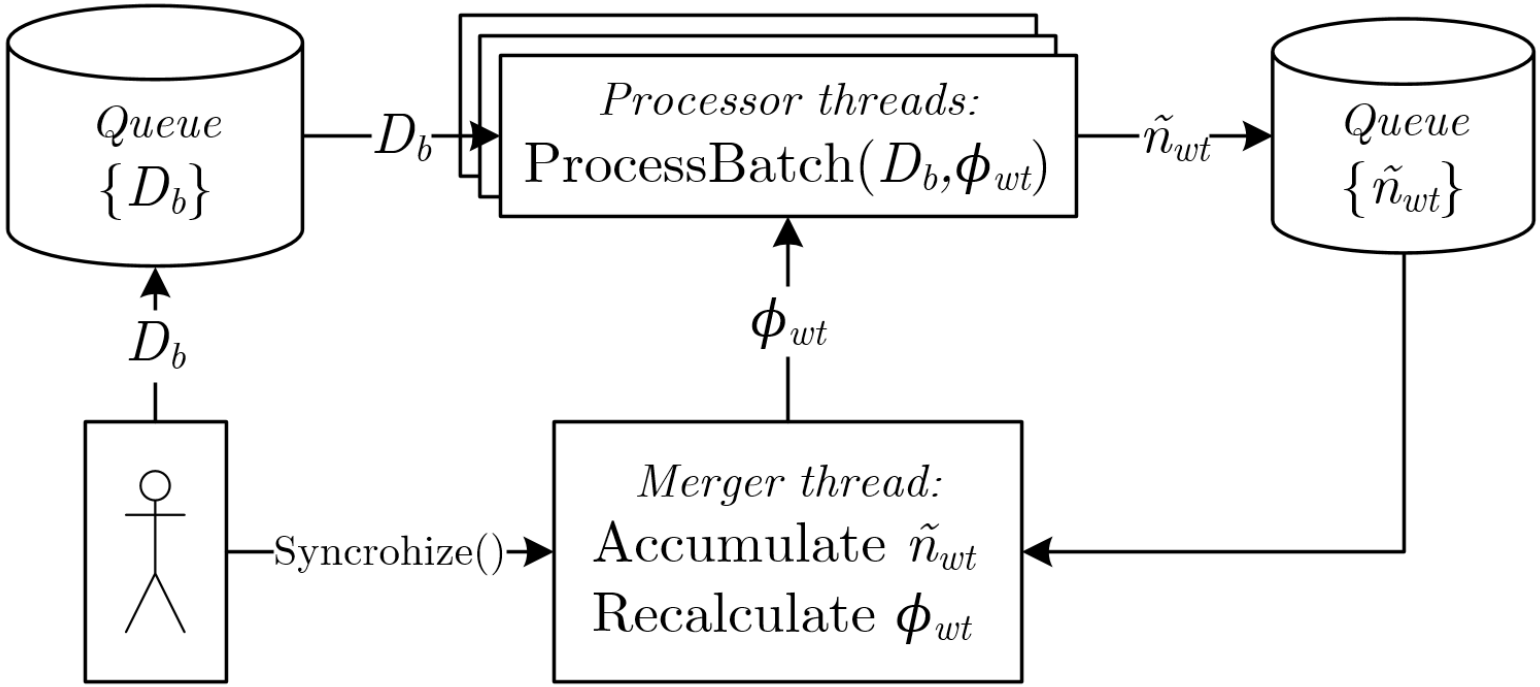
BigARTM community:

- 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

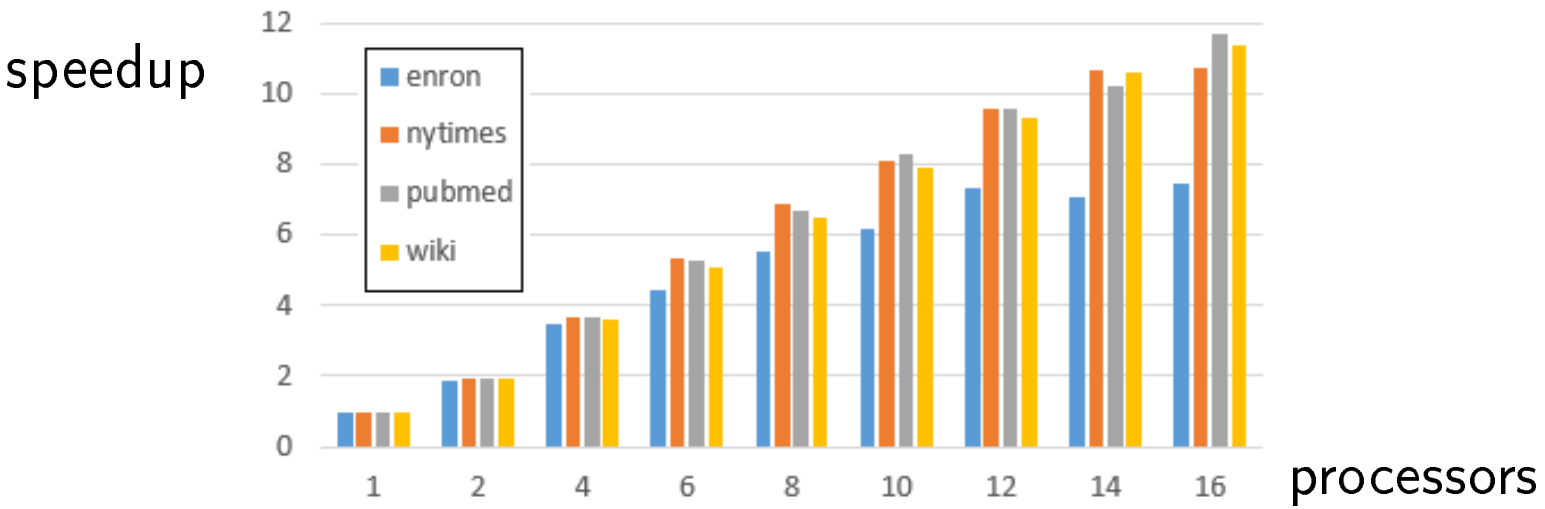
The BigARTM project: parallel architecture



- Concurrent processing of batches $D = D_1 \sqcup \cdots \sqcup D_B$
- 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® Xeon® CPU E5-2670 2.6GHz.

Experiment 2: BigARTM vs Gensim vs Vowpal Wabbit

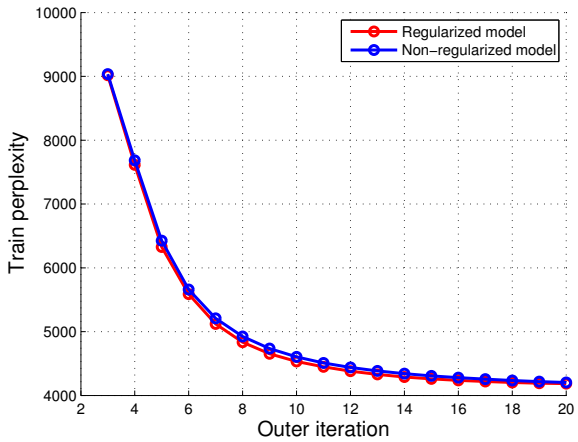
- 3.7M articles from Wikipedia, 100K unique words

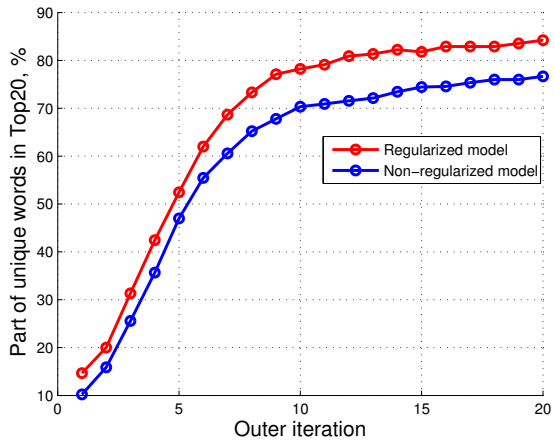
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

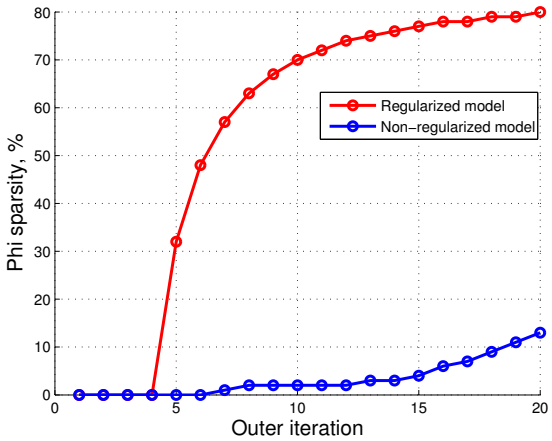
- *procs* = number of parallel threads
- *inference* = time to infer θ_d for 100K held-out documents
- *perplexity* is calculated on held-out documents.

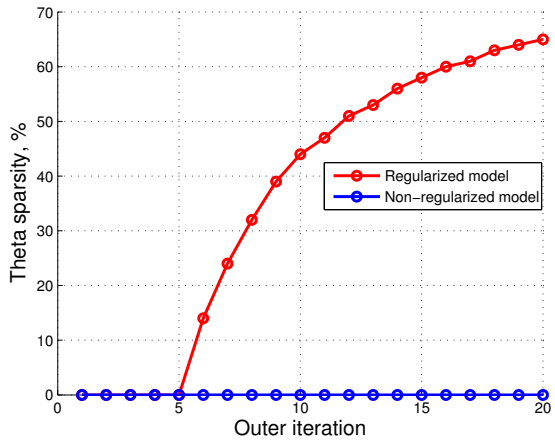
Experiment 3: Running BigARTM with multiple regularizers

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





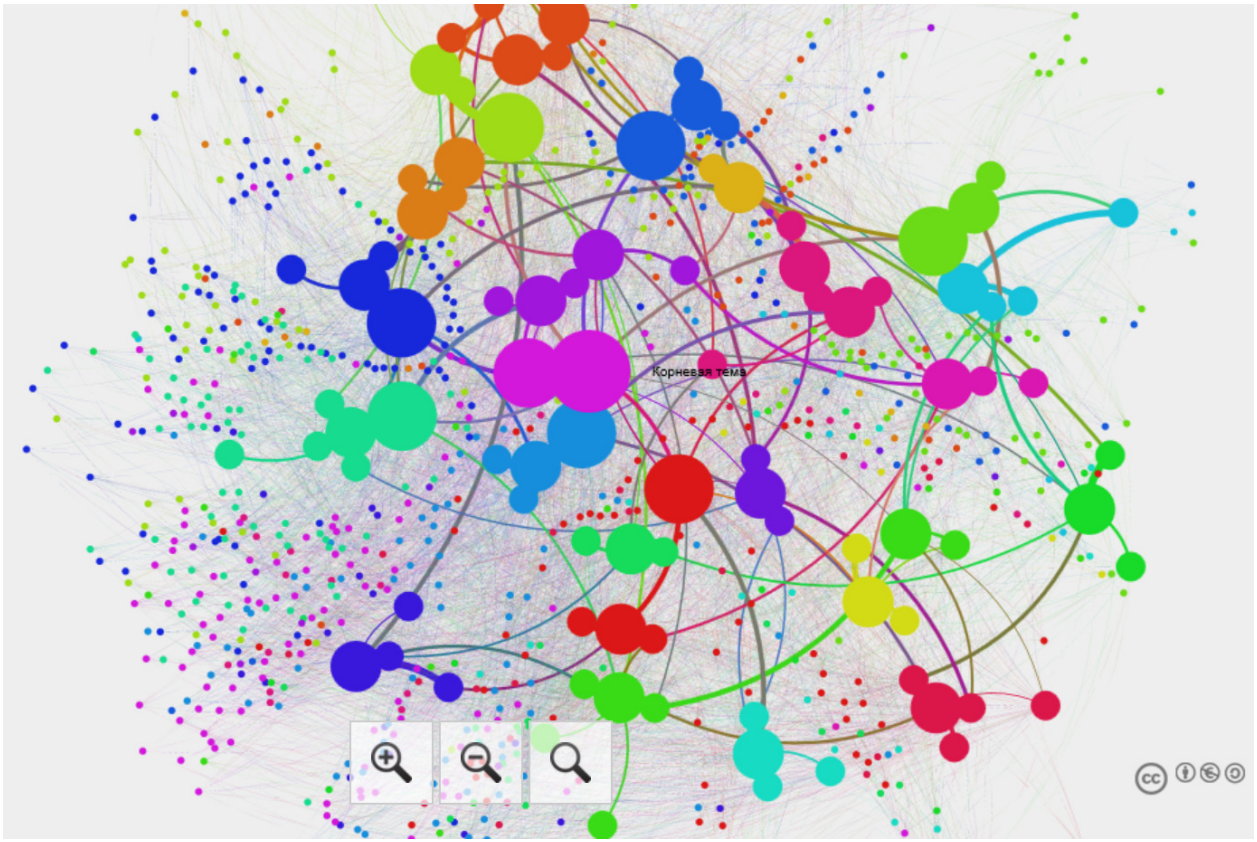




Experiment 4: Hierarchical topic model for MMPR-IIP conferences

$|D| = 865$, $|W| = 42\,000$ n-grams, in Russian

BigARTM is used with 7 regularizers to build 3-level hierarchy.



http://explore-mmro.ru

Experiment 5: The interpretability of n-gram models

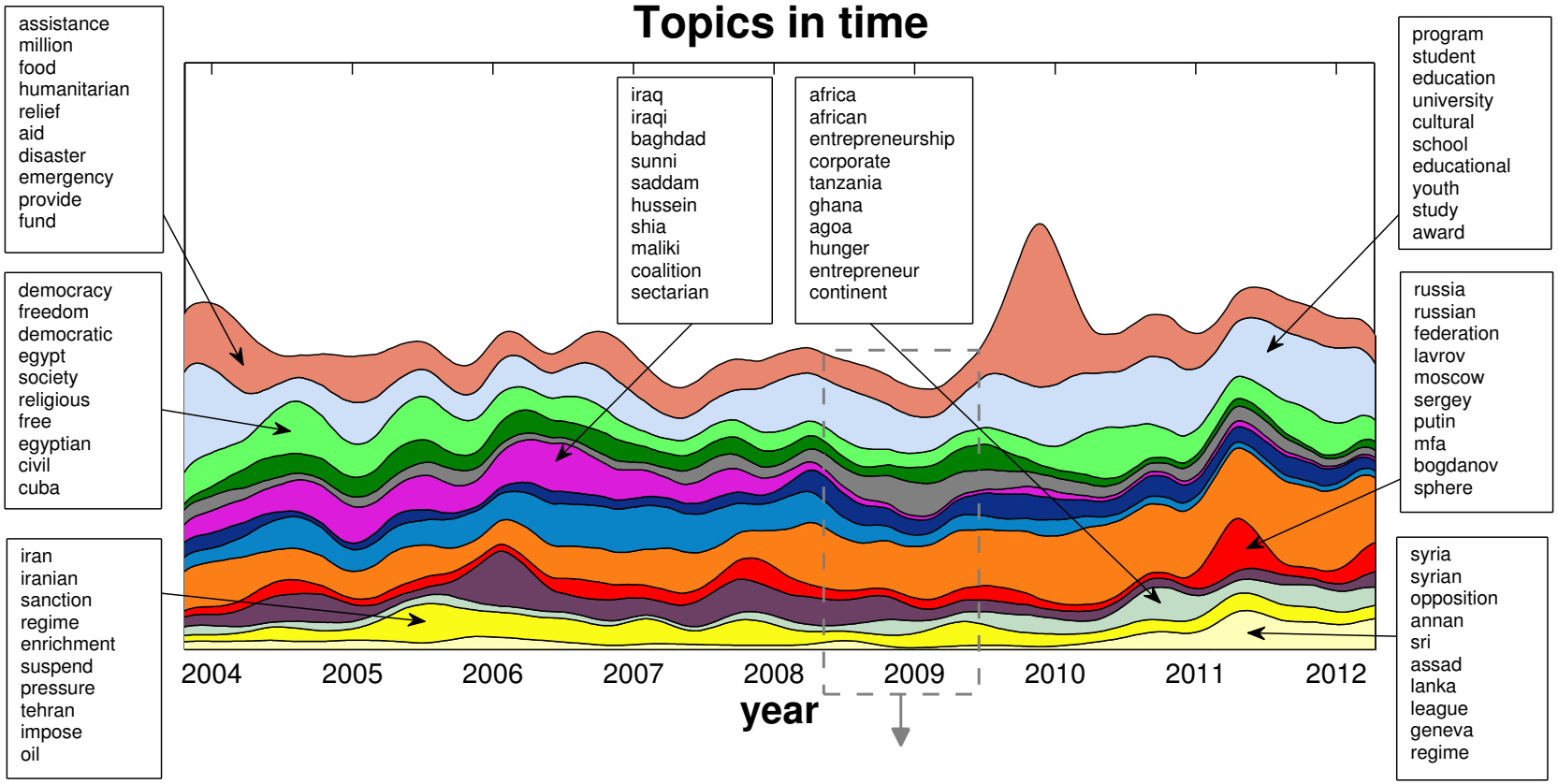
Two modalities — unigrams & bigrams

MMPR-IIP conferences collection, $|D| = 865$, in Russian

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

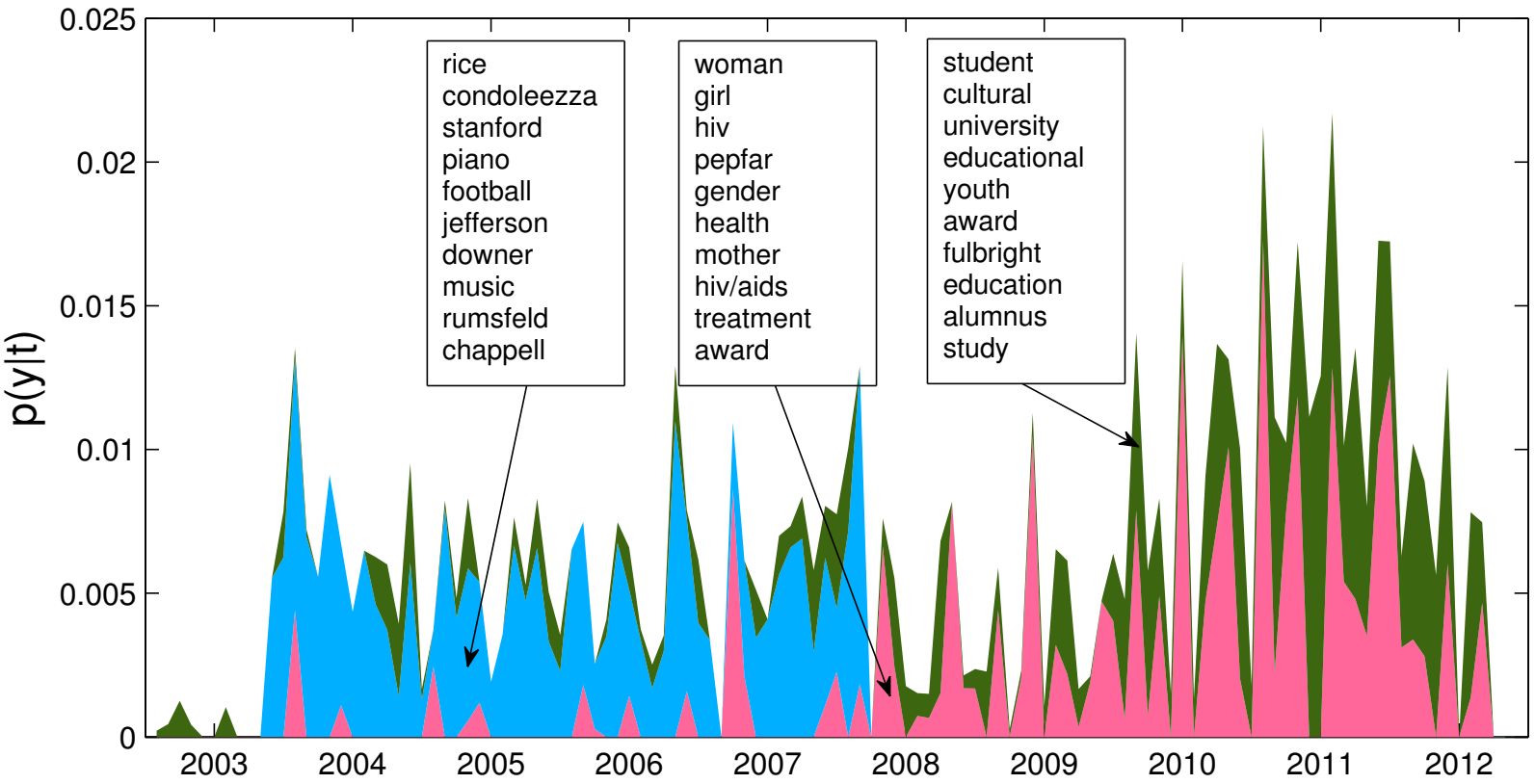
Experiment 6. Temporal topic model of political press-releases

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



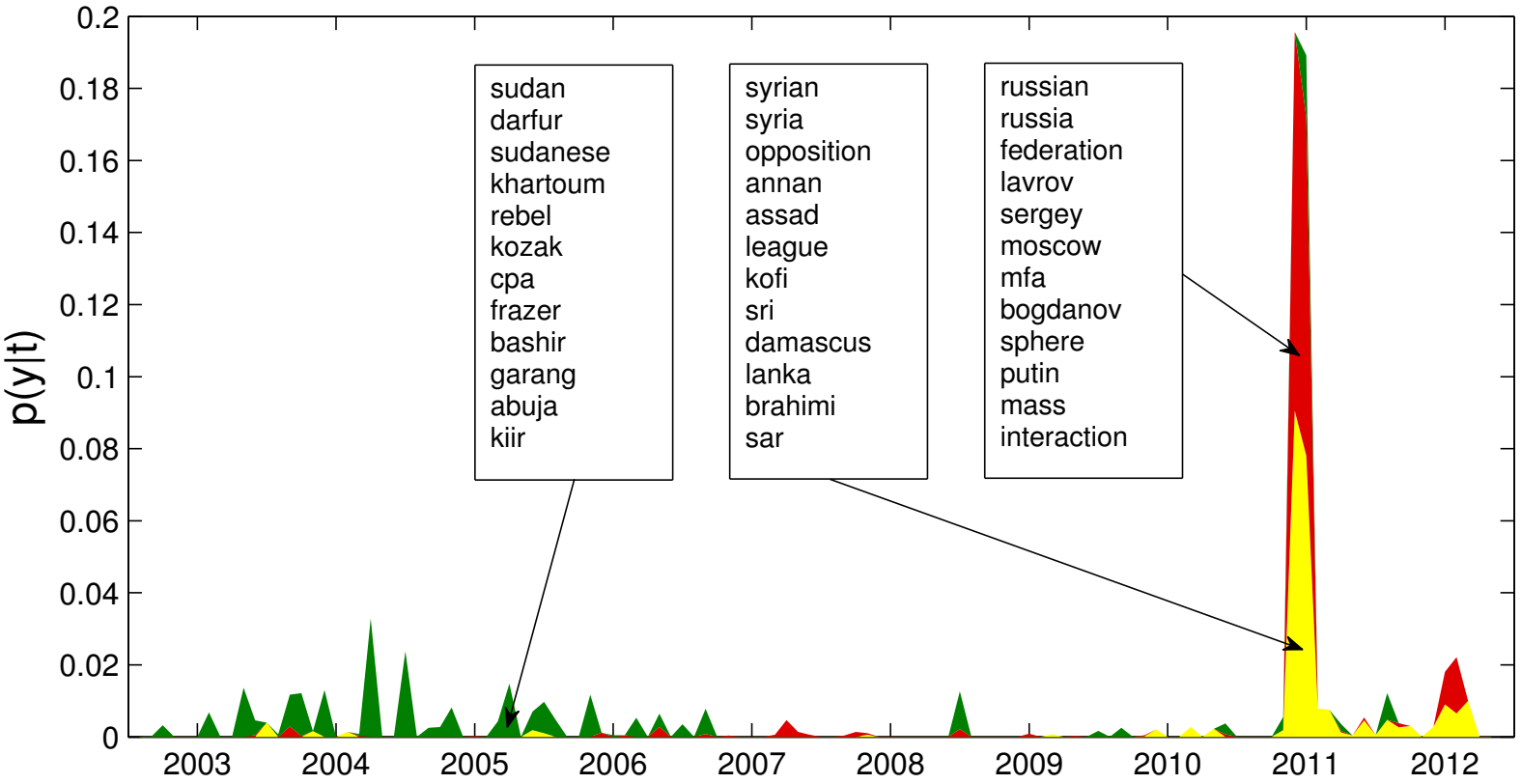
Experiment 6. Temporal topic model of political press-releases

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

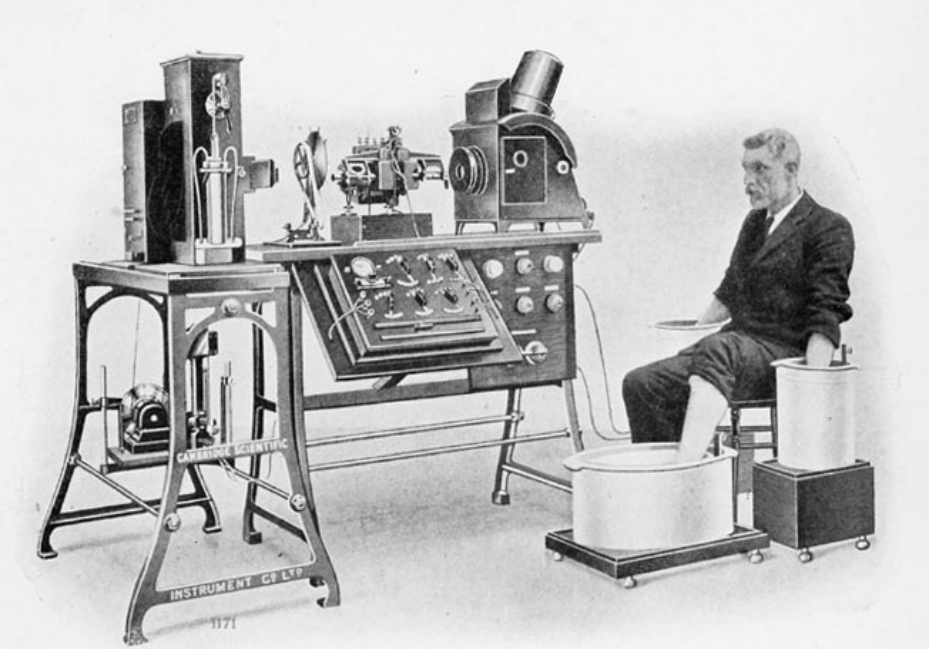


Brief summary

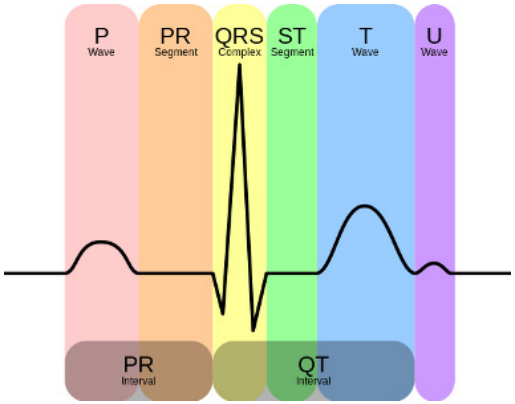
- **Exploratory Search:** a paradigm of Information Retrieval for professionals, researchers, students, and inquisitive persons
- **Multi-criteria Topic Modeling:** a way to meet multiple requirements coming from Exploratory Search
- **ARTM:** a novel non-Bayesian approach for multi-criteria optimization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

http://bigartm.org • Join BigARTM community!

Electrocardiography



PHOTOGRAPH OF A COMPLETE ELECTROCARDIOGRAPH, SHOWING THE MANNER IN WHICH THE ELECTRODES ARE ATTACHED TO THE PATIENT, IN THIS CASE THE HANDS AND ONE FOOT BEING IMMERSED IN JARS OF SALT SOLUTION



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

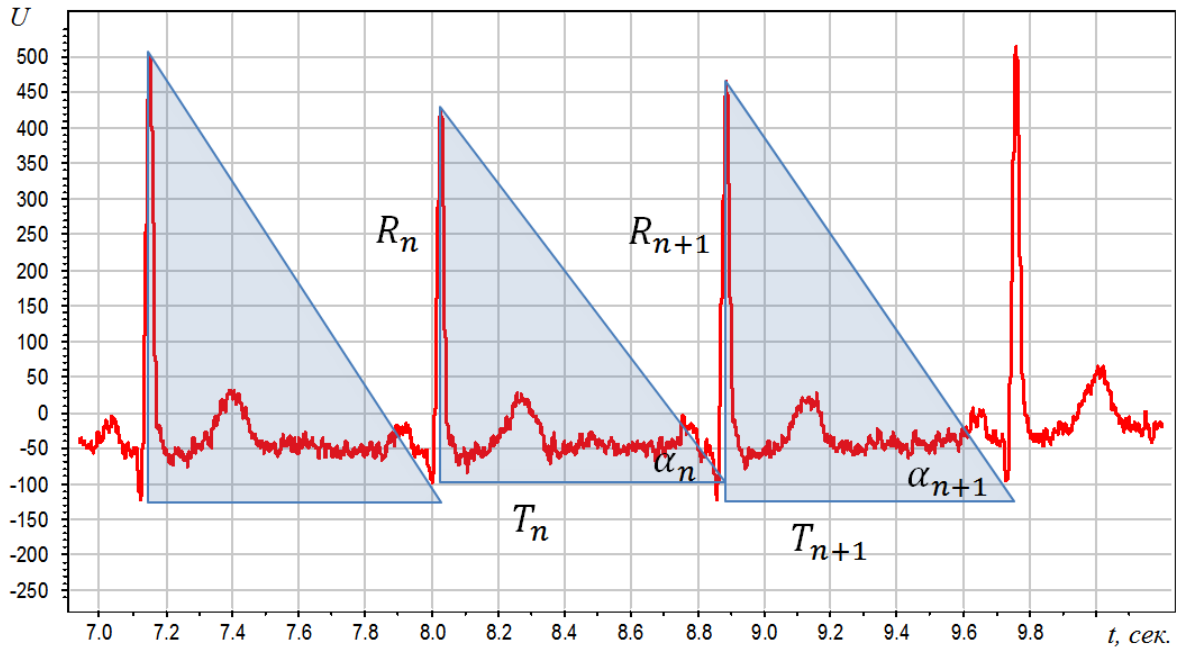
Input: a detailed raw ECG signal (3Mb file)

Output: a sequence of increment signs (225b — 10^4 compression!)

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

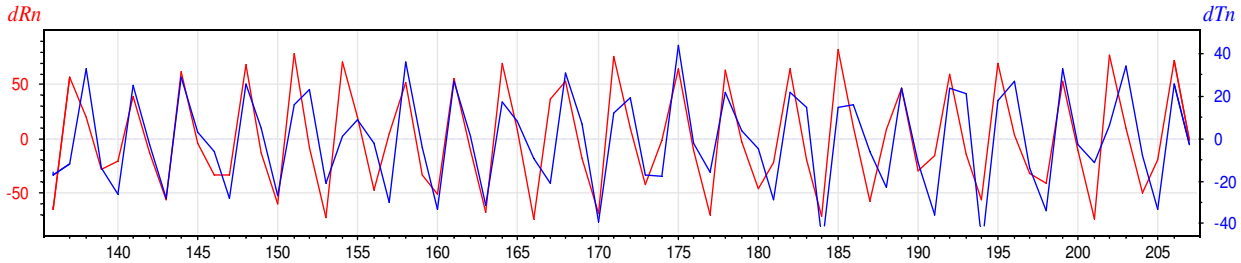
interval $dT_n = T_{n+1} - T_n$

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

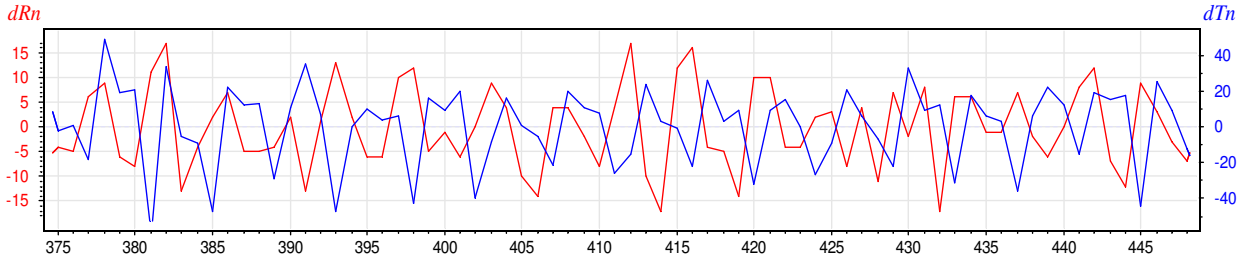


Variability of increments dR_n and dT_n for ill and healthy persons

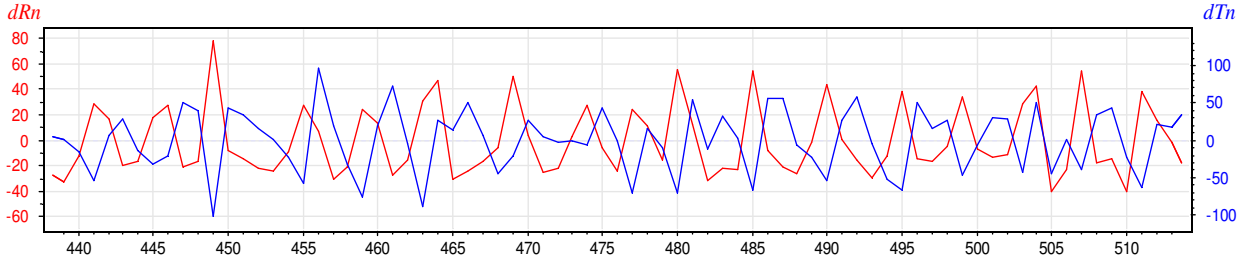
healthy:



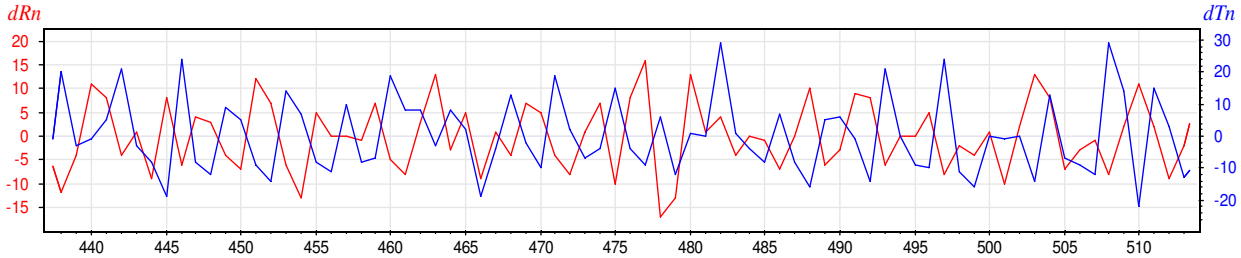
peptic ulcer:



hypertension:



cancer:



Preprocessing step 2: Discretization and symbolic representation

Input: intervals and amplitudes $(T_1, R_1), \ldots, (T_N, R_N)$
Output: *codogram* $x = (s_1, \ldots, s_{N-1})$ — a sequence of symbols from the alphabet $\mathcal{A} = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \mathtt{D}, \mathtt{E}, \mathtt{F}\}$

$$
\begin{array}{llll}
\text{if} & R_n < R_{n+1}, & T_n < T_{n+1}, & \alpha_n < \alpha_{n+1} \quad \text{then} \quad s_n = \mathtt{A} \\
\text{if} & R_n \geqslant R_{n+1}, & T_n \geqslant T_{n+1}, & \alpha_n < \alpha_{n+1} \quad \text{then} \quad s_n = \mathtt{B} \\
\text{if} & R_n < R_{n+1}, & T_n \geqslant T_{n+1}, & \alpha_n < \alpha_{n+1} \quad \text{then} \quad s_n = \mathtt{C} \\
\text{if} & R_n \geqslant R_{n+1}, & T_n < T_{n+1}, & \alpha_n \geqslant \alpha_{n+1} \quad \text{then} \quad s_n = \mathtt{D} \\
\text{if} & R_n < R_{n+1}, & T_n < T_{n+1}, & \alpha_n \geqslant \alpha_{n+1} \quad \text{then} \quad s_n = \mathtt{E} \\
\text{if} & R_n \geqslant R_{n+1}, & T_n \geqslant T_{n+1}, & \alpha_n \geqslant \alpha_{n+1} \quad \text{then} \quad s_n = \mathtt{F}
\end{array}
$$

Preprocessing step 3: Vectorization

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

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

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 $\{\mathtt{AAA}, \mathtt{AAB}, \ldots, \mathtt{FFF}\}$
- 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, $j = 1, \ldots, 216$
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
peptic ulcer	779	94.62 ± 0.10	63.6 ± 2.55

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 decentralized health care services
- Symbolic Dynamics and Topic Modeling can be used for mining diagnostic patterns from biomedical signals

The RSF project

Development of concept and methodology for multi-level monitoring of the state of inter-ethnic relations with the data from social media.

The objectives of Topic Modeling in this project:

1. Identify ethnic topics in social media big data
2. Identify event and permanent ethnic topics
3. Identify geographically located ethnic topics
4. Identify spatio-temporal patterns of the ethnic discourse
5. Sentiment analysis of ethnic discourse
6. Develop the monitoring system of inter-ethnic discourse

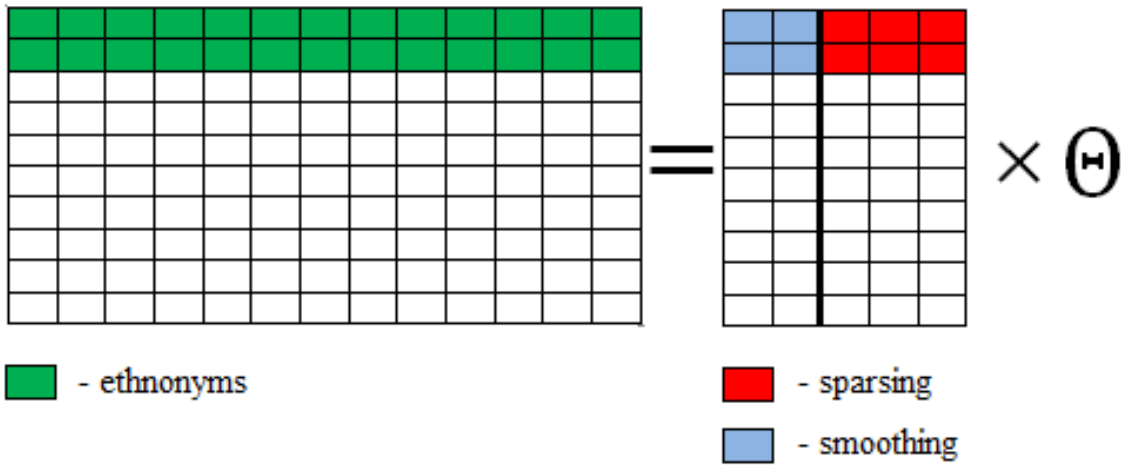
The Russian Science Foundation grant (2015–2017)
(Higher School of Economics, St. Petersburg School of Social Sciences and Humanities, Internet Studies Laboratory LINIS)

Example ethnonyms for semi-supervised topic modeling

османский	русич
восточноевропейский	сингапурец
эвенк	перуанский
швейцарская	словенский
аланский	вепсский
саамский	ниггер
латыш	адыги
литовец	сомалиец
цыганка	абхаз
ханты-мансийский	темнокожий
карачаевский	нигериец
кубинка	лягушатник
гагаузский	камбоджиец

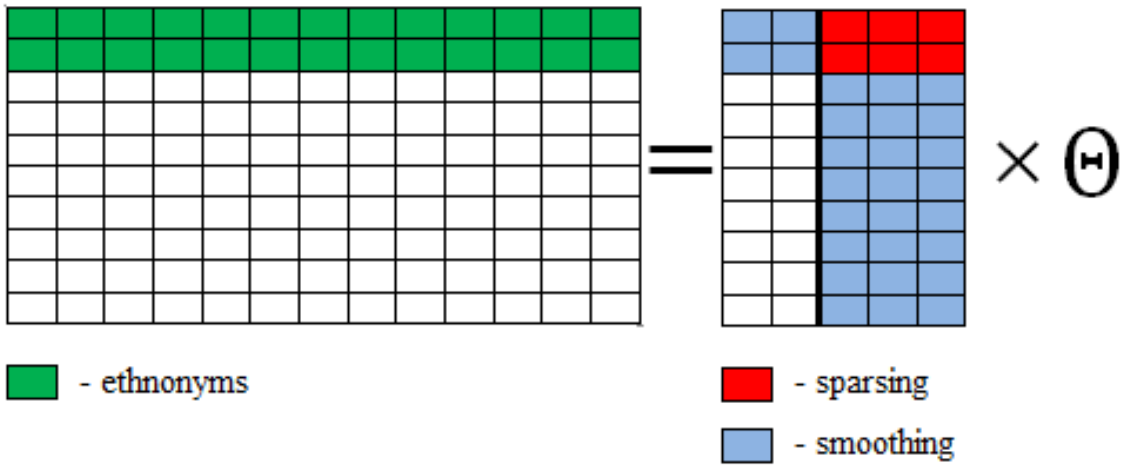
Regularization for finding ethnic topics

- smoothing ethnonyms in ethnic topics
- sparsing ethnonyms in common topics
-
-
-



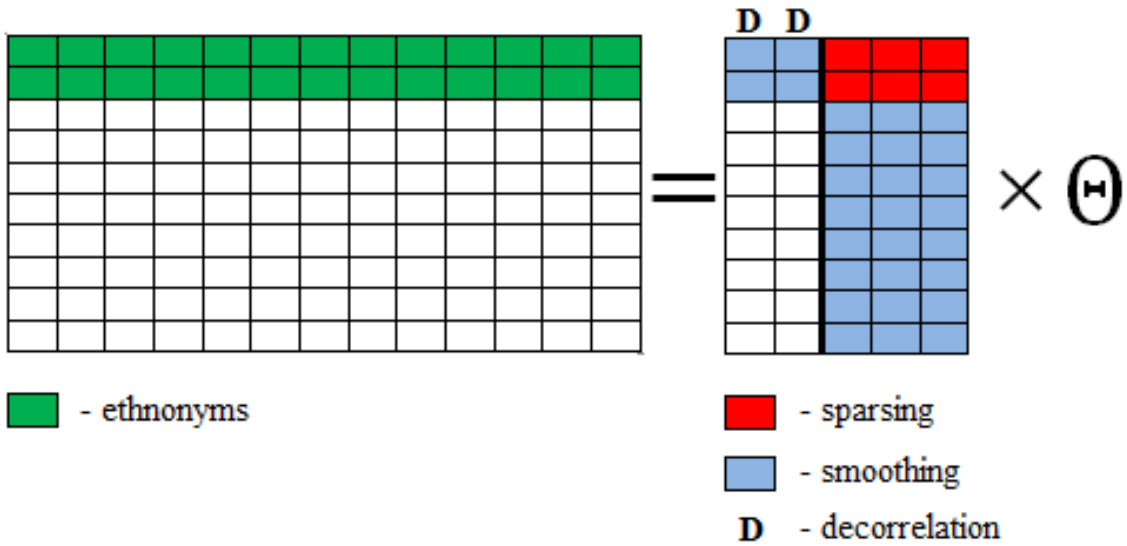
Regularization for finding ethnic topics

- smoothing ethnonyms in ethnic topics
- sparsing ethnonyms in common topics
- smoothing non-ethnonyms for common topics
-
-



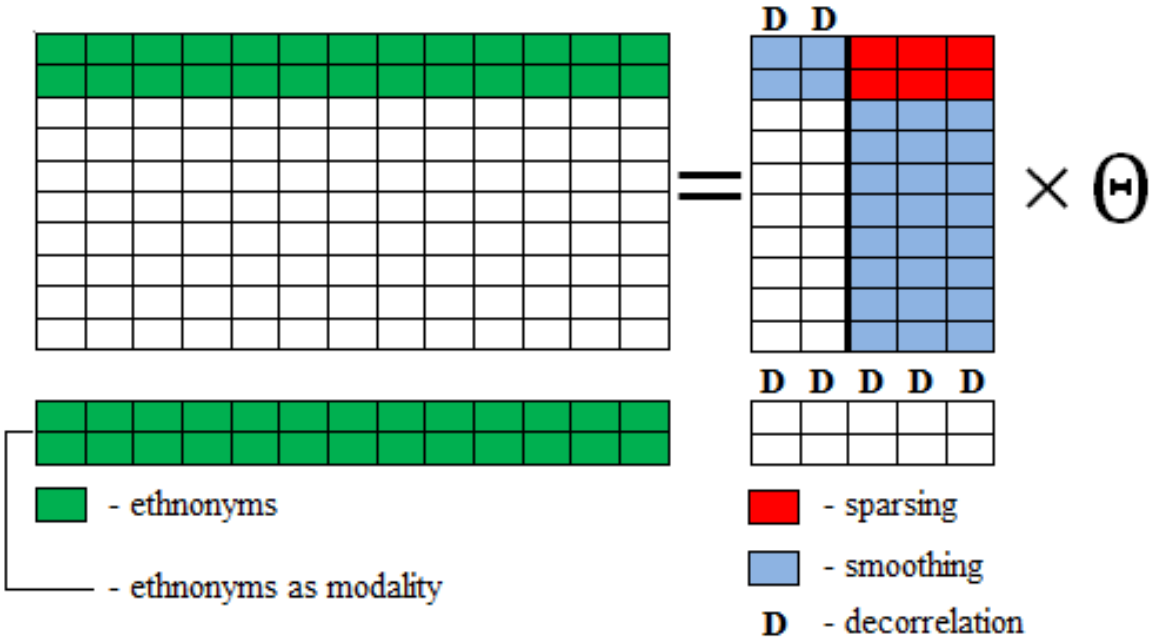
Regularization for finding ethnic topics

- smoothing ethnonyms in ethnic topics
- sparsing ethnonyms in common topics
- smoothing non-ethnonyms in common topics
- decorrelating ethnic topics
-



Regularization for finding ethnic topics

- smoothing ethnonyms in ethnic topics
- sparsing ethnonyms in common topics
- smoothing non-ethnonyms in common topics
- decorrelating ethnic topics
- adding ethnonyms modality and decorrelating their topics



Experiment

- LiveJournal collection: 1.58M of documents
- 860K of words in the raw vocabulary after lemmatization
- 90K of words after filtering out
  - short words with length $\leqslant 2$,
  - rare words with $n_w < 20$ including:
  - non-Russian words, abbreviations, misprints, mangled words, jargon
- 250 ethnonyms

Semi-supervised ARTM for ethnic topic modeling

The number of ethnic topics found by the model:

topic model	ethnic $\|S\|$	common $\|B\|$	++	+−	−+	total
PLSA		300	9	11	18	38
PLSA		400	12	15	17	44
ARTM-6	200	100	18	33	20	71
ARTM-6	250	150	21	27	20	68
ARTM-7	300	100	28	23	23	74
ARTM-7	250	150	22	25	33	80
ARTM-7	250	150	38	42	30	104

- ARTM-6:
  - ethnic topics: sparsing and decorrelating, ethnonyms smoothing
  - common topics: smoothing, ethnonyms sparsing
- ARTM-7:
  - ARTM-6 + ethnonyms as modality

Ethnic topics examples

(русские): русский, князь, россия, татарин, великий, царить, царь, иван, император, империя, грозить, государь, век, московская, екатерина, москва,
(русские): акция, организация, митинг, движение, активный, мероприятие, совет, русский, участник, москва, оппозиция, россия, пикет, протест, проведение, националист, поддержка, общественный, проводить, участие,
(славяне, византийцы): славянский, святослав, жрец, древние, письменность, рюрик, летопись, византия, мефодий, хазарский, русский, азбука,
(сирийцы): сирийский, асад, боевик, район, террорист, уничтожать, группировка, дамаск, оружие, алесио, оппозиция, операция, селение, сша, нусра, турция,
(турки): турция, турецкий, курдский, эрдоган, стамбул, страна, кавказ, горин, полиция, премьер-министр, регион, курдистан, ататюрк, партия,
(иранцы): иран, иранский, сша, россия, ядерный, президент, тегеран, сирия, оон, израиль, переговоры, обама, санкция, исламский, (палестинцы): террорист, израиль, терять, палестинский, палестинец, террористический, палестина, взрыв, территория, страна, государство, безопасность, арабский, организация, иерусалим, военный, полиция, газ,
(ливанцы): ливанский, боевик, район, ливан, армия, террорист, али, военный, хизбалла, раненый, уничтожать, сирия, подразделение, квартал, армейский,
(ливийцы): ливан, демократия, страна, ливийский, каддафи, государство, алжир, война, правительство, сша, арабский, али, муаммар, сирия,
(евреи): израиль, израильский, страна, израил, война, нетаньяху, тель-авив, время, сша, сирия, египет, случай, самолет, еврейский, военный, ближний,

Ethnic topics examples

(американцы): американский, американка, война, россия, военный, страна, вашингтон, америка, армия, конгресс, сирия, союзный, российский, обама, войска, русский, оружие, операция,
(немцы): армия, война, войска, советский, военный, дивизия, немец, фронт, немецкий, генерал, борт, операция, оборона, русский, бог, победа,
(немцы): германий, немец, германский, ссср, немецкий, война, старое, советский, россия, береза, русский, правительство, территория, полный, документ, вопрос, сорт, договор, отношение, франция,
(евреи, немцы): еврей, еврейский, холодный, германий, антисемитизм, гетра, немец, синагога, сша, израиль, малиновского, комиссия, нацбол, документ, война, еврейка, миллион, украина,
(украинцы, немцы): украинский, упс, оун, немец, немецкий, ковальков, хохол, волынский, бандера, организация, россиянин, советский, русский, польский, армия, шухевича, ровенский,
(таджики, узбеки): мигрант, страна, россия, миграция, азия, нелегальный, миграционный, таджикистан, гастарбайтер, гражданка, трудовой, рабочий, фмс, коренево, среднее, узбекистан, таджик, проблема, русский, население,
(канадцы): команда, игра, игрок, канадский, сезон, хоккей, сборная, играть, болельщик, победа, кубок, счет, забирать, хоккейный, выигрывать, хоккеист, чемпионат, шайба,
(японцы): японский, япония, корея, китайский, жилища, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,

Ethnic topics examples

(норвежцы): дитя, ребенок, родиться, детский, семья, воспитанный, право, возраст, отец, воспитание, норвежский, родительский, родить, мальчик, взрослый, опека, сын,
(венесуэльцы): куб, кассир, венесуэла, чавес, президент, уго, мадуро, боливия, фидель, глава, латинский, венесуэльский, лидер, боливарианской, президентский, альенде, гевару,
(китайцы): китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производить, производственный, промышленность, российский, экономический, кнр,
(азербайджанцы): русский, азербайджан, азербайджанец, россия, азербайджанский, таксист, диаспора, анапа, народ, москва, страна, армянин, слово, рынок,
(грузины): грузинский, спецназ, военный, август, баташева, российский, спецназовец, миротворец, операция, румын, бригада, миротворческий, абхазия, группа, войска, русский, цхинвале,
(осетины): конституция, осетия, аминат, русский, осетинский, южный, северный, россия, война, республика, вопрос, алахай, российский, население, конфликт,
(цыгане): наркотик, цыган, цыганка, хороший, место, страна, деньга, время, работать, жизнь, жить, рука, дом, цыганский, наркоманка,

Brief summary

- Semi-supervised topic modeling can find many small topics in big collection (*classifying needles in a haystack*).
- Online BigARTM makes only one pass through a collection, 15 minutes for 1.6M documents on 10 processors.
- Multimodal ARTM is well suitable for learning spatio-temporal topic model, which is a next stage of the project.

References


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.


Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A. Interval semi-supervised LDA: Classifying needles in a haystack // MICAI. Vol. 8265 of Lecture Notes in Computer Science. Springer, 2013. Pp. 265–274.


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, October 19, 2015, Melbourne, Australia.