Additive Regularization for Topic Modeling: Mining Ethnical Discourse in Social Media

Konstantin Vorontsov (CC RAS, MIPT, Moscow, Russia),
Murat Apishev (MSU, Moscow, Russia),
Sergei Koltcov (HSE, St.Petersburg, Russia),
Olessia Koltsova (HSE, St.Petersburg, Russia),
Sergey Nikolenko (Steklov Institute of Mathematics, St.Petersburg, Russia)

IDP-2016

BARCELONA

Intelligent Data Processing: Theory and Applications
Barcelona, Spain • 10–14 October 2016

Contents

What is a "topic" in a text collection

- *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.

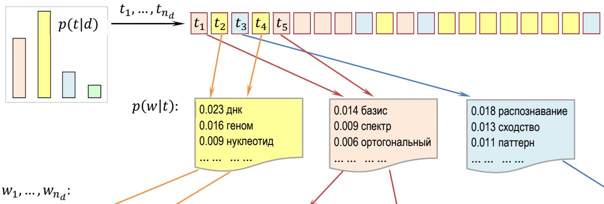
When writing term w in document d author thought of topic t.

Topic model tries to uncover latent topics in a text collection.

Probabilistic Topic Model (PTM) is generating a text collection

PTM explains how terms w appear in documents d from 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

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

This is an ill-posed problem of matrix factorization:

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

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

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

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

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

Maximum a posteriori probability (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

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

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

Maximum log-likelihood with 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

$$\begin{cases} \text{E-step:} & p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\ \text{M-step:} & \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\Big) \\ & \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{cases}$$

Combining topic models by adding their regularizers

Maximum log-likelihood with additive combination of regularizers:

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

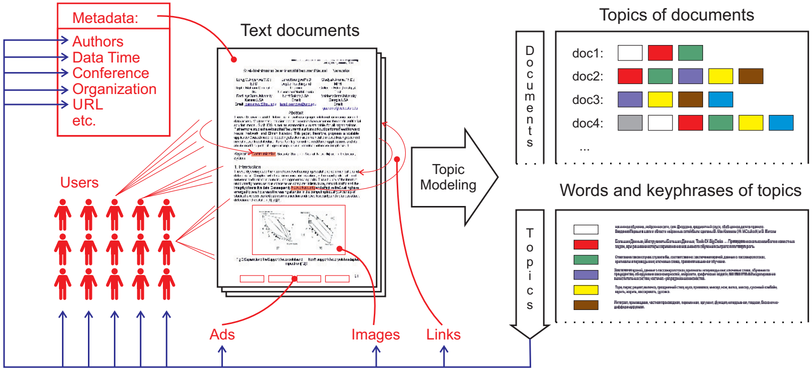
where τ_i are regularization coefficients.

EM-algorithm is a simple iteration method for the system

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \operatorname*{norm}_{t\in T}\big(\phi_{wt}\theta_{td}\big) \\[2ex] \phi_{wt} = \operatorname*{norm}_{w\in W}\Big(\sum\limits_{d\in D} n_{dw} p_{tdw} + \phi_{wt} \sum\limits_{i=1}^{n} \tau_i \frac{\partial R_i}{\partial \phi_{wt}}\Big) \\[2ex] \theta_{td} = \operatorname*{norm}_{t\in T}\Big(\sum\limits_{w\in d} n_{dw} p_{tdw} + \theta_{td} \sum\limits_{i=1}^{n} \tau_i \frac{\partial R_i}{\partial \theta_{td}}\Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical profiles $p(t|d)$, $p(t|w)$, $p(t|\text{author})$, $p(t|\text{time})$, $p(t|\text{category})$, $p(t|\text{tag})$, $p(t|\text{link})$, $p(t|\text{object-on-image})$, $p(t|\text{advertising-banner})$, $p(t|\text{users})$, etc. 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:

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

EM-algorithm is a simple iteration method for the system

$$\begin{cases} \text{E-step:} & p_{tdw} = \operatorname*{norm}_{t \in T} \bigl(\phi_{wt}\theta_{td}\bigr) \\ \text{M-step:} & \phi_{wt} = \operatorname*{norm}_{w \in W^m} \Bigl(\sum_{d \in D} \lambda_{m(w)} n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \Bigr) \\ & \theta_{td} = \operatorname*{norm}_{t \in T} \Bigl(\sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \Bigr) \end{cases}$$

BigARTM project: open source for topic modeling

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

BigARTM simplifies and unifies topic modeling for applications

Stages	Bayesian Inference for PTMs	ARTM	
Requirements analysis:	Requirements analysis	Requirements analysis	
Model formalization:	Generative model design	predefined criteria	user-defined criteria
Model inference:	Bayesian inference for the generative model (VI, GS, EP)	One regularized EM-algorithm for any combination of criteria	
Model implementation:	Researchers coding (Matlab, Python, R)	Production code (C++)	
Model evaluation:	Researchers coding (Matlab, Python, R)	predefined measures	user-defined measures
Deployment:	Deployment	Deployment	

conventions: ::: not unified stages ::: ::: unified stages :::

Bayesian models require maths and coding at each stage.
Therefore practitioners rarely go beyond a basic LDA model.
ARTM breaks this barrier by unifying the modeling process.

Benchmarking BigARTM vs. Gensim and 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.

The set of useful properties that topic models would have

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

- For any property X from the list one can easily find the extensive literature on "X Topic Model"
- For combinations of two properties "X Y Topic Model" the volume of literature is modest
- Publications on combinations of three and more properties are exceptional

Why?

Literature on Topic Modeling is basically Bayesian.
In Bayesian approach, compound models are very hard to construct.

Smoothing, sparsing and decorrelation of topics

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

Smoothing background topics $t \in B \subset T$ makes the model robust:

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

Sparsing subject topics $t \in S = T \backslash B$ makes it more interpretable:

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

Decorrelation make subject topics as different as possible:

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

Semi-supervised learning for topic correction

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

Idea is the same as smoothing, except for the role of β_{wt}, α_{td} parameters. Assessors forms "black" and "white" lists of documents and terms to train each topic in the model individually:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} + \sum_{d \in D} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \to \max .$$

- $\beta_{wt} = \big[w \in W_t^+\big]$, W_t is a *white list* of terms for topic t
- $\alpha_{td} = \big[d \in D_t^+\big]$, D_t is a *white list* of docs for topic t
- $\beta_{wt} = -\big[w \in W_t^-\big]$, W_t is a *black list* of terms for topic t
- $\alpha_{td} = -\big[d \in D_t^-\big]$, D_t is a *black list* of docs for topic t

Semi-supervised learning for finding relevant topics

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

Motivation: we want to find all topics about diseases / disasters / terrorism / inter-ethnic relations / a country / a company / a product / a politician etc. in social media.

We smooth all topics from $T_0 \subset T$ with a set of "seed words" W_0:

$$R(\Phi) = \tau \sum_{t \in T_0} \sum_{w \in W_0} \ln \phi_{wt} \to \max.$$

Paul, M.J., Dredze, M. Discovering health topics in social media using topic models. 2014.

Biterm topic model (BTM) for short texts

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

Short-text topic models are motivated by social media analysis.

We revisit *Biterm Topic Model* as a regularizer in ARTM:

$$R(\Phi) = \tau \sum_{u,w \in W} n_{uw} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{wt} \to \max$$

where n_{uw} is a number of co-occurrences of word pair (u, w) in a short context (sentence or 10-words window).

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng. A Biterm Topic Model for Short Texts // WWW 2013.

Word network topic model (WNTM) for short texts

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

Short-text topic models are motivated by social media analysis.

We revisit *Word Network Topic Model* as a regularizer in ARTM:

$$R(\Phi, \Theta') = \sum_{u,w \in W} n_{uw} \log \sum_{t \in T} \phi_{ut} \theta'_{tw} \to \max_{\Phi, \Theta'}$$

where n_{uw} has the same sense as in Biterm topic model.

Yuan Zuo, Jichang Zhao, Ke Xu. Word Network Topic Model: a simple but general solution for short and imbalanced texts. 2014.

Berlin Chen. Word Topic Models for spoken document retrieval and transcription // ACM Trans., 2009.

The power of multiple modalities

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

All these properties are special cases of modalities.

Example: regularization for building a level of a topical hierarchy:

$$R(\Phi, \Psi) = \sum_{a,w} n_{aw} \ln \sum_{t} \phi_{wt} \psi_{ta} \ \to \ \max_{\Phi,\Psi}$$

where $\psi_{ta} = p(t|a)$ links a subtopic t with parent topics a.
Then, parent level $a \in A$ can be processed as "pseudodocuments".

N. A. Chirkova, K. V. Vorontsov. Additively Regularized Multimodal Topic Hierarchies. JMLDA. 2016 (to appear)

The power of BigARTM

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

BigARTM provides many useful properties out-of-the-box.

Properties to be implemented in the near future:

- *Extendable Topic Model* will create new topics and new vocabulary entries "on-the-fly" (motivated by news flows).
- *Distributed computing* for huge text collections (motivated by Exploratory Search in huge collections of scientific papers).

Topic model for Exploratory Search

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

The main problem of mixing regularizers:
how to determine regularization coefficients τ_i

- greedy coordinate-wide optimization
- fully automatic multicriteria optimization via reinforcement learning (future work)

A. O. Ianina, K. V. Vorontsov Multimodal topic modeling for exploratory search in collective blog. JMLDA. 2016 (to appear)

Mining ethnical discourse in social media

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

The goal of the ongoing research project:
monitoring the inter-ethnic relations from social media data.

The objectives of Topic Modeling in this project:

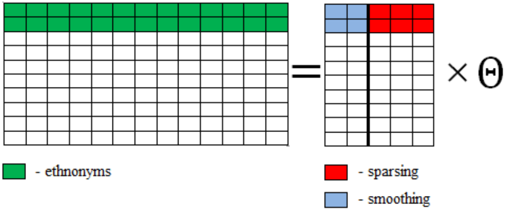
1. Semi-supervised topic learning: identify ethnic topics form a list of seed words (ethnonyms)
2. Spatio-temporal patterns of the ethnic discourse: event-topics, location-topics
3. Spatio-temporal sentiment analysis of the ethnic discourse

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

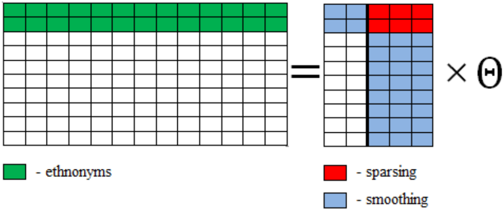
Regularization for finding ethnic topics

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



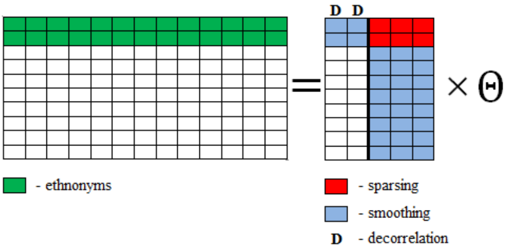
Regularization for finding ethnic topics

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



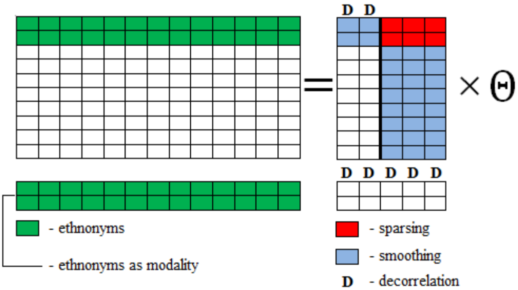
Regularization for finding ethnic topics

- 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 with 6 regularizers:
  - ethnic topics: sparsing and decorrelating, ethnonyms smoothing
  - common topics: smoothing, ethnonyms sparsing
- ARTM-7 with 7 regularizers:
  - 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): куба, кастро, 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,

News flow control for media planning

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

The goals of the ongoing research project are:

1. develop a well-interpretable hierarchical temporal extendable topic model of the news flow
2. develop a solution for filtering and evaluating topic-and-sentiment structure of the news flow
3. incorporate the solution in existing media planning software

Scenario analysis of call center records

interpretable	sparse	robust	decorrelated	multigram
multimodal	multilingual	hierarchical	temporal	spacio-temporal
short-text	sentence	segmentation	relational	sentiment
supervised	classification	semi-supervised	auto-labeled	summarization
fast	online	extendable	parallel	distributed

The goal of the ongoing research project:

1. determine typical scenarios of call-center dialogues between operators and customers
2. elaborate the quantitative measure of how well operator works
3. provide online tips for help operator handle customer's objections

Brief summary

- ARTM theory opens a way to the “topic modeling alchemy”, when you simply specify a set of regularizers and obtain a model with desired properties
- BigARTM is an open source project for topic modeling. Join!
- The number of topic modeling applications is growing rapidly, from expert search and scientific papers mining to media planning and processing call-center records

References

- *K.Vorontsov.* Additive regularization for topic models of text collections. 2014.
- *K.Vorontsov, A.Potapenko.* Additive regularization of topic models. Machine Learning, 2015.
- *K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova, A.Yanina.* Non-bayesian additive regularization for multimodal topic modeling of large collections. CIKM TM workshop, 2015.
- *K.Vorontsov, A.Potapenko, A.Plavin.* Additive regularization of topic models for topic selection and sparse factorization. SLDS 2015.
- *O.Frei, M.Apishev.* Parallel non-blocking deterministic algorithm for online topic modeling. AIST 2016. (to appear)
- *Bodrunova S., Koltsov S., Koltsova O., Nikolenko S. I., Shimorina A.* Interval semi-supervised LDA: Classifying needles in a haystack // MICAI 2013.
- *M.Apishev, S.Koltcov, O.Koltsova, S.Nikolenko, K.Vorontsov.* Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI 2016. (to appear)