Mathematical methods and applications of semantic analysis of text data

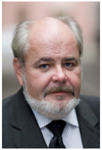
Konstantin Vorontsov
Dr.Sc.(phys.-math.), professor of RAS,
head of Machine Learning & Semantic Analysis laboratory at Institute for Artificial Intelligence, Moscow State University;
head of department "Mathematical Methods of Forecasting", MSU;
head of department "Machine Learning & Digital Humanities", MIPT
voron@mlsa-iai.ru

• New challenges facing artificial intelligence •
Tsinghua University, Beijing, China • March 27, 2023

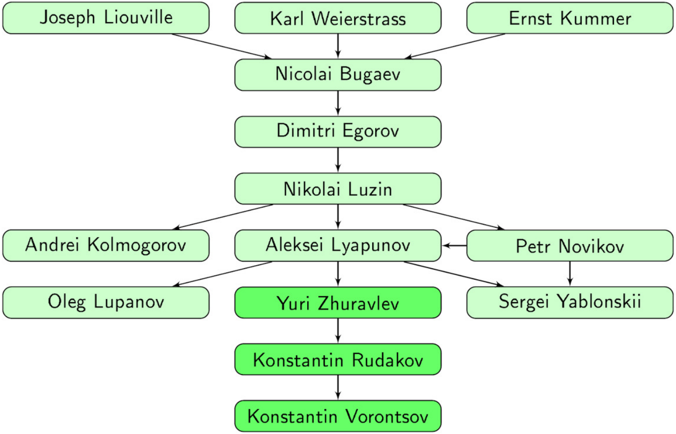
1. Probabilistic Topic Modeling
   - The problem setup and applications
   - Theory of additive regularization (ARTM)
   - BigARTM project and applications
2. Project 1. Knowledge Factory
   - From Exploratory Search to Knowledge Factory
   - Machine aided human summarization (MAHS)
   - Visualizing topics, maps, trends, structures
3. Project 2. News Collider
   - From fake news to information warfare
   - Text markup: digitalizing humanitarian knowledge
   - Model evaluation with inconsistent assessors

The scientific mathematical school at Moscow (MSU, MIPT)

Yury I. Zhuravlev (1935-2022)



Konstantin V. Rudakov (1954-2021)



From AMS Mathematics Genealogy Project (genealogy.math.ndsu.nodak.edu)

Probabilistic Topic Modeling (PTM): the problem setup

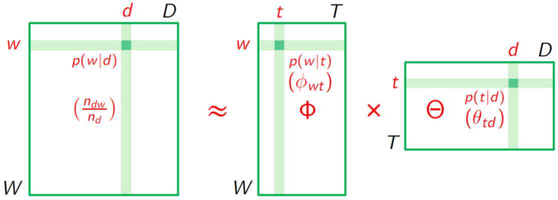
Given: a set of terms (words) W, a set of documents D,
n_{dw} = how many times term w appears in document d

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

$$p(w|d) = \sum_{t \in T} \varphi_{wt}\theta_{td} = \sum_{t \in T} p(w|t)p(t|d).$$

subject to $\varphi_{wt} \geqslant 0$, $\sum_w \varphi_{wt} = 1$, $\theta_{td} \geqslant 0$, $\sum_t \theta_{td} = 1$.

This is a problem of *nonnegative matrix factorization*:



Probabilistic Latent Semantic Analysis (PLSA)

Constrained maximization of the log-likelihood:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta}$$

EM-algorithm is a simple iteration method for the system

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

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

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

Well-posed and ill-posed problems

The problem is *well-posed* in the sense of Hadamard (1923) if the solution

- exists,
- is unique,
- is stable w.r.t. initial conditions.

Jacques Hadamard (1865–1963)

Matrix factorization is an *ill-posed* inverse problem.
If (Φ, Θ) is a solution, then (Φ', Θ') is also the solution:

- $\Phi'\Theta' = (\Phi S)(S^{-1}\Theta)$, $\text{rank} S = |T|$
- $L(\Phi', \Theta') \approx L(\Phi, \Theta)$

Adding regularization criterion is used to obtain an appropriate solution.

Andrey N. Tikhonov (1906–1993)

Latent Dirichlet Allocation (LDA)

Maximize a posteriori probability (MAP) with Dirichlet prior:

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

EM-algorithm is a simple iteration method for the system

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

D.Blei, A.Ng, M.Jordan. Latent Dirichlet Allocation. 2001, 2003.

Additive Regularization for Topic Modeling (ARTM)

Maximize log-likelihood with regularization criteria $R_i(\Phi,\Theta)$:

$$\sum_{d,w} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}; \qquad R(\Phi,\Theta) = \sum_i \tau_i R_i(\Phi,\Theta)$$

EM-algorithm is a simple iteration method for the system of equations with auxiliary variables $p_{tdw} = p(t|d,w)$:

$$\begin{array}{rl} \text{E-step:} & \left\{ \begin{array}{l} p_{tdw} = \underset{t\in T}{\mathrm{norm}}\big(\varphi_{wt}\theta_{td}\big) \\ \varphi_{wt} = \underset{w\in W}{\mathrm{norm}}\Big(\sum\limits_{d\in D} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\Big) \\ \theta_{td} = \underset{t\in T}{\mathrm{norm}}\Big(\sum\limits_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\Big) \end{array} \right. \\ \text{M-step:} & \end{array}$$

K.Vorontsov. Additive regularization for topic models of text collections. 2014.

Bayesian vs classical (non-Bayesian, additive) regularization

Bayesian inference of posterior distribution $p(\Omega|X)$ being usually cumbersome and approximate is used only for Ω point estimate:

$$\mathrm{Posterior}(\Omega|X,\gamma) \propto p(X|\Omega)\,\mathrm{Prior}(\Omega|\gamma)$$

$$\Omega := \arg\max_{\Omega} \mathrm{Posterior}(\Omega|X,\gamma)$$

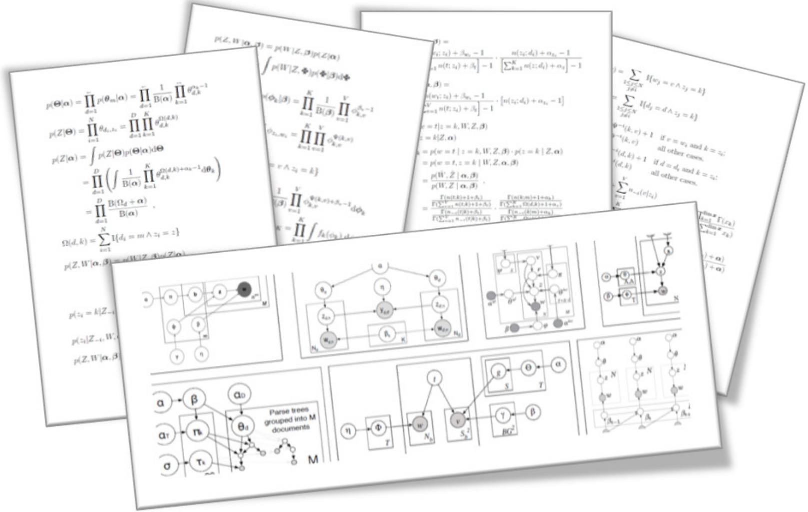
Maximum a posteriori estimation (MAP) gives a point estimate Ω directly without posterior inference:

$$\Omega := \arg\max_{\Omega}\bigl(\ln p(X|\Omega) + \ln \mathrm{Prior}(\Omega|\gamma)\bigr)$$

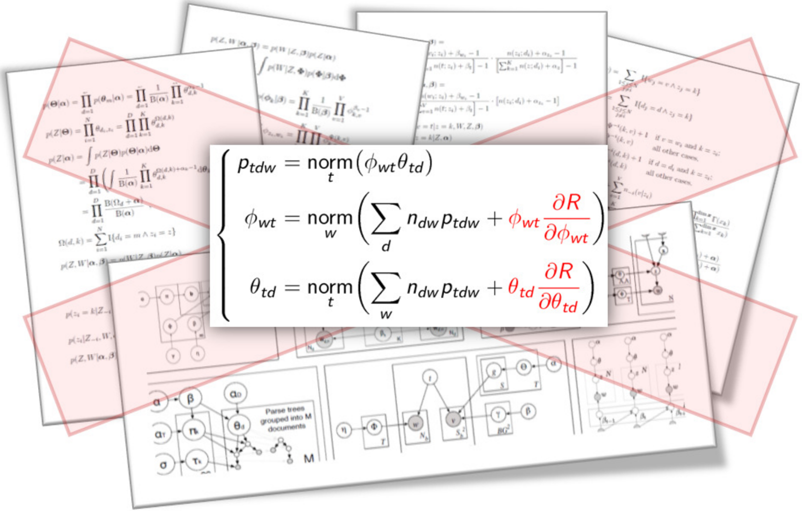
Multicriteria additive regularization (ARTM) generalizes MAP to non-probabilistic regularizers as well as the weighted sum of regularizers, without violating the convergence properties:

$$\Omega := \arg\max_{\Omega}\Bigl(\ln p(X|\Omega) + \sum_{i=1} \tau_i R_i(\Omega)\Bigr)$$

Bayesian learning is overcomplicated for PTM



ARTM: easy way to design, to understand, and to combine PTMs

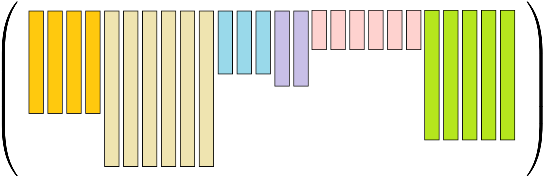


$$\begin{cases} p_{tdw} = \operatorname*{norm}_{t}\big(\phi_{wt}\theta_{td}\big) \\ \phi_{wt} = \operatorname*{norm}_{w}\left(\sum_{d} n_{dw} p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) \\ \theta_{td} = \operatorname*{norm}_{t}\left(\sum_{w} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{cases}$$

Constrained maximization of a function on unit simplices

Let $\Omega = (\omega_j)_{j \in J}$ be a set of non-negative normalized vectors ω_j having dimensions $|I_j|$ respectively, $\omega_j = (\omega_{ij})_{i \in I_j}$:

$\Omega =$



Problem: maximize the function $f(\Omega)$ on unit simplices:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad j \in J; \\ \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J. \end{cases}$$

Necessary extremum conditions and the simple-iteration method

$\underset{i\in I}{\operatorname{norm}}(x_i) = \frac{\max(x_i,0)}{\sum\limits_k \max(x_k,0)}$ is a projection of x vector on unit simplex

Lemma. Let $f(\Omega)$ be continuously differentiable function on Ω.
If ω_j is the local maximum of $f(\Omega)$ and $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} > 0$ for some $i \in I_j$, then ω_j satisfies the system of equations

$$\omega_{ij} = \underset{i\in I_j}{\operatorname{norm}}\left(\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right).$$

- For numerical solution, the simple-iteration method can be used
- Vectors $\omega_j = 0$ are discarded as degenerate solutions
- Iterations are similar to gradient maximization of $f(\Omega)$:

$$\omega_{ij} := \omega_{ij} + \eta\frac{\partial f}{\partial \omega_{ij}},$$

differing in “norm” projection and absence of η parameter

Proof of the Lemma on Maximization on unit simplices

Problem: $f(\Omega) \to \max\limits_{\Omega}; \quad \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j, \quad j \in J.$

The Lagrangian of the optimization problem:

$$\mathscr{L}(\Omega; \mu, \lambda) = f(\Omega) + \sum_{j \in J} \lambda_j \Bigl(\sum_{i \in I_j} \omega_{ij} - 1 \Bigr) - \sum_{j \in J} \sum_{i \in I_j} \mu_{ij} \omega_{ij}.$$

The Karush–Kuhn–Tucker conditions for the vector ω_j:

$$\frac{\partial f(\Omega)}{\partial \omega_{ij}} = \lambda_j - \mu_{ij}, \quad \mu_{ij}\omega_{ij} = 0, \quad \mu_{ij} \geqslant 0.$$

Multiply both sides of the equation by ω_{ij}:

$$A_{ij} \equiv \omega_{ij} \frac{\partial f(\Omega)}{\partial \omega_{ij}} = \omega_{ij} \lambda_j.$$

By the condition of the Lemma, $\exists i$: $A_{ij} > 0$. Then $\lambda_j > 0$.

If $\frac{\partial f(\Omega)}{\partial \omega_{ij}} < 0$ for some i, then $\mu_{ij} > 0 \;\Rightarrow\; \omega_{ij} = 0$.

Thus, $\omega_{ij}\lambda_j = (A_{ij})_+; \; \lambda_j = \sum\limits_i (A_{ij})_+ \;\Rightarrow\; \omega_{ij} = \operatorname*{norm}\limits_i (A_{ij}).$ ∎

Proof of ARTM equations (by Lemma)

Apply the Lemma to the regularized log-likelihood:

$$f(\Phi,\Theta) = \sum_{d,w} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

$$\varphi_{wt} = \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\frac{\partial f}{\partial \varphi_{wt}}\right) = \operatorname*{norm}_{w\in W}\left(\varphi_{wt}\sum_{d\in D} n_{dw}\frac{\theta_{td}}{p(w|d)} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right) =$$

$$= \operatorname*{norm}_{w\in W}\left(\sum_{d\in D} n_{dw} p_{tdw} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right);$$

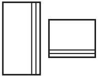
$$\theta_{td} = \operatorname*{norm}_{t\in T}\left(\theta_{td}\frac{\partial f}{\partial \theta_{td}}\right) = \operatorname*{norm}_{t\in T}\left(\theta_{td}\sum_{w\in W} n_{dw}\frac{\varphi_{wt}}{p(w|d)} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) =$$

$$= \operatorname*{norm}_{t\in T}\left(\sum_{w\in d} n_{dw} p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right).$$

■

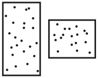
Regularizers for making topics more interpretable

background



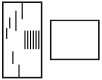
LDA: Smoothing background topics $B \subset T$:
$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w} \beta_w \ln \varphi_{wt} + \alpha_0 \sum_{d} \sum_{t \in B} \alpha_t \ln \theta_{td}$$

sparse



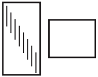
“Anti-LDA”: Sparsing subject domain topics $S = T \backslash B$:
$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w} \beta_w \ln \varphi_{wt} - \alpha_0 \sum_{d} \sum_{t \in S} \alpha_t \ln \theta_{td}$$

seed words



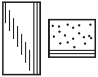
Smoothing relevant topics with seed words vocabulary or query documents

decorrelated



Making topics as different as possible:
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \varphi_{wt} \varphi_{ws}$$

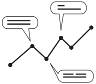
interpretable



Making topics more interpretable by combining regularizers: Decorrelation + Smoothing + Sparsing

Many Bayesian PTMs can be restated as ARTM regularizers

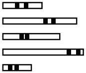
regression



Linear predictive model $\hat{y}_d = \langle v, \theta_d \rangle$ for documents:

$$R(\Theta, v) = -\tau \sum_{d \in D} \Big(y_d - \sum_{t \in T} v_t \theta_{td} \Big)^2$$

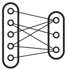
biterm



Using word co-occurrence data n_{uv}:

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \varphi_{ut} \varphi_{vt}$$

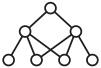
relational



Using document links or citations data n_{dc}:

$$R(\Theta) = \tau \sum_{d,c \in D} n_{dc} \sum_{t \in T} \theta_{td} \theta_{tc}$$

hierarchy

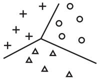


Hierarchical links between topics t and subtopics s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \varphi_{ws} \psi_{st}$$

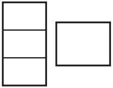
Regularizers for multimodal topic modeling

supervised



The modalities of classes or categories
for text classification or categorization

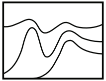
multilanguage



The modalities of languages with translation dictionary
$\pi_{uwt} = p(u|w,t)$ for the $k \to \ell$ language pair:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\varphi_{wt}$$

temporal



Topics dynamics over the modality of time intervals i:

$$R(\Phi) = -\tau \sum_{i\in I} \sum_{t\in T} \bigl|\varphi_{it} - \varphi_{i-1,t}\bigr|$$

geospatial

The modality of geolocations g with proximity $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \Bigl(\frac{\varphi_{gt}}{n_g} - \frac{\varphi_{g't}}{n_{g'}}\Bigr)^2$$

Example 1. Multilingual topic model of Wikipedia

Dataset: 216 175 pairs of parallel Russian–English articles.
Top 10 words and their probabilities $p(w|t)$ in %:

topic #68				topic #79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Assessors evaluated 396 topics from 400 as paired and interpretable

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: open source library for regularized multimodal topic modeling of large collections. 2015.

Example 1. Multilingual topic model of Wikipedia

Dataset: 216 175 pairs of parallel Russian–English articles.
Top 10 words and their probabilities $p(w|t)$ in %:

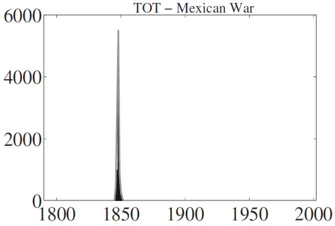
topic #88				topic #251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

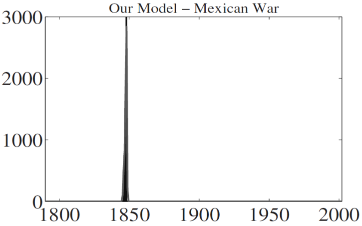
Assessors evaluated 396 topics from 400 as paired and interpretable

K.Vorontsov, O.Frei, M.Apishev, P.Romov, M.Suvorova. BigARTM: open source library for regularized multimodal topic modeling of large collections. 2015.

Example 2. Combining temporal model with n-gram modality

Collection of USA weekly presidential speeches





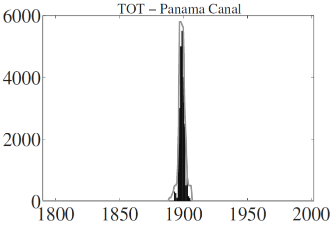
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

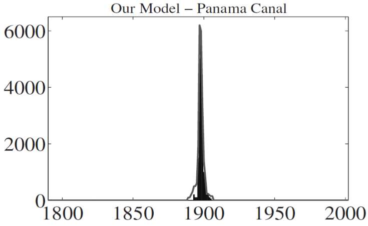
1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An N-gram topic model for time-stamped documents. 2013.

Example 2. Combining temporal model with *n*-gram modality

Collection of USA weekly presidential speeches





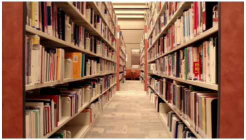
1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An *N*-gram topic model for time-stamped documents. 2013.

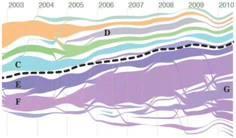
Some of the Topic Modeling applications

exploratory search in digital libraries

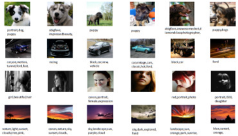


search and recommendation in topical communities

topic detection and tracking in news flows

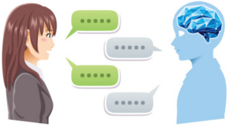


multimodal search for texts and images



mining patterns of customer behavior

dialog management in chatbot intelligence



Topic Model for Digital Humanities applications must be...

1. **Interpretable** so that each topic could tell about itself
2. **Hierarchical** to subdivide topics into subtopics recursively
3. **Temporal** for topic detection and tracking
4. **Multimodal** with authors, categories, tags, links, users, etc.
5. **Multigram** with n-grams being domain concepts
6. **Multilingual** for cross-lingual information retrieval
7. **Segmented** for thematically structured documents
8. **Supervised** for processing expert markups and user logs
9. **Determining number of topics** automatically
10. **Creating and labeling topics** automatically
11. **Online** for fast one-pass data processing
12. **Parallel, distributed** for big data processing

ARTM unifies and simplifies 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 modeling forces new calculus and coding for each model

ARTM introduces the modular “LEGO-style” modeling technology, packing each requirement into a *regularization plug-in*

BigARTM: open source for fast and modular topic modeling

BigARTM features:

- Parallelism + modalities + regularizers + hypergraph
- Out-of-core one-pass processing of large text collections
- Built-in library of regularizers and quality measures

BigARTM community since 2014:

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

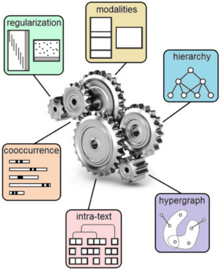
BigARTM license and programming environment:

- 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 cornerstone features of the BigARTM and TopicNet libraries

BigARTM:

- additive regularization
- multimodal data
- topical hierarchy
- intratext regularization
- hypergraph data



TopicNet:

- automated regularization strategies for model selection
- logging experimental conditions and results
- collecting a “topic bank” from multiple modeling runs
- visualization of topic modeling results

V.Bulatov, E.Egorov, E.Veselova, D.Polyudova, V.Alekseev, A.Goncharov, K.Vorontsov. TopicNet: making additive regularisation for topic modelling accessible. LREC-2020

Benchmarking BigARTM vs. Gensim and Vowpal Wabbit

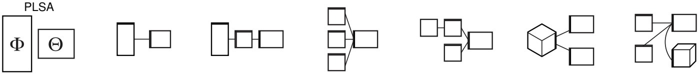
3.7M wiki articles, 100K unique words, $\boxed{\text{time (perplexity)}}$

proc.	$\|T\|$	Gensim	Vowpal Wabbit	BigARTM	BigARTM async
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

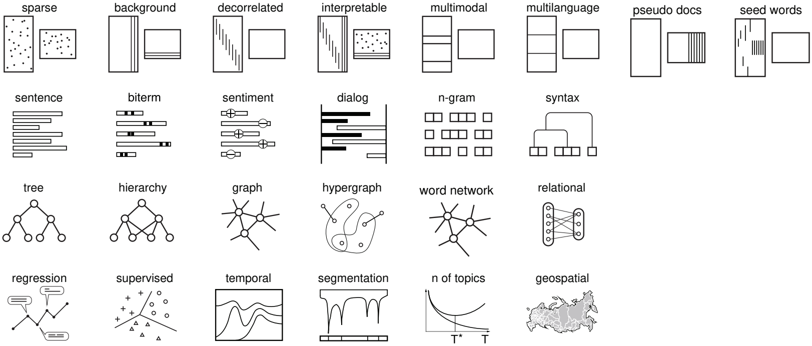
D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.
Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Palette of regularizers in ARTM (the list is open)

Matrix factorization structures:

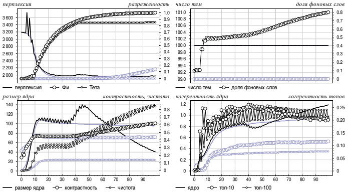


Regularizers to constrain the model or use additional data:



Decorrelation, sparsing and smoothing of topics

Goal: to find a combination of regularizers that improves the interpretability of topics by a set of criteria.



The bag-of-regularizers:

$$\mathscr{L}\Big(\overset{\text{PLSA}}{\Phi\ \ \Theta}\Big) + R\Big(\text{decorrelated}\Big) + R\Big(\text{sparse}\Big) + R\Big(\text{background}\Big) \to \max$$

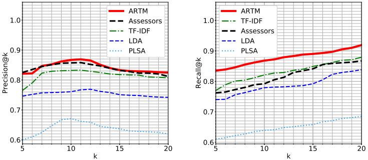
Results:

- topic sparsity $0 \to 95\%$, topic coherence $0.25 \to 0.96$, topic purity $0.14 \to 0.89$, topic contrast $0.43 \to 0.52$,
- without noticeable damage to perplexity: $1920 \to 2020$
- successive regularization strategies have been developed

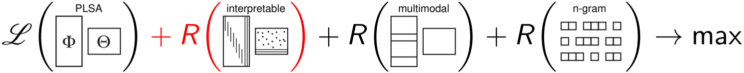
K.Vorontsov, A.Potapenko. Additive regularization of topic models. Mach. Learn., 2015.

Exploratory search in tech news #1

Goal: doc-by-doc exploratory search
— Habr.ru (175K docs)
— TechCrunch.com (760K docs)



The bag-of-regularizers:



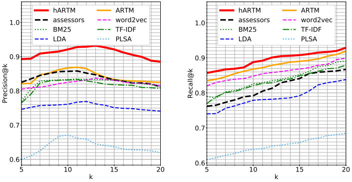
Results:

- Precision and Recall 88% bypass both assessors and baselines (tf-idf, word2vec, PLSA, LDA).
- The topic-based search engine instantly performs the work that people typically complete in about 5–65 minutes.

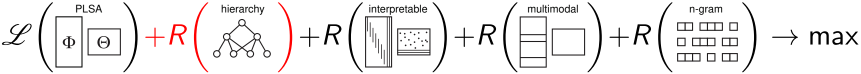
A.Ianina, L.Golitsyn, K.Vorontsov. Multi-objective topic modeling for exploratory search in tech news. AINL, 2017.

Exploratory search in tech news #2

Goal: improving precision and recall of doc-by-doc exploratory search using hierarchical ARTM and cutting off irrelevant topics.



The bag-of-regularizers:



Results:

- Precision and Recall 93% bypass both assessors and baselines (tf-idf, BM25, word2vec, PLSA, LDA, ARTM).
- The optimal dimension of vectors has increased: $200 \to 1400$ (Habr.ru), $475 \to 2800$ (TechCrunch.com).

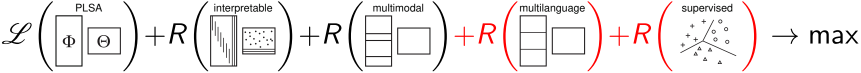
A.Ianina, K.Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT–ISMW, 2019.

Multilingual search and categorization of scientific papers

Goal: multilingual ARTM for 100 languages using multiple library classification systems UDC (УДК), ГРНТИ, ОЭСР, ВАК

модель	ср.ч. УДК	ср.% УДК	ср.ч. ГРНТИ	ср.% ГРНТИ
Базовая ТМ	0.558	0.165	0.536	0.220
XLM-RoBERTa	0.835	0.179	0.832	0.288
ARTM	0.995	0.225	0.852	0.366

The bag-of-regularizers:



Results:

- the accuracy of multilingual search is 94%
- vocabulary reduction to 11K tokens per language (using BPE) results in the model reduction 128 GB → 4.8 GB.

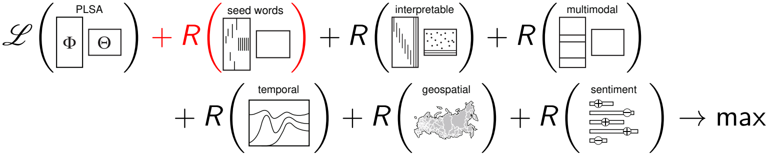
П.Потапова, А.Грабовой, О.Бахтеев, Е.Егоров, Ю.Чехович, К.Воронцов и др. Мультиязыковая автоматическая рубрикация научных документов. 2023 (to appear)

Mining ethnical discourse in social media

Goal: detecting as many topics as possible about nationalities and inter-ethnic relations (using 300 ethnonyms as seed words).

(японцы): японский, япония, корея, китайский, жилиша, авария, фукусиму, цунами, сообщать, океан, станция, хатико, район, правительство, атомный,
(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,

The bag-of-regularizers:



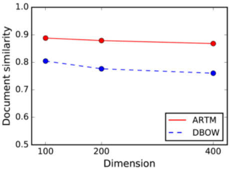
Results: the number of relevant topics 45 (LDA) → 83 (ARTM).

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.

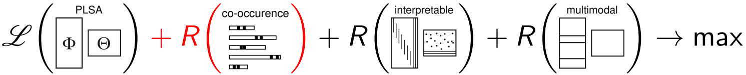
Mining ethnic content online with additively regularized topic models. 2016.

Topic modeling of short texts and probabilistic word embeddings

Goal: sparse interpretable embeddings $p(t|w)$ based on distributional semantics similar to word2vec and WNTM.



The bag-of-regularizers:



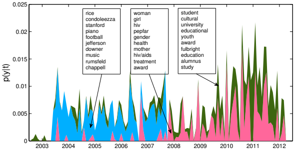
Results:

- Accuracy on document similarity tasks: $0.8 \to 0.9$
- Performance on word similarity tasks: $0.53 \to 0.58$, $0.38 \to 0.61$
- Coherence of topics: $0.08 \to 0.33$
- Modalities improve performance on word similarity tasks

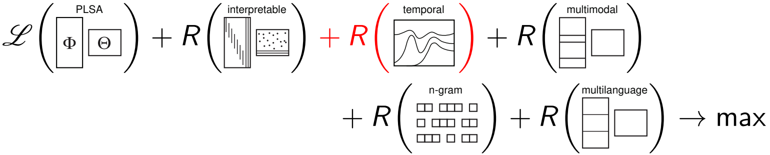
A.Potapenko, A.Popov, K.Vorontsov. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL, 2017.

Topic detection and tracking (TD&T) in news flows

Goal: TD&T in the collection of press releases of the Ministries of Foreign Affairs of 4 countries.



The bag-of-regularizers:



Results:

- classification of topics into permanent and events
- coherence of topics: $5.5 \to 6.5$

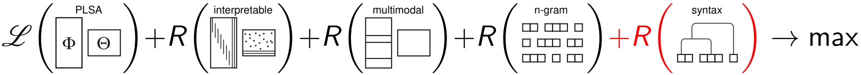
Н.Дойков. Адаптивная регуляризация вероятностных тематических моделей. ВКР бакалавра, ВМК МГУ, 2015.

Unsupervised detection of polarized opinions in political news

Goal: find linguistic-based cues for clustering event topics into polarized opinions

Modalities	*Pr*	*Rec*	*F1*
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

The bag-of-regularizers:



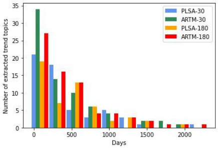
Results:

- detection of opinions within topics: F1-measure = 0.86%
- as a result of the joint use of three modalities: facts as subject–predicate–object (SPO) triplets, semantic roles of words from Fillmore's theory, named entity sentiments.

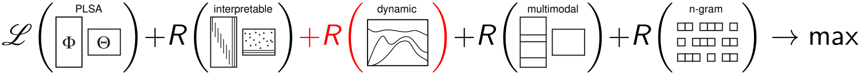
D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Scientific trend detection in big collection of scientific papers

Goal: early detection of trending topics with initial exponential growth in AI/ML research area, 2009–2021.



The bag-of-regularizers:



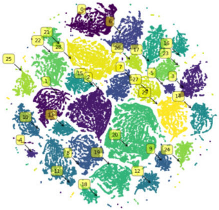
Results:

- automatic detection of 90 from 91 trends in AI/ML area
- 63% of topics are detected in a year, 79% in two years

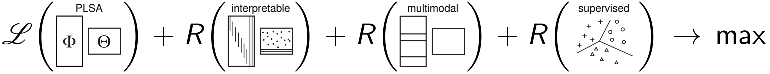
N.Gerasimenko, A.Chernyavskiy, M.Nikiforova, M.Nikitin, K.Vorontsov. Incremental topic modeling for scientific trend detection Doklady RAS, 2022.

Topic modeling of bank transaction data

Goal: reveal patterns of consumer behavior from purchase transaction data;
document = consumer,
word = MCC (Merchant Category Codes).



The bag-of-regularizers:



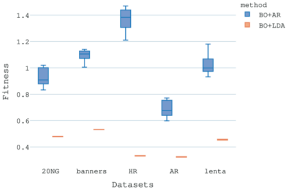
Results:

- topics are interpretable patterns of consumer behavior
- consumer topical behavior profile $p(t|d)$ can be used for predicting gender, age, wealth, interests, etc

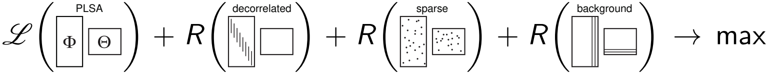
E.Egorov, F.Nikitin, A.Goncharov, V.Alekseev, K.Vorontsov. Topic modelling for extracting behavioral patterns from transactions data. 2019.

Automatic learning of regularization coefficients

Goal: AutoARTM is automatic optimization (AutoML) of hyperparameters such as regularization coefficients, number of iterations, number of topics according to the topic coherence criterion.



The bag-of-regularizers:



Results:

- Significant improvement in topic coherence across 5 datasets
- Genetic algorithm showed the best results

M.Khodorchenko, S.Teryoshkin, T.Sokhin, N.Butakov. Optimization of learning strategies for ARTM-based topic models. LNCS, 2020.

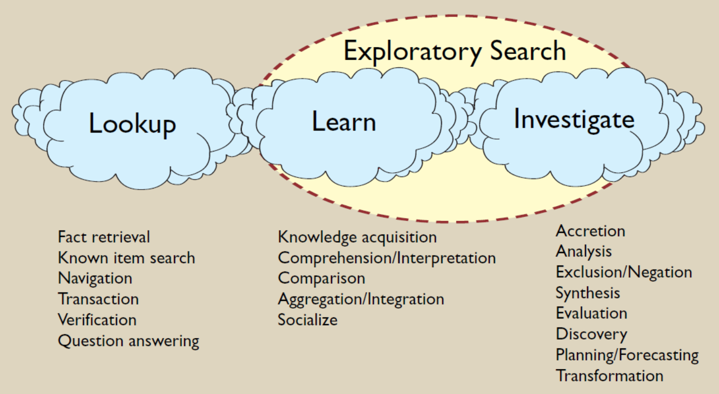
Probabilistic Topic Modeling: conclusions

- 100s of models over 20 years of advances in PTM have been elaborated within overcomplicated Bayesian framework.
- All the while, a high potential of the classical non-Bayesian regularization went almost untested and unnoticed.
- ARTM transforms PTM into «a theory of single Lemma».
- Perhaps, if the community knew about this Lemma, the theory of PTM would not develop within Bayesian framework.
- The Lemma is applicable for learning *Neural Topic Models* as networks with normalized non-negative vector parameters.
- Could this be a way towards interpretable neural networks?

K.V.Vorontsov. Rethinking probabilistic topic modeling from the point of view of classical non-Bayesian regularization. 2023.
http://www.machinelearning.ru/wiki/images/7/76/Voron23rethinking.pdf
Rob Churchill, Lisa Singh. The Evolution of Topic Modeling. November, 2022.

Exploratory Search for learning, knowledge acquisition and discovery

- the user may not know exactly which keywords to use
- the user may not looking for a single answer



Gary Marchionini. Exploratory Search: from finding to understanding. 2006.

The idea of Knowledge Factory

"An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: a depot where knowledge and ideas are received, sorted, summarized, digested, clarified and compared"
— *Herbert Wells, 1940*

The Big Search do not worry about what the user will do with search results.

Knowledge Factory is about to guide the user through further steps of automated knowledge processing and understanding:

- search for collecting
- collect for analyzing
- analyze for understanding
- understand for applying and teaching

Document collection as a principal user's workspace

The document collection is a long-term search interest of the user

Search and recommendation functions:

- searching documents thematically similar to the collection
- detecting new documents relevant to the collection
- generating contextual "see also" recommendations

Analytical functions

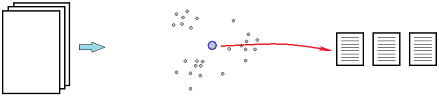
- machine aided human summarization of the collection
- clustering trends, methods, ideas, opinions in the collection
- recommending the reading order of documents in the collection

Communicative functions:

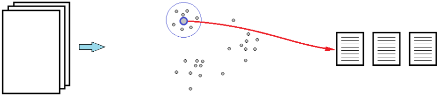
- shared creating, reading and discussing of the collection
- shared visualization and analysis of the collection

Strategies of vector-based document-by-document search

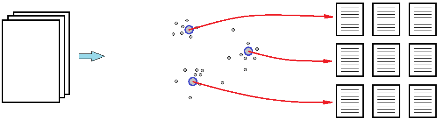
Searching by center vector of the collection (inadequate strategy):



Searching by document from the collection or by document cluster:

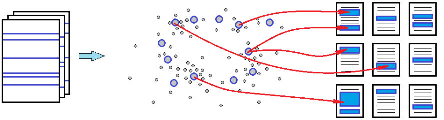


Searching by all cluster vectors of the collection:

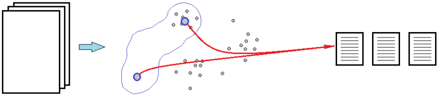


Strategies of vector-based document-by-document search

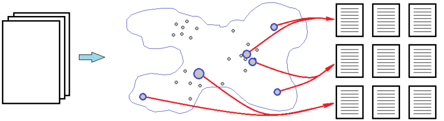
Searching by segment vectors of documents:



Searching by topics adjacent to a part of the collection:



Searching by topics adjacent to a whole collection:

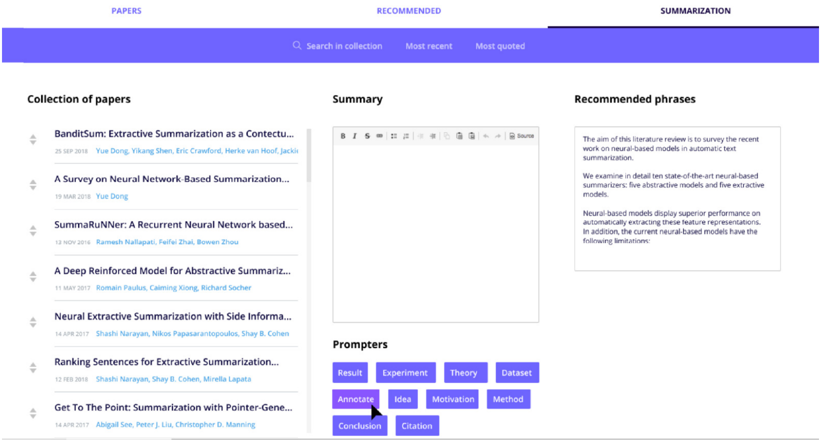


Machine Aided Human Summarization of scientific documents

1. **The MAHS system** recommends the summarization script as an ordered list of documents to be mentioned
2. **The user** corrects the script, bringing it in line with its goals and creative intention
3. In the cycle by all ordered documents in the script:
   - **the user** asks one of the aspect prompters:
     — the main idea of the document,
     — how other authors cite this document,
     — method, — benefit, — flaw, — result, — conclusion etc.
   - **the MAHS prompter** generates a ranked list of phrases
   - **the user** selects a phrase from the ranked list, inserts it into the summary and adjusts it in accordance with his intention

A.Vlasov. Machine aided multi-document summarization of scientific papers. MIPT, 2020.
С.Крыжановская. Технология полуавтоматической суммаризации тематических подборок научных статей. 2022. ВМК МГУ.

Machine Aided Human Summarization of scientific documents



A.Vlasov. Machine aided multi-document summarization of scientific papers. MIPT, 2020.
С.Крыжановская. Технология полуавтоматической суммаризации тематических подборок научных статей. 2022. ВМК МГУ.

Machine Aided Human Summarization of scientific documents

Machine Learning problems for MAHS:

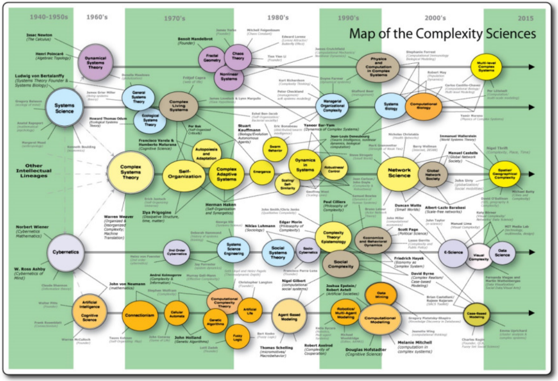
1. The training sample generation: paper $\to$ (refs, survey)
2. Documents ranking for the summarization script
3. Selection of relevant phrases for the prompter
4. Ranking of selected relevant phrases for each prompter
5. Selecting a text fragment relevant context around the link:

Few contextual citation graphs are publicly available. The ACL Anthology Network (AAN) (Radev et al., 2009) is one such contextual citation graph built from the ACL Anthology corpus (Bird et al., 2008), consisting of 24.6K papers manually augmented with citation information. CiteSeer (Giles et al., 1998) provides a large corpus consisting of 1.0M papers with full text and bibliography entries parsed from PDFs. Saier and Farber (2019) introduces a contextual citation graph of approximately 1.0M arXiv papers with full text LaTeX parses where citations are linked to papers in the Microsoft Academic Graph.

M.Yasunaga et al. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. 2019.

An example of a domain map (hand made)

Open problem is to build such maps with topics, trends, authors, links for any given domain area (Y-axis) and time interval (X-axis)



http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences

A source of inspiration: http://textvis.lnu.se

A visual survey of 440 text visualization techniques

Shixia Liu, Weiwei Cui, Yingcai Wu, Mengchen Liu. A survey on information visualization: recent advances and challenges. 2014.
Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // JMLDA, 2015.

Knowledge Factory: conclusions

- We build the Exploratory Search applications upon text vectorization techniques such as automatic term extraction, topic modeling, and transformer deep neural networks
- Machine Aided Human Summarization of scientific documents is a way of non-linear reading, understand big volumes of content, and authoring summaries bringing in line with the user's goals and creative intention
- For scientific Exploratory Search, we develop (jointly with Sber AI team) ruSciBERT, a BERT pre-trained on a collection of Russian scientific papers.
- Near future, we plan to start with a project on multilingual patent search (Russian, English, Chinese, French, Spanish etc.)

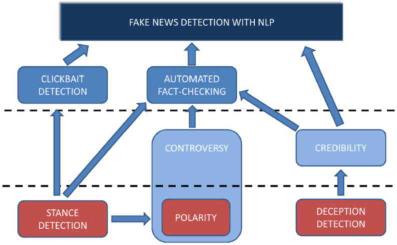
The idea of News Collider

Physicists collide streams of particles for learn more about the structure of matter

We collide news flows for learn more about post-truth and protect society from destructive impacts and information wars

Fake News Detection as academic research area

1. Deception Detection
2. Automated Fact-Checking
3. Stance Detection
4. Controversy Detection
5. Polarization Detection
6. Clickbait Detection
7. Credibility Scores



There are datasets, contests, models... But something is missing

- **Fakes** is not the only instrument of post-truth politics
- **Propaganda** uses also juggling of facts, silencing, emotives
- **InfoWar** attacks the cultural code: ideas, values, attitudes

---

E.Saquete et al. Fighting post-truth using natural language processing: a review and open challenges. Expert Systems With Applications, Elsevier. 2020.

Destruction detection: towards a fuller typology of NLP tasks

Manipulations are common in interpersonal communications
Fakes is not the only instrument of post-truth politics
Propaganda uses also juggling of facts, silencing, emotives etc.
InfoWar attacks the cultural code: ideas, values, attitudes

Impact (manipulation) → fakes → propaganda → infowar

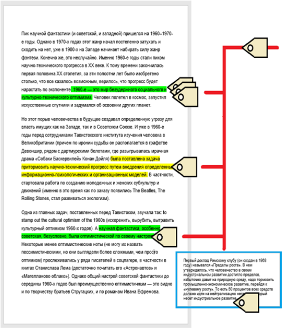
1. ■□■■ manipulation detection
2. ■□■■ silencing / understatement detection
3. □■■■ **deception / rumors / hoaxes detection**
4. □■□□ **clickbait detection**
5. □■□□ **automatic fact-checking**
6. □■■□ **stance / controversy / polarization detection**
7. □□■■ constructs of the worldview: values, ideologems, mythologems
8. ■□■■ reader's emotions detection
9. ■□■■ target audiences detection
10. □■■■ **virality prediction**
11. □■■□ **evaluation of credibility scores**
12. □□□■ detection of destructive influences (threats, recruitment, extremism)

Technically, there are four main types of NLP tasks

1. Classification of a text
   - **deception detection, fact-checking, text credibility**
2. Classification of a pair of texts
   - **stance, controversy, polarization, clickbait detection**
   - identification of disagreements, silencing, understatement
3. Selection and classification of a text fragment
   - **extraction of linguistic-based cues from the text**
   - detection of manipulation techniques
   - detection of worldview constructs, ideologems, mythologems
   - detection of reader's emotions and target audiences
4. Clustering or topic modeling of a text collection
   - **detection of politically polarized opinions**
   - detection of attitudes towards socio-cultural values

Unification of text fragments markup made by experts

The unified markup structure is common for most classical NLP tasks (NER, SentAn, SemRL, SyntPars, etc.) and for many complicated NLU tasks such as detection of semantic errors in academic essays, and known destruction detection tasks.

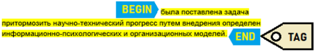


The markup consists of elements:

Element can contain any number of labeled fragments and comments linked together

Labels from a dictionary defined by experts

Fragment is defined by its start/end positions and can have one or multiple labels (or tags):



Comment can be selected from a phrase dictionary or freely authored by the expert, it also can have one or more labels or tags

READ//ABLE tech contest (http://ai.upgreat.one). Technical Regulations. 2019–2022.

The task of READ//ABLE tech contest

Task: detection of semantic errors in schoolchildren essays for USE in Russian, Literature, Social science, History and English.

Contest period: Dec 2019 – Dec 2022

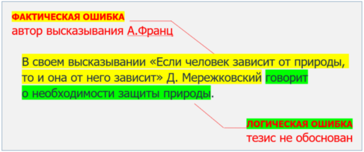
Prize fund:

— 100mln RUB (Russian)
— 100mln RUB (English)

Up Great technology contest READ // ABLE comprehension technology

152 error types:
(R:70 L:16 S:23 H:20 E:23)
236 error subtypes:
(R:112 L:19 S:29 H:26 E:50)

Not only detect the error but also provide an explanation for it.



READ//ABLE tech contest (http://ai.upgreat.one). Technical Regulations. 2019–2022.

The task of Propaganda/Manipulation/Persuasion Detection

Simple markup: «fragment, class label»

Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitanis Garumna flumen, a Belgis Matrona et Sequana dividit. Horum omnium fortissimi sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos animos pertinent important, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proeliis cum Germanis contendunt, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language

Attack on Reputation: Smears

Manipulative Wording: Exaggeration

Justification: Appeal to Values

Commissio
PopulusQue
Europaea

Simplified markup: «short text, class label»

Advanced markup: «persuasion-fragment, target-fragment, label»

SemEval-2023 task 3. Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup.
https://propaganda.math.unipd.it/semeval2023task3

G.Martino, P.Nakov et al. A survey on computational propaganda detection. 2020.

Unification of evaluation techniques

- Consist(A, B) is consistency between markups A and B based on the optimal matching of their elements
- ACAM (Average Consistency of Algorithmic Markup) is defined by averaging Consist(A, E) between model A and expert E markups
- ACEM (Average Consistency of Expert Markup) is defined by averaging Consist(E_1, E_2) between expert markups E_1 and E_2
- RCAM (Relative Consistency of Algorithmic Markup) is defined as the ratio ACAM / ACEM; if it is greater than 100%, then the algorithm outperforms experts

In READ//ABLE tech contest, exceeding the technological barrier RCAM $>$ 100% was a condition for finishing the competition.

READ//ABLE tech contest (http://ai.upgreat.one). Technical Regulations. 2019–2022.

News Collider: conclusions

- Confronting destructive ideological pressure and imacts of information warfare is a mission and challenge for AI & DH interdisciplinary scientific community
- Last years, Large-scale Pre-trained Language Models demonstrate the impressive ability to solve increasingly difficult NLU tasks such as detecting destruction and other threats in the media information space
- The markup of text fragments is a mainstream way towards formalization of humanitarian knowledge in psycholinguistics, history, political science and other areas
- We are moving towards standardization of a pipeline «markup $\to$ modeling $\to$ evaluation» for difficult NLU tasks