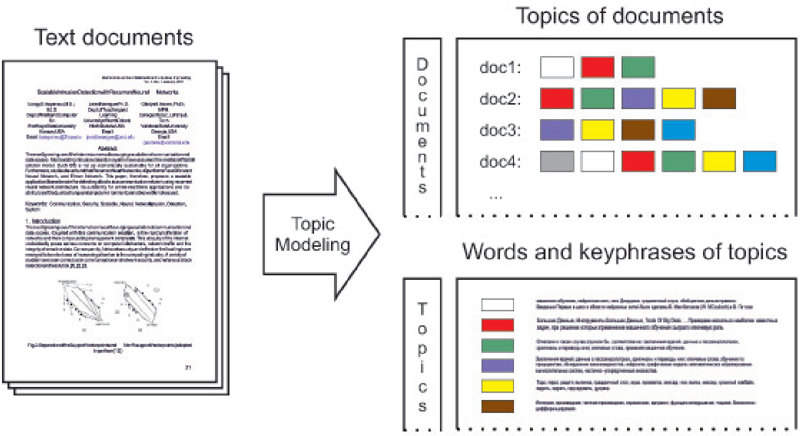
Additive Regularization of Topic Models

Anna Potapenko

National Research University Higher School of Economics,
Yandex School of Data Analysis, Moscow, Russia

November 28, 2017
ETH Zurich

Topic modeling – revealing a hidden thematic structure of a text collection – soft clustering of words and documents:



Probabilistic statement of the problem

Given:

- D — a collection of documents; W — a vocabulary of words;
- $p(w|d)$ — frequencies of words w in documents d

Find:

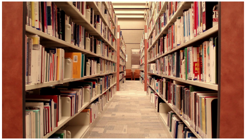
- $\phi_{wt} = p(w|t)$ — a distribution over words for each topic
- $\theta_{td} = p(t|d)$ — a distribution over topics for each document

Basic assumptions:

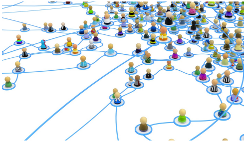
- A *topic* is a set of coherent words that often co-occur in the documents.
- A document is a *bag of words*.
- Each *observed* word in a document has a *latent* topic.

Topic modeling applications

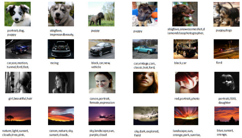
exploratory search
in digital libraries



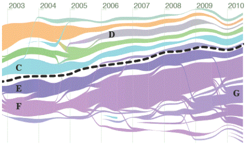
personalized search
in social media



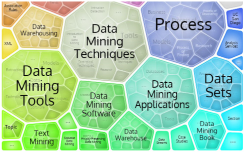
multimodal search
for texts and images



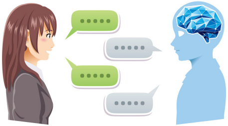
topic detection and
tracking in news flows



navigation in big
text collections

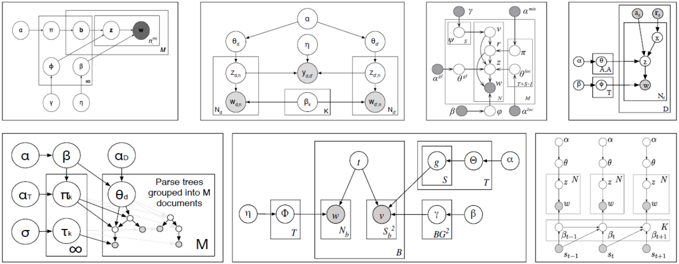


dialog manager in
chatbot intelligence



Probabilistic Topic Modeling: milestones and mainstream

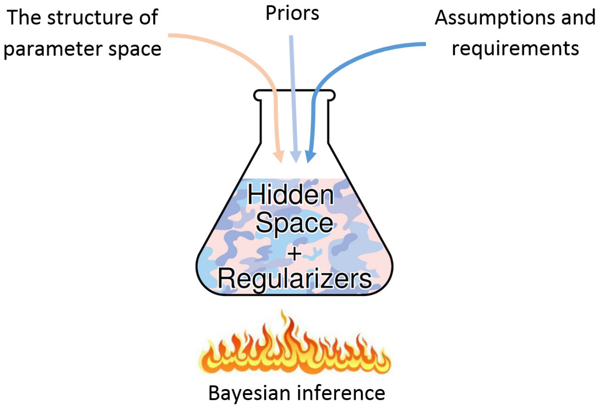
1. PLSA — Probabilistic Latent Semantic Analysis (1999)
2. LDA — Latent Dirichlet Allocation (2003)
3. 100s of PTMs based on Graphical Models & Bayesian Inference



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

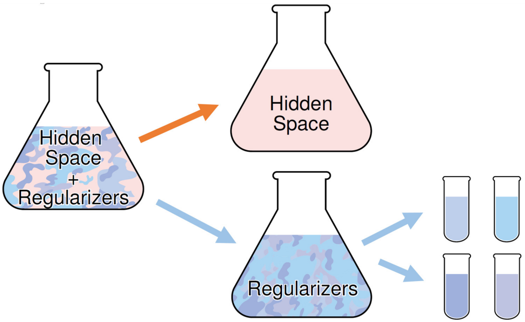
Bayesian approach in Topic Modeling

The *generative process* encapsulates all our knowledge about the hidden space structure, prior distributions, and requirements



Non-Bayesian regularization for Topic Modeling

- A *simple generative process* describes the hidden space
- Regularizers describe most of the requirements and assumptions
- Regularizers can be additively mixed and interchanged



PLSA: Probabilistic Latent Semantic Analysis

Generative model explains terms w in documents d by topics t:

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

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

$$\sum_{d,w} n_{dw} \ln \sum_t \phi_{wt}\theta_{td} \;\to\; \max_{\Phi,\Theta},$$

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

Solution is obtained via iterations of EM-algorithm.

ARTM: Additive Regularization of Topic Model

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

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

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

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

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

Voronsov K.V., Potapenko A. A. — Additive Regularization of Topic Models // Machine Learning Journal, 2014. DOI: 10.1007/s10994-014-5476-6.

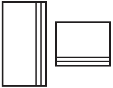
Regularized EM-algorithm

Theorem. If Φ, Θ is the solution of the regularized likelihood maximization problem, then it satisfies the following system of equations with auxiliary variables $p_{dwt} = p(t|d,w)$:

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

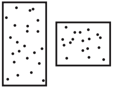
where $\underset{t\in T}{\operatorname{norm}}\, x_t = \frac{\max\{x_t,0\}}{\sum\limits_{s\in T}\max\{x_s,0\}}$ is non-negative normalization;

Regularizers for the interpretability of topics

background


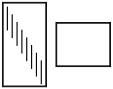
Smoothing background topics $B \subset T$:

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

sparse


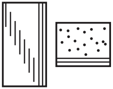
Sparsing subject domain topics $S = T \setminus B$:

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

decorrelated


Making topics as different as possible:

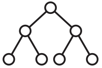
$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \phi_{wt} \phi_{ws}$$

interpretable


Making topics more interpretable
by combining the above regularizers

Revisiting Bayesian topic models

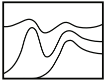
hierarchy



Hierarchical links between topics t and subtopics s:

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

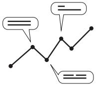
temporal



Topics dynamics over the modality of time intervals i:

$$R(\Phi) = -\tau \sum_{i\in I} \sum_{t\in T} \left|\phi_{it} - \phi_{i-1,t}\right|.$$

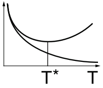
regression



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

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

n of topics

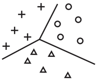


Sparsing $p(t)$ for topic selection:

$$R(\Theta) = -\tau \sum_{t\in T} \frac{1}{|T|} \ln p(t), \quad p(t) = \sum_{d} p(d)\theta_{td}.$$

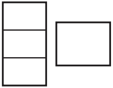
Special cases of the multimodal topic modeling

supervised



The modalities of classes or categories
for text classification and categorization.

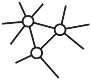
multilanguage



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

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

graph



The modality of graph vertices v with doc sets D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v) \in E} S_{uv} \sum_{t \in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|} \right)^2.$$

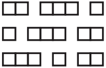
geospatial

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

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

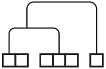
Beyond the "bag-of-words" restrictive hypothesis

n-gram



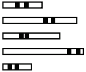
The modalities of n-grams, collocations, named entities

syntax



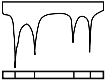
The modality of n-grams after SyntaxNet preprocessing

coherence



Modeling co-occurrence data n_{uv} for biterms (u, v):
$R(\Phi) = \tau \sum_{u,v} n_{uv} \ln \sum_{t} n_t \phi_{ut} \phi_{vt}$

segmentation



E-step regularization affecting $p(t|d, w)$ distributions for segmentation and sentence topic models

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

Benchmarking BigARTM vs. Gensim and Vowpal Wabbit

- 3.7M articles from Wikipedia, 100K unique words

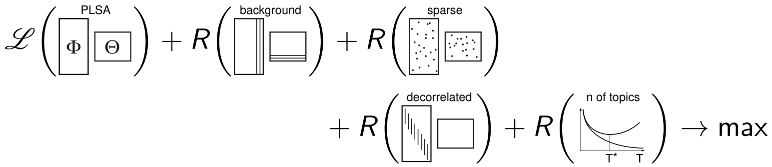
	procs	train	inference	perplexity
BigARTM	1	35 min	72 sec	4000
Gensim.LdaModel	1	369 min	395 sec	4161
VowpalWabbit.LDA	1	73 min	120 sec	4108
BigARTM	4	9 min	20 sec	4061
Gensim.LdaMulticore	4	60 min	222 sec	4111
BigARTM	8	4.5 min	14 sec	4304
Gensim.LdaMulticore	8	57 min	224 sec	4455

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

Experiment (on NIPS papers dataset)

Goal: Improve interpretability of topics and do topics selection.

Lego of regularizers:



Multiple quality measures:

- fitting the data: perplexity
- interpretability: topics coherence
- diversity: topics purity, contrast

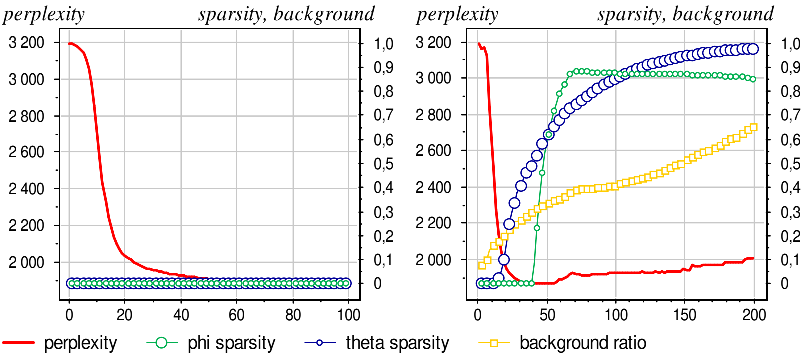
Vorontsov K.V., Potapenko A.A. and Plavin A.V. — Additive Regularization of Topic Models for Topic Selection and Sparse Factorization. SLDS 2015, UK.

Topics examples (top-20 words)

- **PLSA:** face, images, faces, recognition, set, image, based, hme, facial, representation, view, figure, model, experts, network, human, expert, space, examples, system
- **ARTM, domain:** face, faces, facial, cottrell, pentland, gesture, lane, emotion, person, steering, appearance, baluja, setpoint, camera, tracking, pose, pomerleau, mouth, darrell
- **ARTM, background:** model, data, models, parameters, noise, neural, mixture, prediction, set, gaussian, likelihood, networks, test, figure, training, performance, network, number

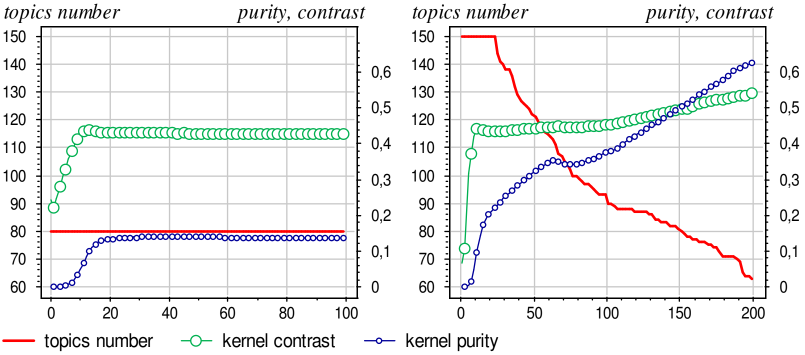
We type in red those words that are included into kernels.

Lego of regularizers



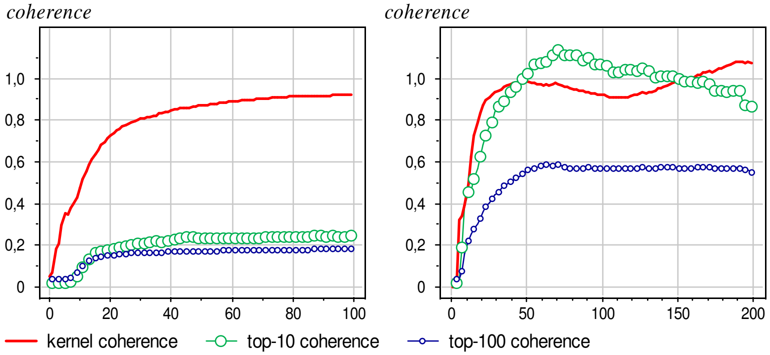
1. We achieve extremely high sparsity of Φ and Θ matrices.
2. Perplexity deteriorates mainly due to topics number decreasing.

Lego of regularizers



1. The number of topics gradually decreases from 150 to 60 by eliminating the most insignificant topics at each iteration.
2. We can stop at the appropriate moment by other criteria.
3. The proposed model achieves high topics purity and contrast.

Lego of regularizers



1. All kinds of coherence measures increase although it has not been explicitly incorporated into the optimization problem.

Bridging the gap between topic models and word embeddings

- Topic models (e.g. PLSA) learn probabilities of words in topics ϕ_w and topics in documents θ_d:

$$p(w|d) = \sum_t p(w|t)p(t|d) = \langle \phi_w, \theta_d \rangle$$

- Word embedding models (e.g. SGNS) learn real-value vectors of words ϕ_w and contexts θ_c (usually also words):

$$p(w|c) = \frac{\exp \langle \phi_w, \theta_c \rangle}{\sum_{w \in W} \exp \langle \phi_w, \theta_c \rangle}$$

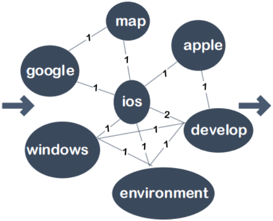
Anna Potapenko, Artem Popov, and Konstantin Vorontsov — Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. AINL 2017.

Bridging the gap between topic models and word embeddings

Define pseudo-documents based on word co-occurrences in local contexts and apply ARTM approach.

Combine the best of two worlds: good performance for word similarity task and interpretability of the components.

doc1	Google Map for iOS
doc2	iOS is developed by Apple
doc3	iOS develop environment for Windows



word	pseudo-document
map	google ios
google	map ios
apple	develop ios
ios	google map apple windows environment develop develop
windows	environment ios develop
develop	apple ios ios windows environment
environment	ios windows develop

Zuo Y. et al. — Word Network Topic Modeling, 2016.

Combining the best of two worlds: word similarity task

Data: English Wikipedia dump
Quality: Spearman correlation with ranked list of word pairs
Embedings size: 400 for all the models in this slide

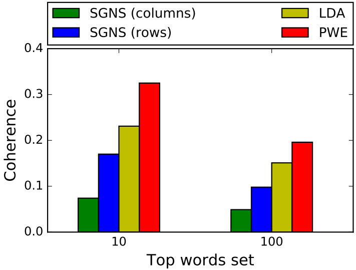
	sim	WordSim Sim.	WordSim Rel.	WordSim	Bruni MEN	Rad. Turk
LDA	hel	0.553	0.478	0.493	0.583	0.51
SVD (PPMI)	cos	0.711	0.648	0.672	0.236	0.616
SGNS	cos	**0.752**	0.633	0.665	0.**744**	**0.661**
ARTM (offline)	dot	0.71	0.62	0.65	0.67	0.59
ARTM (online)	dot	0.723	**0.675**	**0.682**	0.672	0.642
ARTM (sparse)	dot	0.728	0.672	0.68	0.675	0.635

ARTM embeddings perform on par with SGNS on word similarity task and are highly sparse (93% of zeros).

Combining the best of two worlds: interpretability of the components

Quality: Topic coherence (known to correlate with human scores):

$$Coherence(t) = \frac{2}{k(k-1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} PPMI(w_i, w_j)$$



art	arbitration
painting	ban
museum	requests
painters	arbitrators
gallery	noticeboard
sculpture	block
painter	administrators

ARTM embeddings have higher interpretability than SGNS or LDA.

Combining the best of two worlds: cross-modality similarities

Data: Russian news (lenta.ru) with several *modalities*: title, text, date, and category.

Star wars release 2015-12-18	The Oscars 2016-02-29	Victory Day 2015-05-09
jedi	statuette	great
sith	award	anniversary
fett	nomination	normandy
anakin	linklater	parade
chewbacca	oscar	demonstration
film series	birdman	vladimir
hamill	win	celebration
prequel	criticism	concentration
awaken	director	auschwitz
boyega	lubezki	photograph

Combining the best of two worlds: word analogies task

Quality: Cherry-picked examples!

Query	ARTM	SGNS
king + girl – boy	queen, princess, lord, prince	queen, princess, regnant, kings
moscow + spain – russia	madrid, barcelona, aires, buenos	madrid, barcelona, valladolid, malaga
ruble + india – russia	rupee, birbhum, pradesh, madhaya	rupee, rupiah, devalued, debased
better + bad – good	really, something, thing, nothing	worse, easier, prettier, funnier
computer + cars – car	computers, software, servers, implementations	computers, software, hardware, microcomputers

Conclusions and discussion

- Additive Regularization for Topic Models (ARTM) is a general framework that makes it easy to design topic models.
- BigARTM is a fast open source implementation.
- ARTM approach can be used to build interpretable sparse embeddings for words and documents. What's next?

Contacts:

- Anna Potapenko: anna.a.potapenko@gmail.com
- Prof. Konstantin Vorontsov: voron@forecsys.ru