

Regularization of Topic Models for Question Answering

Konstantin Vorontsov, Anna Potapenko

(MIPT, Yandex • Moscow, Russia)

DeepHack.Q&A

2016/01/31 – 2016/02/06 • Moscow, Russia
DeepHackLab • Moscow Institute of Physics and Technology

- 1 Question Answering**
 - The Allen AI Science Challenge
 - Statistical Approach to Question Answering
 - Combining Generative and Discriminative Approaches

- 2 Probabilistic Topic Modeling**
 - Basic topic models PLSA and LDA
 - ARTM — Additive Regularization for Topic Modeling
 - BigARTM open source project

- 3 Probabilistic Topic Modeling for Question Answering**

Demotivation. Can a model be smarter than an 8th grader?

The goal of **Aristo project** is to demonstrate the ability of AI to consistently understand and correctly answer general questions about the world.

But AI2-Kaggle QA Challenge is a toy problem:

- Dataset is not true test for general human knowledge because the 4-answers-test ignores important aspects of Human Intelligence
- Working methods are not true Artificial Intelligence because BigData, computation and simple information retrieval heuristics are sufficient
- Your implementation will not true Question Answering system because real-life QA systems retrieve information from the web and rank thousands potential answers

Basic statistical approach

- 1 Download external collections: wiki, textbooks, etc.
- 2 Preprocess texts and generate vocabulary
- 3 Split each text into (approximately equal) segments s
- 4 (Optionally) expand questions q by synonyms
- 5 $\text{Rel}(q|s)$: the relevance of a question q to a segment s
- 6 $\text{Rel}(a|s)$: the relevance of an answer a to a segment s
- 7 The relevance of an answer to a question:

$$\text{Rel}(a|q) = \sum_s \text{Rel}(a|s) \text{Rel}(q|s)$$

- 8 Use several relevance measures as features of (q, a) pairs
- 9 Learn a classifier on a training set of (q, a) pairs

Options and variants

Step 2: Preprocess texts and generate vocabulary

- use stemming
- use stop-word filtration
- use TF-IDF word filtration
- use Named Entity Recognition or Term Extraction
- use all n -grams extracted from questions ($n = 2, 3$)

Step 3: Split each text into (approximately equal) segments s

- segment is one or more paragraphs, length from n_1 to n_2 words
- segment is a sequence of sentences, length from n_1 to n_2 words
- use half-overlapping segments

Options and variants

Step 4: Question expansion

- use synonyms from WordNet and other linguistic resources
- use coherent words
- use topic model, word2vec or other word embedding technique

Step 5, 6: Relevance $\text{Rel}(x|s)$ of a text x to a segment s

- compare x and s as bag of words (or terms)
- compare x and s as distributions over vocabulary
- compare x and s as averaged word vector representations

Relevance $\text{Rel}(x|s)$ of a text x to a segment s (the bigger the better)

- *Jaccard similarity* compares x and s as subsets of vocabulary:

$$\text{Rel}(x|s) = \frac{|x \cap s|}{|x \cup s|}$$

- *KL-divergence similarity* compares x and s as distributions over W :

$$\text{Rel}(x|s) = e^{-\gamma \text{KL}(x||s)}, \quad \text{KL}(x||s) = \sum_w x(w) \log \frac{x(w)}{s(w)}$$

- *JS-divergence similarity* compares x and s as distributions over W :

$$\text{Rel}(x|s) = e^{-\gamma \text{JS}(x,s)}, \quad \text{JS}(x,s) = \text{KL}(x||\frac{x+s}{2}) + \text{KL}(s||\frac{x+s}{2})$$

- *Cosine similarity* compares x and s as vectors over W :

$$\text{Rel}(x|s) = \frac{\sum_w x(w)s(w)}{\sqrt{\sum_w x^2(w)}\sqrt{\sum_w s^2(w)}}$$

Using relevance measures as features of a/q pairs

Combination of options & variants produces > 360 features:

- 6 or more preprocessing options
- 3 or more document split variants
- 5 or more question expansion variants
- 4 or more similarity measures

Many machine learning techniques to learn a classifier:

- SVM — support vector machine
- GBM — gradient boosting machine
- ANN — artificial neural network, etc.

What is “topic”?

Topic is a specific terminology of a particular domain area

Topic is a set of coherent terms that often co-occur in documents.

Topic model uncovers latent semantic structure of a text collection:

- *topic* t is a probability distribution $p(w|t)$ over terms w
- *document* d is a probability distribution $p(t|d)$ over topics t

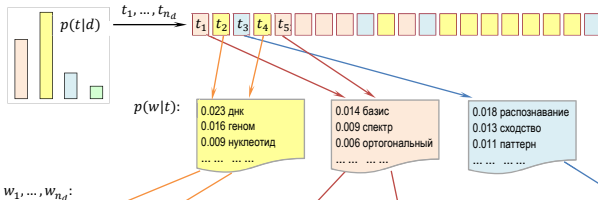
Motivations to use Probabilistic Topic Models (PTM) for QA:

- semantic similarity is more powerful than the lexical one
- topics gather synonyms automatically
- regularization enables to learn PTM with desired properties

PTM is a generative model of a text collection

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

$$p(w|d) = \sum_t p(w|t)p(t|d)$$



Разработан спектрально-аналитический подход к выявлению размытых протяженных повторов в **геномных** последовательностях. Метод основан на разномасштабном оценивании сходства **нуклеотидных** последовательностей в пространстве коэффициентов разложения фрагментов кривых GC- и GA-содержания по классическим **ортогональным базисам**. Найлены 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 **синтезии** при сравнении пары **геномов**. Его можно использовать для детального изучения фрагментов **хромосом** (поиска размытых участков с умеренной длиной повторяющегося **паттерна**).

Inverse problem: text collection \rightarrow PTM

Given: D is a set (collection) of documents

W is a set (vocabulary) of terms

n_{dw} = how many times term w appears in document d

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

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

under nonnegativity and normalization constraints

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

The ill-posed problem of matrix factorization:

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

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

PLSA — Probabilistic Latent Semantic Analysis [Hofmann, 1999]

Constrained maximization of the log-likelihood:

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

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

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

where $\text{norm}_{t \in T} x_t = \frac{\max\{x_t, 0\}}{\sum_{s \in T} \max\{x_s, 0\}}$ is vector normalization.

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

Maximum a posteriori (MAP) **with Dirichlet prior**:

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

EM-algorithm is a simple iteration method for the system

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

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

Maximum log-likelihood with additive regularization criterion R :

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

EM-algorithm is a simple iteration method for the system

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

Many Bayesian PTMs can be reinterpreted as regularizers in ARTM

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

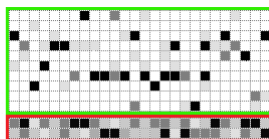
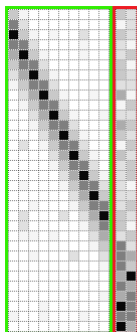
Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models // Machine Learning. Volume 101, Issue 1 (2015), Pp. 303-323.

Assumptions: what topics would be well-interpretable?

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

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

ϕ_{wt} terms \times topics θ_{td} topics \times documents



Smoothing (rethinking LDA)

The non-sparsity assumption for background topics $t \in B$:

ϕ_{wt} are similar to a given distribution β_w ;

θ_{td} are similar to a given distribution α_t .

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

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

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

$$\phi_{wt} = \text{norm}_w(n_{wt} + \beta_0 \beta_w), \quad \theta_{td} = \text{norm}_t(n_{td} + \alpha_0 \alpha_t),$$

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

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

Sparsing (further rethinking LDA)

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

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

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

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

$$\phi_{wt} = \text{norm}_w(n_{wt} - \beta_0 \beta_w)_+, \quad \theta_{td} = \text{norm}_t(n_{td} - \alpha_0 \alpha_t)_+.$$

Varadarajan J., Emonet R., Odobez J.-M. A sparsity constraint for topic models — application to temporal activity mining // NIPS-2010 Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions.

Decorrelation

The dissimilarity assumption:

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

Maximize covariances between column vectors ϕ_t :

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

The regularized M-step makes columns of Φ more distant:

$$\phi_{wt} = \text{norm}_w \left(n_{wt} - \tau \phi_{wt} \sum_{s \in S \setminus t} \phi_{ws} \right)_+.$$

Tan Y., Ou Z. Topic-weak-correlated latent Dirichlet allocation // 7th Int'l Symp. Chinese Spoken Language Processing (ISCSLP), 2010. — Pp.224–228.

Topic selection

Assumption: infrequent topics are not well-interpretable.

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

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

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

$$\theta_{td} = \text{norm}_t \left(n_{td} - \tau \frac{n_d}{n_t} \theta_{td} \right)_+.$$

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

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

Combining topic models

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

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

where τ_i are regularization coefficients.

EM-algorithm is a simple iteration method for the system

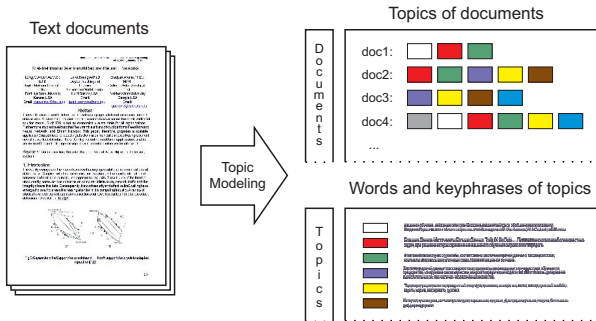
$$\begin{cases} \text{E-step:} & p_{tdw} = \text{norm}_{t \in T}(\phi_{wt} \theta_{td}) \\ \text{M-step:} & \begin{cases} \phi_{wt} = \text{norm}_{w \in W} \left(\sum_{d \in D} n_{dw} p_{tdw} + \phi_{wt} \sum_{i=1}^n \tau_i \frac{\partial R_i}{\partial \phi_{wt}} \right) \\ \theta_{td} = \text{norm}_{t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

Multimodal Probabilistic Topic Modeling

Given a text document collection *Probabilistic Topic Model* finds:

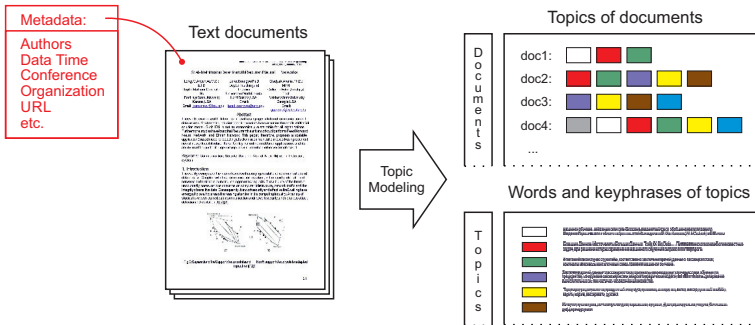
$p(t|d)$ — topic distribution for each document d ,

$p(w|t)$ — term distribution for each topic t .



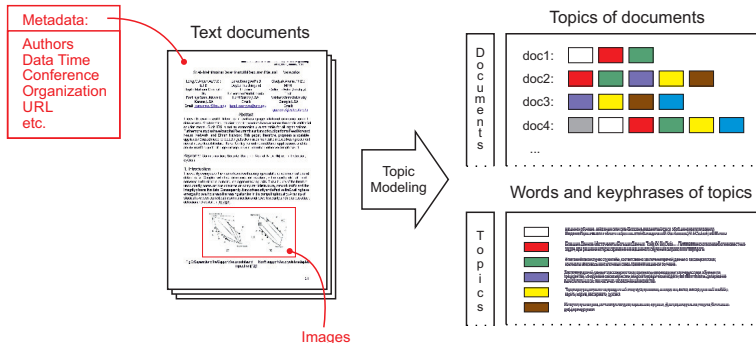
Multimodal Probabilistic Topic Modeling

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



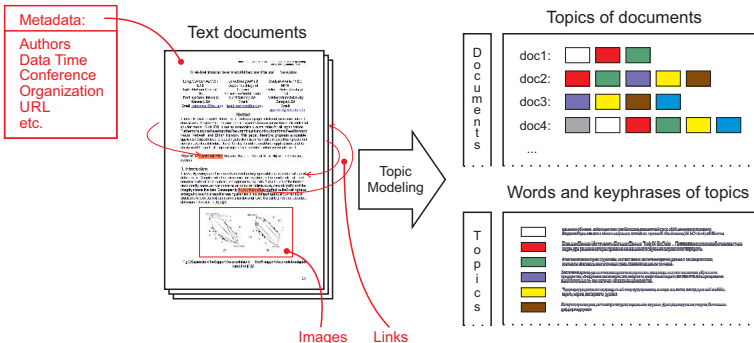
Multimodal Probabilistic Topic Modeling

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



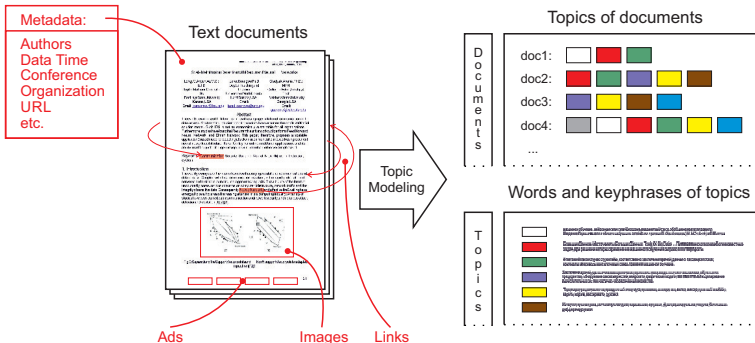
Multimodal Probabilistic Topic Modeling

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



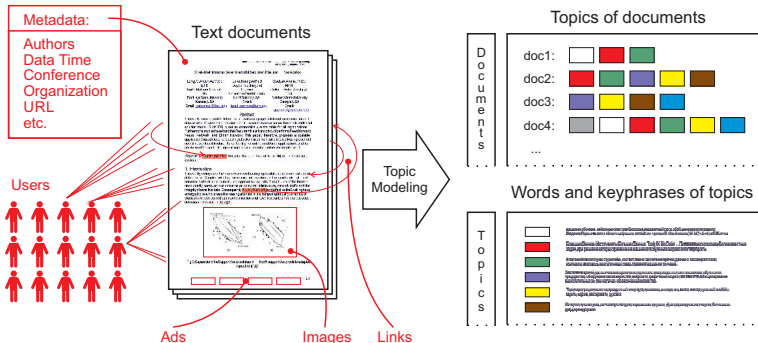
Multimodal Probabilistic Topic Modeling

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



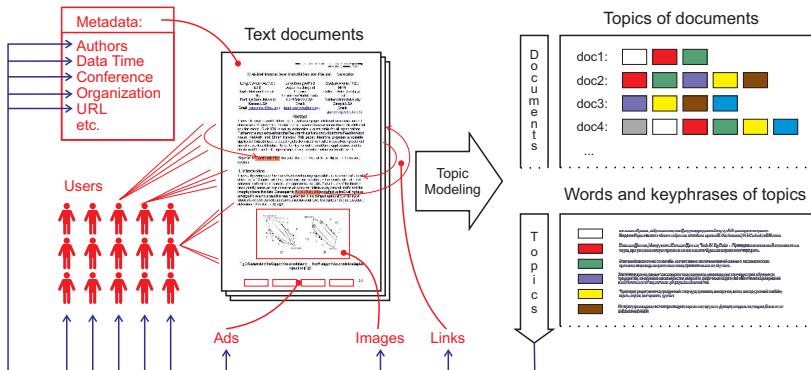
Multimodal Probabilistic Topic Modeling

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



Multimodal Probabilistic Topic Modeling

Multimodal Topic Model finds topical distribution for terms $p(w|t)$, authors $p(a|t)$, time $p(y|t)$, objects on images $p(o|t)$, linked documents $p(d'|t)$, advertising banners $p(b|t)$, users $p(u|t)$, and binds all these modalities into a single topic model.



Multimodal extension of ARTM

W^m is a vocabulary of tokens of m -th modality, $m \in M$

$W = W^1 \sqcup \dots \sqcup W^M$ is a joint vocabulary of all modalities

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

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

EM-algorithm is a simple iteration method for the system

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

Bayesian learning is overcomplicated for PTMs

$$p(\Theta|\alpha) = \prod_{d=1}^D p(\theta_{d,:}|\alpha) = \prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K \alpha_k^{\theta_{d,k}-1}$$

$$p(Z|\Theta) = \prod_{d=1}^D p(z_d|\Theta_d) = \prod_{d=1}^D \prod_{k=1}^K \theta_{d,k}^{I(d,k)}$$

$$p(Z|\alpha) = \int p(Z|\Theta)p(\Theta|\alpha)d\Theta$$

$$= \prod_{d=1}^D \left(\int \frac{1}{B(\alpha)} \prod_{k=1}^K \alpha_k^{\theta_{d,k}+I(d,k)-1} d\theta_d \right)$$

$$= \prod_{d=1}^D \frac{B(I(d,:)+\alpha)}{B(\alpha)}$$

$$B(d,k) = \sum_{n=1}^N \mathbb{1}\{d_n = m \wedge A_{1n} = z\}$$

$$p(Z,W|\alpha,\beta) = \prod_{d=1}^D \prod_{i=1}^V \prod_{j=1}^V p(z_i, w_j | z_{-i}, w_{-j}, \alpha, \beta)$$

$$p(z_i, w_j | z_{-i}, w_{-j}, \alpha, \beta) = \frac{p(z_i, w_j, z_{-i}, w_{-j})}{p(z_{-i}, w_{-j})}$$

$$= \frac{p(z_i, w_j) \prod_{k=1}^K \theta_{d,k}^{I(d,k)-1}}{\prod_{k=1}^K \theta_{d,k}^{I(d,k)-1}}$$

$$= \frac{p(z_i, w_j)}{\prod_{k=1}^K \theta_{d,k}^{I(d,k)-1}}$$

$$= \frac{p(z_i, w_j)}{\prod_{k=1}^K \frac{1}{B(\beta)} \prod_{l=1}^V \beta_l^{\theta_{d,l}-1}}$$

$$= \frac{p(z_i, w_j) \prod_{l=1}^V \beta_l^{\theta_{d,l}-1}}{\prod_{k=1}^K \prod_{l=1}^V \beta_l^{\theta_{d,l}-1}}$$

$$= \frac{p(z_i, w_j) \prod_{l=1}^V \beta_l^{\theta_{d,l}-1}}{\prod_{k=1}^K \prod_{l=1}^V \beta_l^{\theta_{d,l}-1}}$$

$$= \frac{p(z_i, w_j) \prod_{l=1}^V \beta_l^{\theta_{d,l}-1}}{\prod_{k=1}^K \prod_{l=1}^V \beta_l^{\theta_{d,l}-1}}$$

Graphical models showing nodes and dependencies, including a parse tree grouped into M documents.

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

$$p(\theta|\alpha) = \prod_{d=1}^D p(\theta_{d,:}|\alpha) = \prod_{d=1}^D \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k - 1}$$

$$p(\theta|\beta) = \prod_{k=1}^K p(\theta_{:,k}|\beta) = \prod_{k=1}^K \frac{1}{B(\beta)} \prod_{d=1}^D \theta_{d,k}^{\beta_d - 1}$$

$$p(Z|\alpha) = \int p(Z|\theta)p(\theta|\alpha)\theta$$

$$= \prod_{d=1}^D \left(\int \frac{1}{B(\alpha)} \prod_{k=1}^K \theta_{d,k}^{\alpha_k - 1} \theta_{d,:} \right)$$

$$= \prod_{d=1}^D \frac{B(\Omega_d + \alpha)}{B(\alpha)}$$

$$\Omega(d,k) = \sum_{m=1}^M 1\{d_m = m \wedge z_m = k\}$$

$$p(Z,W|\alpha, \beta) = \prod_{d=1}^D \prod_{i=1}^N p(z_i, w_i | z_{-i}, w_{-i}, \alpha, \beta)$$

$$p(z_i = k | Z_{-i}, W_{-i}, \alpha, \beta)$$

$$p(w_i = j | z_i = k, Z_{-i}, W_{-i}, \alpha, \beta)$$

$$p_{tdw} = \text{norm}_t(\phi_{wt}\theta_{td})$$

$$\phi_{wt} = \text{norm}_w \left(\sum_d n_{dw} p_{tdw} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$$

$$\theta_{td} = \text{norm}_t \left(\sum_w n_{dw} p_{tdw} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right)$$

Graphical models showing nodes for α , β , θ , ϕ , z , w , and M . Nodes are grouped into M documents.

BigARTM project

BigARTM features:

- **Parallel + Online** + Multimodal + Regularized Topic Modeling
- Out-of-core one-pass processing of Big Data
- Built-in library of regularizers and quality measures

BigARTM community:

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



BigARTM license and programming environment:

- Freely available for commercial usage (BSD 3-Clause license)
- Cross-platform — Windows, Linux, Mac OS X (32 bit, 64 bit)
- Programming APIs: command-line, C++, and Python

Fast online EM-algorithm for regularized multimodal PTMs

Input: collection D split into batches D_b , $b = 1, \dots, B$;

Output: matrix Φ ;

- 1 initialize ϕ_{wt} for all $w \in W$, $t \in T$;
- 2 $n_{wt} := 0$, $\tilde{n}_{wt} := 0$ for all $w \in W$, $t \in T$;
- 3 **for all** batches D_b , $b = 1, \dots, B$
- 4 iterate each document $d \in D_b$ at a constant matrix Φ :
 $(\tilde{n}_{wt}) := (\tilde{n}_{wt}) + \mathbf{ProcessBatch}(D_b, \Phi)$;
- 5 **if** (synchronize) **then**
- 6 $n_{wt} := n_{wt} + \tilde{n}_{dw}$ for all $w \in W$, $t \in T$;
- 7 $\phi_{wt} := \mathop{\text{norm}}_{w \in W^m} \left(n_{wt} + \phi_{wt} \frac{\partial R}{\partial \phi_{wt}} \right)$ for all $w \in W^m$, $m \in M$, $t \in T$;
- 8 $\tilde{n}_{wt} := 0$ for all $w \in W$, $t \in T$;

Fast online EM-algorithm for Multi-ARTM

ProcessBatch iterates documents $d \in D_b$ at a constant matrix Φ .

matrix $(\tilde{n}_{wt}) := \text{ProcessBatch}$ (set of documents D_b , matrix Φ)

- 1 $\tilde{n}_{wt} := 0$ for all $w \in W$, $t \in T$;
- 2 **for all** $d \in D_b$
- 3 initialize $\theta_{td} := \frac{1}{|T|}$ for all $t \in T$;
- 4 **repeat**
- 5 $p_{tdw} := \text{norm}_{t \in T}(\phi_{wt}\theta_{td})$ for all $w \in d$, $t \in T$;
- 6 $n_{td} := \sum_{w \in d} \lambda_{m(w)} n_{dw} p_{tdw}$ for all $t \in T$;
- 7 $\theta_{td} := \text{norm}_{t \in T}(n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}})$ for all $t \in T$;
- 8 **until** θ_d converges;
- 9 $\tilde{n}_{wt} := \tilde{n}_{wt} + \lambda_{m(w)} n_{dw} p_{tdw}$ for all $w \in d$, $t \in T$;

Brief summary

ARTM...

- reduces barriers to entry into PTMs for practitioners
- allows them to concentrate on the task, rather than maths
- has very simple inference: $\frac{\partial R}{\partial \phi_{wt}}$, $\frac{\partial R}{\partial \theta_{td}}$
- uses the same general EM-algorithm for all models
- covers PLSA, LDA, and 100s of known Bayesian PTMs
- allows to combine modalities and regularizers easily
- has linear time complexity $O(n \cdot |T| \cdot \text{iterations})$
- has online EM algorithm, which runs the entire collection once
- is implemented in BigARTM open source project

Experiment 1: BigARTM vs Gensim vs Vowpal Wabbit

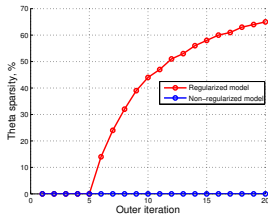
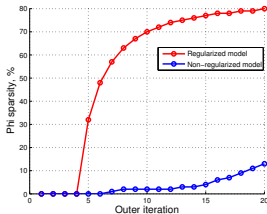
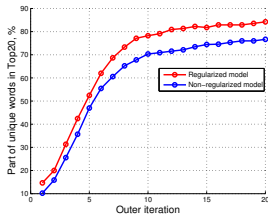
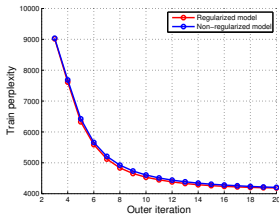
- 3.7M articles from Wikipedia, 100K unique words

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

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

Experiment 2: Running BigARTM with multiple regularizers

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



Experiment 3: The interpretability of n -gram models

Two modalities — unigrams & bigrams

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

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

ARTM for Question Answering: regularization ideas

- Use modalities for 2- and 3-grams extracted from questions
- Use background topics for separate common words from specific topics
- Decorrelate Φ for specific topics
- Smooth Φ for background topics
- Smooth Θ for background topics
- Sparse Φ for specific topics (optional)
- Sparse Θ for specific topics (optional)

Topic model based similarities

- Average JS-divergence similarity between topics:

$$\text{Rel}_1(x|s) = \frac{1}{|T|} \sum_{t \in T} e^{-\gamma \text{JS}(x_t, s_t)}$$

where $x_t = \text{norm}_w \left(p(w|t)[w \in x] \right)$ is the normalized projection of $\phi_{wt} = p(w|t)$ distribution on the subset of terms x

- JS-divergence similarity between θ -vectors of documents:

$$\text{Rel}_2(x|s) = e^{-\gamma \text{JS}(\theta_x, \theta_s)}$$

- The aggregated similarity:

$$\text{Rel}(x|s) = \text{Rel}_1(x|s) \text{Rel}_2(x|s)$$







Celikyilmaz A., Hakkani-Tur D., Tur G. LDA Based Similarity Modeling for Question Answering. NAACL HLT Workshop on Semantic Search. 2010.

Alternative idea

- Each question corresponds to a topic, $q \leftrightarrow t$.
The predefined interpretation of topics has the advantage that you can verify manually how PTM performs Question Expansion
- Smooth each column of Φ matrix by $\beta_{wt} = p(w|t)$, a known distribution of terms in the question q corresponding to t :

$$R(\Phi) = \tau \sum_{t \in T} \sum_{w \in W} \beta_{wt} \ln \phi_{wt} \rightarrow \max$$

References

-  *Hofmann T.* Probabilistic Latent Semantic Indexing. ACM SIGIR, 1999.
-  *Blei D., Ng A., Jordan M.* Latent Dirichlet Allocation. Journal of Machine Learning Research, 2003.
-  *Vorontsov K. V., Potapenko A. A.* Additive regularization of topic models. Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications, 2015.
-  *Vorontsov K. V., Frei O. I., Apishev M. A., Romov P. A., Suvorova M. A., Yanina A. O.* Non-Bayesian Additive Regularization for Multimodal Topic Modeling of Large Collections. Topic Models: Post-Processing and Applications. 2015.
-  *Celikyilmaz A.* A Semantic Question Answering System Using Topic Models. NIPS Workshop on Applications for Topic Models: Text and Beyond. 2009.
-  *Celikyilmaz A., Hakkani-Tur D., Tur G.* LDA Based Similarity Modeling for Question Answering. NAACL HLT Workshop on Semantic Search. 2010.