Linguistic Regularization of Topic Models

Anna Potapenko (anya_potapenko@mail.ru)
Yandex School of Data Analysis, Higher School of Economics, Moscow, Russia

Topic Modeling

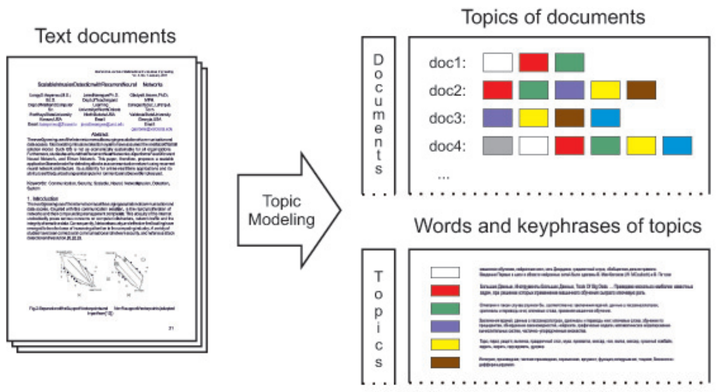
Given a collection of documents, **assume** that each observable word w in document d refers to some latent topic t:

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

Find:

- $\phi_{wt} \equiv p(w|t)$ — words for each topic,
- $\theta_{td} \equiv p(t|d)$ — topics for each document,

resulting in $p(w|d)$ close to $\hat{p}(w|d) \propto n_{dw}$ – frequencies of words in documents.



Contributions

- Overcome bag-of-words limitation in ARTM – a framework for combining topic models.
- Improve quality of a topic model by optimizing both global and local context criteria.

Local contexts

- Let a regularizer depend on position of word in document, e.g.

$$R = \sum_{d\in D}\sum_{i=1}^{n_d} -\tau_{dw_i} KL(\alpha_{tdi}||p_{tdw_i}) \to \max,$$

where α_{tdi} are topic distributions for local contexts, e.g. weighted p_{tdw} for words in window.

- Transfer Φ, Θ modifications to p_{tdw}:

$$\tilde{p}_{tdw} = (1-\tau_{dw})p_{tdw} + \tau_{dw}\frac{\sum_{i=1}^{n_d}[w_i = w]\alpha_{tdi}}{n_{dw}}$$

- Apply them effectively on E-step when scanning documents word by word.

Additive Regularization

Regularized Likelihood:

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

Maximization is performed via modified EM-algorithm – iterative process, alternating:

E-step (Bayes' Rule for $p(t|d,w) \equiv p_{tdw}$):

$$p_{tdw} \propto \phi_{wt}\theta_{td}$$

M-step (parameters estimates):

$$\phi_{wt} \propto \left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+$$

$$\theta_{td} \propto \left(\sum_{w\in W} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+$$

Implementation:

bigARTM.org – open source library of additively regularized topic models.

Multi-criteria evaluation

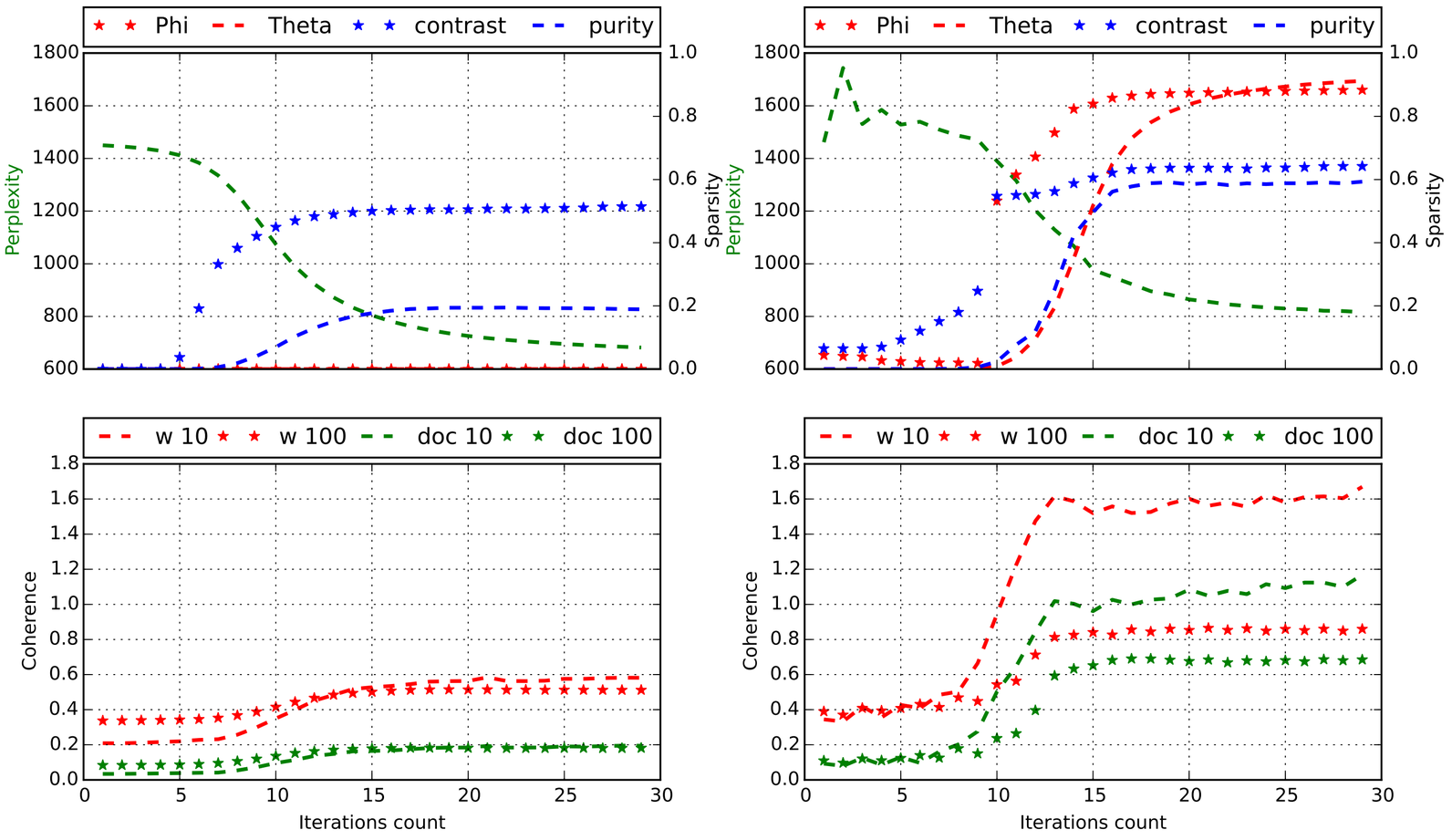


Fig. 1. LDA (left) vs ARTM (right). Trade off perplexity and interpretability: coherence, purity and contrast.

Combination of requirements (regularizers):

- local contexts coherence (adjust topic distributions $p(t|d,w)$)
- topics diversity (minimize correlation between columns in Φ)
- sparsity of domain topics (plus several smooth topics for general lexis)

Quality measures:

- **Perplexity:** $P = \exp\left(-\frac{1}{n}L(\Theta,\Phi)\right)$
- **Coherence**: $C = \frac{2}{k(k-1)}\sum_{j=2}^{k}\sum_{i=1}^{j}\mathrm{PMI}(w_i,w_j)$

  based on document and window (h=100) frequencies, for top-10 and top-100 words.
- **Sparsity**: proportion of zeros in Φ, Θ
- **Topic kernels**: $W_t = \{w\colon p(t|w) > 0.25\}$, contrast: $\frac{1}{|W_t|}\sum_{w\in W_t} p(t|w)$, purity: $\sum_{w\in W_t} p(w|t)$

Looking into topic segmentation

Problems:

- The model assumes that the proportion of any topic is constant during the document. Therefore, topical shifts can't be detected.
- General lexis words are assigned to the most probable topics of the document. Therefore, the key words are buried among them.

Solution:

- Weaken the influence of Θ – topic distribution for a document (e.g. by strong smoothing) to enable topics to evolve during the document.
- Use background topics for a token if its topical identity according to $p(t|d,w)$ is not consistent with the topics of the local context.

Fig. 2. Key words of the blue topic: "people, attack, terrorist, american, state, war, world, life, terrorism". PLSA (left) doesn't catch the change from assistance topic to terrorism. ARTM (right) retains terrorism-related terms in the blue topic, assigning assistance-related terms to the yellow one and general lexis to the grey one.

Reference: Vorontsov K. V., Potapenko A. A. Additive Regularization of Topic Models – Machine Learning. Special Issue "Data Analysis and Intelligent Optimization with Applications": Volume 101, Issue 1 (2015), Pp. 303-323.