

LINGUISTIC REGULARIZATION OF TOPIC MODELS

ANNA POTAPENKO (ANYA_POTAPENKO@MAIL.RU)

YANDEX SCHOOL OF DATA ANALYSIS, HIGHER SCHOOL OF ECONOMICS, MOSCOW, RUSSIA

TOPIC MODELING

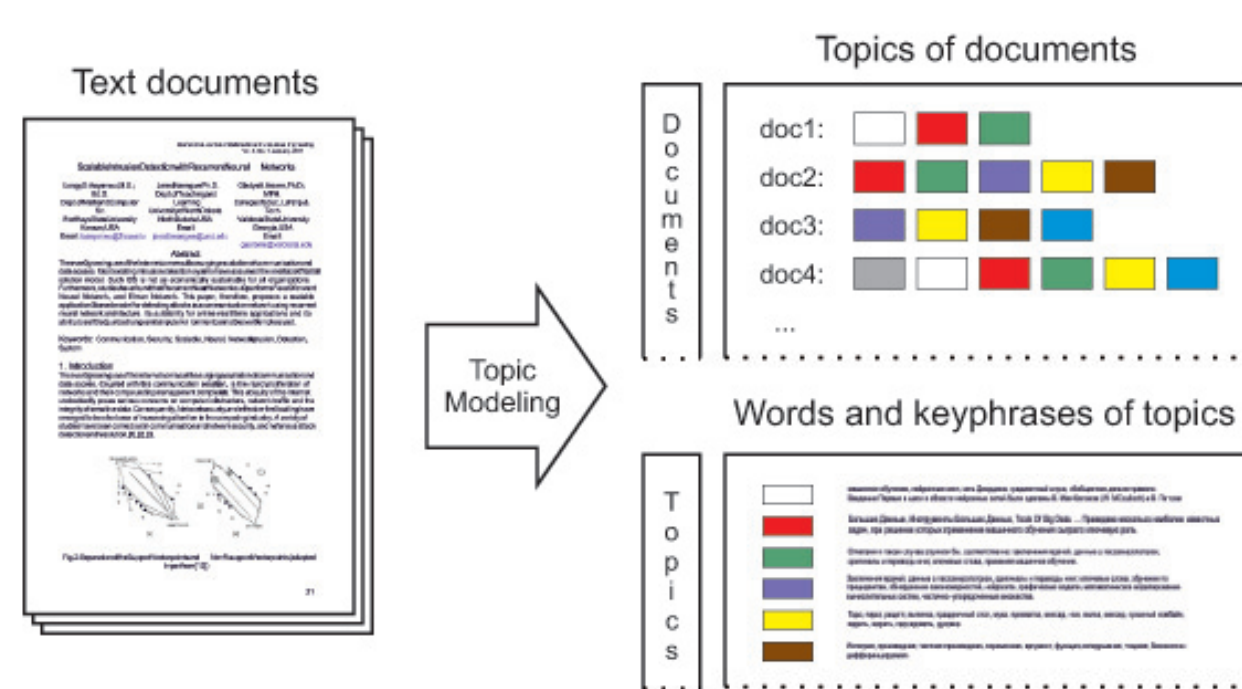
Given a collection of documents, **assume** that each observable word w in document d refers to some latent topic t :

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

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.

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

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:

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

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

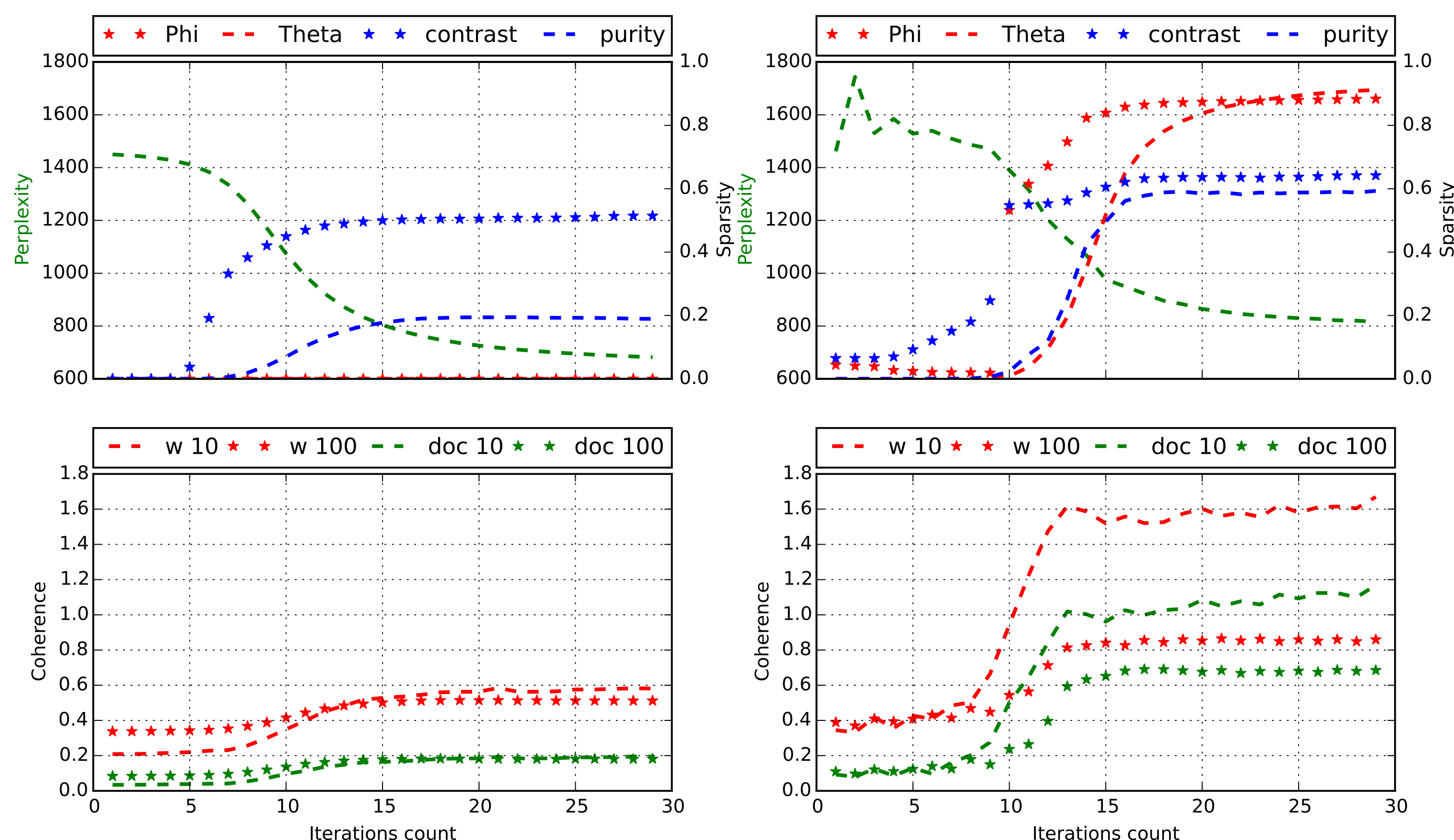


Fig. 1. LDA (left) vs ARTM (right). Trade off perplexity and interpretability: coherence, purity and contrast.

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:** $P = \exp(-\frac{1}{n} L(\Theta, \Phi))$
- **Coherence:** $C = \frac{2}{k(k-1)} \sum_{j=2}^k \sum_{i=1}^j \text{PMI}(w_i, w_j)$
based on document and window (h=100) frequencies, for top-10 and top-100 words.
- **Sparsity:** proportion of zeros in Φ, Θ
- **Topic kernels:** $W_t = \{w: p(t|w) > 0.25\}$,
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.

Air Canada's planes assisted in the evacuation. Israel sent tents and mineral water and medical supplies. Italy has sent beds and sheets and blankets and inflatable rafts to help with rescue efforts. Kuwait has pledged \$400 million in oil and a hundred million dollars in humanitarian aid. Qatar and the UAE has pledged \$100 million each. Sri Lanka, one of the world's most impoverished nations that is struggling to overcome the effects of the tsunami, has sent a donation of \$25,000. In all, more than a hundred countries have stepped forward with offers of assistance, and additional pledges of support are coming in every day. To every nation in every province and every local community across the globe that is standing with the American people, and with those who hurt on the Gulf Coast, our entire nation thanks you for your support. Four years ago, the American people saw a similar outpouring of sympathy and support when another tragedy struck our nation, the terrorist attacks of September the 11th, 2001. This Sunday, Americans will mark the fourth anniversary of that terrible day when nearly 3,000 innocent people were murdered. The attacks took place on American soil, yet they left grieving families on virtually every continent. Citizens from dozens of nations were killed on September the 11th. Innocent men and women and children of every race and every religion. And in the four years since the September the 11th attacks, the terrorists have continued to kill -- in Madrid and Istanbul and Jakarta and

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.

Air Canada's planes assisted in the evacuation. Israel sent tents and mineral water and medical supplies. Italy has sent beds and sheets and blankets and inflatable rafts to help with rescue efforts. Kuwait has pledged \$400 million in oil and a hundred million dollars in humanitarian aid. Qatar and the UAE has pledged \$100 million each. Sri Lanka, one of the world's most impoverished nations that is struggling to overcome the effects of the tsunami, has sent a donation of \$25,000. In all, more than a hundred countries have stepped forward with offers of assistance, and additional pledges of support are coming in every day. To every nation in every province and every local community across the globe that is standing with the American people, and with those who hurt on the Gulf Coast, our entire nation thanks you for your support. Four years ago, the American people saw a similar outpouring of sympathy and support when another tragedy struck our nation, the terrorist attacks of September the 11th, 2001. This Sunday, Americans will mark the fourth anniversary of that terrible day when nearly 3,000 innocent people were murdered. The attacks took place on American soil, yet they left grieving families on virtually every continent. Citizens from dozens of nations were killed on September the 11th. Innocent men and women and children of every race and every religion. And in the four years since the September the 11th attacks, the terrorists have continued to kill -- in Madrid and Istanbul and Jakarta and

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.