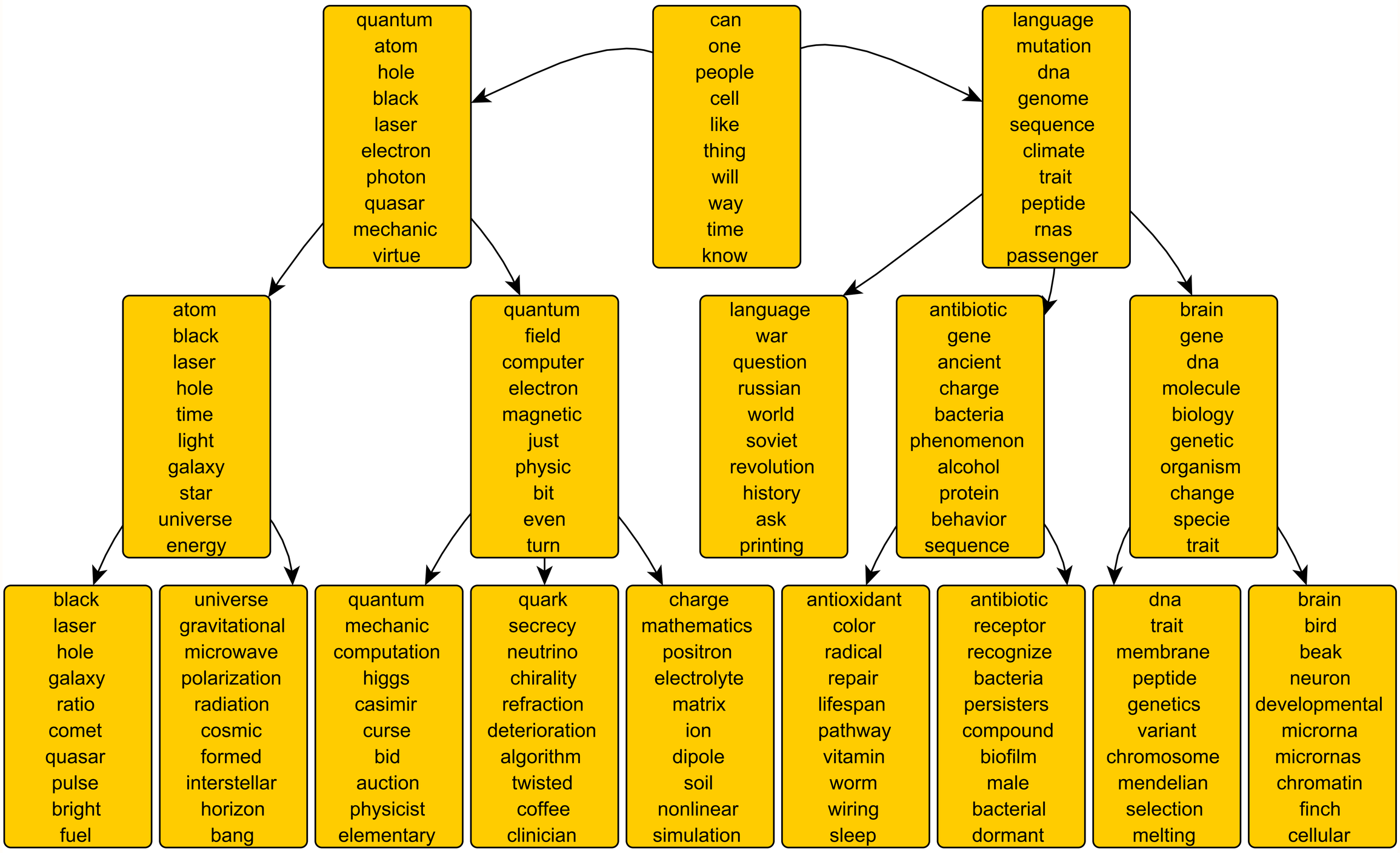
Topic Hierarchies with Additive Regularization

Nadia Chirkova
Lomonosov Moscow State University



Part of topic hierarchy built on serious-science.org materials.

Introduction

Topic modeling is a popular technique of semantic analysis of text documents corpora. A topic is defined as a distribution over words. In topic model each document is assigned its distribution over topics that specifies what topics describe this document.

Additive Regularization of Topic Models (ARTM) [1] is a powerful approach to topic modeling that allows building clear to people, interpretable topics. It has an effective open-source implementation BigARTM (bigartm.org).

In standard problem formulation all topics are equivalent. But people used to hierarchical topic structure where each topic is split into subtopics. Such hierarchical representation helps to navigate throw corpora.

The goal of the research is to introduce an approach to building topic hierarchies based on ARTM and to implement it in BigARTM. Proposed algorithm should produce interpretable hierarchical structure and scale to large data.

ARTM

Data. Let D be documents set, W be words set. Text corpora is represented by document-term count matrix $F = \{n_{dw}\}_{W \times D}$ used to estimate $p(w|d)$.

Model. Let S be a set of topics, $|S|$ is fixed. The topic model is a low-rank factorization of F:

$$p(w|d) \approx \sum_{s \in S} p(w|s)p(s|d) = \sum_{s \in S} \phi_{ws}\theta_{sd} \Leftrightarrow F \approx \Phi\Theta$$

with parameters

$\Phi = \{\varphi_{ws}\}_{W \times S}$ (topic distributions over words),
$\Theta = \{\theta_{sd}\}_{S \times D}$ (document distributions over topics).

Model learning. Maximize regularized log-likelihood:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{s \in S} \varphi_{ws}\theta_{sd} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi,\Theta}$$

$$\text{s.t.} \sum_{w \in W} \phi_{ws} = 1; \phi_{ws} \geqslant 0; \quad \sum_s \theta_{sd} = 1; \theta_{sd} \geqslant 0.$$

Regularizers R_i with weights τ_i impose subject-specific criteria. Thus, model hyperparameters are $|S|$ and $\{\tau_i\}$.

E. g., Φ sparsing regularizer [1] $R_1(\Phi) = -\sum_s \sum_w \ln \phi_{ws}$ encourages topics to have less words with $p(w|s) > 0$.

Russian Summer School in Information Retrieval, August 22-26, 2016, Saratov

Regularizer to build hierarchy

Hierarchy definition. We define topic hierarchy as oriented multipartite (multilevel) graph of topics so that edges connect only topics from neighboring levels. If there is edge $t \to s$ then t is parent topic and s is children.

Interlevel regularizer. Let's build topic hierarchy top down, level by level, each level is a topic model. The first level is built as usual.

Suppose we have already built $\ell \geqslant 1$ levels and want to build $(\ell+1)$-th level. Denote T and Φ^p are ℓ-th level's topics set and Φ. The key problem is to establish parent-children relationship between topics. To do this we introduce additional matrix factorization problem:

$$p(w|t) = \sum_{s \in S} p(w|s)p(s|t) \Leftrightarrow \Phi^p \approx \Phi\Psi \qquad (1)$$

with new parameter matrix called interlevel matrix:

$$\Psi = \{\psi_{st}\}_{S \times T},\ \psi_{st} = p(s|t).$$

If similarity measure in (1) is KL-divergence then new optimization task with Φ regularizer is

$$R_2(\Phi) = \sum_t \sum_w \phi^p_{wt} \ln \sum_s \phi_{ws}\psi_{st},$$

$$\sum_{d,w} n_{dw} \ln \sum_s \varphi_{ws}\theta_{sd} + \sum_i \tau_i R_i(\Phi,\Theta,\Psi) \to \max_{\Phi,\Theta,\Psi}.$$

The key property of this regularizer is that optimization problem is equivalent to adding $|T|$ virtual documents, each document is Φ^p column multiplied by τ_2. Then Ψ is part of Θ!

Hyperparameters of hierarchy are levels count, each level topics count and other regularizers weights.

Topic graph sparsing regularizer

With our hierarchy definition topics are allowed to have several parents. We want them to have just $1-2$ parents. Then the interlevel matrix should be sparse.

The way to achieve it is to maximize KL-divergence between uniform distribution and $p(t|s)$:

$$R_3(\Psi) = \sum_{s,t} \frac{1}{|T|} \ln \frac{1/|T|}{p(t|s)} = \text{Const} - \sum_{s,t} \ln p(t|s).$$

Applying Bayes formula to $p(t|s)$

$$R_3(\Psi) = -\sum_s \sum_t \left(\ln(\psi_{st}p_t) - \ln \sum_t (\psi_{st}p_t)\right),$$

p_t is computed from parent level so it is fixed. Generally the idea is similar to R_1 regularizer's concept.

Two regularizers together

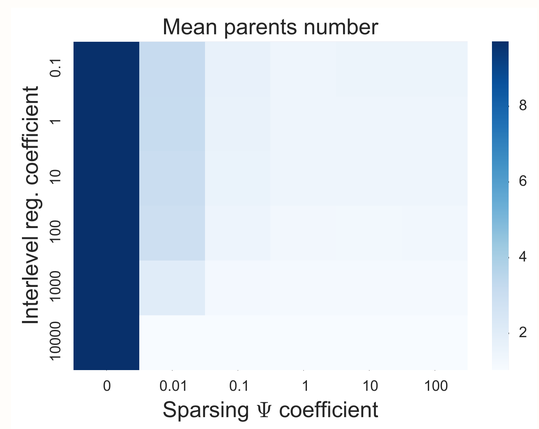
Corpora: Postnauka (postnauka.ru) materials: $|D| = 1728$, $|W| = 38467$, 1-11-31 topics hierarchy.

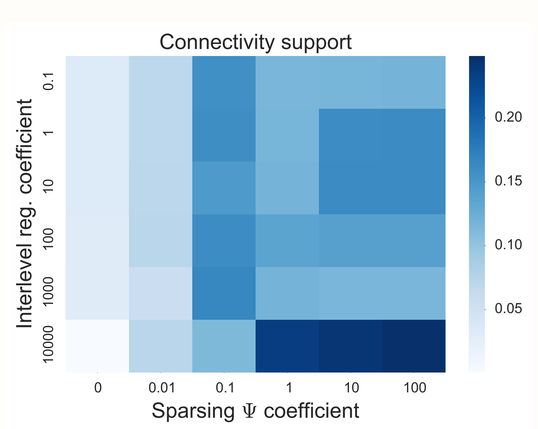
Quality measures:

1. Mean parents count: $\frac{1}{|S|}\sum_s [\psi_{st} > 0]$;
2. Connectivity support: $\min_s \max_t \psi_{st}$ (if low, some topics have no parents);
3. Hellinger distance $h(\Phi^p, \Phi\Psi)$;
4. Coherence: $\frac{1}{45}\sum_{i=1}^{10}\sum_{j>i} \mathrm{PMI}(w_i, w_j)$ (popular interpretability measure)

The first two metrics measure graph structure quality.

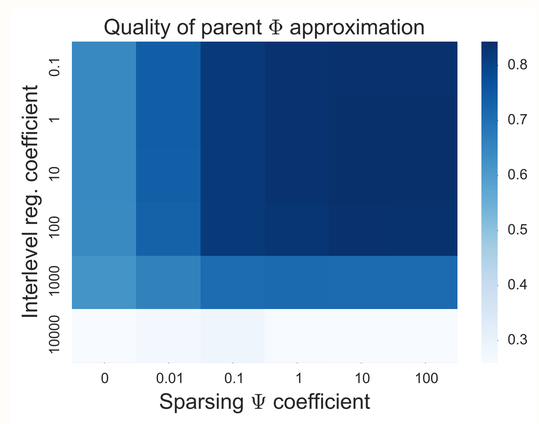
Experiment. When first level is built, several values of interlevel reg. τ_2 and sparsing Ψ reg. τ_3 were iterated (τ_1 is fixed). For each model measures were computed (↑: the higher the better):





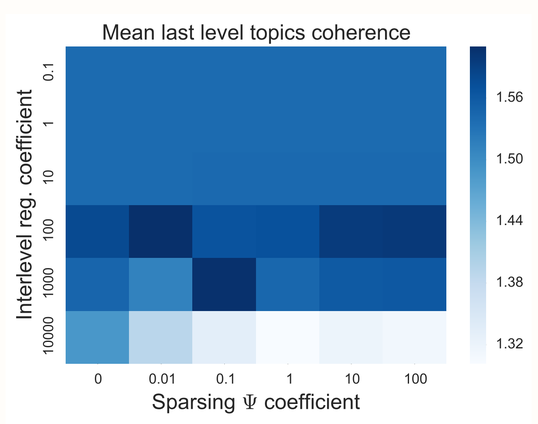
Mean parents count ↓

Graph is sparse with large sparsing weight τ_3 or small τ_3 and large τ_2.



Connectivity support ↑

There is optimal τ_3 for the majority of τ_2.



Φ^p approximation ↓

Approximation (1) is successful only with large τ_2. Sparsing worsens approximation.

Coherence ↑

Topics are bad interpreted with large τ_2, there is other optimal value.

The better approximation, the worse topics, tradeoff point is $\tau_2 = 10^3, \tau_3 = 0.1$.

On the first page similar hierarchy built for serious-science.org, English project of Postnauka, is presented.

References

[1] *Vorontsov K. V., Potapenko A. A.* Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. — Analysis of Images, Social Networks, and Texts (AIST-2014). — LNCS, Springer.