Additive Regularization for Hierarchical Multimodal Topic Modeling

N. A. Chirkova[1,2], K. V. Vorontsov[3]

[1]JSC Antiplagiat, [2]Lomonosov Moscow State University
[3]Federal Research Center "Computer Science and Control" of RAS

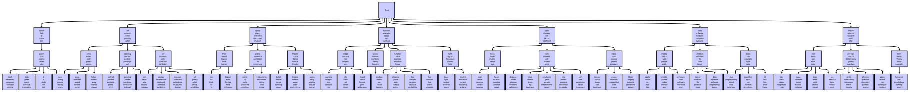
October 14, 2016

Topic hierarchies for automatic text categorization

How to overview a large text collection in a few minutes?

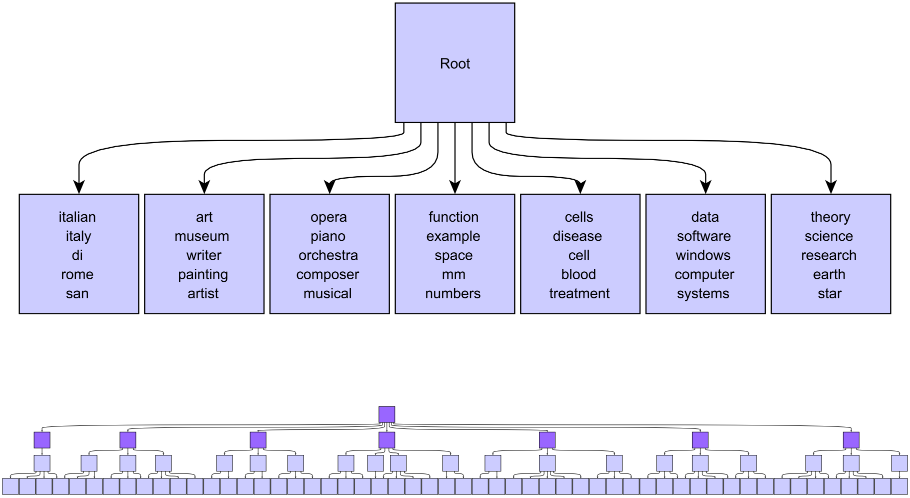
Topic hierarchy:

- soft hierarchical documents clustering into topics;
- topics are described by specific terminology.

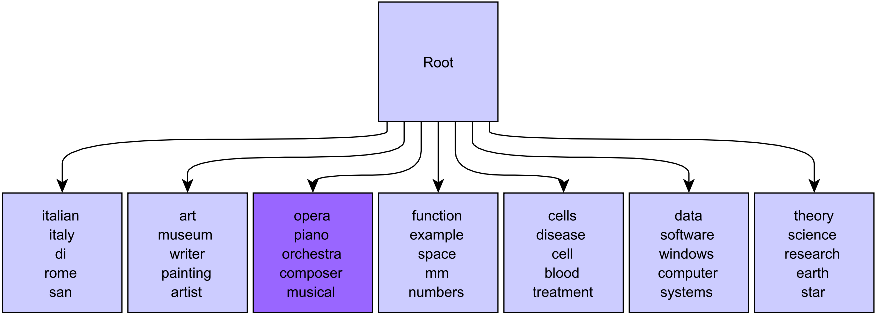


A fragment of English Wikipedia topic hierarchy

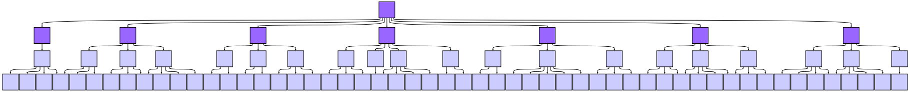
Topic hierarchies for automatic text categorization



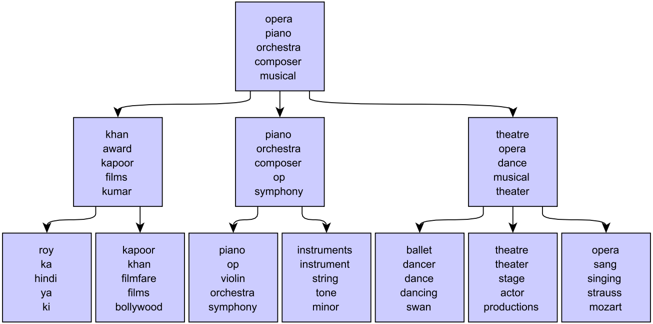
Topic hierarchies for automatic text categorization

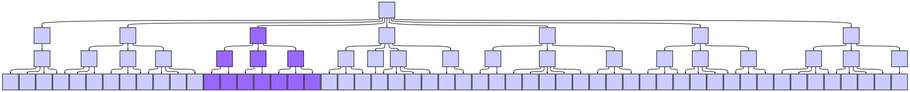


Topic articles: Toccata and Fugue, F major, E minor, Carl Friedrich Abel, List of compositions by Frédéric Chopin by genre, Piano quintet, F minor...

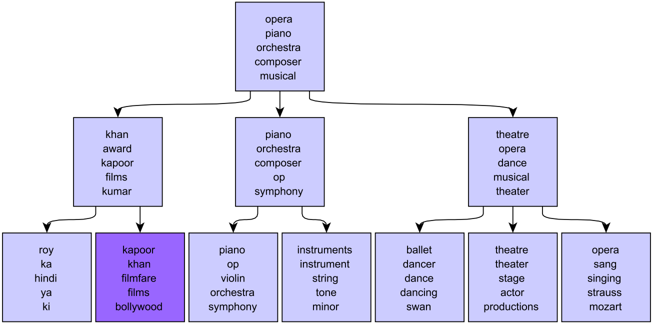


Topic hierarchies for automatic text categorization

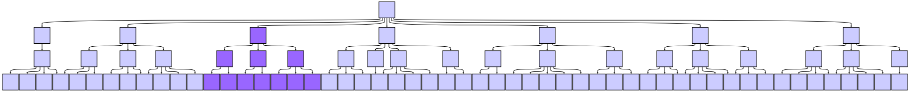




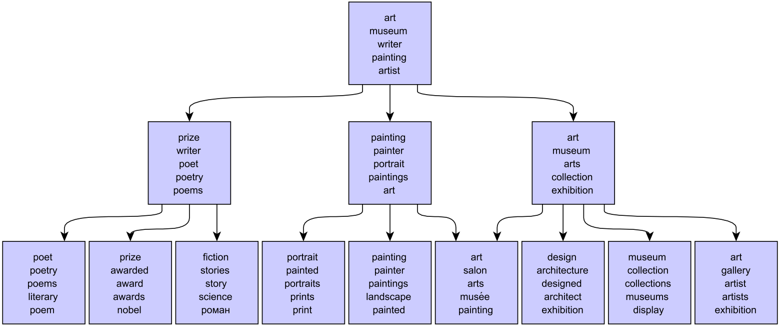
Topic hierarchies for automatic text categorization

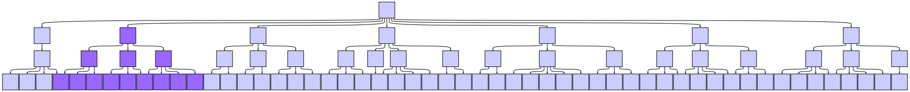


Topic articles: Filmfare Award for Best Actor, Filmfare Award for Best Film, Karisma Kapoor, Rishi Kapoor, Arjun Rampal, Shammi Kapoor...

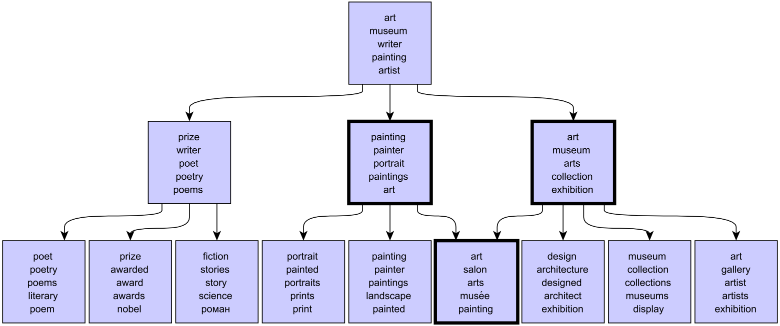


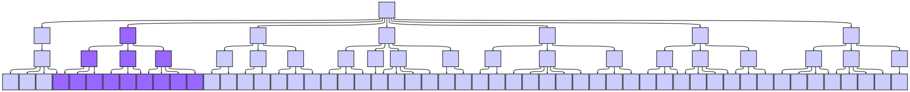
Topic hierarchies for automatic text categorization



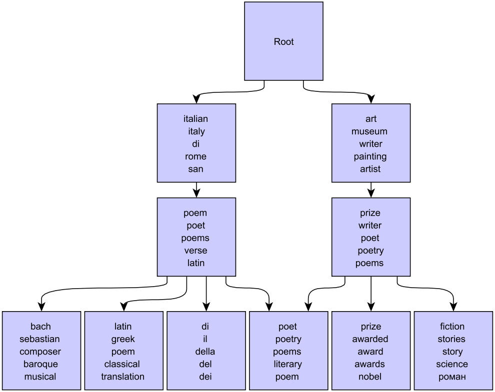


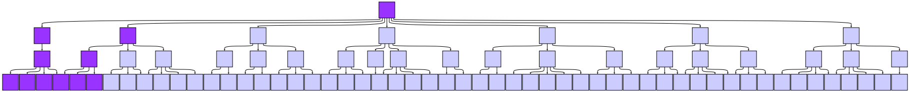
Topic hierarchies for automatic text categorization



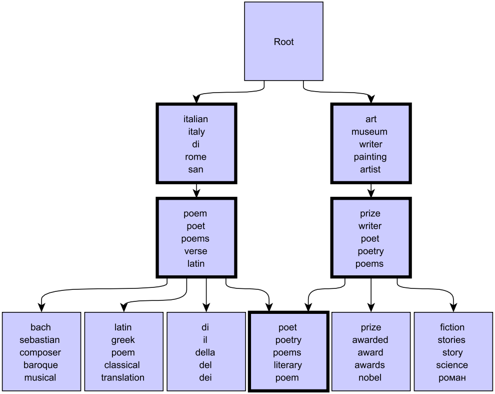


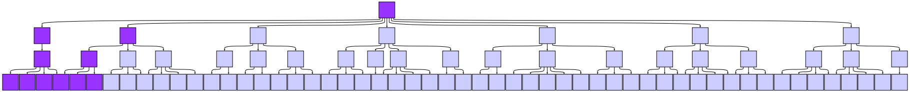
Topic hierarchies for automatic text categorization



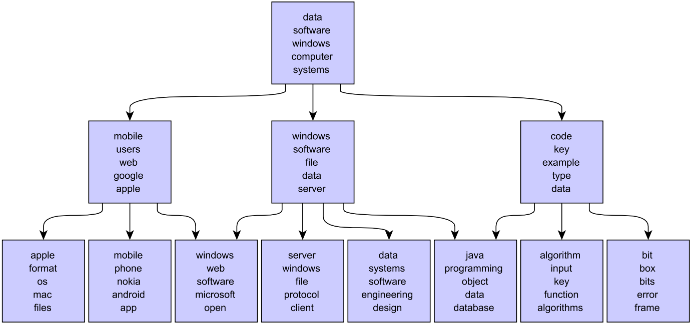


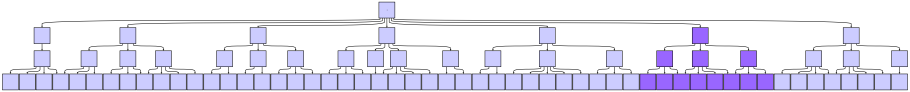
Topic hierarchies for automatic text categorization



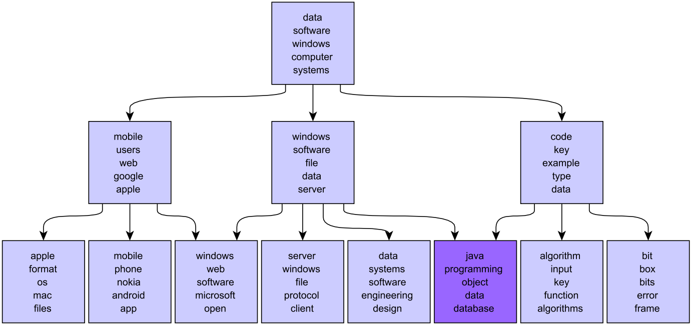


Topic hierarchies for automatic text categorization

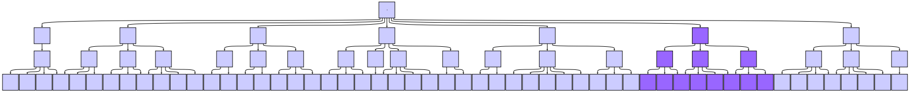




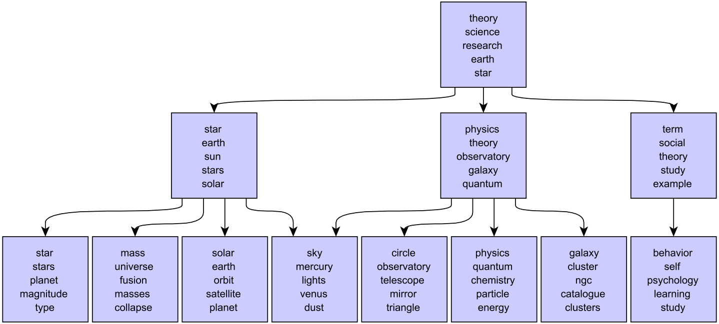
Topic hierarchies for automatic text categorization

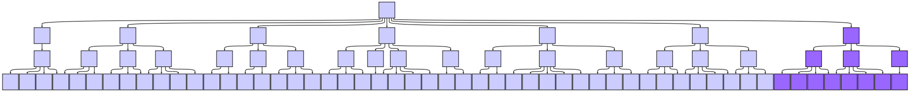


Topic articles: Functional (C++), SQL/CLI, SQL/JRT, Constructor (object-oriented programming), Static cast, Copy constructor, C++/CX, Java Persistence Query Language...



Topic hierarchies for automatic text categorization





Applications of topic hierarchies

- Navigation through large text collection
- Harmonization of existing categorizations
  - duplicate categories detection
  - splitting of miscellaneous topics
- Searching of semantically similar documents
- News filtering

$\Rightarrow$ The need for **automatic** learning of topic hierarchies.

Applications of topic hierarchies: real world tasks

- Navigation through large multilingual, multisource, multilmodal text collection
- Harmonization of existing categorizations
  - duplicate categories detection
  - miscellaneous categories splitting
  - detecting of relations between categories
- Personalized searching for semantically similar documents
- News filtering with respect to geography and time

$\Rightarrow$ The need for **automatic** learning of flexible topic hierarchies.

Topic hierarchies in ARTM

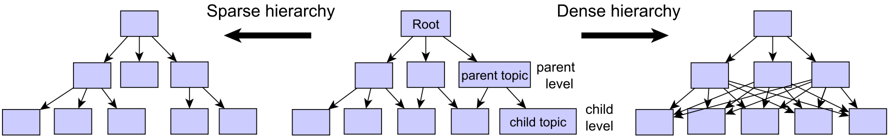
Additive Regularization of Topic Models:

- Modeling fixed number of topics from a set of multimodal documents:
  - text, tags, authors, categories, geotags ans timestamps, commented users, etc $\rightarrow$ flexibility
- Regularization to satisfy additional requirements:
  - topics sparsity, decorrelation, interpretability; consistency with partial markup, etc $\rightarrow$ flexibility
- Scalable open-source implementation: BigARTM.org

The goal of the research: to extend ARTM to learn topic hierarchies and to implement approach in BigARTM.

Topic hierarchies in ARTM: key features

Topic hierarchy is a multipartite (multilevel) graph of topics:



The flexibility of hierarchical structure:

- multiple inheritance (a topic may have several parent topics);
- control over hierarchy sparsity.

$\Rightarrow$ Automatic determination of children topics number.

Topic hierarchies in ARTM: approach

1. Each level (except *Root*) is a flat topic model with its own regularizers.
2. When learning topics of ℓ-th level we use specific regularier to find parent topics from $(\ell - 1)$-th level.
3. We propose a regularizer to control hierarchy sparsity.

ARTM: a flat topic model

Given:

- documents set $d \in D$,
- modalities $m \in M$,
- modalities disjoint dictionaries $W = \bigsqcup_{m \in M} W^m$ of tokens $w \in W$,
- document-token counters matrix n_{dw} used to estimate $p(w|d)$:

$$p(w|d) = \frac{n_{dw}}{\sum_{w' \in W^m} n_{dw'}}$$

Flat topic model for each modality m**:**

$$p(w|d) \approx \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td} \quad d \in D,\, w \in W^m,$$

with topics set T and model parameters
$\Phi^m = \{\phi_{wt}\}_{W^m \times T}$ with $p(w|t)$ and $\Theta = \{\theta_{td}\}_{T \times D}$ with $p(t|d)$ values,
$\Phi = \bigsqcup_{m \in M} \Phi^m$

Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A. Non-bayesian additive regularization for multimodal topic modeling of large collections

ARTM: flat model learning

Optimization task:

$$\underbrace{\sum_{m\in M}\kappa_m \sum_{d\in D}\sum_{w\in W^m} n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}}_{Log-Likelihood} + \underbrace{\sum_i \tau_i R_i(\Phi,\Theta)}_{Regularizers} \to \max_{\Phi,\Theta}$$

$$\sum_{w\in W^m}\phi_{ws} = 1; \phi_{ws} \geqslant 0\,\forall m; \quad \sum_s \theta_{sd} = 1; \theta_{sd} \geqslant 0$$

EM-algorithm for topic model training:

$$\boxed{\operatorname*{norm}_{i\in I}[y_i] = \frac{\max\{y_i,0\}}{\sum_{i'\in I}\max\{y_{i'},0\}}}$$

$$\text{E-step}: \quad p(t|d,w) = \operatorname*{norm}_{t\in T}[\phi_{wt}\theta_{td}]$$

$$\text{M-step}: \quad \phi_{wt} = \operatorname*{norm}_{w\in W^m}\left[n_{wt} + \frac{\partial R}{\partial \phi_{wt}}\phi_{wt}\right], \quad n_{wt} = \sum_{d\in D} n_{dw}p(t|d,w)$$

$$\theta_{td} = \operatorname*{norm}_{t\in T}\left[n_{td} + \frac{\partial R}{\partial \theta_{td}}\theta_{td}\right], \quad n_{td} = \sum_{w\in W} n_{dw}p(t|d,w)$$

Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A. Non-bayesian additive regularization for multimodal topic modeling of large collections

ARTM: regularizers example

The goal: distributions $p(w|t)$ and $p(t|d)$ should be sparse.

- Θ sparsing:

$$R_1(\Theta) = -\sum_{d \in D} \sum_{t \in T} \frac{1}{|T|} \ln \theta_{td}$$

Updated M-step:

$$\theta_{td} = \underset{t \in T}{\mathrm{norm}} \left[n_{td} - \frac{\tau_2}{|T|} \right]$$

- Φ sparsing:

$$R_2(\Phi^m) = -\sum_{t \in T} \sum_{w \in W^m} \frac{1}{|W^m|} \ln \phi_{wt}$$

Updated M-step:

$$\phi_{wt} = \underset{w \in W^m}{\mathrm{norm}} \left[n_{wt} - \frac{\tau_1}{|W^m|} \right]$$

Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A. Non-bayesian additive regularization for multimodal topic modeling of large collections

hARTM: Φ interlevel regularizer

Already learned: levels $1, \dots, \ell$,
ℓ-th level: topics set $a \in A$, parameters $\Phi^\ell \in \mathbb{R}^{W \times A}$ and $\Theta^\ell \in \mathbb{R}^{A \times D}$.
Level to learn: topics set $t \in T$, parameters $\Phi \in \mathbb{R}^{W \times T}$ and $\Theta \in \mathbb{R}^{T \times D}$.
The goal: to establish parent-child relations "t is a child of a".

Hypothesis: parent topic is a mixture of children topics

$$p(w|a) = \sum_{t \in T} p(w|t)p(t|a), \quad w \in W^m,\ a \in A.$$

Φ regularization criteria with new parameters $\Psi = \{\psi_{ta}\}_{T \times A}$, $\psi_{ta} = p(t|a)$:

$$\Phi^\ell \approx \Phi\Psi$$

$$R_3(\Phi, \Psi) = \sum_{m \in M} \sum_{a \in A} \sum_{w \in W^m} n_{wa} \ln \sum_{t \in T} \phi_{wt} \psi_{ta}$$

Implementation: $|A|$ pseudodocuments with n_{wa} (counted on M-step).

hARTM: Θ interlevel regularizer

Already learned: levels $1, \dots, \ell$,
ℓ-th level: topics set $a \in A$, parameters $\Phi^\ell \in \mathbb{R}^{W \times A}$ and $\Theta^\ell \in \mathbb{R}^{A \times D}$.
Level to learn: topics set $t \in T$, parameters $\Phi \in \mathbb{R}^{W \times T}$ and $\Theta \in \mathbb{R}^{T \times D}$.
The goal: to establish parent-child relations "t is a child of a".

Hypothesis:

$$p(a|d) = \sum_{t \in T} p(a|t)p(t|d), \quad a \in A,\ d \in D.$$

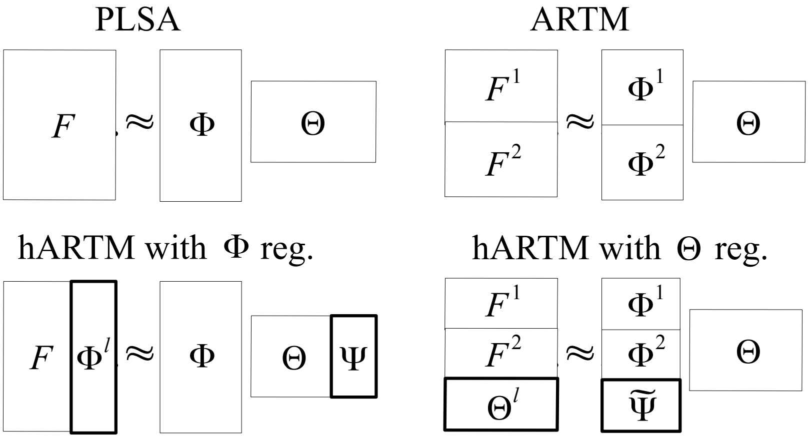
Θ regularization criteria with new parameters $\widetilde{\Psi} = \{\tilde{\psi}_{at}\}_{A \times T}$, $\tilde{\psi}_{at} = p(a|t)$:

$$\Theta^\ell \approx \widetilde{\Psi}\Theta$$

$$R_4(\Theta, \widetilde{\Psi}) = \sum_{a \in A} \sum_{d \in D} n_{ad} \ln \sum_{t \in T} \tilde{\psi}_{at}\theta_{td}$$

Implementation: new modality with tokens corresponding to $a \in A$.

hARTM: interlevel regularizers illustration



$$F = \bigsqcup_{m \in M} F^m, \; F^m = \{f_{dw}\}_{W^m \times T}, \; f_{dw} = \mathrm{norm}_{w \in W^m}[n_{dw}]$$

hARTM: hierarchy sparsing with Θ interlevel regularizer

The goal: topics have small number of parent topics
$\Leftrightarrow p(a|t)$ is sparse.

- Entropy sparsing regularizer:

$$R_5(\widetilde{\Psi}) = -\sum_{t \in T} \sum_{a \in A} \frac{1}{|A|} \ln \tilde{\psi}_{at}$$

Updated M-step:

$$\tilde{\psi}_{at} = \operatorname*{norm}_{a \in A} \left[n_{at} - \frac{\tau_5}{|A|} \right]$$

Drawback: the possibility of $p(a|t) = 0 \, \forall a$

- Power sparsing regularizer:

$$R_5(\widetilde{\Psi}) = \frac{1}{q} \sum_{t \in T} \sum_{a \in A} \tilde{\psi}_{at}^q, \; q > 1$$

Updated M-step:

$$\tilde{\psi}_{at} = \operatorname*{norm}_{a \in A} \left[n_{at} + \tau_5 \tilde{\psi}_{at}^q \right]$$

hARTM: hierarchy sparsing with Φ interlevel regularizer

The goal: topics have small number of parent topics
$\Leftrightarrow p(a|t)$ is sparse.

- Entropy sparsing regularizer:

$$R_5(\Psi) = \sum_{t \in T} \sum_{a \in A} \frac{1}{|A|} \ln p(a|t) = \frac{1}{|A|} \sum_{a} \sum_{t} \ln \frac{\psi_{ta}\, p(a)}{\sum_{a'} \psi_{ta'}\, p(a')}$$

Updated M-step:

$$\psi_{ta} = \underset{t \in T}{\mathrm{norm}} \left[n_{ta} - \tau_5 \left(\frac{1}{|A|} - p(a|t) \right) \right]$$

At any time $\forall t\, \exists a : p(a|t) > 0$.

hARTM in BigARTM

Key BigARTM concepts:

- Documents set is split into *batches* and stored on disk
- 1 EM-step = a pass through batches $\times$ iterating over each batch
- Storing Φ permanently, retraining Θ for any loaded batch

Φ interlevel regularizer implementation:

1. Learn levels $\ell = 1, 2, 3 \ldots$
2. For levels $\ell > 1$ add 1 extra batch composed from $(\ell - 1)$-th level's Φ
3. Extract Ψ as Θ corresponding to extra batch

Θ intervelel regularizer implementation:

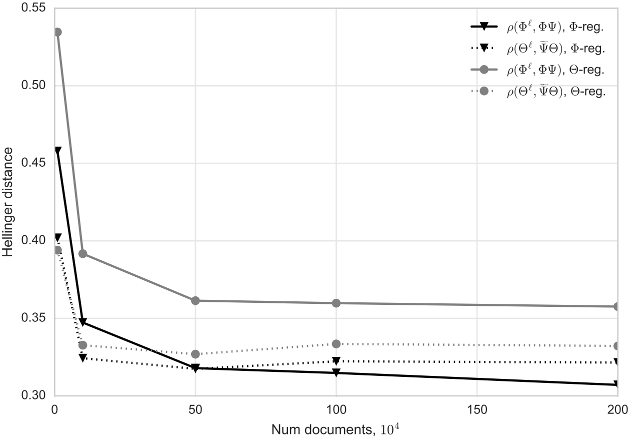
1. Learn levels $\ell = 1, 2, 3 \ldots$
2. For levels $\ell > 1$ modify all batches: add extra modality composed from $(\ell - 1)$-th level's Θ
3. Extract $\widetilde{\Psi}$ as Φ corresponding to extra modality

Experiments: comparison of Φ and Θ interlevel regularizers

Wikipedia: $D = 3.6 \cdot 10^6$, $W = 10^5$.

Learning 2^{nd} level, $|A| = 50$, $|T| = 250$, vary number of batches.

Measuring the quality of approximation $\Phi^\ell \approx \Phi\Psi$ and $\Theta^\ell = \widetilde{\Psi}\Theta$.



Approximation is quite the same with both regularizers, Φ-reg. is better.

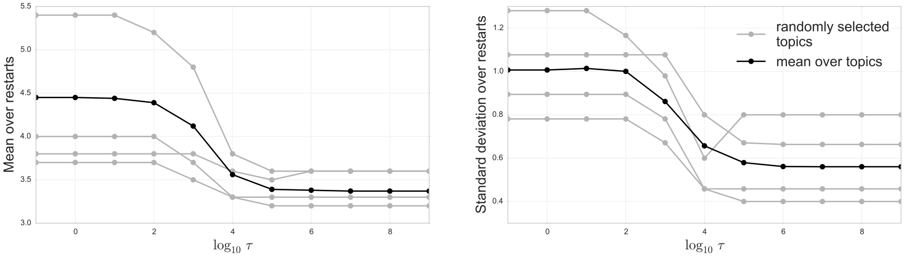
Experiments: children number study

Postnauka: $D = 1728$, $W = 38467$.

Learning 2[nd] level with Φ-reg., $|A| = 10$, $|T| = 30$, vary hierarchy sparsing reg. τ_5.

Measuring the mean and standard deviation of estimated subtopics count over 10 restarts.

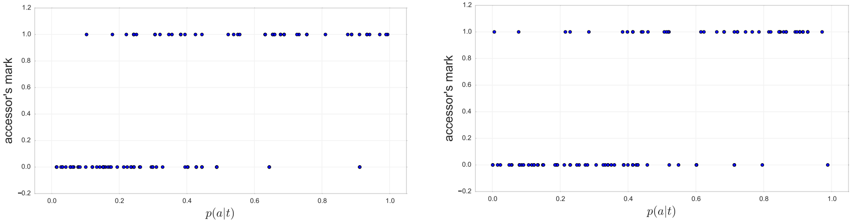
t is a child of a if $p(t|a) >$ threshold.



The bigger τ_5, the more sparse the hierarchy. For large τ_5 subtopics count estimation is robust (std < 1).

Experiments: parent-child relations study

Postnauka: $D = 1728$, $W = 38467$.
Learning $10 - 30 - 90$ topics hierarchy with Φ-reg. Generating 100 pairs topic-subtopic, asking an expert to mark a pair as "relation exists" or not.



no sparsing $\qquad$ Ψ sparsing

When using the hierarchy sparcing, we can impose a threshold with minimum errors.

Summary

Contributions:

- An approach to learn topic hierarchies from multimodal data with additional requirements.
- A method to control hierarchy sparsity.
- Open-source implementation in BigARTM with friendly interface.

Ongoing projects with hARTM:

- Creating a user-friendly navigator through Postnauka.ru materials.
- Developing a system for online news flow filtration.