Иерархические тематические модели для интерактивной навигации по коллекциям текстовых документов

Чиркова Надежда Александровна
Научный руководитель: Воронцов Константин Вячеславович

05.05.16

Аддитивная регуляризация тематических моделей

Дано: D — коллекция текстовых документов, W — множество терминов, n_{dw} — матрица частот слов в документах.

$$F_{wd} = p(w|d) = \frac{n_{dw}}{n_d}$$

Модель: S — множество тем

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

с параметрами $\Phi = \{\varphi_{ws}\}_{W \times S}$ и $\Theta = \{\theta_{sd}\}_{S \times D}$:
$\varphi_{ws} = p(w|s)$ — распределение слов в теме s;
$\theta_{sd} = p(s|d)$ — распределение тем в документе d.

Обучение модели:

оптимизация регуляризованного логарифма правдоподобия:

$$\mathscr{L}(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{s \in S} \varphi_{ws}\theta_{sd} + \sum_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta} \qquad (1)$$

$$\sum_{w \in W} \phi_{ws} = 1; \phi_{ws} \geqslant 0; \quad \sum_s \theta_{sd} = 1; \theta_{sd} \geqslant 0. \qquad (2)$$

Обучение APTM

Теорема (Воронцов К. В., Потапенко А. А., 2014) — EM-алгоритм обучения модели

Если $R(\Phi, \Theta)$ непрерывно дифференцируема по своим параметрам, то стационарная точка задачи (1)—(2) удовлетворяет следующей системе уравнений:

E-шаг: $p(s|d,w) = \underset{s\in S}{\text{norm}}(\phi_{ws}\theta_{sd})$;

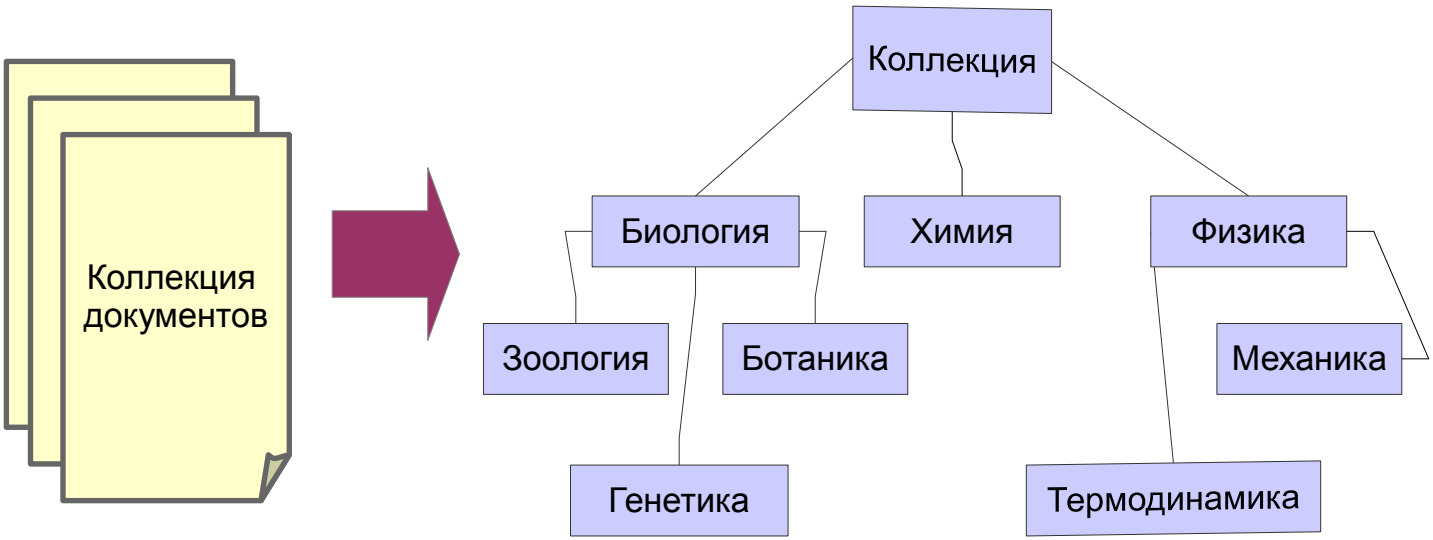
M-шаг: $\varphi_{ws} = \underset{w\in W}{\text{norm}}\left(n_{ws} + \varphi_{ws}\dfrac{\partial R}{\partial \varphi_{ws}}\right)$; $n_{ws} = \sum_{d\in D} n_{dw} p(s|d,w)$;

$$\theta_{sd} = \underset{s\in S}{\text{norm}}\left(n_{sd} + \theta_{sd}\dfrac{\partial R}{\partial \theta_{sd}}\right) ; \; n_{sd} = \sum_{w\in W} n_{dw} p(s|d,w) \; .$$

$\underset{t\in T}{norm}(x_t) = \max\{x_t, 0\}/(\sum_{t\in T} \max\{x_t, 0\})$.

Обучение модели — метод простой итерации.

Иерархические тематические модели



Тематическая иерархия — многодольный граф тем с увеличивающимся количеством тем на каждом уровне (доле).

Существующие подходы к построению иерархий

- *Blei et al.* Hierarchical topic models and the nested Chinese restaurant process (2003)
- *Mimno et al.* Mixtures of Hierarchical Topics with Pachinko Allocation (2007)
- *Zavitsanos et al.* Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes (2011)
- *Wang et al.* A Phrase Mining Framework for Recursive Construction of a Topical Hierarchy (2013)
- *Wang et al.* Scalable and Robust Construction of Topical Hierarchies (2014)
- *Wang et al.* Constructing topical hierarchies in heterogeneous information networks (2014)
- *Pujara J., Skomoroch P.* Large-Scale Hierarchical Topic Models (2012)

Недостатки существующих подходов:

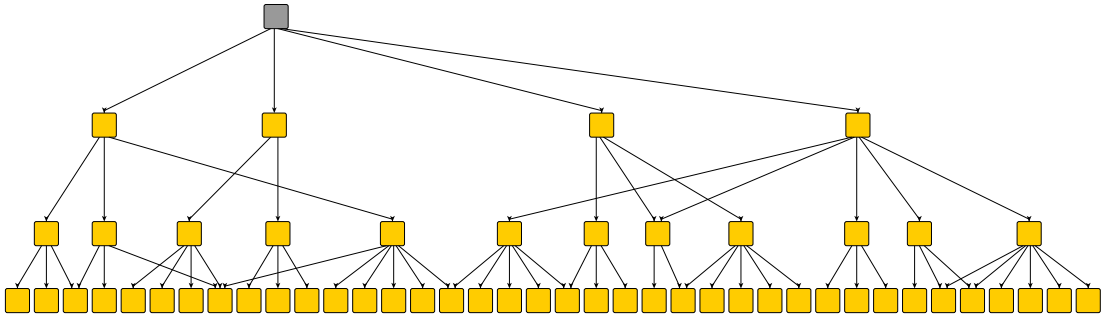
- сложно вводить предметно-ориентированные ограничения на параметры модели;
- большое количество гиперпараметров модели;
- не учитывают удобство пользователя.

Цель работы

Цель работы: предложить способ построения иерархической тематической модели:

1. которую удобно комбинировать с другими модификациями тематических моделей,
2. требующую минимальной настройки под конкретную коллекцию документов,
3. на основе которой будет удобно создавать тематический навигатор для пользователя.

Послойное нисходящее построение с регуляризатором Θ



T тем (построено): Φ^p, Θ^p

S тем: Φ, Θ

Обычная модель:

параметры Φ, Θ, $\varphi_{ws} = p(w|s)$, $\theta_{sd} = p(s|d)$, цель: $F \approx \Phi\Theta$

Добавим новую матрицу параметров $\Psi \in R^{|T|\times|S|}$:

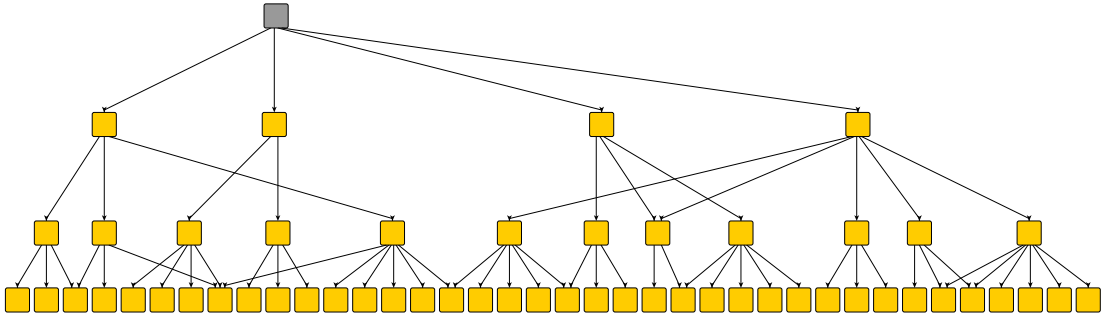
$$\psi_{ts} = p(t|s), \quad \Theta^p \approx \Psi\Theta.$$

Новая оптимизационная задача:

$$\sum_{d,w} n_{dw} \ln \sum_{s\in S} \varphi_{ws}\theta_{sd} + \lambda \underbrace{\sum_{t\in T}\sum_{d\in D} \theta^p_{td} \ln \sum_{s\in S} \psi_{ts}\theta_{sd}}_{\text{регуляризатор } \Theta} + R(\Phi,\Theta,\Psi) \to \max_{\Phi,\Theta,\Psi}$$

Введение регуляризатора Θ равносильно добавлению в коллекцию $|T|$ виртуальных **слов**, соответствующих родительским темам.

Послойное нисходящее построение с регуляризатором Φ



T тем (построено): Φ^p, Θ^p

S тем: Φ, Θ

Аналогично вводится регуляризатор Φ: $\widetilde{\Psi} \in R^{|S|\times|T|}$:

$$\tilde{\psi}_{st} = p(s|t), \quad \Phi^p \approx \Phi\widetilde{\Psi}.$$

Новая оптимизационная задача:

$$\sum_{d,w} n_{dw} \ln \sum_{s\in S} \varphi_{ws}\theta_{sd} + \lambda \underbrace{\sum_{w\in W}\sum_{t\in T} \phi^p_{wt} \ln \sum_{s\in S} \phi_{ws}\tilde{\psi}_{st}}_{\text{регуляризатор } \Phi} + R(\Phi,\Theta,\Psi) \to \max_{\Phi,\Theta,\Psi}$$

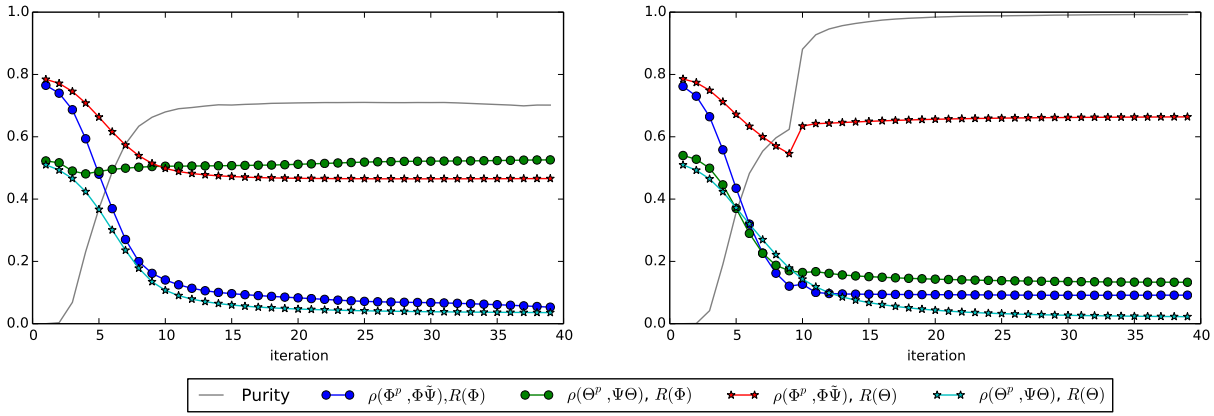
Введение регуляризатора Φ равносильно добавлению в коллекцию $|T|$ виртуальных **документов** — столбцов Φ^p.

Экспериментальное сравнение двух регуляризаторов

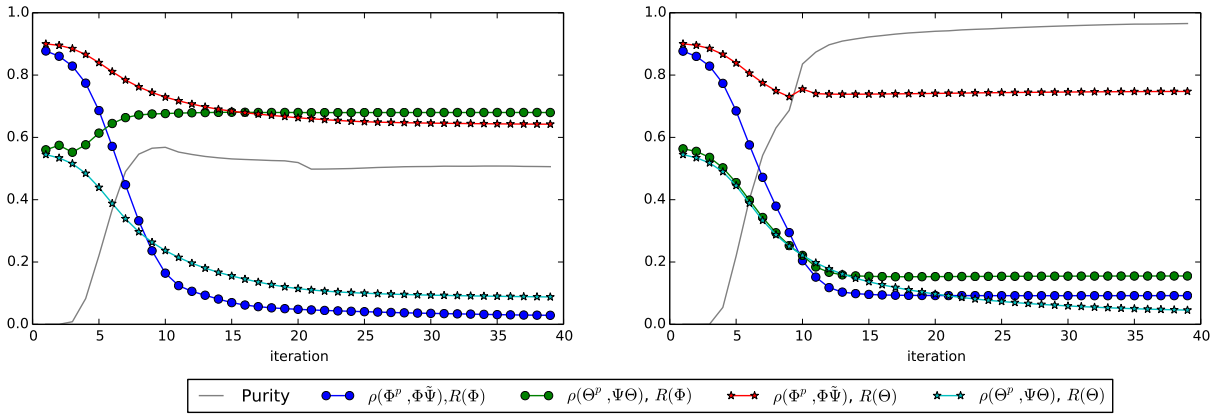
По оси x — номер итерации, по оси y — качество приближения Θ^p и Φ^p.

Разреживание с 1-й итерации **Разреживание с 10-й итерации**

Переход между 1-м и 2-м уровнями:



Переход между 2-м и 3-м уровнями:



Вывод: регуляризатор Φ выполняет функцию обоих регуляризаторов.

Одновременное построение всех уровней иерархии

Дано: n_{dw} — матрица частот слов в документах.

$$F_{wd} = p(w|d) = \frac{n_{dw}}{n_d}$$

Модель:

$$p(w|d) = \sum_{\ell} \sum_{s^\ell,\ldots,s^L} \phi^{\ell}_{ws^\ell} \psi^{\ell}_{s^\ell,s^{\ell+1}} \ldots \psi^{L-1}_{s^{L-1},s^L} \theta_{s^L d} \eta_{\ell d},$$

- $\phi_{ws^\ell} = p(w|s^\ell)$ — слов в теме s^ℓ,
- $\psi^{\ell}_{s^\ell,s^{\ell+1}} = p(s^\ell|s^{\ell+1})$ — распределение надтем темы $s^{\ell+1}$,
- $\theta_{s^L d}$ — распределение тем в документе d,
- $\eta_{\ell d}$ — распределение уровней иерархии в документе d.

Параметры модели: $\Phi^1, \ldots, \Phi^L$; $\Psi^1, \ldots, \Psi^{L-1}$; Θ; H.

Модель в виде матричного разложения:

$$F \approx \Phi^L \Theta^L \Lambda^L + \Phi^{L-1} \Theta^{L-1} \Lambda^{L-1} + \ldots + \Phi^1 \Theta^1 \Lambda^1,$$

$$\Lambda^\ell = \operatorname{diag}\{\eta_{\ell d}\}_{d \in D}, \quad \Theta^\ell = \Psi^\ell \ldots \Psi^{L-1} \Theta;$$

Обучение иерархической модели

Метод максимального правдоподобия:

$$\ln L = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{\ell} \sum_{s^\ell, \ldots, s^L} \phi^\ell_{ws^\ell} \psi^\ell_{s^\ell, s^{\ell+1}} \ldots \psi^{L-1}_{s^{L-1}, s^L} \theta_{s^L d} \eta_{\ell d} \to \max_{\{\Phi\}, \{\Psi\}, \Theta, H} \tag{3}$$

$$\phi^\ell_{ws^\ell} \geqslant 0, \sum_w \phi^l_{ws^l} = 1; \quad \psi^l_{s^l, s^{l+1}} \geqslant 0, \sum_{s^l} \psi^l_{s^l, s^{l+1}} = 1;$$

$$\theta_{s^L d} \geqslant 0, \sum_{s^L} \theta_{s^L d} = 1; \quad \eta_{ld} \geqslant 0, \sum_l \eta_{ld} = 1. \tag{4}$$

Обучение иерархической модели

Теорема

Стационарная точка задачи (3)—(4) удовлетворяет следующей системе уравнений:

$$\begin{aligned}
\phi'_{ws^\ell} &= \underset{w}{\text{norm}} \frac{\partial \ln L}{\partial \phi^\ell_{ws^\ell}} \phi^\ell_{ws^\ell} \\
\psi^\ell_{s^\ell, s^{\ell+1}} &= \underset{s^\ell}{\text{norm}} \frac{\partial \ln L}{\partial \psi^\ell_{s^\ell, s^{\ell+1}}} \psi^\ell_{s^\ell, s^{\ell+1}} \\
\theta_{s^L d} &= \underset{s^L}{\text{norm}} \frac{\partial \ln L}{\partial \theta_{s^L d}} \theta_{s^L d} \\
\eta_{\ell d} &= \underset{\ell}{\text{norm}} \frac{\partial \ln L}{\partial \eta_{\ell d}} \eta_{\ell d}
\end{aligned}$$

Обучение модели — метод простой итерации.

Трансформация распределений для разреживания модели

Разреженное дискретное распределение: большинство элементов носителя имеют вероятность 0.
Примеры:

- разреживание $p(s^\ell|s^{\ell+1})$ — каждая тема имеет мало надтем;
- разреживание тем $p(s^\ell, \ell|w)$ — слово относится к 1–2 темам.

Пусть $\alpha = (\alpha_1, \ldots, \alpha_K), \alpha_i \geqslant 0, \sum_{i=1}^K \alpha_i = 1$ — дискретное распределение, например $p(s^\ell|s^{\ell+1})$.
Разреживающая трансформация распределения:

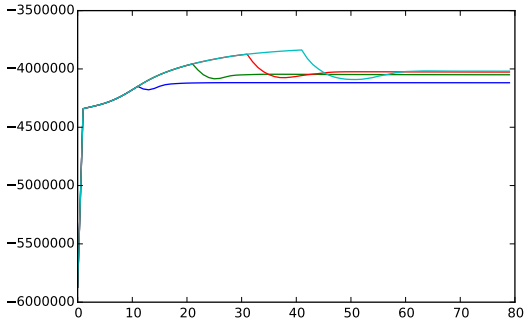
$$\alpha_i \leftarrow \underset{i}{\text{norm}}\, \alpha_i^p$$

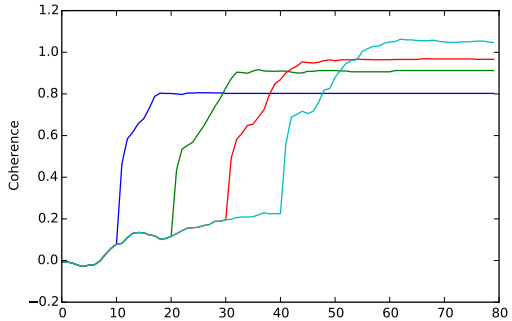
$\leftarrow$ — присваивание, $p > 1$ — гиперпараметр. Трансформация чередуется с итерациями алгоритма обучения модели.

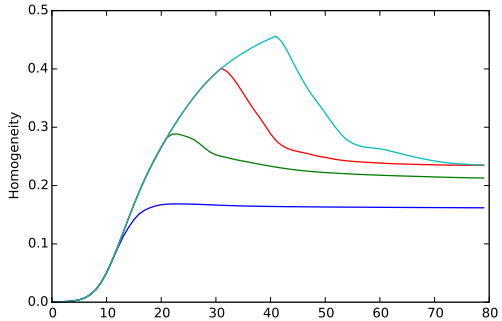
Эксперименты с разреживанием тем

По оси x — номер итерации:

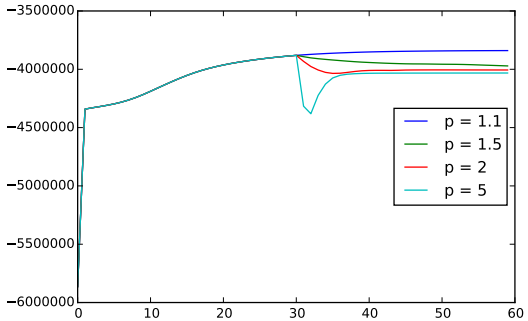
Разная стартовая итерация разреживания

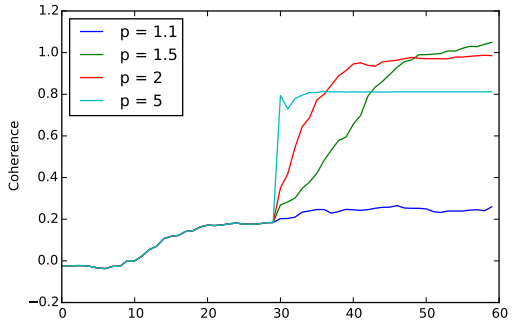


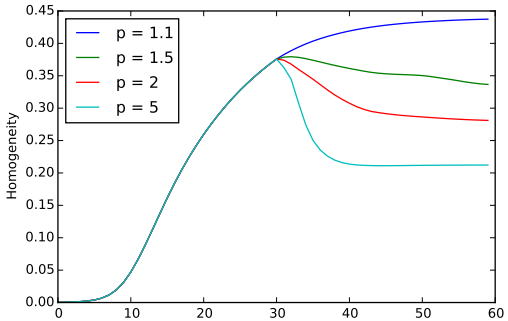




Разный параметр разреживания







$\ln L$

Мера интерпретируемости тем

Мера качества кластеризации документов

Вывод: Лучше начинать разреживание позже, можно брать параметр по умолчанию $p = 2$ или $p = 1.5$.

Примеры тем

Примеры тем при разных стартовых итерациях разреживания:

10	**при**, энергия, изменение, магнитный, заряд, волна, **свойство**, линия, расстояние	дело, император, идея, орган, **там**, **поздно**, николай, философский, **читать**	a, x, b, **следующий**, формула, x, прямой, f, ноль
20	**при**, **можно**, энергия, **через**, направление, поле, магнитный, заряд, волна	дело, **где**, император, смерть, реформа, **хороший**, сын, орган, **там**	a, x, b, получать, пример, уравнение, x, равный, **следующий**
30	энергия, направление, поле, магнитный, заряд, электрический, волна, линия, проводник	царь, император, смерть, александр, реформа, сын, восстание, **там**, дело	a, x, b, равный, выражение, формула, дробь, корень, ноль
40	поле, магнитный, заряд, электрический, волна, направление, линия, проводник, электромагнитный	царь, император, смерть, александр, реформа, церковь, правление, сын, восстание	a, b, выражение, дробь, корень, sin, степень, формула, cos

Экспериментальное сравнение послойного и одновременного построения иерархий

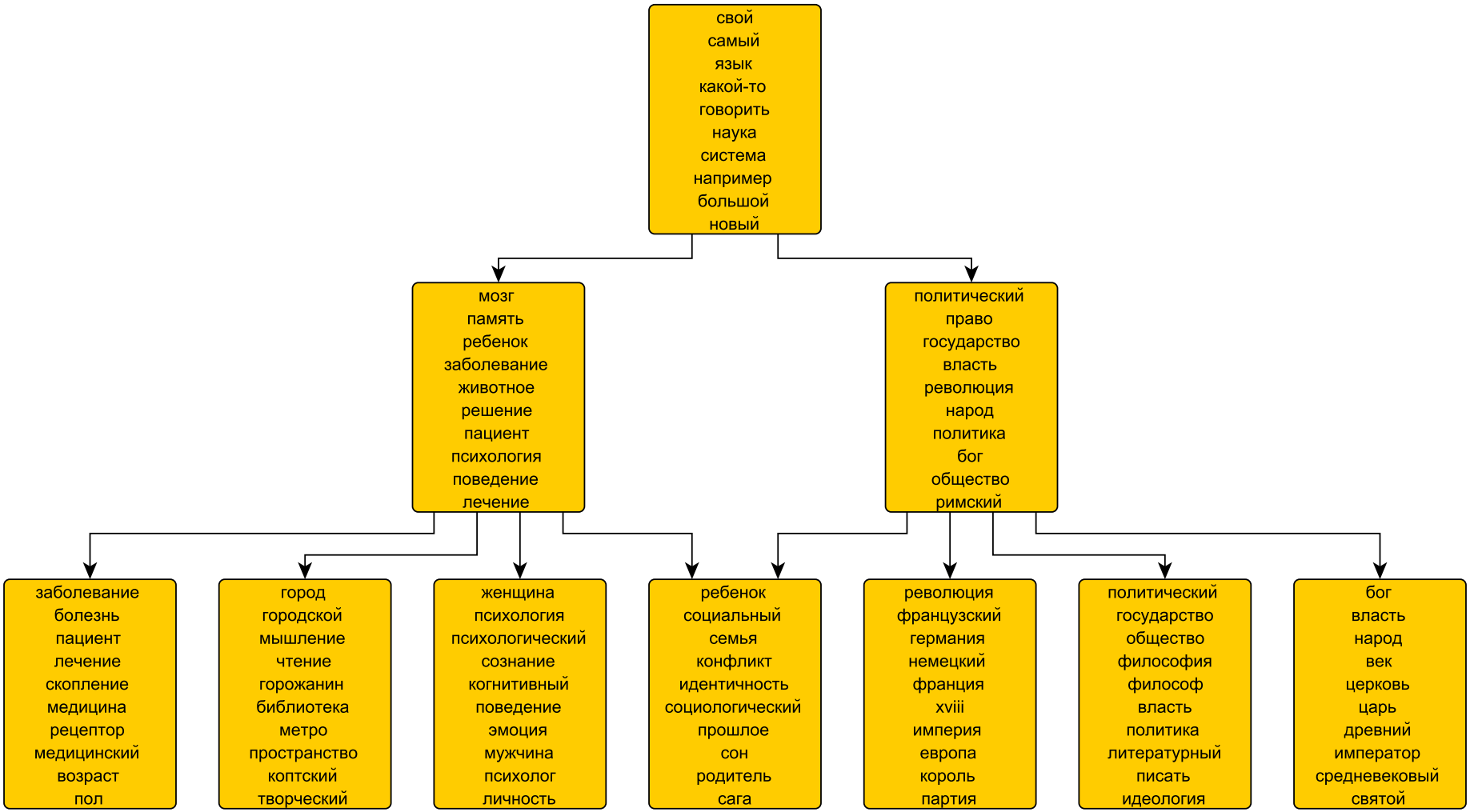
Сравнение на иерархии $1 - 10 - 30$ тем.
Коллекция школьных конспектов ($|D| = 1491$, $|W| = 27263$):

	Интер-ть тем	Плотность графа	Время обучения
ARTM с рег. Ф	1.166	**0.129**	**10 мин. 22 с**
Единая модель	**1.223**	0.433	30 мин. 34 с

Коллекция текстовых записей видеолекций сайта Постнаука ($|D| = 1728$, $|W| = 38467$):

	Интер-ть тем	Плотность графа	Время обучения
ARTM с рег. Ф	**1.253**	**0.146**	**17 мин. 14 с**
Единая модель	1.007	0.46	31 мин. 59 с

Фрагмент иерархической модели с рег. Ф коллекции «Постнака».



На защиту выносятся следующие результаты:

1. Предложены две нисходящие стратегии построения тематической иерархии с использованием регуляризатора связи уровней.
2. Предложена иерархическая вероятностная тематическая модель текстовой коллекции, основанная на сумме матричных разложений.
3. Предложена методология оценивания качества тематических иерархий.