Построение и оценка качества гетерогенных иерархических тематических моделей

Селезнева Мария Сергеевна

Московский физико-технический институт
(государственный университет)

Научный руководитель: Воронцов К. В., д.ф.-м.н., ВЦ РАН

2018 год

Решаемая проблема

Цель

Разработать метод автоматического построения агрегатора контента из разнородных (гетерогенных) источников.

Задачи

1. Построить интерпретируемую иерархическую тематическую модель коллекции текстов, собранных из гетерогенных источников.
2. Предложить способ автоматичекой оценки качества тематических иерархий.

Метод

Аддитивная регуляризация тематических моделей (ARTM).

Актуальность проблемы

Текущее положение дел

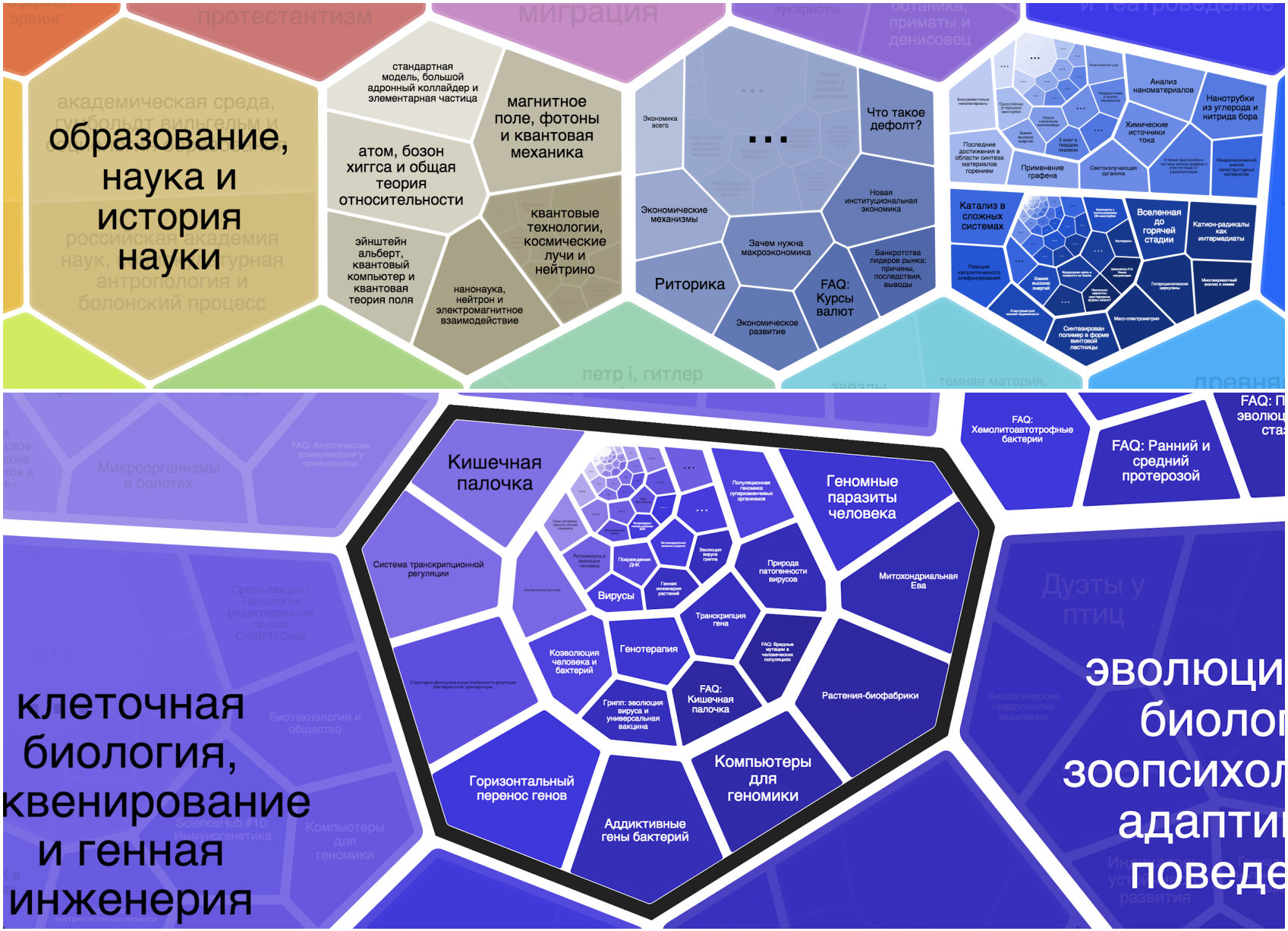
1. Тематические модели давно и успешно используются для визуализации корпусов научных текстов.
2. Иерархические ТМ предоставляют дополнительные возможности для визуализации больших гетерогенных коллекций и навигации по ним.

Проблемы

1. Общепризнанного подхода к измерению качества иерархических моделей **не существует**.
2. Существующие методы построения тематических моделей **не учитывают разнородность** коллекций.

Пример тематической иерархии

Пример тематической иерархии



Постановка задачи тематического моделирования

Дано: W — словарь терминов(токенов)

D — коллекция текстовых документов $d = \{w_1, ..., w_{n_d}\}$

Матрица $F = \{n_{dw}\}_{W \times D}$

n_{dw} — сколько раз w встретилось в документе d

T — множество тем

Найти: Матрицы $\Phi = \{\phi_{wt}\}_{W \times T}$, $\Theta = \{\theta_{td}\}_{T \times D}$

$\phi_{wt} = p(w|t)$ — вероятность термина w в теме t

$\theta_{td} = p(t|d)$ — вероятность темы t в документе d

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

Это задача матричного разложения $F = \Phi\Theta$!

Постановка задачи тематического моделирования

- У этой задачи существует бесконечно много решений вида $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$, где S — матрица ранга $|T|$
- Можно вводить регуляризацию матриц Φ и Θ

  PLSA: $R(\Phi, \Theta) = 0$

  LDA: $R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$

  ARTM: $R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$

Задача оптимизации ARTM

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

при условиях $\phi_{wt} \geq 0,\ \theta_{td} \geq 0,\ \sum_{w \in W} \phi_{wt} = 1,\ \sum_{t \in T} \theta_{td} = 1.$

Иерархическая модель ARTM

Дано: Φ^l, Θ^l — параметры $l-$того уровня иерархии

A – множество тем $l-$того уровня

Задача матричного разложения:

$\Phi^l = \Phi^{l+1}\Psi^{l+1}$, где

$\Phi^{l+1} = \{p(w|t)\}_{W\times T}$, $\Psi^{l+1} = \{p(t|a)\}_{T\times A}$

Формула регуляризатора

$$R(\Phi,\Psi) = \sum_{a\in A}\sum_{w\in W} n_{wa} \ln \sum_{t\in T} \phi_{wt}\psi_{ta} \to \max_{\Phi,\Psi}.$$

Введение регуляризатора эквивалентно добавлению в коллекцию $|A|$ псевдодокументов. Ψ образует $|A|$ дополнительных столбцов Θ.

Метрики качества плоских моделей

Измерение качества тем плоской модели

$$\text{Quality}(t) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(w_i^{(t)}, w_j^{(t)}),$$

где $f(w_i^{(t)}, w_j^{(t)})$ – мера совстречаемости топ токенов $w_i \in t$ и $w_j \in t$.

В литературе встречаются разные варианты $f(w_i, w_j)$:

Newman et al, 2010: $\ln \frac{p(w_i,w_j)}{p(w_i)p(w_j)}$

Mimno et al, 2011: $\ln \frac{d(w_i,w_j)+\epsilon}{d(w_i)}$

Nikolenko et al, 2015: $\ln \frac{\sum_d \text{tfidf}(w_i,d)\text{tfidf}(w_j,d)+\epsilon}{\sum_d \text{tfidf}(w_i,d)}$

Nikolenko et al, 2016: $\langle v_{w_i}, v_{w_j} \rangle$

Метрики качества иерархических моделей

Измерение качества тем плоской модели

$$\text{Quality}(t) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(w_i^{(t)}, w_j^{(t)}),$$

где $f(w_i^{(t)}, w_j^{(t)})$ – мера совстречаемости топ токенов $w_i \in t$ и $w_j \in t$.

Не существует принятой метрики качества иерархических моделей. **Предлагается ввести отдельно метрики качества связей иерархии.**

Измерение качества связей иерархии

$$\text{Quality}_e(a, t) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(w_i^{(a)}, w_j^{(t)}),$$

где $f(w_i^{(a)}, w_j^{(t)})$ – мера совстречаемости топ токенов $w_i \in a$ и $w_j \in t$.

Предлагаемые метрики качества связей (a, t)

EmbedSim:
$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle v(w_i^{(a)}), v(w_j^{(t)})\rangle,$$
$v(w)$ — векторное представление токена w.

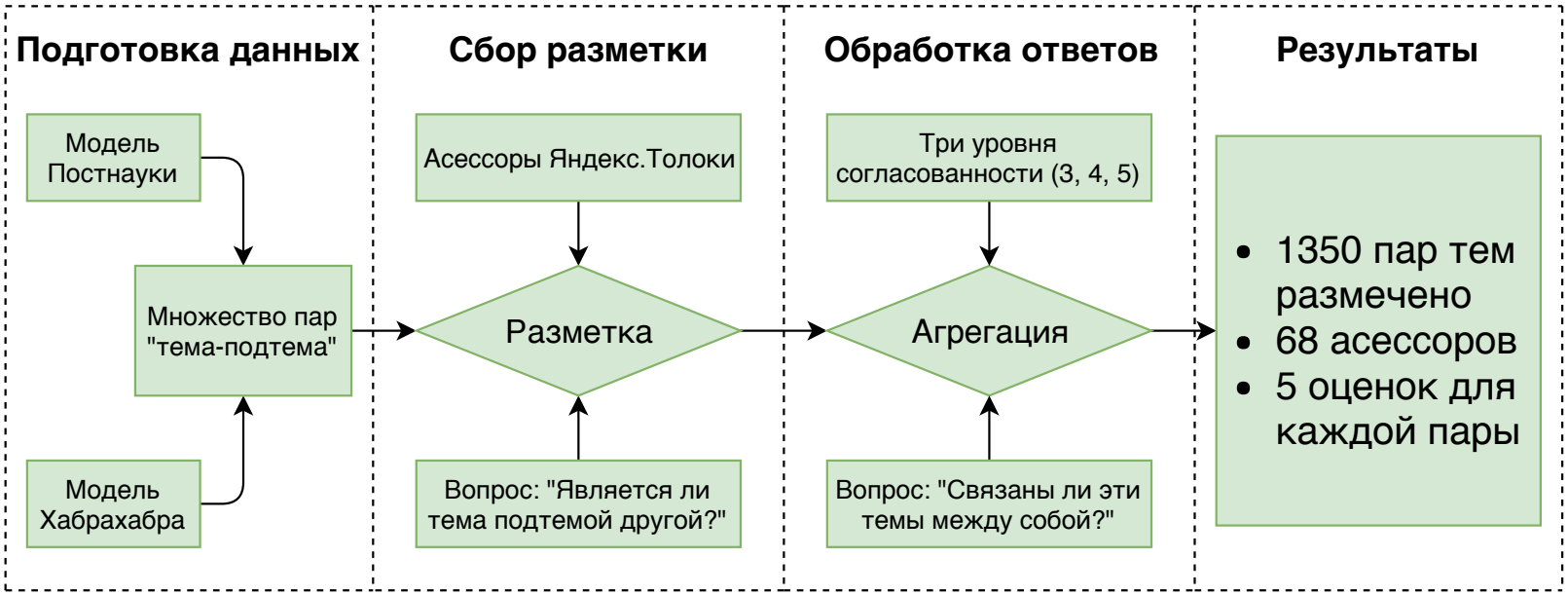
CoocSim:
$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ln\frac{d(w_i^{(a)}, w_j^{(t)})+\varepsilon}{d(w_j^{(t)})},$$
$d(w_i, w_j)$ — совстречаемость токенов w_i и w_j.

HellingerSim:
$$1-\frac{1}{\sqrt{2}}\|\sqrt{p(w|a)}-\sqrt{p(w|t)}\|_2$$

KLSim:
$$-D_{KL}(p(w|a)\|p(w|t))$$

Сбор асессорской разметки

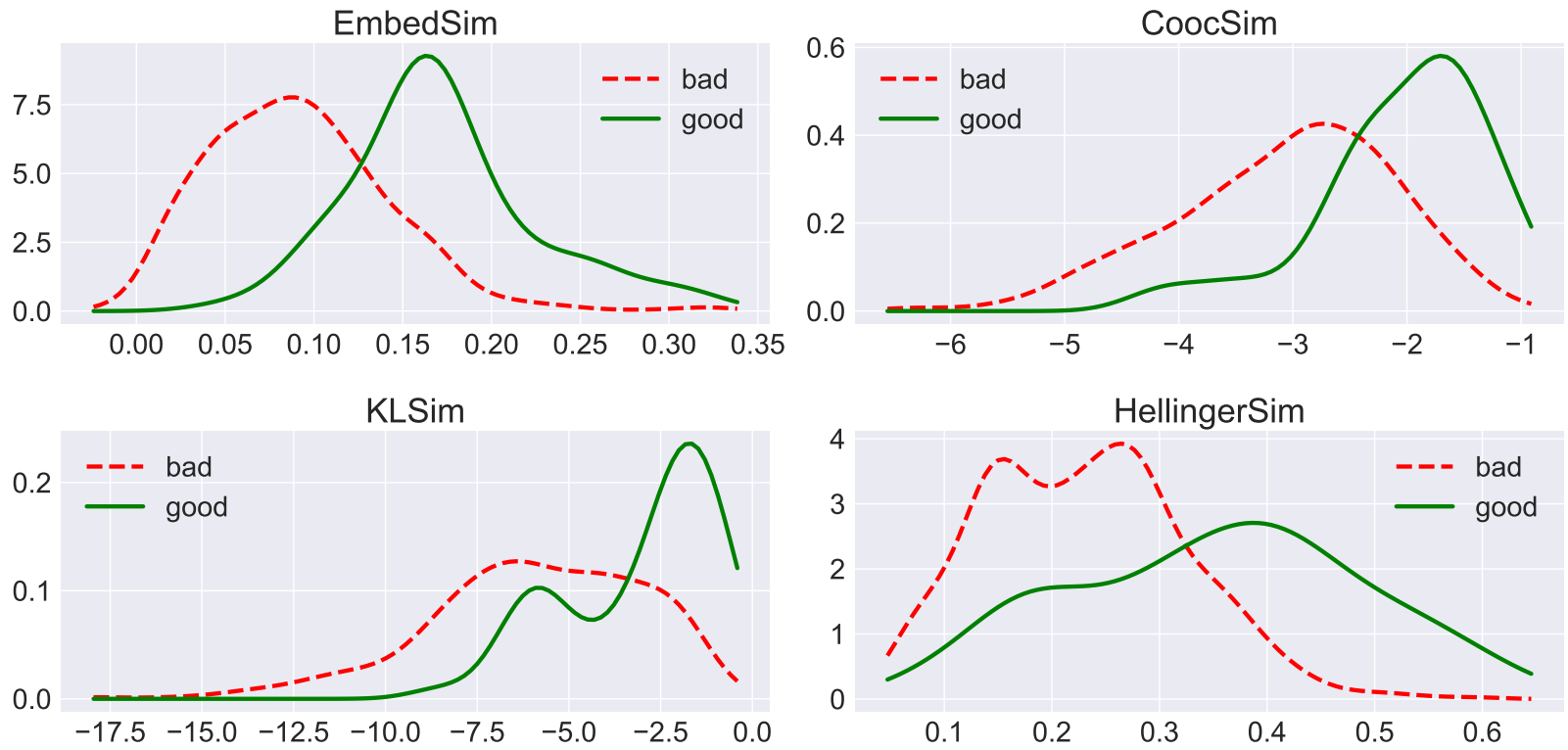
Сбор оценок асессоров осуществлялся по следующей схеме:



В итоге получен датасет ребер (пар тем соседних уровней), помеченных как «хорошие» (не менее 4 из 5 асессоров считает, что темы связаны) и «плохие» (остальные).

Сравнение метрик с асессорскими оценками

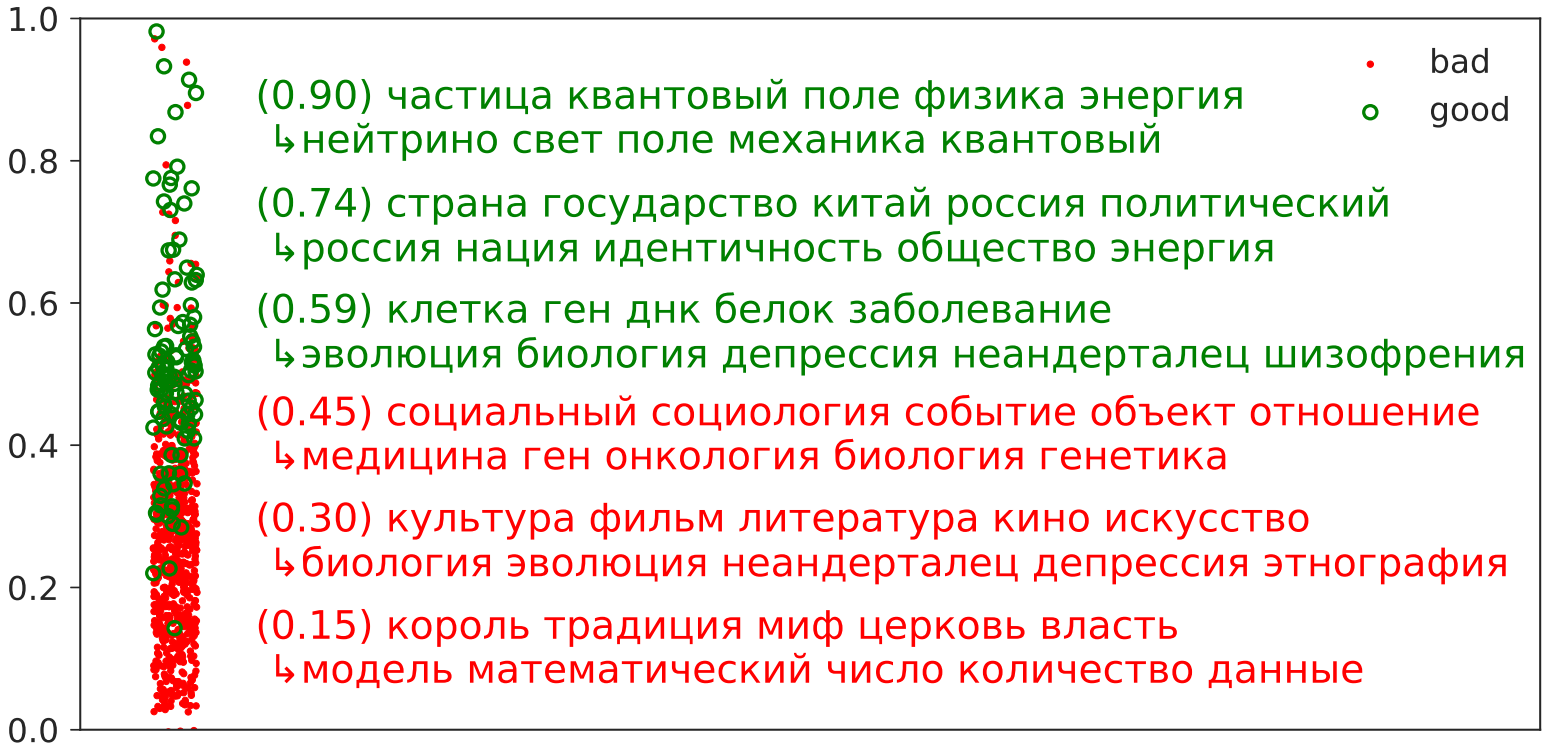
На графиках представлены плотности распределения значений предложенных метрик для «плохих» и «хороших» ребер.



Лучше всего разделяет «плохие» и «хорошие» ребра метрика EmbedSim (AUC=0.878), поэтому используем ее в дальнейшем.

Интерпретируемость метрик на примере EmbedSim

Несколько примеров пар тем, которые оценивали асессоры и соответствующие значения метрики EmbedSim.



«Хорошие» пары – те, для которых не менее 4 асессоров согласились, что между темами есть связь.

Данные

- ПостНаука: 2976 документов, 43196 слов, 1799 тэгов
- Хабрахабр: 81076 документов, 588400 слов (35640 совпадают с ПостНаукой), 77102 тэгов (673 совпадают с ПостНаукой)

Гетерогенность коллекций

- Размер коллекции Хабрахабра во много раз превышает размер коллекции ПостНауки.
- Тематические структуры коллекций существенно различные: ПостНаука содержит больше тем.

Базовый алгоритм

В качестве базового алгоритма рассматриваем построение модели по объединенной коллекции.

Проблемы базового алгоритма

- почти все темы модели содержат более 90% документов из Хабрахабра;
- не выделяются специфические для ПостНауки темы;
- построение модели по большому корпусу занимает много времени.

Базовый алгоритм не решает поставленной задачи!

Предлагаемый алгоритм

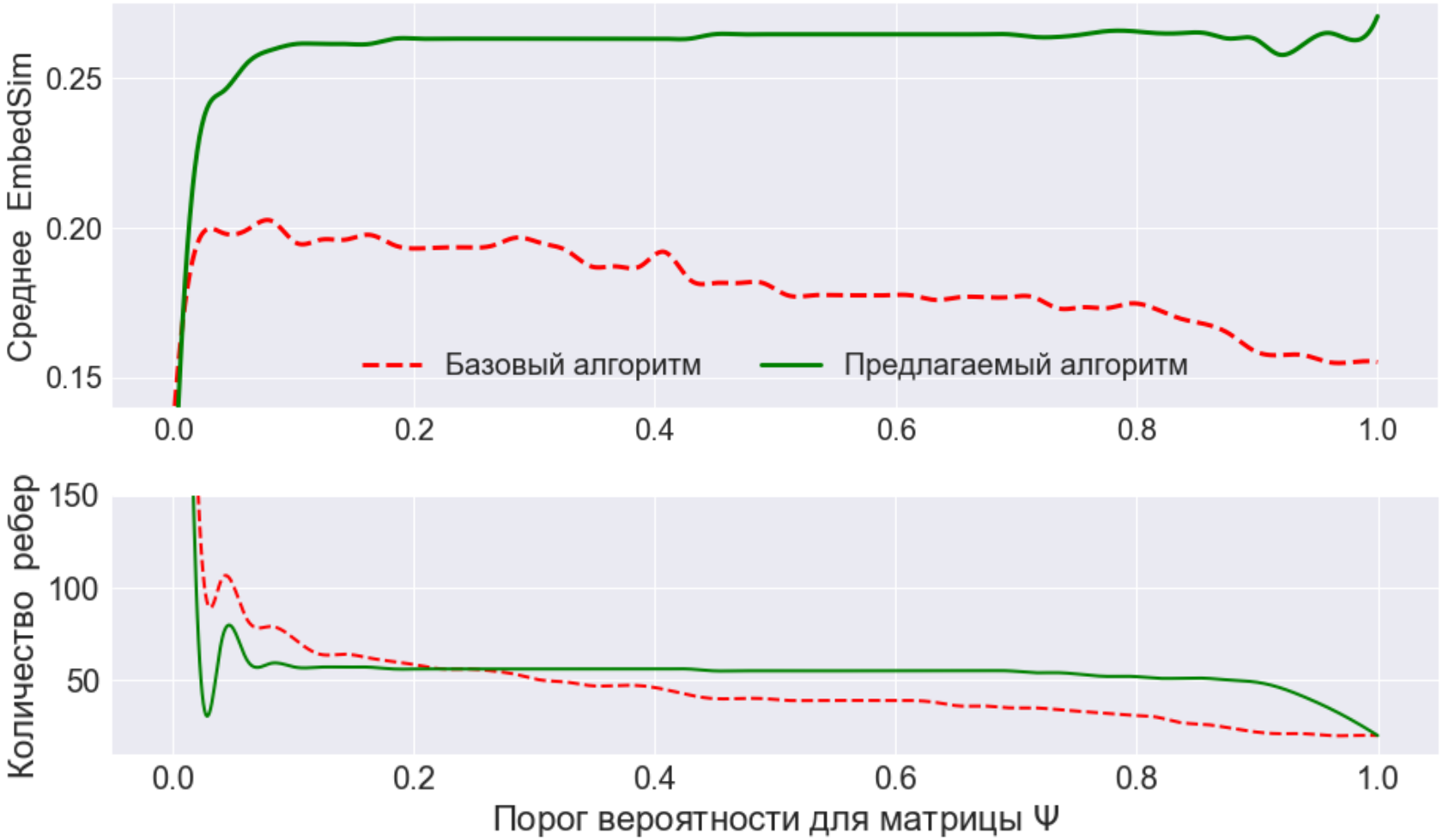
Φ_0^1 – матрица модели ПостНауки, D_1 – новая (доливаемая) коллекция, в нашем случае Хабрахабр.

Фильтрация D_1 – ранжирование документов по степени «похожести» на исходную коллекцию.
Для всех i = 1, ..., N:

1. Добавление в коллекцию новых документов , отранжированных с наибольшими значениями при фильтрации.
2. Инициализация матрицы Φ_i^1 матрицей Φ_0^1.
3. Построение модели hARTM.

Сравнение алгоритмов: среднее качество

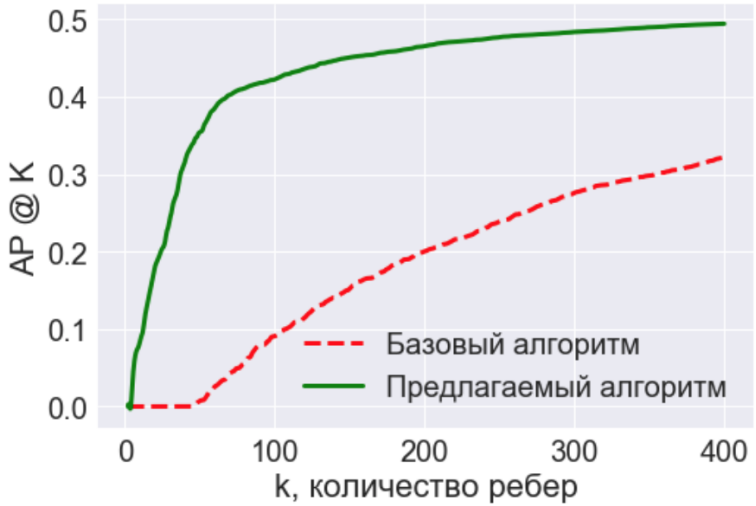
Среднее качество ребер иерархии по метрике EmbedSim.

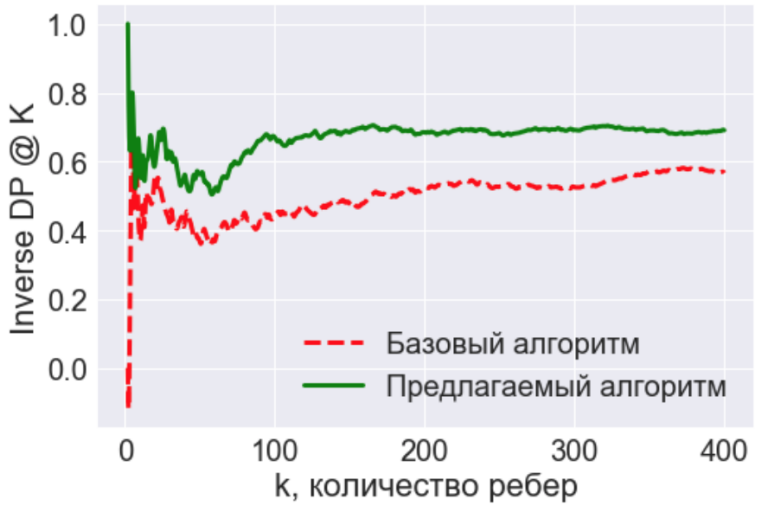


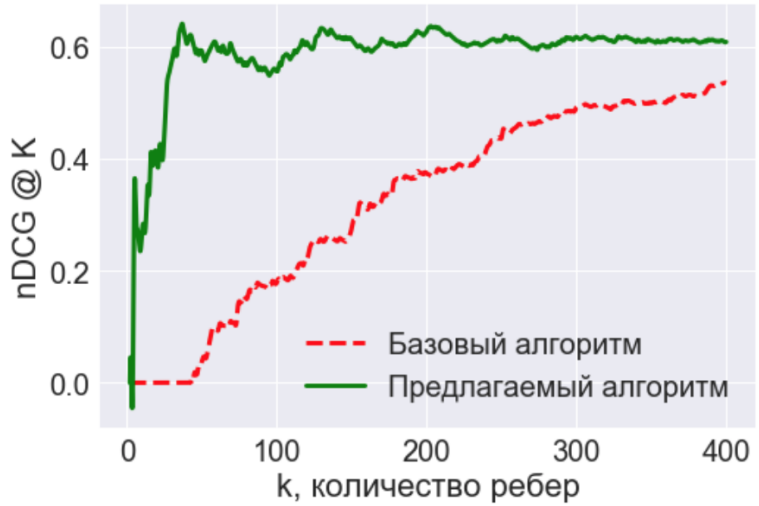
Предлагаемый алгоритм дает лучшее качество ребер иерархии равномерно по порогу $\Psi(t|a)$.

Сравнение алгоритмов: качество ранжирования

Метрики качества ранжирования ребер по значениям Ψ базового и предлагаемого алгоритмов. Правильное ранжирование задается метрикой EmbedSim.







Заключение

1. Предложены метрики качества связей иерархии, согласованные с человеческим представлениям о связности тем.
2. Предложен алгоритм итеративного построения модели, учитывающий гетерогенную структуру источников.
3. Показано, что предлагаемый алгоритм превосходит базовый по качеству.

Публикации и выступления

1. Belyy, A. V., Seleznova, M. S., Sholokhov, A. K., & Vorontsov, K. V. (2018). Quality evaluation and improvement for hierarchical topic modeling. In *Computational Linguistics and Intellectual Technologies (pp. 110-123)*.
2. Селезнева, М. С., Белый, А. В. & Шолохов, А. К. (2017). Агрегирование гетерогенных текстовых коллекций в иерархической тематической модели русскоязычного научно-популярного контента. *Труды 60-й Всероссийской научной конференции МФТИ. Прикладная математика и информатика (с. 90-91)*.