Вероятностное тематическое моделирование несбалансированных текстовых коллекций

Панкратов Виктор Владимирович

Московский физико-технический институт
Факультет упарвления и прикадной математики
Кафедра интеллектуальных систем

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

Москва,
2021г

Постановка задачи: вероятностная модель

Заданы три множества:

- D — множество документов
- W — множество слов
- T — множество тем

n_{wd} — число вхождений слова $w \in W$ в документ $d \in D$

Требуется найти матрицы Φ, Θ: $\phi_{wt} = p(w|t)$, $\theta_{td} = p(t|d)$

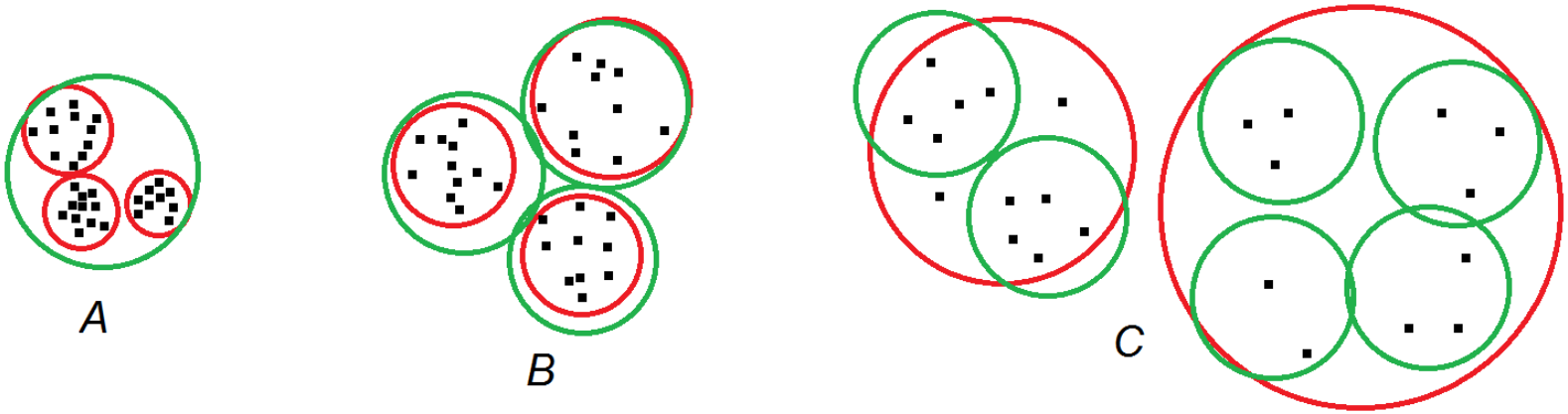
Критерий — максимизация регуляризованного правдоподобия:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_{i} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}, \tag{1}$$

где R_i — регуляризаторы, τ_i — коэффициенты регуляризации
Решение (1) осуществляется с помощью EM алгоритма.

Potapenko A., Vorontsov K. Additive regularization of topic model. 2015

Проблема несбалансированности

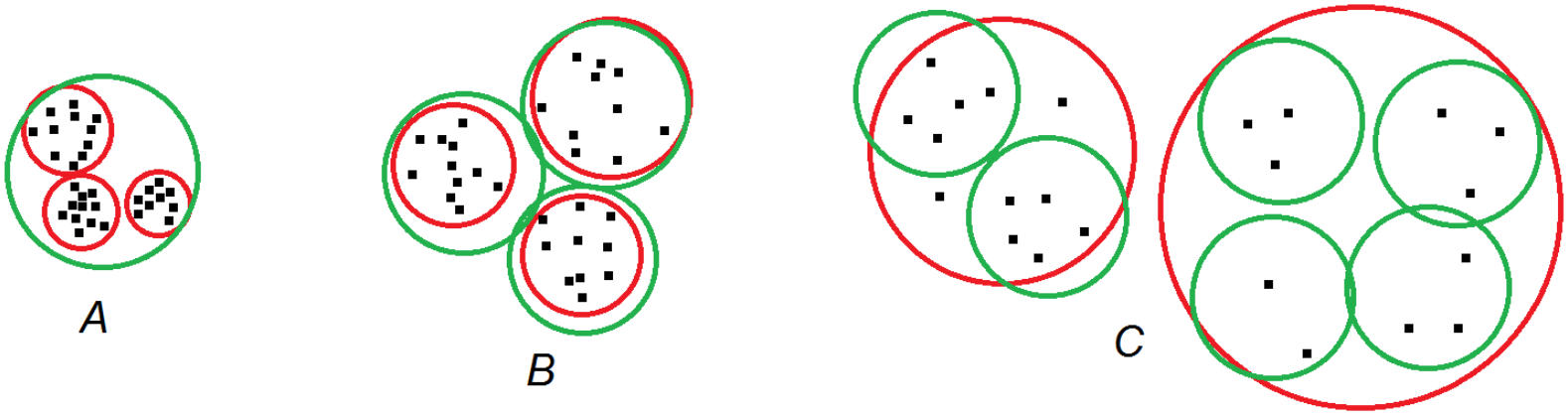


В случае тематически несбалансированной коллекции матричное разложение путём максимизации правдоподобия может приводить к дроблению крупных тем (A) и слиянию мелких (C).

Цель работы:

Предложить и экспериментально проверить решение проблемы несбалансированности с помощью регуляризатора

Семантическая неоднородность



Гипотеза условной независимости:
$p(w|t) = p(w|d,t); p(w,d|t) = p(w|t)p(d|t)$
Проверка - статистика семантической неоднородности:

$$S_t = KL(p(w,d|t)||p(w|t)p(d|t))$$

Тема — кластер размерности $|W|$, центр которого - $p(w|t)$.
S_t — удаленность $p(w|d,t)$ от центра кластера.

E.Veselova, K.Vorontsov. Topic balancing with additive regularization of topic models. 2020

Регуляризатор семантической неоднородности

Статистика семантической неоднородности

$$\mathrm{S_t} = \mathrm{KL}(\hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})||\mathrm{p}(\mathrm{w}|\mathrm{t})\mathrm{p}(\mathrm{d}|\mathrm{t})) = \sum_{\mathrm{d}\in\mathrm{D}}\sum_{\mathrm{w}\in\mathrm{d}} \hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})\ln\frac{\hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})}{\mathrm{p}(\mathrm{w}|\mathrm{t})\mathrm{p}(\mathrm{d}|\mathrm{t})}$$

Здесь $\hat{\mathrm{p}}$ — частотные оценки вероятности
Преобразовывая и суммируя по всем темам:

$$\sum_{\mathrm{t}\in\mathrm{T}}\mathrm{S_t} = \sum_{\mathrm{d}\in\mathrm{D}}\sum_{\mathrm{w}\in\mathrm{d}}\left(\sum_{\mathrm{t}\in\mathrm{T}}\hat{\mathrm{p}}(\mathrm{w},\mathrm{d}|\mathrm{t})\right)\ln\frac{\hat{\mathrm{p}}(\mathrm{w}|\mathrm{d})}{\mathrm{p}(\mathrm{w}|\mathrm{d})}$$

В результате преобразований получаем регуляризатор:

$$\mathrm{R} = \sum_{\mathrm{t}\in\mathrm{T}}\mathrm{S_t} = \sum_{\mathrm{d}\in\mathrm{D}}\sum_{\mathrm{w}\in\mathrm{d}}\beta_{\mathrm{dw}}\mathrm{n_{dw}}\ln\sum_{\mathrm{t}\in\mathrm{T}}\phi_{\mathrm{wt}}\theta_{\mathrm{td}},\quad \beta_{\mathrm{dw}} = \sum_{\mathrm{t}\in\mathrm{T}}\frac{\mathrm{p}(\mathrm{t}|\mathrm{d},\mathrm{w})}{\mathrm{p}(\mathrm{t})}$$

Эксперимент: генерация коллекции

Для эксперимента будем генерировать синтетическую коллекцию:

- Генерируются столбцы Φ_0, Θ_0, из симметричных распределений Дирихле
- Генерируется одна фоновая тема из распределения Ципфа
- Фоновая тема — столбец в Φ_0
- Фоновая тема — константная строка в Θ_0
- Из полученных матриц Φ_0,Θ_0 генерируются документы

$$t_i \sim \mathrm{Dir}(t|d_i),\ w_i \sim \mathrm{Dir}(w|t_i),\ i = 1\ldots \tag{2}$$

David M. Blei, Andrew Y. Ng and Michael I. Jordan Latent Dirichlet Allocation. 2003

Эксперимент: оценивание качества решения

- Φ_0 — исходная матрица вероятностей $p(w|t)$
- Φ — матрица вероятностей $p(w|t)$, найденная алгоритмом

Для всех пар i, j будем проверять равенства:

$$\arg\min_k \ (\mathrm{dist}(\Phi[i], \Phi_0[k])) = j \tag{3}$$

$$\arg\min_k \ (\mathrm{dist}(\Phi[k], \Phi_0[j])) = i \tag{4}$$

dist — заданная функция расстояния между столбцами Φ, Φ_0

Критерии качества восстановления матриц — число тем:

- взаимно близких: (3),(4) выполнены для некоторых i, j
- невосстановленных: $\Phi_0[j]$, если (3) не выполнено для всех i
- ложных: $\Phi[i]$, если (4) не выполнено для всех j

Эксперимент: параметры

Параметры генерации синтетической коллекции:

- Число тем $|\mathrm{T}| = 100$
- Число документов $|\mathrm{D}| = 2000$
- Число слов в каждом документе $|\mathrm{w} \in (\mathrm{d} \in \mathrm{D})| = 1000$
- Коэффициент регуляризации $\tau = 0.5$

Степень несбалансированности коллекции:

- Отношение минимальной и максимальной сумм вероятностей:

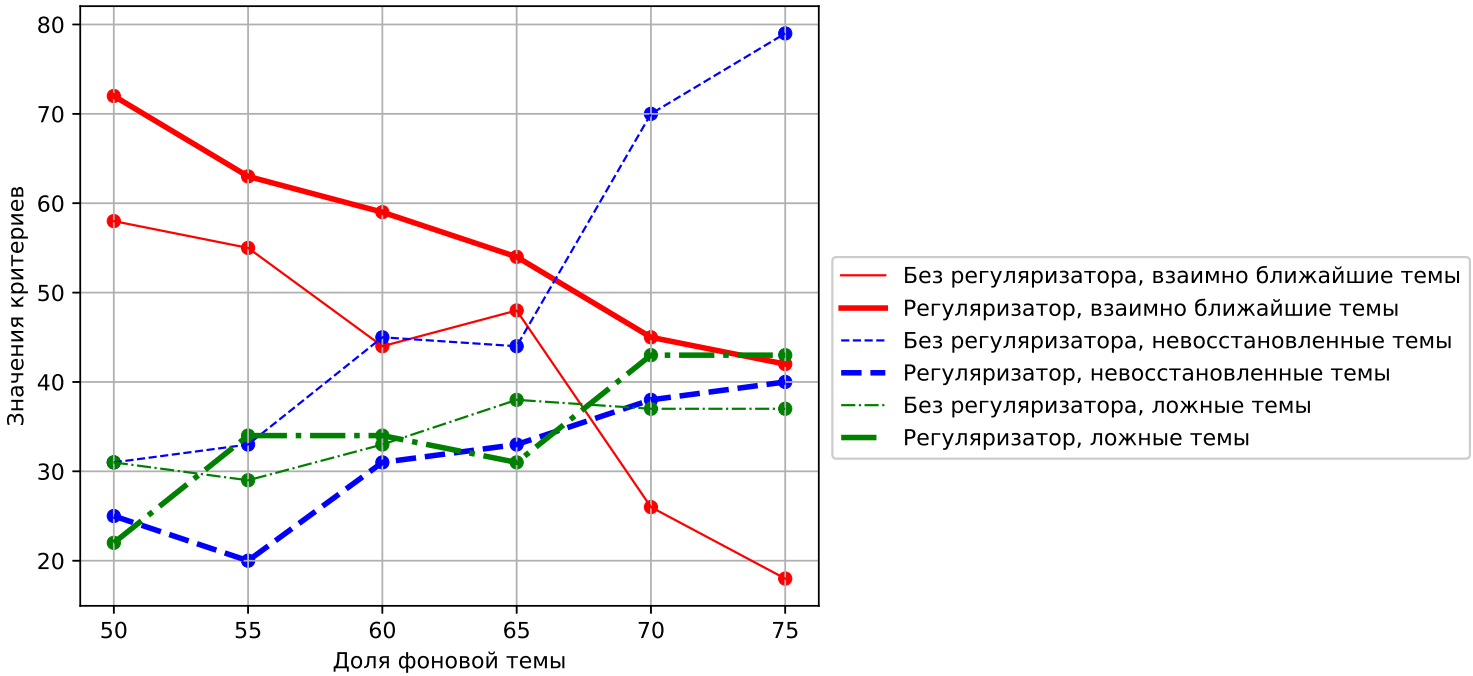
$$\mathrm{S}_1 = \frac{\max\limits_{\mathrm{t}\in\mathrm{T}} \sum_{\mathrm{d}} \theta_{\mathrm{td}}}{\min\limits_{\mathrm{t}\in\mathrm{T}} \sum_{\mathrm{d}} \theta_{\mathrm{td}}}$$

- Отклонение столбцов Θ от равномерного распределения:

$$\mathrm{S}_2 = \left|\Theta - \tfrac{1}{|\mathrm{T}|}\right|$$

Эксперимент: сбалансированные коллекции

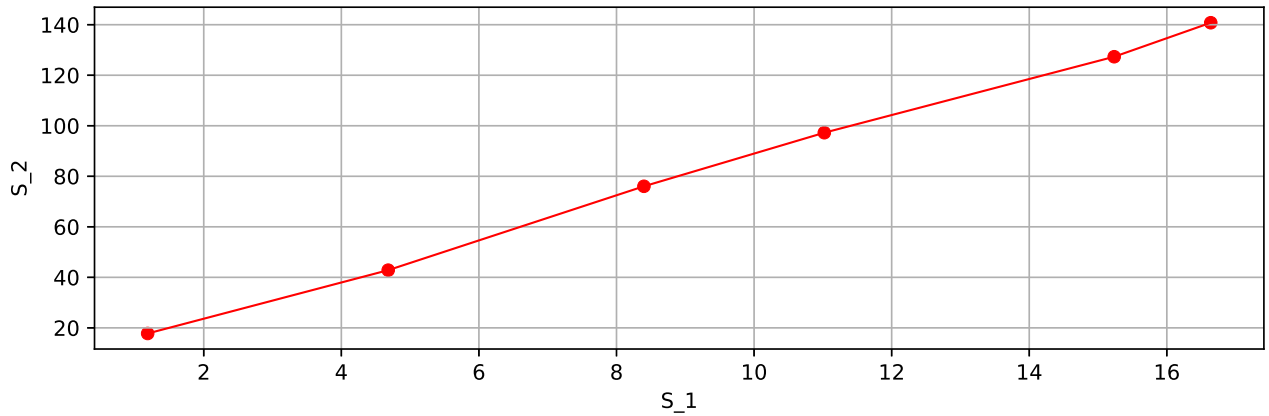
$S_1 \approx 1.2$, $S_2 \approx 20$, изменяем доли фоновых тем



- Добавление регуляризатора улучшает восстановление тем
- С ростом доли фоновой темы темы восстанавливаются хуже

Эксперимент: несбалансированные коллекции

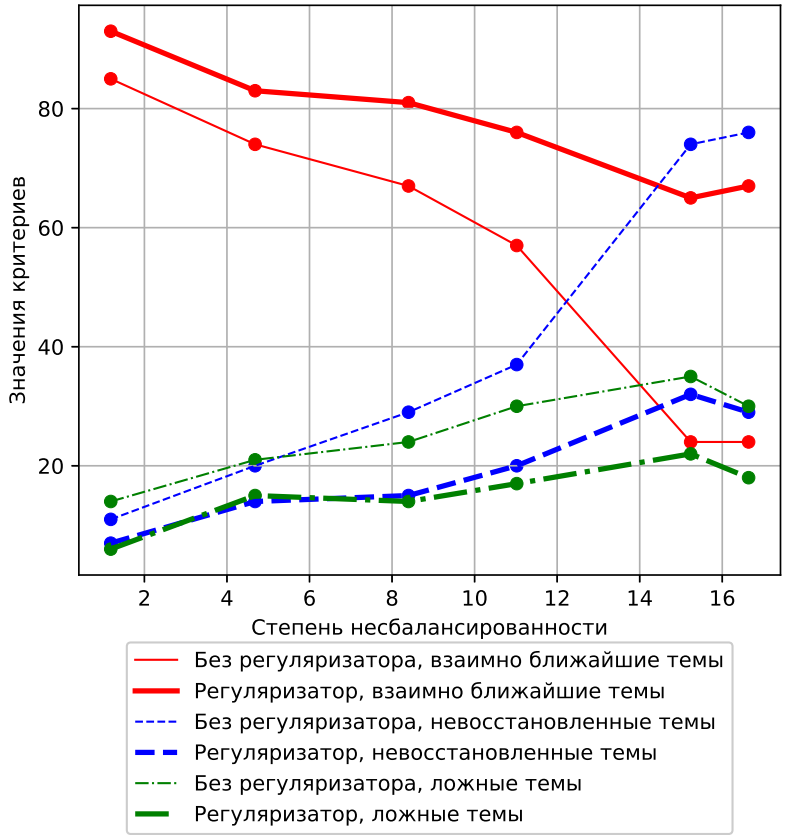
Генерируется шесть коллекций с одной темой, мощность которой во много раз превышает мощность остальных.
Степени несбалансированности шести синтетических коллекций:

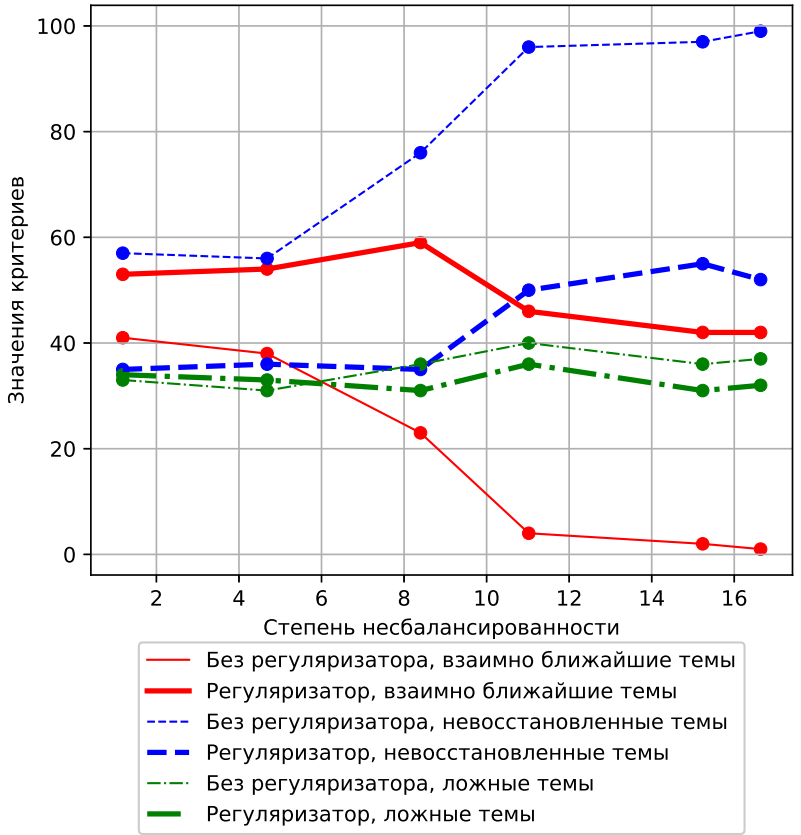


Для сбалансированных коллекций: $S_1 \approx 1.2,\ S_2 \approx 20$

Эксперимент: несбалансированные коллекции

Доля фоновой темы: 0.5 — первый эксперимент, 0.7 — второй

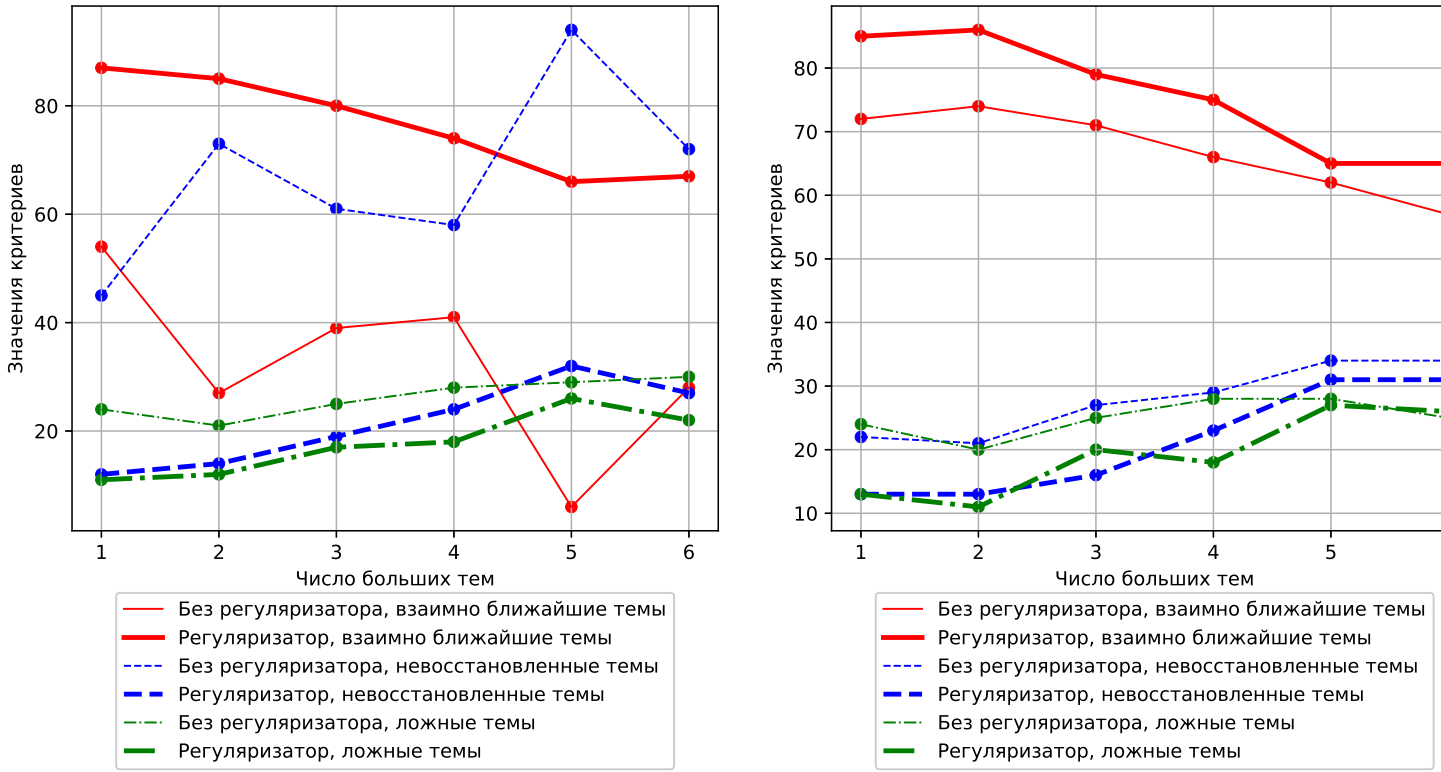




- Модель с регуляризатором — темы лучше восстанавливаются
- Темы восстанавливаются хуже при увеличении S_1, S_2
- Темы восстанавливаются хуже с ростом доли фоновой темы

Эксперимент: несбалансированные коллекции

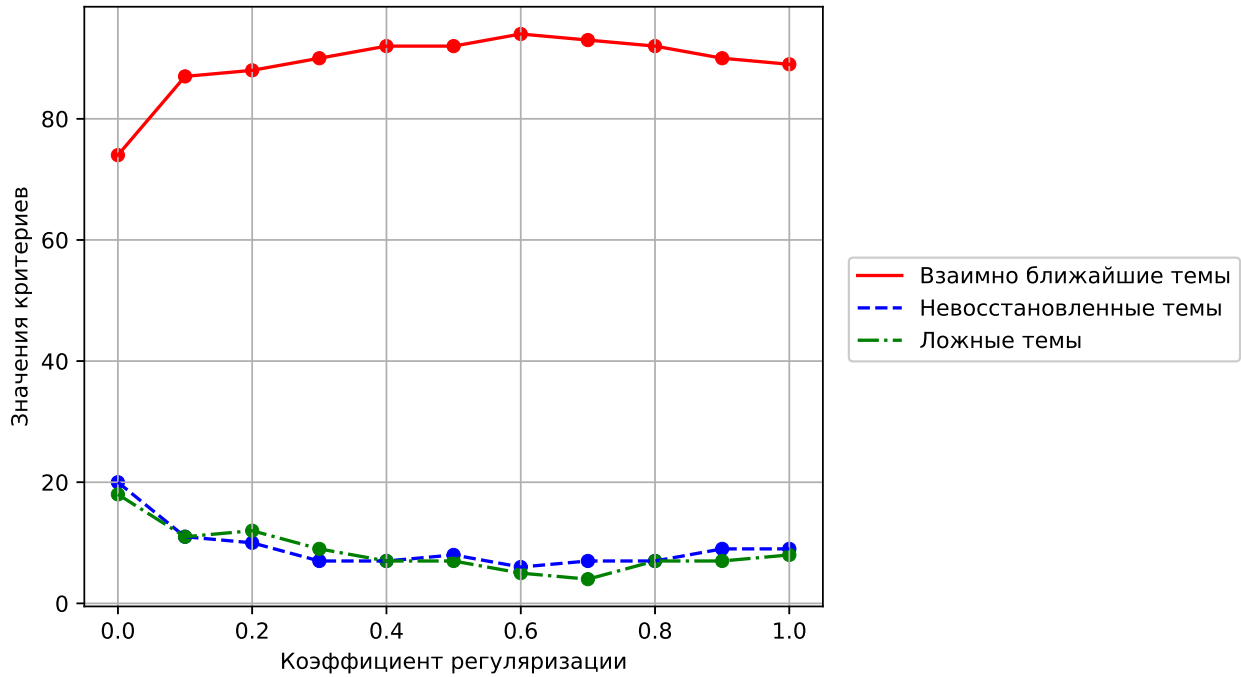
Изменяем число больших тем в коллекции
$S_1 \approx 3.5,\ S_2 \approx 25 \cdot$ (число больших тем $+1$)
Расстояния: косинусное (слева), Йенсена-Шеннона (справа)



- Расстояние Йенсена-Шеннона даёт более устойчивые результаты
- Качество восстановления тем падает с ростом S_2

Эксперимент: несбалансированные коллекции

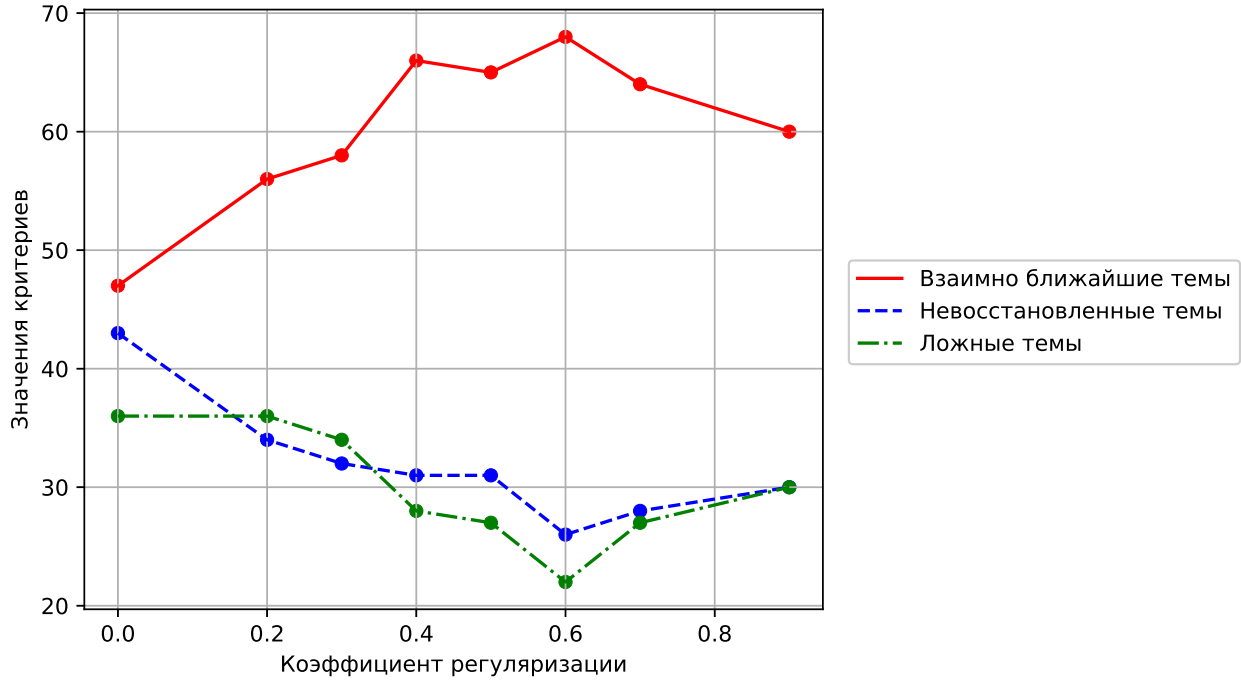
Коллекция: половина крупных, половина мелких тем.
$S_1 = 1.3$, $S_2 = 107$, изменяем коэффициент регуляризации



- Большая S_2 — не всегда несбалансированная коллекция
- Оптимум в окрестности $\tau \approx 0.6$

Эксперимент: несбалансированные коллекции

Коллекция: по $\frac{1}{20}$ крупных и мелких тем, остальные равномощны.
$S_1 = 1.65$, $S_2 = 48$, изменяем коэффицент регуляризации



- Малое значение S_2 — не всегда сбалансированная колелкция
- Оптимум в окрестности $\tau \approx 0.6$

Результаты, выносимые на защиту

- На синтетических данных показано, что тематическая несбалансированность коллекции приводит к дроблению крупных тем и слиянию мелких
- Предложен алгоритм устранения проблемы несбалансированности путем добавления регуляризатора на основе семантической неоднородности тем
- В экспериментах показано, что предложенный регуляризатор отчасти устраняет проблему несбалансированности