Вероятностное тематическое моделирование несбалансированных текстовых коллекций

Панкратов Виктор Владимирович

Московский физико-технический институт
Кафедра интеллектуальных систем

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

16/12/2023

Постановка задачи: вероятностная модель

Заданы три множества:
D - множество документов,W - множество слов,T - множество тем
Задано n_{wd} - число вхождений слова w в документ d.

Предположение о порождении коллекции

Появление слова $w \in W$ в документе $d \in D$ описывается двумя матрицами: Φ, Θ.

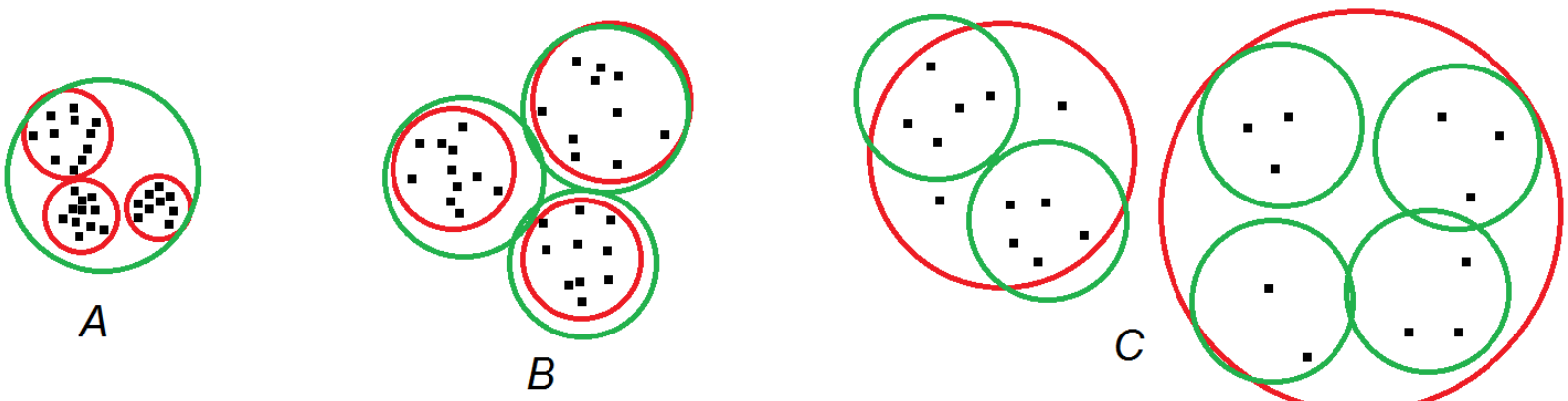
$$\phi_{wt} = p(w|t) \qquad \theta_{td} = p(t|d)$$

Задача: восстановить Φ, Θ.
Критерий качества:

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

максимизация правдоподобия, используется EM-алгоритм

Проблема несбалансированности

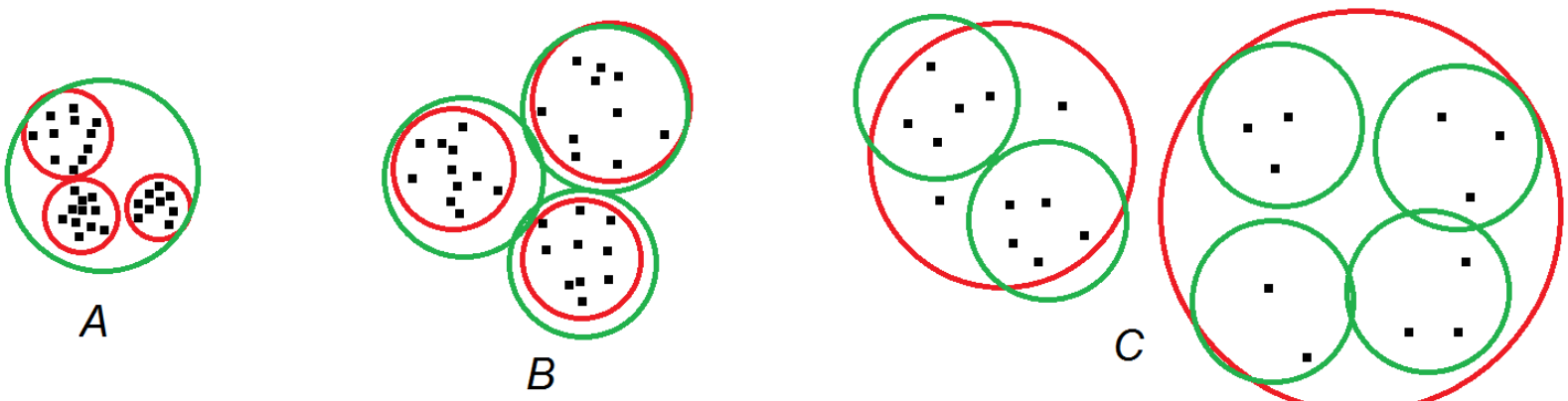


Проблема несбалансированности: максимизация правдоподобия приводит к дроблению крупных тем(A) и слиянию мелких(C).

Цель работы

Предложить и экспериментально проверить решение проблемы несбалансированности с помощью регуляризатора.

Семантическая неоднородность



Гипотеза условной независимости:

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

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.

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:

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

Используется регуляризатор, полученный из статистики семантической неоднородности:

$$\mathrm{R} = \sum_{\mathrm{d}\in\mathrm{D}}\sum_{\mathrm{w}\in\mathrm{d}}\beta_{\mathrm{dw}}\mathrm{n_{dw}}\ln\sum_{\mathrm{t}\in\mathrm{T}}\phi_{\mathrm{wt}}\theta_{\mathrm{td}},\quad \beta_{\mathrm{dw}} = \sum_{\mathrm{t}\in\mathrm{T}}\frac{\mathrm{p}(\mathrm{t}|\mathrm{d},\mathrm{w})}{\mathrm{p}(\mathrm{t})} \qquad (2)$$

Сравнение двух моделей

Пусть на одной коллекции построены две тематические модели,

- Φ_1 - матрица вероятностей p(w|t), полученная первой моделью
- Φ_2 - матрица вероятностей p(w|t), полученная второй моделью

Для всех пар i, j проверяются равенства:

$$\arg\min_{k} \ (\mathrm{dist}(\Phi_1[i], \Phi_2[k])) = j \tag{3}$$

$$\arg\min_{k} \ (\mathrm{dist}(\Phi_1[k], \Phi_2[j])) = i \tag{4}$$

Здесь dist – косинусное расстояние.
Взаимно близкие темы: (3),(4) выполнены для некоторых i, j.

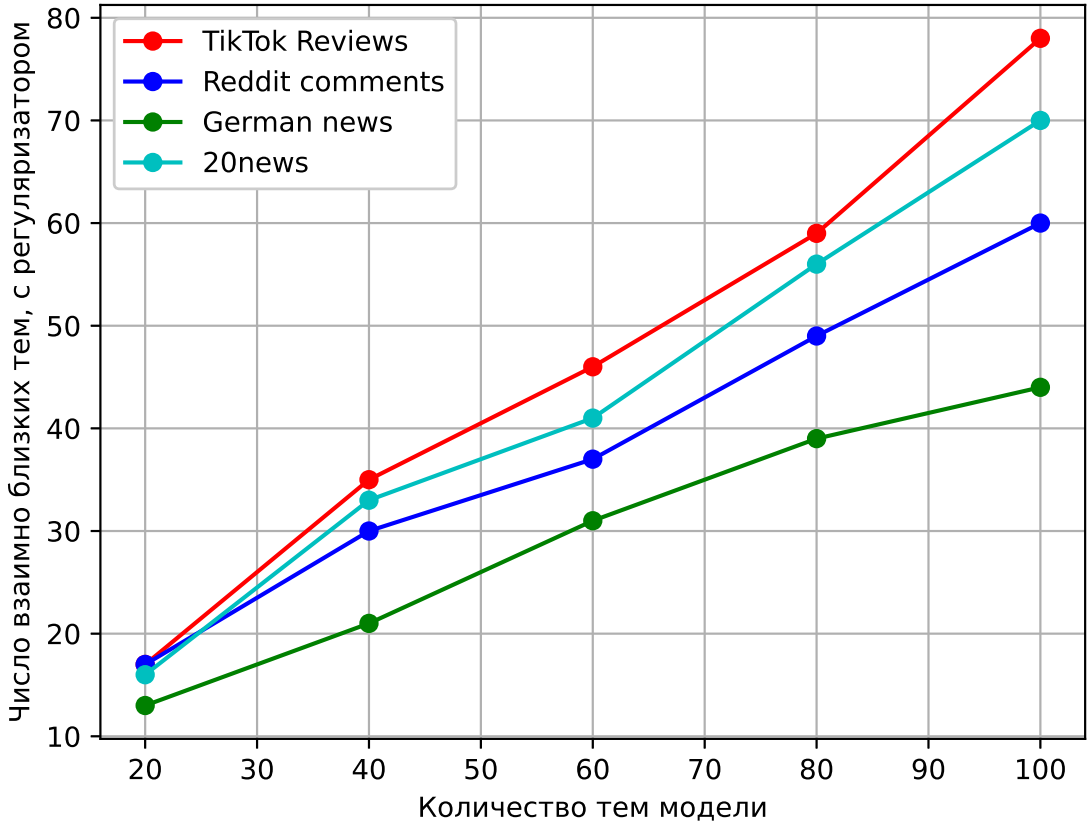
Подготовка данных

Для экспериментов использовалась коллекция 20newsgroups[1]. Она преобразовывалась согласно следующему алгоритму

- Составляется матрица n_{dw}
- Удаляются не монотематические документы. Для этого строится произвольная тематическая модель, для каждого документа d считается $t_d = \text{argmax } p(t|d)$ и проверяется $\frac{p(t_d|d)}{p(t_i|d} > 2 \; \forall t_i \neq t_d$
- Для каждого генерируемого документа выбирается его тема
- Документ генерируется как множество случайно выбранных сочетаний из 10 подряд идущих слов в исходных документах соответствующей темы

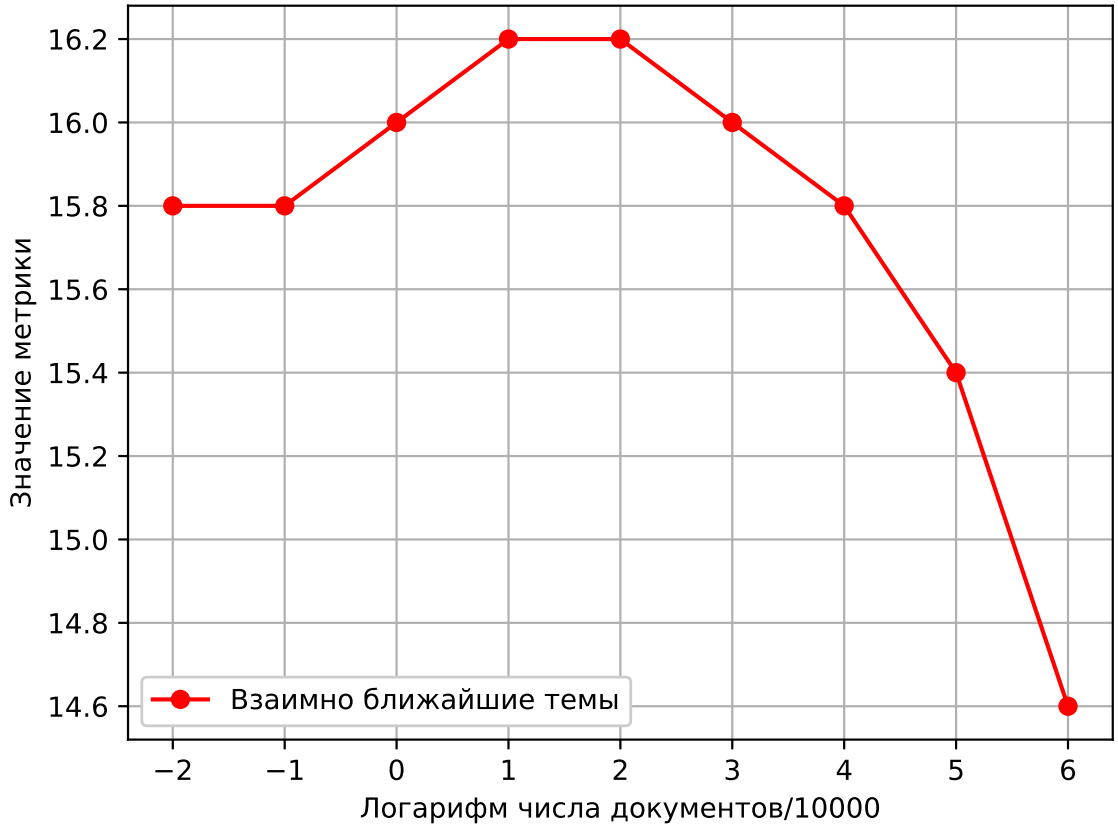
[1] http://qwone.com/ jason/20Newsgroups/

Эксперимент, другие датасеты



Полученные в предыдущем семестре результаты повторяются на других датасетах - выделить все взаимно близкие к исходным темы при подборе регуляризаторов не удается.

Эксперимент, число документов



Коэффициент регуляризации зависит от размера коллекции