

# 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.

07/06/2023

# 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:  $\Phi, \Theta$ .

$$\phi_{wt} = p(w|t)$$

$$\theta_{td} = p(t|d)$$

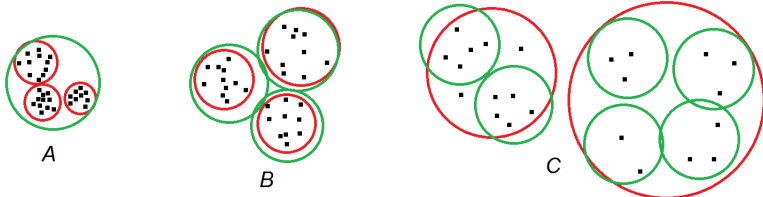
Задача: восстановить  $\Phi, \Theta$

Критерий решения:

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

максимизация правдоподобия, осуществляется ЕМ-алгоритмом

# Проблема несбалансированности

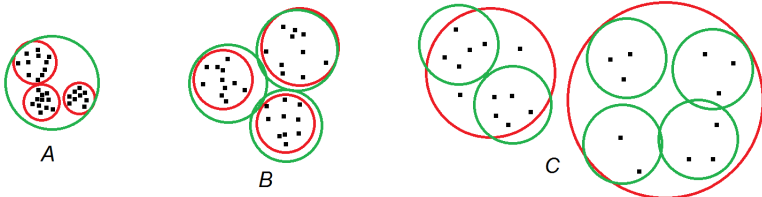


Проблема несбалансированности: максимизация правдоподобия приводит к дроблению крупных тем(A) и слиянию мелких(C)

## 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.

$$S_t = \text{KL}(p(w, d|t) || p(w|t)p(d|t))$$

Тема - кластер размерности  $|W|$ , центр которого -  $p(w|t)$

$S_t$  - удаленность  $p(w|d, t)$  от центра кластера.

# Семантическая неоднородность

Статистика семантической неоднородности

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

Здесь  $\hat{p}$  - частотные оценки вероятности

Преобразовывая и суммируя по всем темам:

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

Используется регуляризатор, полученный из статистики семантической неоднородности

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

# Сравнение двух моделей

- $\Phi_1$  - матрица вероятностей  $p(w|t)$ , полученная первой моделью
- $\Phi_2$  - матрица вероятностей  $p(w|t)$ , полученная второй моделью

Для всех пар  $i, j$  проверяются равенства:

$$\arg \min_k (\text{dist}(\Phi_1[i], \Phi_2[k])) = j \quad (3)$$

$$\arg \min_k (\text{dist}(\Phi_1[k], \Phi_2[j])) = i \quad (4)$$

Здесь  $\text{dist}$  – косинусное расстояние.

Взаимно близкие темы: (3),(4) выполнены для некоторых  $i, j$

# Эксперимент: подготовка данных

Для эксперимента использовалась коллекция 20newsgroups. Она преобразовывалась согласно следующему алгоритму

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

# Эксперимент: параметры

Степень несбалансированности коллекции:

- Отношение минимальной и максимальной сумм вероятностей:

$$\rho_{imb} = \frac{\max_{t \in T} \sum_d \theta_{td}}{\min_{t \in T} \sum_d \theta_{td}}$$

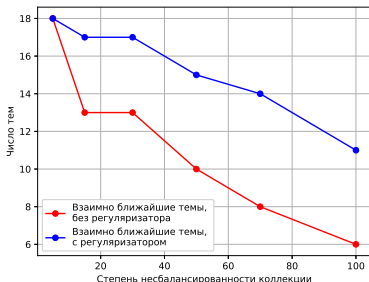
Параметры генерации коллекции на основе реальной:

- Число тем  $|T| = 20$
- Число документов  $|D| \in [2000, 5000]$
- Мощность темы - число документов, сгенерированных по этой теме
- Число слов в каждом документе  $|w \in (d \in D)| = 500$



# Эксперимент: демонстрация проблемы несбалансированности

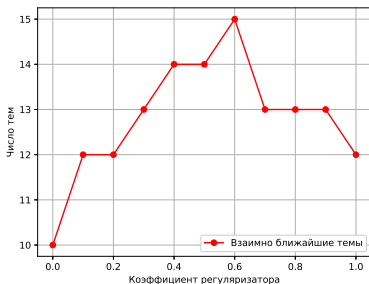
На основе 20newsgroups построены коллекции с разными  $\rho_{imb}$ .  
Мощность одной темы 3000, остальных -  $3000/\rho_{imb}$ .



При возрастании степени несбалансированности уменьшается число взаимно близких тем. Уменьшение сильнее для модели без регуляризатора

# Эксперимент: подбор коэффициента регуляризатора семантической неоднородности

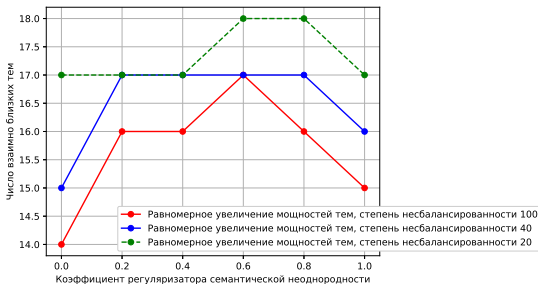
Построена зависимость для коллекции, сгенерированной на основе 20newsgroups. Мощность четырех тем - 500, остальных - 10.



Исходная модель выдала 10 тем взаимно близких к искомому. Подбор коэффициента регуляризации улучшил соответствие найденных и искомых тем.

# Эксперимент: подбор коэффициента регуляризатора семантической неоднородности

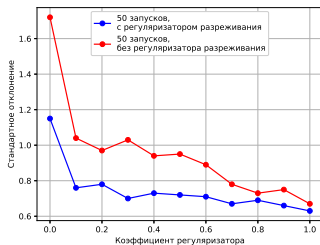
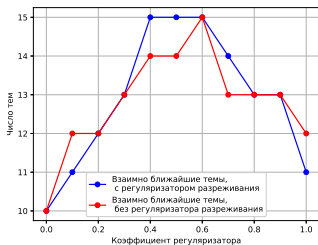
Построена зависимость для коллекций, сгенерированной на основе 20newsgroups. Мощность темы  $i$  равна  $10 \cdot (k_i + 1)$ ,  $k \in \{1, 2, 5\}$



Наибольшее число взаимно близких тем в каждом эксперименте было для модели с регуляризатором с коэффициентом 0.6

# Эксперимент: добавление регуляризатора разреживания

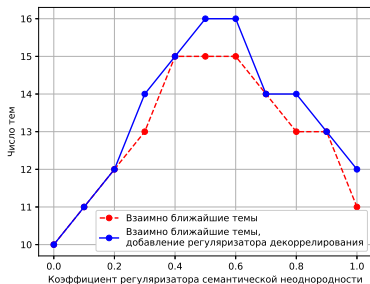
Добавим для первого эксперимента регуляризатор: каждые 10 итераций обнуляются 0.1 от всех элементов каждой строки  $\Phi$  и каждого столбца  $\Theta$ . Не обнуляются элементы больше  $1/10000$ . Проводится 50 экспериментов



Стандартное отклонение результатов становится меньше, но результаты не улучшаются

# Эксперимент: добавление регуляризатора декоррелирования

В модель добавлен регуляризатор декоррелирования ( $\tau = 1$ )



Число взаимно близких тем при значении коэффициента 0.6 является одним из максимумов, как и в прошлых экспериментах. При таком значении результаты улучшились.

# Эксперимент: слияние и расщепление тем

Построена тематическая модель на коллекции 20newsgroups. 20 тем: оставлены 10 документов. 20 тем: без изменений

Построено две модели - без регуляризатора семантической неоднородности и с добавлением регуляризатора ( $\tau = 0.6$ )

Исходная модель	Модель без регуляризатора	Модель с регуляризатором
void include null int code	application window use get problem code	function include void return null
menu mouse button application window	application user window server program	button menu user command server
firearm legal citizen law crime	wrong society person life crime	murder person crime moral society
wrong society orientation partner relationship life	wrong society person life crime	orientation wife partner relationship marriage

Слияние тем - 2 исходные темы соответствуют одной полученной моделью при обработке без регуляризатора. Добавление регуляризатора устранило проблему

# Эксперимент: слияние и расщепление тем

Построена тематическая модель на коллекции 20newsgroups. 20 тем: оставлены 10 документов. 20 тем: без изменений

Построено две модели - без регуляризатора семантической неоднородности и с добавлением регуляризатора ( $\tau = 0.6$ )

Исходная модель	Модель без регуляризатора	Модель с регуляризатором
size frame pixel quality image	user graphic support software animation	animation interactive graphic draw image
size frame pixel quality image	display image library animation screen	animation interactive graphic draw image
spirit soul eternal life heaven	death sin church doctrine god	soul holy god divine revelation
spirit soul eternal life heaven	death faith heaven sin believe	soul holy god divine revelation

Расщепление тем - исходная тема соответствует двум полученным моделью при обработке без регуляризатора. Добавление регуляризатора устранило проблему.

# Заключение

- Показано, что тематическая несбалансированность коллекции приводит к дроблению крупных и слиянию мелких тем
- Предложен алгоритм устранения проблемы несбалансированности, заключающийся в добавлении регуляризации на основе семантической однородности тем
- Проведены эксперименты, демонстрирующие возможные модификации модели и показывающие, что модель улучшает интерпретируемость результатов обработки несбалансированных коллекций