

# 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.

18/05/2023

# Постановка задачи: вероятностная модель

Заданы три множества:

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

Для каждого  $w \in W$ ,  $d \in D$  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  $\Phi, \Theta$

# Проблема несбалансированности

Ставится оптимизационная задача максимизации функции правдоподобия

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

Задача (1) некорректно поставлена: множество ее решений бесконечно. Чтобы его уменьшить, применяют регуляризацию

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta} \quad (2)$$

Задача (1), (2) решается с помощью ЕМ алгоритма. Это приводит к дроблению крупных тем и слиянию мелких, что получило название проблема несбалансированности.

Цель работы: предложить и экспериментально проверить на реальных данных решение проблемы несбалансированности с помощью регуляризатора

# Семантическая неоднородность

Гипотеза условной независимости :

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

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.

# 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

$$R = \sum_{d \in D} \sum_{w \in d} \beta_{dw} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}, \quad \beta_{dw} = \sum_{t \in T} \frac{p(t|d, w)}{p(t)} \quad (3)$$

# Эксперимент: оценивание качества решения

Опишем способ оценивания качества восстановления матриц

- $\Phi_0$  - исходная матрица вероятностей  $p(w|t)$
- $\Phi$  - матрица вероятностей  $p(w|t)$ , найденная алгоритмом

Для всех пар  $i, j$  будем проверять равенства:

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

$$\arg \min_k (\text{dist}(\Phi[k], \Phi_0[j])) = i \quad (5)$$

Здесь  $\text{dist}$  – произвольная метрика, в работе подсчитывалось несколько вариантов.

Взаимно близкие темы: (4),(5) выполнены для некоторых  $i, j$

Невосстановленная тема:  $\Phi_0[j]$ , если (4) не выполнено для всех  $i$

Ложная тема:  $\Phi[i]$ , если (5) не выполнено для всех  $j$

# Эксперимент: данные

Для эксперимента использовались реальные коллекции новостей. Основные - BBC(BBC), ted talks(TED), описаний книг(books) и 20newsgroups(news). Ниже приводится статистика по каждой коллекции:

**BBC** 1926 документов, 10924 слов в коллекции. Среднее число слов документа 141

**TED** 3000 документов, 19608 слов в коллекции. Среднее число слов документа 211

**BOOKS** 4825 документов, 27902 слов в коллекции. Среднее число слов документа 505

**NEWS** 18846 документов, 30511 слов в коллекции. Среднее число слов документа 59

# Эксперимент: подготовка данных

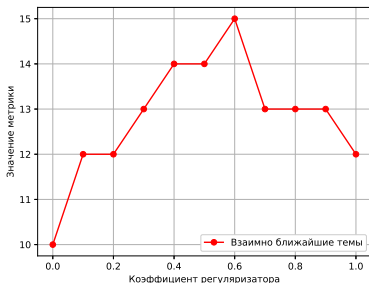
Данные из описанных ранее коллекций преобразуются согласно следующему алгоритму

- Составляется матрица  $n_{dw}$
- Удаляются не монотематические документы. Для этого строится произвольная тематическая модель, для каждого документа  $d$  считается  $t_d = \operatorname{argmax} p(t|d)$  и проверяется  $\frac{p(t_d|d)}{p(t_i|d)} > 2 \forall t_i \neq t_d$
- Удаляются редко встречающиеся слова:  $w : \sum_d n_{dw} < C$ ,  $C \in \mathbb{N}$  – параметр эксперимента
- Для каждого генерируемого документа выбирается его тема
- Документ генерирует как множество случайно выбранных сочетаний из  $k$  подряд идущих слов в исходных документах соответствующей темы



# Эксперимент: демонстрация проблемы и решения с помощью регуляризатора

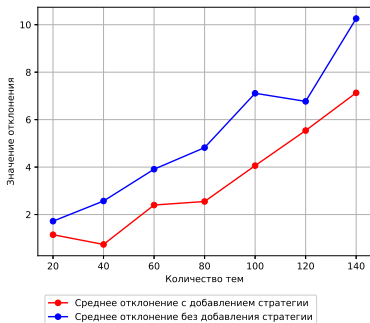
Построена зависимость числа взаимно близких тем от коэффициента регуляризатора семантической неоднородности.



Исходная модель смогла выделить только 10 тем взаимно близких к искомому. В то же время подбор коэффициента регуляризации значительно улучшил соответствие найденных и исходных тем.

# Эксперимент: добавление стратегии разреживания

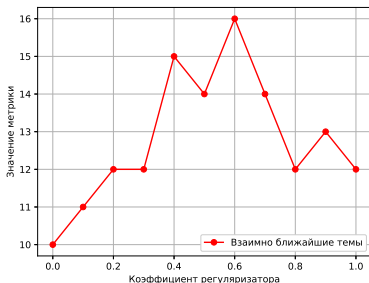
Включим в модель стратегию разреживания на основе работы Глушаченкова. Её эффект оценен на примере отклонений между результатами работы моделей на коллекции новостей ВВС



Результаты при добавлении стратегии получаются более устойчивыми

# Эксперимент: добавление регуляризатора декоррелирования

Проведем эксперимент аналогичный первому с добавлением регуляризатора декоррелирования



Добавление регуляризатора декоррелирования незначительно улучшило результаты

# Заключение

- На примере коллекции 20newsgroups и сгенерированных на её основе коллекций показано, что тематическая несбалансированность коллекции приводит к дроблению крупных и слиянию мелких тем
- Предложен алгоритм устранения проблемы несбалансированности путем добавления регуляризации на основе семантической однородности тем
- Проведены эксперименты, демонстрирующие возможные модификации модели и показывающие, что модель устраняет проблему несбалансированности