

# 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.

17/05/2024

# 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:  $\Phi, \Theta$ .

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

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

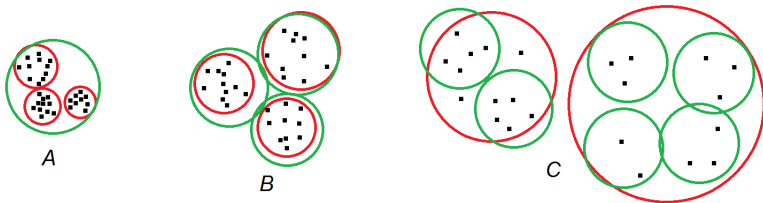
Задача: восстановить  $\Phi, \Theta$ .

Критерий качества:

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

максимизация правдоподобия, используется ЕМ-алгоритм

# Проблема несбалансированности

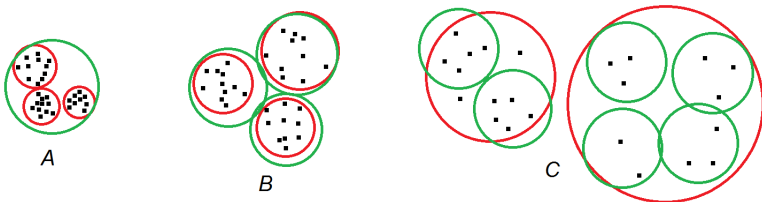


Проблема несбалансированности: максимизация правдоподобия приводит к дроблению крупных тем(A) и слиянию мелких(C).

## 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.

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

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

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

# Регуляризатор семантической неоднородности

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

Пусть на одной коллекции построены две тематические модели,

- $\Phi_1$  - 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  $i, j$ .

# Подготовка данных

Для экспериментов использовалась коллекция 20newsgroups<sup>1</sup>. 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  $\frac{p(t_d|d)}{p(t_i|d)} > 2 \forall t_i \neq t_d$
- Для каждого генерируемого документа выбирается его тема
- Документ генерируется как множество случайно выбранных сочетаний из 10 подряд идущих слов в исходных документах соответствующей темы

---

<sup>1</sup><http://qwone.com/~jason/20Newsgroups/>

# Эксперимент

