

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.

18/05/2022

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: Φ, Θ .

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

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

Задача: восстановить Φ, Θ

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

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

Для всех пар i, j будем проверять равенства:

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

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

Здесь 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

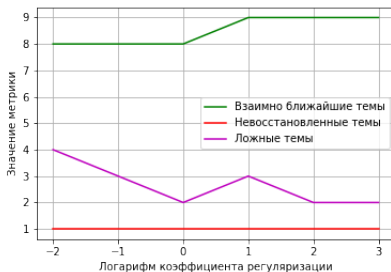
Эксперимент: данные

Данные для эксперимента генерируются на основе реальной коллекции 20newsgroups:

- Составляется матрицу n_{dw}
- Удаляются редко встречающиеся слова: $w : \sum_d n_{dw} < C$,
 $C \in \mathbb{N}$ – параметр эксперимента
- Произвольным образом выбирается подмножество документов, образующее несбалансированную коллекцию и запоминаются их темы
- Удаляются слова, встречающиеся в не более чем одном документе

Эксперимент: коэффициент регуляризации

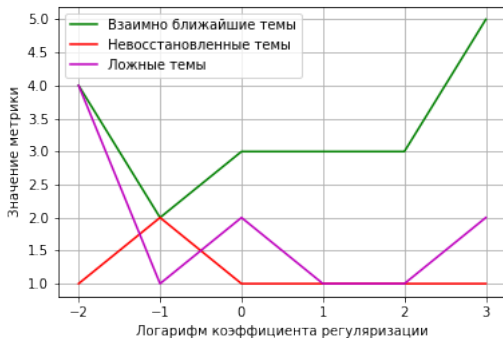
Изучим зависимость числа восстановленных тем от коэффициента регуляризации. В данном эксперименте восстанавливалось 10 тем:



Как и в случае синтетических коллекций оптимальное τ принадлежит $[0.5, 1]$

Эксперимент: коэффициент регуляризации

Повторим предыдущий эксперимент с метриками, вычисленными по расстоянию Йенсена-Шеннона:



Эксперимент: коэффициент регуляризации

Рассчитаем среднюю долю восстановленных документов для каждой темы и убедимся еще раз в выводе об оптимальном τ :



Эксперимент: восстановление разного числа тем

Проведем эксперимент по восстановлению нескольких тем из 20 исходных и посчитаем метрики в зависимости от их количества:

