Итеративное улучшение тематической модели с обратной связью от пользователя

Алексей Ильич Горбулев

Московский физико-технический институт

Курс: Моя первая научная статья/Группа Б05-021а

Эксперт: д. ф.-м. н. К. В. Воронцов

Консультант: В. А. Алексеев

2023

Цель исследования

Мотивация: тематические модели неустойчивы, неполны
Цель исследования: получить *интерпретируемую* тематическую модель за некоторое число итераций
Метод: итеративное улучшение тематической модели с использованием регуляризаторов пользовательской разметки тем на релевантные, нерелевантные и «мусорные»

Литература

- Alekseev V. et al. "TopicBank: Collection of coherent topics using multiple model training with their further use for topic model validation"
- Victor Bulatov, Evgeny Egorov, Eugenia Veselova, Darya Polyudova, Vasiliy Alekseev, Alexey Goncharov, Konstantin Vorontsov. “TopicNet: Making Additive Regularisation for Topic Modelling Accessible”
- Воронцов К.В. "Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект BigARTM"

О тематическом моделировании и ARTM

- D — множество (коллекция) документов
- W — множество термов

  Термами могут быть слова в нормальной форме, словосочетания.

  Каждый документ $d \in D$ представляет собой последовательность термов.
- T — множество тем

  Как правило, количество тем $|T|$ заранее задано.

При построении вероятностной тематической модели

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d) - \sum_{t \in T} \varphi_{wt} \theta_{td}$$

в подходе ARTM происходит максимизация log правдоподобия с k регуляризаторами R_i:

$$\sum_{d,w} n_{dw} \ln \sum_{t} \varphi_{wt} \theta_{td} + \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

О тематическом моделировании и ARTM

Далее применяется EM-алгоритм.

- E-шаг:

$$p_{tdw} = \underset{t \in T}{\text{norm}}(\varphi_{wt}\theta_{td})$$

- M-шаг:

$$\varphi_{wt} = \underset{w \in W}{\text{norm}}(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}})$$

$$\theta_{td} = \underset{t \in T}{\text{norm}}(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}})$$

Постановка задачи

Пусть D — коллекция документов, количество тем $|T|$ задано заранее.
После обучения базовой тематической модели M_0 каждая из тем $t \in T$ отнесена пользователем в одну из трёх категорий:

- T_+ (релевантные, *имеющие отношение к исследовани.*)
- T_0 (нерелевантные, *дублирующие релевантные*)
- T_- («мусорные», *не имеющие отношение к исследованию*)

После обучения новой тематической модели M_1 $|T_+|$ должно увеличиться, и должно быть сохранено как можно больше тем из T_+, а $|T_-|$ должно уменьшиться. Процесс продолжается итеративно.

Решение

На каждой итерации:

- С помощью регуляризатора сглаживания

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T_0} \sum_{w \in W} \beta_{wt} \ln \varphi_{wt} + \alpha_0 \sum_{d \in D_0} \sum_{t \in T} \alpha_{td} \ln \theta_{td}$$

  зафиксировать столбцы матрицы Φ, соответствующие релевантным темам, используя с достаточно большим коэффициентом

- Для выявления новых релевантных тем использовать регуляризатор декоррелирования, действующий на столбцы матрицы Φ, соответствующих темам из T_-:

$$R(\Phi_{new}) = -\frac{\tau}{2} \sum_{t \in (T_-^i)} \sum_{s \in T_-^i \setminus t} \sum_{w \in W} \varphi_{wt} \varphi_{ws}$$

Вычислительный эксперимент

В качестве коллекции текстов используется набор из 16 449 новостей, опубликованных на сайте Lenta.ru в период с мая по август 2008-го года. Предполагается разделение на 100 тем.

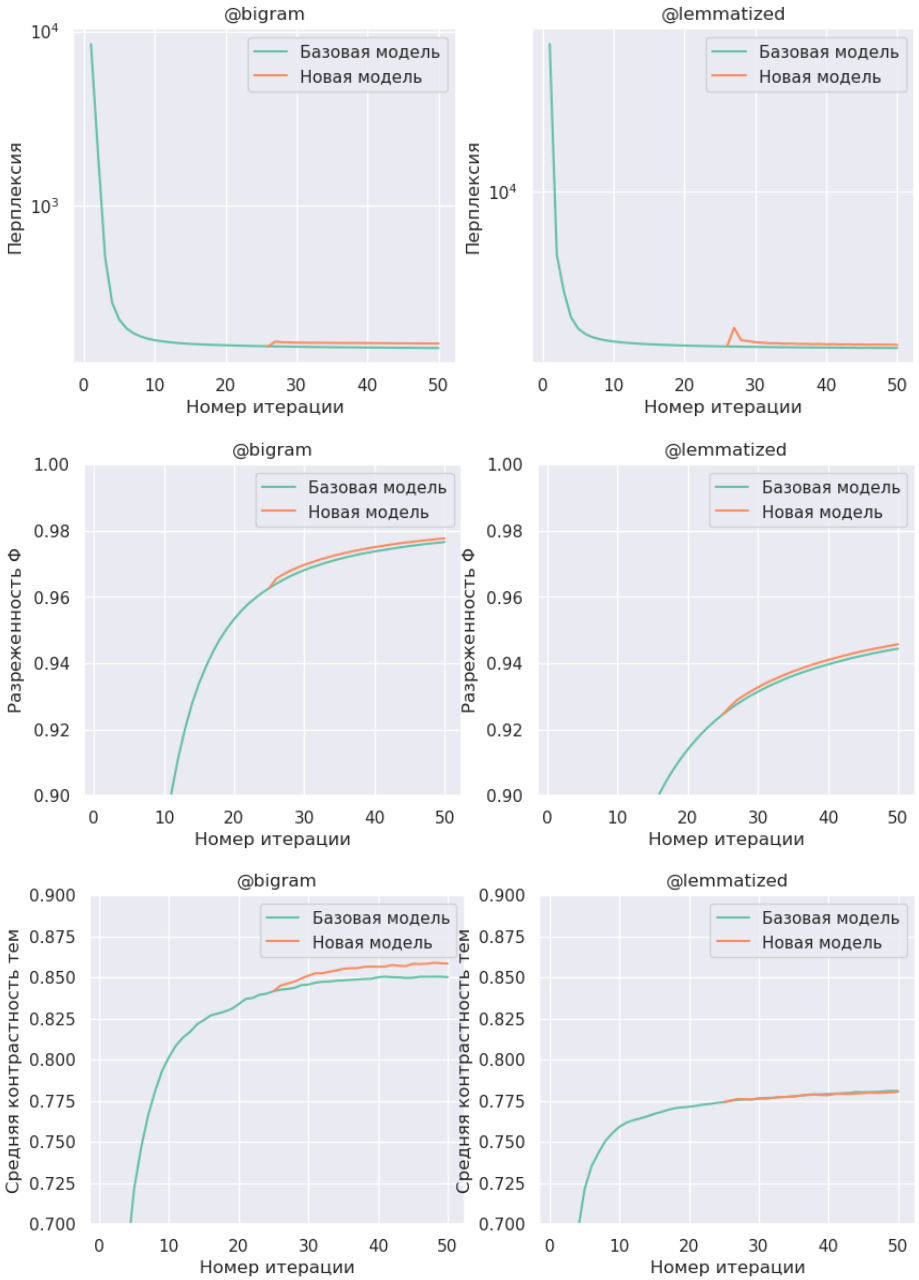
Предобработка:

- Заголовок и текст каждой новости разбиваются на токены, далее происходит лемматизация.
- Далее по PMI отбирается 10 000 биграмм, которые характеризуют коллекцию текстов.

Базовая модель: TopicNet, 100 тем, без регуляризаторов на темы

Новая модель: TopicNet, 100 тем, регуляризатор сглаживания для тем T_+ с $\tau = 10^{10}$, регуляризатор декоррелирования для тем из T_-, а также регуляризатор улучшения когерентности и битермов

Вычислительный эксперимент



Критерии качества:
перплексия, разреженность Φ, средняя контрастность тем, размеры групп T_+ и T_-.

На базовой модели удалось выделить одну релевантную тему, все остальные оказались «мусорными».

На данный момент работы продолжаются, пока после обучения новой модели разметка осталась той же.

Заключение

Результаты:

- предложен метод итеративного улучшения тематической модели,
- показано, как использовать регулятор сглаживания для сохранения тем из T_+,
- предложен регуляризатор декоррелирования тем из T_-.