Автоматическое выделение терминов для тематического моделирования

Никитина Мария
nikitina.mariia@phystech.edu

1 марта 2023 г.

Аннотация

В настоящее время каждый день появляются новые научные термины. Необходимо научиться находить их в коллекции документов. Делать это вручную долго и дорого, потому что нужно привлекать узкоспециализированных специалистов. В данной статье рассматриваются два решения этой проблемы: метод выделения коллокаций (TopMine) в сочетании с модульной технологией тематического моделирования (с использованием библиотеки BigARTM) и современные методы, основанные на нейросетевых моделях языка. Эти два решения ранее не сравнивались.

Ключевые слова: тематическое моделирование · TopMine · BigARTM · выделение коллокаций

1 Введение

Тематическое моделирование – это технология обработки естественного языка, направленная на определение тем, к которым относится текстовый документ из коллекции, и какие слова каждую тему образуют. Иначе говоря, тематическая модель осуществляет *мягкую кластеризацию*, выбирая для документа кластеры-темы.

Вероятностная тематическая модель определяет вероятности тем в каждом документе и вероятности слов в каждой теме. Большим отличием такой модели от глубоких нейронных сетей типа BERT [2] или GPT-3 [1] является простота организации и свойство интерпретируемости в ущерб качеству предсказания вероятности появления слов в документе. Векторное представление тяжёлой нейросети всё ещё не удалось интерпретировать, в то время как тематический эмбединг – это вектор вероятностей тем.

Основной задачей и новизной данной статьи является сравнение этих двух подходов. Нейросеть использована готовая. Для построения же тематической модели требуется составление *регуляризаторов* – критериев, учитывающих специфические особенности

данных или предметной области, от подбора которых значительно зависит качество определения основных тем документов (*Аддитивная регуляризация тематических моделей, ARTM*). К основному алгоритму регуляризаторы подключаются как модули с помощью библиотеки BigARTM [3] с открытым кодом.

Перед выполнением кластеризации необходимо выделить из коллекции документов ключевые слова и словосочетания и отбросить те, что не несут основной смысловой нагрузки. Поиск составных терминов является нетривиальной и трудоёмкой задачей. Для её решения используется метод поиска коллокаций TopMine, использующий информацию о частоте и совстречаемости слов в коллекции [5].

С учётом интерпретируемости и простоты тематическая модель может стать хорошей заменой нейросети. Предшествующие исследования предлагаемого подхода показали хорошие результаты как по полноте, так и по вычислительной эффективности. Однако они до сих пор не сравнивались с нейросетевыми моделями. Важно понять, насколько хорошо тематическая модель выполняет рассчитанную под неё задачу по сравнению со сложной нейронной сетью.

2 Постановка задачи

Задача называется *корректно поставленной* по Адамару, если её решение существует, единственно и устойчиво. В общем случае построение тематической модели – некорректно поставленная задача, её нужно дополнить регуляризаторами. В практических задачах автоматической обработки текстов существует очень много критериев и ограничений.

Пусть $p_{\omega d}$ – вероятность появления терма ω в документе d, $\varphi_{\omega t}$ – вероятность того, что терм ω относится к теме t, θ_{td} – вероятность встречи темы t в документе d. Тогда $P = (p_{\omega d})_{W \times D}$ – матрица частот термов в документах, $\Phi = (\varphi_{\omega t})_{W \times T}$ – матрица термов тем, $\Theta = (\theta_{td})_{T \times D}$ – матрица тем документов. W, D, T – множества всех термов, документов и тем соответственно.

Аддитивная регуляризация тематических моделей основана на максимизации логарифма правдоподобия и регуляризаторов $R_i(\Phi, \Theta)$ с неотрицательными коэффициентами регуляризации τ_i, $i = 1, ..., k$ [4]:

$$\sum_{d \in D} \sum_{w \in d} \ln \sum_{t \in T} \varphi_{\omega t} \theta_{td} + R(\Phi, \Theta) \to \max_{\Phi, \Theta}; \qquad R(\Phi, \Theta) = \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta); \tag{1}$$

при ограничениях неотрицательности и нормировки:

$$\sum_{w \in W} \varphi_{\omega t} = 1; \quad \varphi_{\omega t} \geqslant 0; \qquad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geqslant 0. \tag{2}$$

Решается задача нахождения разложения $P = \Phi\Theta$ при достижении максимума. Для её решения применяется EM-алгоритм. Таким образом, основной проблемой становится поиск регуляризаторов $R_i(\Phi, \Theta)$, подходящих под нашу задачу поиска терминов в коллекции документов.

3 Вычислительный эксперимент

4 Анализ ошибки

5 Заключение

Список литературы

[1] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *ArXiv*, abs/2005.14165, 2020.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018.

[3] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. BigARTM: Open source library for regularized multimodal topic modeling of large collections. In *Communications in Computer and Information Science*, pages 370–381. Springer International Publishing, 2015.

[4] Воронцов К.В. Вероятностное тематическое моделирование: теория, модели, алгоритмы и проект bigartm. 2020.

[5] Николай Шаталов. Методы обучения без учителя для автоматического выделения составных терминов в текстовых коллекциях. 2019.