Аддитивная регуляризация тематических моделей в задаче анализа этносоциального дискурса

Апишев Мурат Азаматович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: great-mel@yandex.ru

Тематическое моделирование является мощным инструментом статистического анализа текстов. Оно основано на приближённом представлении матрицы частот слов в документах в виде произведения двух матриц: матрицы Φ вероятностей слов в темах и матрицы Θ вероятностей тем в документах. Темы являются скрытыми переменными, которые оцениваются в процессе обучения модели. Круг задач, решаемых с помощью тематического моделирования, широк, и включает в себя информационный поиск, создание рекомендательных систем, анализ данных новостных потоков и социальных сетей.

Аддитивная регуляризация тематических моделей (АРТМ) [1] позволяет вводить любое число дополнительных требований к модели, комбинируя их с помощью взвешенной суммы регуляризаторов.

В описываемой работе рассматривается приложение АРТМ к задаче выявления этносоциального дискурса в данных социальных сетей, т.е. выделения тем, связанных с обсуждением национальностей и смежных вопросов.

Основным инструментом АРТМ в данной задаче является регуляризатор для частичного обучения, позволяющий учесть экспертную информацию, представленную словарём этнонимов, т.е. терминов, характеризующих искомую тематику. Множество тем модели разбивается на две группы: предметные и фоновые. Все предметные темы в матрице Φ сглаживаются по содержимому словаря этнонимов, все фоновые разреживаются по тем же словам. Такая регуляризация поощряет появление этничных тем среди предметных и прочих тем — среди фоновых. Для повышения разнообразия среди предметных тем к ним применяется разреживающий регуляризатор декорреляции тем. Фоновые темы при этом слабо равномерно сглаживаются, что позволяет регуляризатору декорреляции «уводить» в них не-этнонимы, способствуя получению более качественных этничных тем.

Качество модели также можно повысить, введя регуляризацию матрицы Θ. Для этого все строки матрицы, соответствующие пред-

метным темам, равномерно сглаживаются, а строки, соответствующие фоновым — разреживаются.

Следующая модификация связана с использованием мультимодальных моделей. Модальностями в текстах являются виды слов: теги, имена авторов, метки категорий и т.п. Каждой вводимой модальности соответствует своя матрица вероятностей слов в темах Φ_m. В указанной выше модели APTM в качестве дополнительной модальности продублированы этнонимы из словаря. К матрице Φ_m также применяется собственный регуляризатор декорреляции тем.

Все эксперименты производились на коллекции постов LiveJournal (1.36 млн. документов). Метриками качества являлись tf-idf когерентность [2] (мера, отражающая интерпретируемость) и экспертные оценки социологов. Были настроены пять моделей: базовая модель без регуляризации PLSA, модель с равномерным сглаживанием LDA, модель с регуляризатором частичного обучения и регуляризацией Θ, та же модель с декорреляцией и модель со всеми описанными модификациями. Последняя модель оказалась наилучшей по всем характеристикам, т.е. находящей наибольшее количество разнообразных и интерпретируемых этничных тем.

Эксперименты были проведены с использованием библиотеки BigARTM [3], которая является на текущий момент самым эффективным инструментом для тематического моделирования, поддерживающим APTM и мультимодальные тематически модели [4].

Работа поддержана грантом РНФ н. 15-18-00091.

Литература

1. Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization // AIST'2014, Analysis of Images, Social networks and Texts. — Springer International Publishing Switzerland, 2014. Communications in Computer and Information Science (CCIS). Vol. 436. pp. 29–46.
2. S. I. Nikolenko, O. Koltsova, and S. Koltsov. Topic modelling for qualitative studies. Journal of Information Science, 2015.
3. Страница библиотеки BigARTM: `http://bigartm.org`
4. K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Suvorova, and A. Yanina. Non-bayesian additive regularization for multimodal topic modeling of large collections. In Proceedings of the 2015 Workshop on Topic Models: Post-Processing and Applications, TM '15, pages 29–37, New York, NY, USA, 2015. ACM.