# Технический отчет "Тематические модели для выделения этнорелевантных тем в социальных сетях"

***И. В. Ефимова***
efimova@phystech.edu
Московский физико-технический институт, ФУПМ, Кафедра интеллектуальных систем

Известные вероятностные алгоритмы тематического моделирования не предназначены для работы с короткими документами, наиболее частыми в социальных медиа. С другой стороны, тематические модели дают интерпретируемые матричные представления слов и документов. Векторные модели, опирающиеся на локальные контексты слов, исследуют их семантическую близость, и координаты векторных представлений слов тематически не интерпретируемы. В работе предлагается вероятностная тематическая модель, основанная на идее локально-контекстного подхода.

**Ключевые слова**: *вероятностное тематическое моделирование; дистрибутивная семантика; аддитивная регуляризация; социальные сети; лемматизация*

## 1 Проблема коротких текстов

В предшествующих работах для решения задачи поиска этнорелевантных тем использовались методы вероятностного тематического моделирования с частичным обучением, в основе которых лежит модель LDA (Latent Dirichlet Allocation) [1]. В [2] показано, что использование комбинация из семи регуляризаторов разрежевания, сглаживания и декоррелирования в рамках теории ARTM (Additive Regularization of Topic Models) позволяет находить больше хорошо интерпретируемых релевантных тем при настройке параметров регуляризации.

Особенностью исследуемой коллекции является малая длина документов, если считать каждый пост отдельным документом. Для вероятностных тематических моделей это проблема. На данный момент существует множество подходов к её решению [3–5].

Однако наиболее лингвистически обоснованным подходом к проблеме коротких текстов представляется использование векторных моделей дистрибутивной семантики [6]. Эти модели используют только локальные контексты слов, поэтому длина сообщений для них не так важна, как для тематических моделей. Модели векторных представлений слов (word embeddings) и тематические модели (topic models)имеют схожие математические постаноки и сводятся к построению матричных разложений [7]. Но координаты векторных представлений слов не имеют тематиечской интерпретации. Поэтому в данной работе предлагается тематическая модель, объединяющая достоинства двух подходов к статистическому моделированию семантики. Она позволяет использовать регуляризаторы для частичного обучения и выделения этнорелевантных тем, одновременно решая проблему коротких сообщений.

Эксперименты проводились на данных социальной сети Вконтакте с использованием библиотеки bigARTM. Построенная модель сравнивалась с PLSA по следующим критериям: разреженность, средняя контрастность тем ядра , средняя чистота тем ядра и когерентность. В результате получили, что использование нового подхода улучшает качество.

## 2 Выкачка этно-обогащенного контента

При выкачке данных случайных пользователей и групп из социальной сети Вконтакте возникла проблема наличия небольшого количества этничного контента (1%), что затрудняет построение тематических моделей для выявления этничных тем. Для решения этой

проблемы на первоначальном этапе были выкачаны записи групп, содержащие этнонимы в метаописании. Далее для поиска этно-релевантных групп планируется использовать тематические модели.

## 3 Пополнение словаря этнонимов

Одной из целей работы является разработка технологии мониторинга этно-релевантных тем. Для мониторинга необходим инструмент, который будет всё время пополнять ранжированный словарь слов, подозрительных на этничность. Поэтому с помощью тематических моделей планируется пополнять словарь этнонимов новыми словами, которые встречаются рядом с этнонимами или попадают в топы этничных тем или похожи на этнонимы по контекстам.

## 4 Лемматизация

Лемматизация — одна из важнейших задач при обработке текстов, от которой сильно зависит результат работы тематических моделей. В особенности актуальна проблема омонимии. Планируется сделать анализ качества и скорости лемматизаторов русского текста Pymorphy2, Mystem3.0, МетаФраз. На основе этих лемматизаторов и их анализа планируется разработка новых.

## 5 Заключение

Был предложен локально-контекстный подход к построению подходящей коллекции для вероятностных тематических моделей из исходной коллекции коротких документов. Эксперименты на данных социальной сети показали, что данный подход дает лучшее качество по сравнению со стандартными вероятностными тематическими моделями.

## Литература

[1] *Blei D. M.* Latent dirichlet allocation // Mach. Learn. Res., 2003.

[2] *Vorontsov K., Potapenko A.* Additive regularization of topic models // Machine Learning Journal, 2014.

[3] *Hong L., Davison B. D.* Empirical study of topic modeling in twitter // Proceedings of the First Workshop on Social Media Analytics., 2010.

[4] *Zhao W. X., Jiang J., Weng J., He J., Lim E.-P. et al.* Comparing twitter and traditional media using topic models // Advances in Information Retrieval., 2011.

[5] *Yan X., Guo J., Lan Y., Cheng X.* A biterm topic model for short texts // Proceedings of the 22nd International Conference on World Wide Web., 2013.

[6] *Turney P. D, Pantel P.* From frequency to meaning: Vector space models of semantics // CoRR, 2010.

[7] *Levy O., Goldberg Y., Dagan I.* Improving distributional similarity with lessons learned from word embeddings // Transactions of the Association for Computational Linguistics, 2015.