Секция «Социология, экономика и право СМИ»

Использование инструментов тематического моделирования для анализа медиатекстов научно-популярного издания «N+1»

Научный руководитель – Аникина Мария Евгеньевна

Толчина Мария Сергеевна
Студент (магистр)
Московский государственный университет имени М.В.Ломоносова, Факультет журналистики, Кафедра социологии массовых коммуникаций, Москва, Россия
E-mail: tomars01@mail.ru

Медиатексты активно производят и потребляют в цифровом пространстве. По данным Mediascope, 85% россиян «пользуются Интернетом хотя бы раз в месяц» [1], при этом 54% их активности в сети приходится на потребление ресурсов, содержащих информационный контент [1]. Кроме того, за последнее десятилетие значительно возросло количество текстов, которые генерируются ежеминутно [3]. Подобная ситуация затрагивает и контент, популяризирующий современные достижения науки, которыми интересуют 85% россиян [2]. Однако возникает вопрос, что именно относится к научно-популярным материалам.

О критериях, по которым текст можно отнести к научно-популярному, писали многие отечественные исследователи (Э.А. Лазаревич, 1981; А.А. Тертычный, 2013; М.В. Литке, 2021 и др.). Однако к единому мнению они так и не пришли. Формулирование подобных маркеров невозможно без понимания того, о чем пишут научно-популярные издания, а точнее какие тематические блоки они затрагивают в своих текстах. Некоторые медиа подобной направленности отличаются высокой публикационной активностью, выпуская ежедневно по несколько расширенных новостных материалов. К таким ресурсам относится и интернет-издание «N + 1». Предварительный анализ сайта показал, что за 2024 год суммарно получилось только 22 дня без публикации. При этом в остальное время количество новостей могло доходить до 15 в сутки. Таким образом, число опубликованных материалов за прошлый год составило 2646. Для такого массива оптимальный способ изучения — интеллектуальный анализ данных (Data Mining). Один из инструментов, который может помочь при определении тематических блоков интернет-текстов, — вероятностное тематическое моделирование. Это автоматизированная процедура, которая позволяет определить «скрытые» темы в коллекции документов, а также сформировать набор слов, наиболее точно характеризующих элементы датасета. В работе предпринимается попытка определить возможности использования тематического моделирования для анализа научно-популярных медиатекстов.

Методика исследования

Для реализации вероятностного тематического моделирования был использован пакет на основе Java для статистической обработки естественного языка Mallet. Его работа основана на реализации латентного распределение Дирихле, распределения Пачинко и иерархического LDA [5]. В рамках этого метода коллекция текстов представляется как мешок слов, в котором грамматические свойства лексических единиц и их пространственно-плоскостное варьирование не учитываются. Mallet раскладывает слова на «вероятностные корзины» [4] до момента, пока алгоритм не придет к наиболее вероятному распределению. Ввод команд для программного решения осуществляется через терминал. Следует учитывать, что число тем, по которым распределяются слова, задается самостоятельно исследователем. Для определения оптимального количества топиков Mallet предлагает вычислить логарифмическую вероятность, которая указывает на степень информационной энтропии $(-\infty, 0]$.

Материалом для вероятностного тематического моделирования послужили медиатексты на сайте научно-популярного интернет-издания «N + 1» за 2024 год (N=2646). Для реализации сплошного исследования было принято оптимизировать сбор данных при помощи парсинга (использовалось расширение Instant Data Scraper и функция IMPORTXML). После этого необходимо было провести предварительную обработку текста — перевести табличные данные в txt-формат, лемматизировать и удалить стоп-слова. Последнее осуществлялось при помощи дополнительно загруженного в Mallet словаря. Для лемматизации использовалась программа морфологического анализа MyStem [6], а процесс обработки документов с выгрузкой в одну директорию осуществлялся при помощи библиотек os и subprocess. После обработки данных Mallet выводит документ с распределением слов по ключевым топикам, которые отображаются в терминале, вес каждого слова в топике, а также вес темы в документе.

Результаты

Проведенный анализ медиатекстов научно-популярного интернет-издания «N+1» показал, что при выделении 10 топиков Mallet сформирует тематические модели, в которых четко прослеживаются направления, связанные с экспериментами на мышах, ботаникой, описанием исследовательских процедур, робототехникой, историей и археологией, медициной, космосом, физикой и химией. Они частично перекликаются с тегами, которые иногда указываются после материалов. В корпусе встретилось 32 темы, из которых наиболее распространенными оказались «Медицина» (N=702), «Зоология» (N=364), «Астрономия» (N=264), «Археология» (N=248), «Экология и климат» (N=189). В этом плане распределение тематического моделирования оказалось близко к тем маркерам, которые указывают авторы.

Рассчитанная логарифмическая вероятность относительно каждого варианта распределения топиков показала, что при расширении и уточнении словаря стоп-слов степень энтропии понижается. Это может быть связано с тем, что в научно-популярных текстах присутствуют фигуры речи, характерные для научного стиля (вставные конструкции, вводные слова, «загородки» и т.д.). Кроме того, некоторые частотные слова в топиках указывают на то, что словарь стоп-слов еще нужно дорабатывать. Например, можно удалить глаголы, связанные с исследовательской деятельностью, а также синонимы к «ученым» и «исследователям».

Таким образом, тематическое моделирование с помощью программы Mallet может быть полезно для формирования представления о том, какие топики составляют ландшафт научно-популярных медиа. Однако некоторый инструментарий этого пакета для статистической обработки требует дополнения.

Источники и литература

1) Медиатренды 2024 // Mediascope: https://mediascope.net/upload/iblock/82a/azh2s3pelef0ddsbug69odhl1nulmihy/Mediascope_Медиатренды%202024.pdf
2) Российская наука сегодня // ВЦИОМ Новости: https://wciom.ru/analytical-reviews/analiticheskii-obzor/rossiiskaja-nauka-segodnja
3) Data Never Sleeps 10.0 // DOMO: https://www.domo.com/data-never-sleeps
4) Getting Started with Topic Modeling and MALLET. // Pogramming Historian: https://programminghistorian.org/en/lessons/topic-modeling-and-mallet
5) Mallet: MAchine Learning for LanguagE Toolkit: https://mimno.github.io/Mallet/
6) Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine: https://cachev2-m9-3.cdn.yandex.net/download.yandex.ru/company/iseg-las-vegas.pdf?lid=246