УДК 004.89

СЕРВИС АГРЕГАЦИИ ТЕКСТОВ НАУЧНЫХ СТАТЕЙ НА РУССКОМ ЯЗЫКЕ С ИСПОЛЬЗОВАНИЕМ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ (НА ПРИМЕРЕ УГЛЕРОДНОЙ ТЕМАТИКИ)

Кобылкина А.А. (ТюмГУ), **Марочкина В.В.** (ТюмГУ)
Научный руководитель – кандидат технических наук, доцент Глазкова А.В.
(ТюмГУ)

Введение. Сервис агрегации текстов научных статей на русском языке с использованием тематического моделирования представляет собой информационную систему, обеспечивающую автоматизированную классификацию и анализ большого массива научной литературы. Актуальность разработки подобного сервиса обусловлена постоянным ростом количества научных статей, где традиционные способы систематизации (ручной отбор статей, составление списков литературы) становятся слишком трудоёмкими, поэтому требуются решения для систематизации данных и быстрого анализа. В качестве демонстрации универсальности подхода рассматривается углеродная тематика, однако методика может быть адаптирована к любым другим областям знаний [1].

Основная часть. В рамках сервиса решаются следующие задачи:

1. Автоматическая классификация текстов.
    1) Анализ большого корпуса научных статей и выделение основных тематических кластеров.
    2) Сокращение времени поиска за счёт группировки публикаций по содержательному сходству.
    3) Выявление наиболее активных областей исследования и отслеживание изменения их популярности в разные периоды времени.
2. Интерпретация и наименование тем.
    1) Использование LLM для автоматизированного присвоения осмысленных названий темам.
    2) Облегчение анализа за счёт формирования понятных формулировок и снижения роли субъективных факторов при описании результатов тематического моделирования.

В рамках исследования была сформирована база из 905 статей по углеродной тематике, для сбора данных экспертами-экологами был предоставлен список тематических разделов, связанных с углеродной тематикой. Для каждой публикации сохраняются основные метаданные и ссылка на источник. На основе собранного корпуса текстов проведено тематическое моделирование, для которого рассматривалось использование двух подходов — LDA и BERTopic [2, 3]. Перед построением тематических моделей тексты проходили предварительную обработку, включающую приведение текста к нижнему регистру, удаление неалфавитных символов, токенизацию с использованием библиотеки razdel, удаление стоп-слов (предлогов, союзов и т.д.) на базе библиотеки *stop*_words, а также лемматизацию при помощи pymorphy3. Процесс выбора модели состоял из двух этапов. На первом этапе строились несколько вариантов тематических моделей с различными параметрами, а их качество оценивалось с помощью метрик когерентности (c_v), перплексии, показателей разделимости и перекрытия тем[4]. По итогам этого анализа было отобрано N моделей с наилучшими значениями метрик. На втором этапе эксперты-экологи оценивали интерпретируемость тем и их смысловую наполненность, исходя из списков ключевых слов, сформированных каждой моделью[5]. В результате экспериментов модель LDA продемонстрировала когерентность 0.4721 и экспертную оценку 0.85, в то время как BERTopic показала когерентность 0.6224 и экспертную оценку 0.95. Кроме того, эксперты отметили, что модели с меньшим числом тем описывают более общие направления без достаточной детализации, а модели BERTopic обеспечивает более высокую интерпретируемость благодаря

использованию n-грамм, что позволяет лучше понять контекст исследуемых тем. С учётом рекомендаций экспертов по количеству тем и высоких значений метрик, в качестве итогового решения была выбрана модель BERTopic. Формируемые ею кластеры обеспечивают как адекватный уровень детализации, так и удобную для человека интерпретацию ключевых терминов. Также разработан веб-сервис, который позволяет визуализировать кластеры в многомерном пространстве, отражая их взаимное расположение, относительный размер и иерархические связи. Каждому кластеру присвоены названия, автоматически сформированные на основе ключевых слов, выявленных тематической моделью. Пользователь имеет возможность детально просматривать и сравнивать выявленные темы, а также отслеживать их динамику, что особенно важно для аналитики тенденций в научных исследованиях. Дополнительно в систему интегрирован поиск по базе научных статей, где можно найти публикации по нужным критериям. Наличие ссылок на исходные источники упрощает процесс сбора библиографии и дальнейшего изучения конкретных работ.

Выводы. Проведено сравнение моделей тематического моделирования LDA и BERTopic с использованием метрик и экспертной оценки. Разработан сервис, который предоставляет инструменты для быстрой систематизации данных, наглядного анализа тематических направлений и углублённого изучения отдельных групп публикаций.

Список использованных источников:

1. Митрофанова О.А., Атугодаге М.М. Динамическое тематическое моделирование русскоязычного корпуса юридических документов // Terra Linguistica. 2023. Т. 14. № 1. С. 70–87. DOI: 10.18721/JHSS.14107
2. Коршунов, Антон Тематическое моделирование текстов на естественном языке / Антон Коршунов, Андрей Гомзин // Труды ИСП РАН : электронный журнал. – URL: https://www.ispras.ru/proceedings/docs/2012/23/isp_23_2012_215.pdf. – Дата публикации: 2012. – ISSN 2220-6426
3. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure //arXiv preprint arXiv:2203.05794. – 2022
4. Милкова, М.А. ТЕМАТИЧЕСКИЕ МОДЕЛИ КАК ИНСТРУМЕНТ «ДАЛЬНЕГО ЧТЕНИЯ» / М.А. Милкова // Цифровая экономика : электронный журнал. – URL: http://digital-economy.ru/arkhiv-zhurnala/ds. – Дата публикации: 07.05.2019.
5. Нокель, М. А. ТЕМАТИЧЕСКИЕ МОДЕЛИ: ДОБАВЛЕНИЕ БИГРАММ И УЧЕТ СХОДСТВА МЕЖДУ УНИГРАММАМИ И БИГРАММАМИ / М. А. Нокель, Н. В. Лукашевич // Вычислительные методы и программирование. – 12.03.2015. – Т. 16, № 2. – С. 215–234.