УДК 004.89

Разработка модуля постобработки данных для системы анализа научных публикаций в сфере искусственного интеллекта

Гогуев К.М. (ИТМО)

Научный руководитель – кандидат технических наук, доцент Гусарова Н.Ф. (ИТМО)

Введение. В условиях стремительного роста числа научных публикаций в сфере искусственного интеллекта (ИИ) возрастает необходимость автоматизированных инструментов анализа, структурирования и интерпретации наукометрических данных. Особую актуальность приобретает задача постобработки метаданных публикаций, направленная на нормализацию, параметризацию, выявление тематических сущностей и расширенный анализ авторского состава. Эффективная постобработка позволяет повысить точность идентификации публикаций по тематике ИИ, обеспечить корректную классификацию авторов и сформировать аналитические показатели человеческого капитала в области искусственного интеллекта.

Основная часть. Исследование выполнено в рамках разработки сервиса SapphireCrawler – интеллектуального агента наукометрического анализа, осуществляющего регулярный сбор публикаций российских авторов в сфере ИИ из открытых источников.

Ключевой задачей работы является разработка модуля постобработки данных, обеспечивающего:

1. Нормализацию и валидацию метаданных публикаций, полученных краулерами.
2. Параметризацию авторских и аффилиационных данных.
3. Выделение тематических сущностей ИИ на основе эталонной КДРМ ИТМО.
4. Расширенный анализ авторов.
5. Подготовку данных для аннотирования и аналитических запросов.

В рамках модуля реализованы алгоритмы определения типа аффилиации на основе ключевых слов, а также алгоритмы классификации авторов по категориям.

Для повышения качества тематической фильтрации публикаций внедрён модуль выделения сущностей ИИ на основе эталонной концептуальной модели КДРМ ИТМО. Это позволяет формировать структурированные признаки публикаций и обеспечивать точность идентификации статей, связанных с тематикой ИИ.

Разработанный модуль интегрирован в архитектуру сервиса, включающую иерархию краулеров, специализированные парсеры атрибутов и пайплайны обработки данных с последующей записью в базу данных и предоставлением доступа через API.

Выводы. Разработан модуль постобработки данных для системы анализа научных публикаций в сфере ИИ, реализующий нормализацию метаданных, выделение тематических сущностей, классификацию авторов и анализ их динамики развития. Интеграция модуля в архитектуру сервиса обеспечивает формирование структурированной базы данных публикаций и создает основу для интеллектуального анализа научной активности в области искусственного интеллекта, включая задачи медицинского направления.

Список использованных источников:

1. Bayesian variational transformer: A generalizable model for rotating machinery fault diagnosis // URL: https://www.sciencedirect.com/science/article/abs/pii/S0888327023008440 (дата обращения: 09.10.2025).
2. Cornell University. BayesFormer: Transformer with Uncertainty Estimation // URL: https://arxiv.org/abs/2206.00826 (дата обращения: 09.10.2025)
3. 1 Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System – 2016. – DOI: 10.1145/2939672.2939785 // https://www.researchgate.net/publication/310824798_XGBoost_A_Scalable_Tree_Boosting_System (дата обращения: 14.02.2026).