УДК 004.932.2

ПАРСИНГ СТАТЕЙ КАК МЕТОД СБОРА ДАННЫХ В ХИМИИ

Лузанова А.М. (ИТМО)**, Щербакова Е.А.** (ИТМО)
Научный руководитель – доктор химических наук, профессор Скорб Е.В. (ИТМО)

Введение. Современная научная область химии требует доступа к широкому массиву данных для поддержки исследований и разработок. За последние десятилетия количество научных статей в химических журналах стремительно возросло, что создает необходимость в эффективных методах обработки этой информации. Парсинг статей [1], основанный на алгоритмах искусственного интеллекта, предоставляет возможность автоматизированного сбора и анализа данных из научных публикаций [2].

Основная часть. В данной работе представлена инновационная нейронная сеть, предназначенная для автоматизированной разметки данных в платформе Label Studio. Для обучения данной нейронной сети была проведена тщательная ручная разметка данных, взятых из реальных научных статей, опубликованных в журналах с Impact Factor (IF) больше 5. Мы выбирали статьи из разных годов, чтобы учесть, как с течением времени меняется качество изображений. Разработанная нейронная сеть представляет собой высокоэффективный инструмент для автоматизированной разметки данных. Она обладает возможностью идентификации и классификации ключевых элементов в химических реакциях, схемах и рисунков в научной статье, что делает ее идеальным средством для оптимизации процесса разметки данных. Первым этапом создания данного инструмента служил тщательный процесс ручной разметки данных. Это позволило создать базовый набор данных для обучения нейронной сети, которая сделала процесс сбора данным полуавтоматическим.

Выводы. В результате исследования, направленного на разработку и обучение нейронной сети для разметки данных в Label Studio, мы достигли следующих ключевых результатов. Применение разнообразных статей позволило создать адаптивную модель, способную эффективно обрабатывать различные форматы данных. Использование платформы Label Studio улучшило процесс обучения и интеграции нейронной сети. Разработанная система обладает потенциалом значительно оптимизировать процессы разметки данных, повышая эффективность исследовательской работы.

Список использованных источников:

1. Xu Y. et al. MolMiner: You only look once for chemical structure recognition //Journal of Chemic al Information and Modeling. – 2022. – Т. 62. – №. 22. – С. 5321–5328.
2. Rajan K. et al. DECIMER. ai-An open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications. – 2023.