Отзыво прохождении

Учебной практики (научно-исследовательская работа)

|  |  |
| --- | --- |
| *Студент* | ***Мурадян Денис Степанович*** |
| Дата | *08 июня 2025* |

Перед Мурадяном Денисом Степановичем была поставлена задача Разработка универсальной системы парсинга веб-страниц на базе больших языковых моделей (LLM) для автоматизированного извлечения структурированных данных с сайтов различной сложности. Актуальность обусловлена необходимостью значительного ускорения сбора данных из открытых онлайн-источников и оптимизации затрат на разработку индивидуальных парсеров.

Обучающийся Мурадян Денис Степанович в ходе прохождения практики своевременно и качественно выполнил следующие задачи:

1. Изучил существующие методы автоматического извлечения данных с веб-сайтов.
2. Разработал модуль получения и предобработки веб-страниц с автоматическим определением необходимости JS-рендеринга и выбором между статическими запросами (requests) и динамическим рендерингом через Selenium.
3. Спроектировал и реализовал две стратегии очистки HTML-разметки:

* Строгая очистка для режима Structuring;
* Частичная очистка для режима Codegen.

1. Интегрировал систему с LLM (MistralAI) для двух режимов парсинга: Structuring (непосредственная генерация структурированных данных в формате JSON) и Codegen (автоматическая генерация и исполнение Python-скриптов), обеспечив кэширование результатов в SQLite и ChromaDB.
2. Организовал вывод результатов через пользовательские интерфейсы: Gradio, FastAPI и веб-фронтенд на Jinja2/JavaScript.

В ходе работы обучающийся Мурадян Денис Степанович активно взаимодействовал с научным руководителем, своевременно выполнял поставленные задачи, проявлял самостоятельность, оперативно устранял вовремя выявленные замечания к работе.

* Система продемонстрировала высокую точность извлечения данных при тестировании на разнообразном наборе сайтов (Gismeteo, официальный сайт СПбГУ, Habr, Яндекс.Финанс (валютные курсы), RussianFood).
* Время запуска на ранее неизвестном сайте (без предварительного кэша) в режиме Codegen варьировалось от 17 до 75 с, тогда как при повторном запуске с кэшем — от 4,6 до 17,8 с.
* Режим Structuring обрабатывал страницы в среднем за 7,3–32,7 с.
* Интеграция с ChromaDB и семантическое кэширование обеспечили значительное ускорение повторных запросов при сохранении высокой точности.

Считаю, что обучающийся Мурадян Денис Степанович за учебную практику заслуживает оценку «зачтено» в системе ECTS «A».

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Научный руководитель практики, старший преподаватель кафедры информатики | / \_\_\_\_\_\_\_\_\_ / | Бушмелев Ф.В. |