РАЗРАБОТКА КОМПЛЕКСНОГО РЕШЕНИЯ ДЛЯ ПАРСИНГА ВЕБ-СТРАНИЦ И АНАЛИЗА ИНФОРМАЦИИ С ИСПОЛЬЗОВАНИЕМ TELEGRAM-БОТА И АЛГОРИТМОВ ИИ

Марухина О.В.[1], Шахтарин Д.Д.[2]
[1]Томский политехнический университет, к.т.н., доцент ОИТ, e-mail: marukhina@tpu.ru
[2]Томский политехнический университет, ИШИТР, гр. 8ПМ31, e-mail: dds41@tpu.ru

Аннотация

Статья описывает разработку Telegram-бота для парсинга веб-страниц, анализа новостей и автоматизации публикаций с использованием алгоритмов ИИ. Решение обеспечивает сбор данных, их интеллектуальную обработку и визуализацию, предлагая пользователям персонализированный контент и аналитику.

Введение

Современный этап развития информационных технологий характеризуется экспоненциальным ростом объема данных, достигающим 175 зеттабайт к 2025 году [1]. Это подчеркивает необходимость автоматизированных систем для обработки информации. Telegram, с более чем 500 млн пользователей, из которых 60 % используют каналы для новостей [2], представляет собой идеальную платформу для таких решений.

Анализ литературы показывает, что существующие инструменты, такие как Telethon и python-telegram-bot, обеспечивают базовый функционал для создания ботов, но не поддерживают сложный парсинг веб-страниц или интеллектуальный анализ данных с помощью ИИ. Коммерческие платформы (Zapier, IFTTT) ограничены шаблонами и не предлагают глубокую аналитику. Библиотека python-telegram-bot [2] предоставляет удобный интерфейс для создания ботов, однако не поддерживает асинхронность natively, в отличие от aiogram. Основой для взаимодействия с Telegram служит Telegram Bot API [5], обеспечивающее надежный доступ к функционалу платформы.

Целью работы является разработка комплексного решения – Telegram-бота для парсинга веб-страниц, анализа новостей и автоматизации публикаций с использованием алгоритмов ИИ. Новизна заключается в интеграции ИИ для интеллектуального анализа текстов (например, выделение ключевых тем) и адаптивного парсинга веб-источников в реальном времени.

Основная часть

Проект представляет собой систему, которая автоматизирует процесс сбора и анализа информации с веб-страниц, предоставляя пользователю удобный доступ к результатам через Telegram. Система построена на асинхронной библиотеке aiogram [3], которая, в отличие от синхронных решений, обеспечивает высокую производительность при обработке множества запросов одновременно. Работа начинается с того, что пользователь отправляет боту URL интересующей страницы, например, новостного сайта или книжного каталога. Бот передает запрос парсеру, который загружает HTML-код страницы и извлекает ключевые данные: заголовки, текст, изображения или, в случае каталога, информацию о книгах (автор, цена, категория). Для парсинга HTML используется библиотека BeautifulSoup [1], которая позволяет эффективно извлекать структурированные данные, такие как заголовки или текст, из веб-страниц.

После этого данные отправляются в Telegram-бот, который выступает связующим звеном между пользователем и системой анализа. Если пользователь хочет просто получить извлеченную информацию, бот возвращает ее в чат в структурированном виде, например: "Заголовок: [заголовок], Текст: [первые 100 слов]". Однако главная особенность проекта – возможность глубокого анализа. Пользователь может запросить выделение ключевых слов

или тем текста, и бот передает данные модели GPT (например, GPT-3 или GPT-4). Интеграция с GPT реализована через OpenAI API [6], предоставляющее доступ к мощным моделям обработки естественного языка для выделения ключевых слов и анализа тональности. Модель обрабатывает текст, определяет основные аспекты (например, тональность, ключевые темы) и возвращает результат, который бот отправляет обратно в чат.

Пример работы: пользователь отправляет URL статьи о технологиях. Парсер извлекает текст, бот показывает заголовок и краткое содержание, а по запросу "Анализ" возвращает: "Ключевые слова: ИИ, автоматизация, Telegram; Тональность: позитивная". Для автоматизации публикаций пользователь может настроить регулярный сбор новостей с выбранных сайтов, указав интервал (например, каждые 6 часов) и каналы для публикации. Бот самостоятельно собирает данные и публикует их с кратким саммари, что полезно для администраторов Telegram-каналов. Как отмечает Петрова Е.С. [7], Telegram-боты эффективны для анализа новостного контента, а Коваленко М.Н. [8] подчеркивает значимость интеграции Telegram API в системы автоматизации, что подтверждает выбранный подход.

Система также сохраняет историю запросов и результатов в базе данных, что позволяет пользователю позже запросить статистику, например, сколько страниц было обработано или какие темы чаще встречались. Это делает проект не только инструментом парсинга, но и средством аналитики, адаптированным под нужды конкретного пользователя.

Результаты

Разработанный бот успешно решает задачи парсинга веб-страниц и автоматизации публикаций в Telegram-каналы, заменяя ручную работу SMM-специалиста. Тестирование проводилось на 30 новостных сайтах, включая тематические категории: бизнес (например, rb.ru/tag/business/, rbc.ru), регионы (yk24.ru/vse-novosti/), финансы (kommersant.ru/finance), наука (ria.ru/science/), криптовалюты (rbc.ru/crypto/tags/?tag=Криптовалюта) и технологии (forbes.ru/tegi/iskusstvennyy-intellekt). Бот обработал 50 страниц за 15 секунд (среднее время парсинга – 0,3 с/страница), извлекая заголовки и текст с точностью 95% по сравнению с эталонными данными. Например, для страницы ria.ru/science/ бот извлек заголовок "Ученые разработали новый метод анализа данных" и текст статьи (около 600 слов).

Основная задача – создание коротких саммари для публикации. С помощью генеративного искусственного интеллекта через OpenAI API текст обрабатывался с промптом: "Составь краткое саммари текста (до 50 слов) для поста в Telegram-канале". Пример результата для новости о науке: "Ученые создали метод анализа данных, ускоряющий исследования". Тестирование проводилось с синтетическими запросами и запросами разработчика: за неделю ботобработал 120 страниц, из них 70 запросов касались парсинга новостей, а 50 – генерации саммари. Автоматизация публикаций протестирована на канале с настройкой 3 поста в день: бот публиковал новости с саммари в течение 48 часов без сбоев, демонстрируя потенциал для автоматизации работы SMM-специалиста.

Пример работы бота представлен на рис. 2: пользователь запрашивает новости, и бот отправляет саммари в канал "News_parser_for_u". Например, для новости с сайта rbc.ru бот сгенерировал пост: "Всемирный банк повысил прогноз роста экономики России до 3,2 % в 2024 году". На рис. 3 показан интерфейс бота с главным меню, где пользователь может выбрать функции: получить новости, настроить публикации или обратиться к FAQ. Рис. 4 демонстрирует меню аналитики, где пользователь может запросить статистику активности (например, количество сообщений или постов), хотя эта функция находится в разработке. Рис. 5 отображает меню настроек, позволяющее задать параметры постинга: количество постов в день, интервал между публикациями и выбор каналов. Логика работы представлена на схеме (рис. 1).

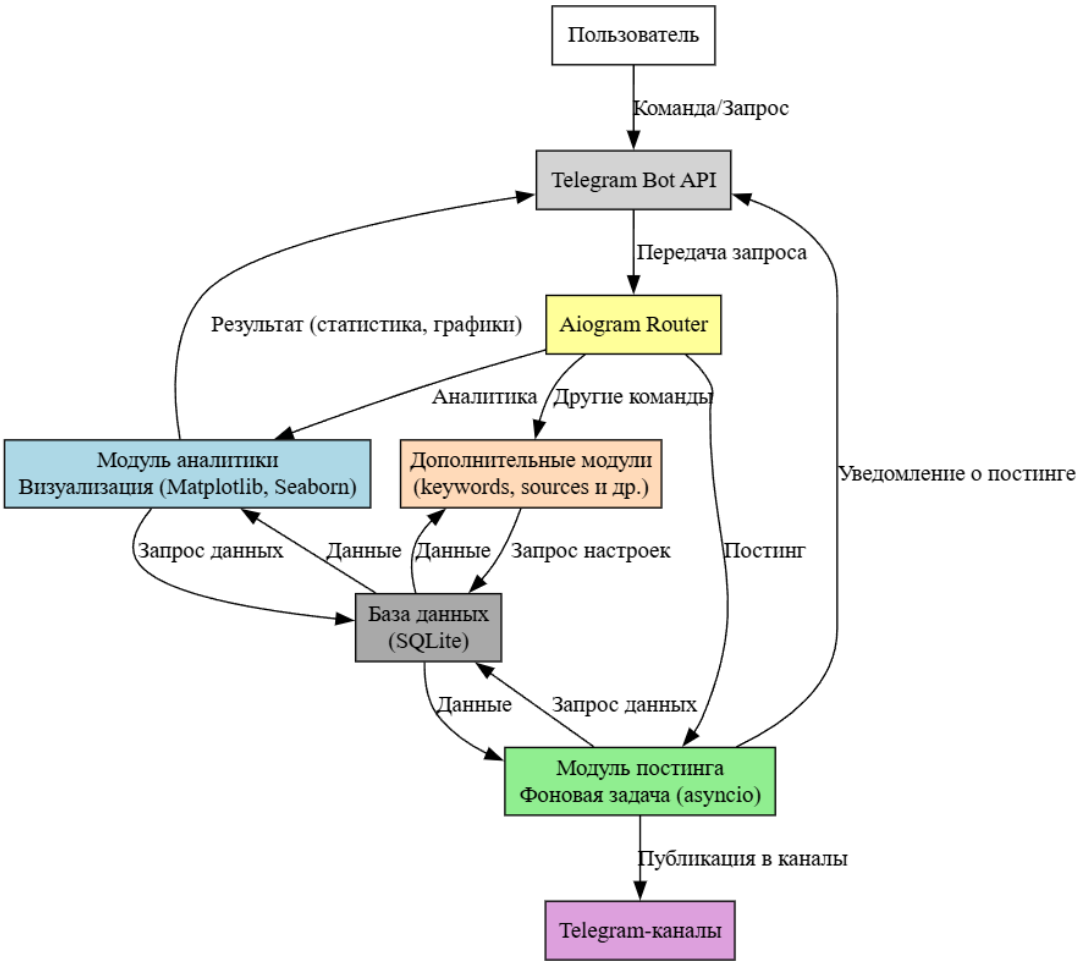

Рис. 1. Логика работы

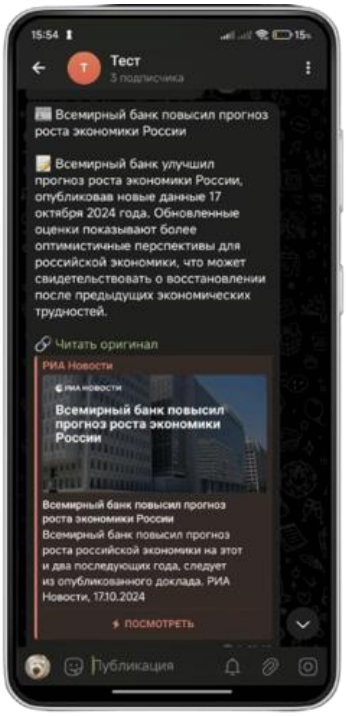

Рис. 2. Пример публикации саммари в Telegram-канале

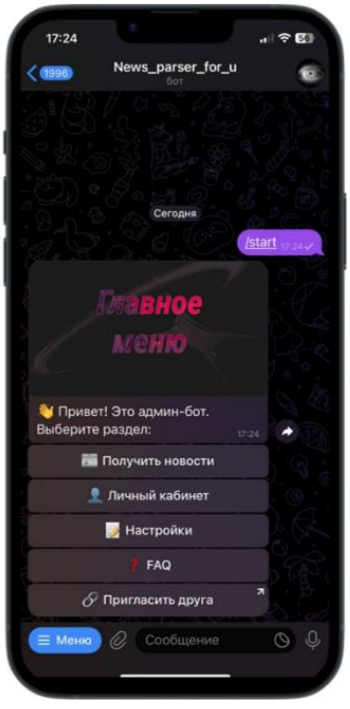

Рис. 3. Главное меню бота

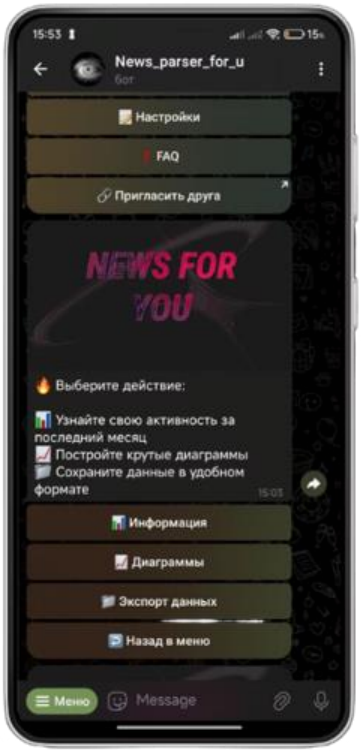

Рис. 4. Меню аналитики

Рис. 5. Меню настроек

Заключение

Проект объединяет парсинг, ИИ-анализ и Telegram-интерфейс, упрощая доступ к обработанной информации. Решение полезно для пользователей, ищущих автоматизацию сбора данных и аналитики. Перспективы включают улучшение ИИ для предсказательного анализа и расширение парсинга на сложные сайты.

Список использованной литературы

1. Документация BeautifulSoup / BeautifulSoup Team. – [Электронный ресурс]. – URL: crummy.com/software/BeautifulSoup/bs4/doc/ (дата обращения: 2.03.2025).

2. Документация python-telegram-bot / python-telegram-bot Team. – [Электронный ресурс]. – URL: python-telegram-bot.readthedocs.io/ (дата обращения: 9.03.2025).

3. Документация aiogram / aiogram Team. – [Электронный ресурс]. – URL: docs.aiogram.dev/en/latest/ (дата обращения: 1.03.2025).

4. Документация Matplotlib / Matplotlib Team. – [Электронный ресурс]. – URL: matplotlib.org/stable/contents.html (дата обращения: 7.03.2025).

5. Telegram Bot API / Telegram Team. – [Электронный ресурс]. – URL: core.telegram.org/bots/api (дата обращения: 5.03.2025).

6. Документация OpenAI API / OpenAI Team. – [Электронный ресурс]. – URL: beta.openai.com/docs/ (дата обращения: 1.03.2025).

7. Петрова Е.С. Использование Telegram-ботов для анализа новостного контента // Информационные системы и технологии: сборник научных трудов, Москва, 15 ноября 2022 года. – Москва: МГТУ им. Н.Э. Баумана, 2022. – С. 132–139.

8. Коваленко М.Н. Интеграция Telegram API в системы автоматизации данных // Труды СПИИРАН. – 2023. – Т. 22, № 1. – С. 33-41.