# Техническое задание

# Общие сведения

**Название системы.** Система интеллектуальной обработки и классификации веб-сайтов с последующей интеграцией результатов в образовательную среду.

**Основания для создания системы.** Необходимость автоматизации процессов сбора, предобработки и классификации веб-контента для последующего использования в образовательной среде. Система предназначена для ускорения поиска релевантной информации студентами и преподавателями, а также для обеспечения доступа к иностранным научным источникам посредством машинного перевода.

# Цели и назначение создания автоматизированной системы

Целью создания системы является повышение эффективности образовательного процесса за счет автоматизации поиска и классификации веб-сайтов, сокращение временных затрат на сбор и анализ данных, обеспечение доступности иностранной научной литературы через машинный перевод, увеличение точности и актуальности классификаций веб-страниц.

# Характеристика объектов автоматизации

Объектом автоматизации является процесс обработки и классификации веб-сайтов, а также взаимодействие студентов и преподавателей с базой данных через Telegram-бот.

# Требования к автоматизированной системе

**Функциональные характеристики.** Парсер должен уметь переходить на сайты по URL-адресам и извлекать заголовки (title) и описания (description) из html кода. До обучения модели нейронной сети на достаточном количестве данных, должен формироваться датасет на основе поисковой выдачи Яндекса по введенным мной запросам.

**Предобработка данных.** Удаление пунктуации, стоп-слов, нормализация и лемматизация текста.Приведение данных к векторному.

**Модель нейронной сети.** Архитектура LSTM для анализа последовательностей и многоклассовой классификации веб-страниц.Парсеры должны постоянно пополнять базу новыми данными, что даст возможность переобучения модели на обновленных данных.

**База данных**. Хранит в себе url-адреса, заголовки (title), описание (description) и метки классификации веб-сайта (lebel)

**Telegram-бот**. Интерфейс для взаимодействия студентов и преподавателей с базой данных. Возможность отправки запросов на интересующие темы и получение результатов поиска по сформированной базе данных. Оценка релевантности найденных материалов и обратная связь от пользователей.

**Многоязыковая поддержка**. Автоматическое определение языка веб-страницы перед ее обработкой и классификацией. Интеграция современных моделей машинного перевода для повышения качества переводов.

**Требования к надежности.** Система должна обеспечивать стабильную работу при высоких нагрузках.Включает в себя механизмы обнаружения и устранения ошибок, включая логирование и мониторинг состояния системы.Резервное копирование данных с периодичностью 1 раз в 2 недели.

**Требования к эргономике и техническому обслуживанию.** Удобство использования Telegram-бота для студентов и преподавателей.Подробное руководство пользователя и документация.

**Требования к совместимости.** Совместимость с современными операционными системами и браузерами.

**Требования к производительности.** Максимальная задержка ответа на запрос пользователя не более 5 секунд при поиске по базе. Если требуется машинный перевод, то время увеличивается. Способность обрабатывать не менее 30 запросов в минуту.

**Используемые технологии и инструменты**. Библиотеки Python: Selenium для парсинга, NLTK для предобработки текста, TensorFlow/Keras для построения модели нейронной сети, aiogram для разработки Telegram-бота. Базы данных: PostgreSQL или MongoDB. Готовые модели машинного перевода с Hugging Face

# Состав и содержание работ по созданию автоматизированной системы

**Анализ требований и проектирование системы:**

Продолжительность этапа: 2 месяц.

Результаты: Функциональные и технические требования, архитектурное решение.

**Реализация системы:**

Продолжительность этапа: 3 месяца.

Результаты: Работоспособные компоненты системы, включая парсер, модули предобработки, модель нейронной сети, базу данных и Telegram-бот.

**Тестирование и внедрение:**

Продолжительность этапа: 1 месяц.

Результаты: Протестированная и готовая к эксплуатации система.

**Эксплуатация и сопровождение:**

Результаты: Обеспечение бесперебойной работы системы, регулярное обновление и улучшение.

# Порядок разработки автоматизированной системы

**Анализ требований и проектирование системы.** Определение функциональных и нефункциональных требований.Проектирование архитектуры системы.

**Реализация системы.** Разработка компонентов системы, включая парсер для сбора данных, модули предобработки, модель нейронной сети, базу данных и Telegram-бот.

**Тестирование и внедрение.** Проведение комплексного тестирования системы.

**Эксплуатация и сопровождение.** Обеспечение технической поддержки и сопровождения системы.Регулярное обновление и улучшение системы.

# Порядок контроля и приемки автоматизированной системы

Контроль и приемка системы осуществляются руководителем дипломной работы на каждом этапе создания системы. Приемочные испытания проводятся после завершения реализации и тестирования системы. Критерии приемки включают:

Соответствие функциональным и техническим требованиям.

Надежность и безопасность системы.

Производительность системы.

Качество и полнота документации.

# Требования к составу и содержанию работ по подготовке объекта автоматизации к вводу автоматизированной системы в действие

Необходимо провести подготовку сервера, на котором будет располагаться система. Подготовка включает в себя установку необходимых программных модулей, настройку серверов и баз данных

# Требования к документированию

Разработка и предоставление полного комплекта эксплуатационной и программной документации. Описание всех функций и характеристик системы. Руководство пользователя, административное руководство, техническое описание системы и другие необходимые документы.

# Источники разработки

Источником финансирования проекта является разработчик системы. Технические и методические материалы предоставляются из открытых источников информации на основе анализа существующих аналогов и лучших практик в области информационных систем для образования.