**МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ**

**БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ**

**ФАКУЛЬТЕТ РАДИОФИЗИКИ И КОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ**

**Кафедра системного анализа и компьютерного моделирования**

САВЧУК
Юрий Игоревич

# РАЗРАБОТКА ПРОГРАММНОГО МОДУЛЯ АВТОМАТИЗИРОВАННОГО ПОИСКА ТЕКСТОВЫХ ДАННЫХ В РАЗНОФОРМАТНЫХ ИСТОЧНИКАХ НА ОСНОВЕ МЕТОДА ГЕНЕРАЦИИ С ДОПОЛНЕННОЙ ВЫБОРКОЙ

Аннотация (реферат) к дипломной работе

Научный руководитель:
канд.физ.-мат.наук,
доцент В.В. Скакун

Допущена к защите

«___» _____2025 г.

Зав. кафедрой системного анализа
и компьютерного моделирования
кандидат физ.-мат. наук, доцент В.В. Скакун

Минск, 2025

# Реферат

Дипломная работа, 54 страниц, 4 таблицы, 6 рисунков, 1 формула, 9 источников.

Ключевые слова: автоматизированный поиск, текстовые данные, разноформатные источники, Retrieval-Augmented Generation (RAG), LLaMA 3.1, семантический поиск, обработка естественного языка.

Объект исследования — методы автоматизированного поиска текстовых данных в разноформатных источниках.

Предмет исследования — применение метода Retrieval-Augmented Generation (RAG) на основе модели LLaMA 3.1 для повышения релевантности и точности поиска в структурированных (JSONL, SQLite), полуструктурированных (HTML, DocBook) и неструктурированных (PDF, ODT) данных.

Цель работы — разработка программного модуля, реализующего автоматизированный поиск информации в разнородных источниках с использованием метода RAG и модели LLaMA 3.1, а также сравнительный анализ его эффективности с традиционными и семантическими методами поиска.

В работе изучены современные методы извлечения и поиска текста, включая традиционные (TF-IDF, BM25) и семантические (DPR) подходы. Разработанный модуль включает компоненты для индексации данных, генерации эмбеддингов, обработки запросов и тестирования. Проведено тестирование модуля на реальных данных с оценкой качества поиска по метрикам Precision, Recall, EM.

Ключевые результаты:

- Повышение релевантности поиска за счёт комбинации семантического извлечения и генерации контекстно-адаптированных ответов.
- Унификация обработки разнородных данных (PDF, HTML, JSONL) в единой системе.
- Снижение зависимости от ручной разметки данных.

Перспективы развития включают расширение поддержки форматов (HTML, JSONL), улучшение обработки многозвычных данных и оптимизацию параметров RAG для повышения точности и скорости работы.

# Рэферат

Дыпломная праца, 54 старонак, 4 табліцы 6 малюнкаў, 1 формула, 9 крыніц.

Ключавыя словы: аўтаматызаваны пошук, тэкставыя дадзеныя, рознафарматныя крыніцы, Retrieval-Augmented Generation (RAG), LLaMA 3.1, семантычны пошук, апрацоўка натуральнай мовы.

Аб'ект даследавання — метады аўтаматызаванага пошуку тэкставых дадзеных у рознафарматных крыніцах.

Прадмет даследавання — прымяненне метаду Retrieval-Augmented Generation (RAG) на аснове мадэлі LLaMA 3.1 для павышэння рэлевантнасці і дакладнасці пошуку ў структураваных (JSONL, SQLite), часткова структураваных (HTML, DocBook) і неструктураваных (PDF, ODT) дадзеных.

Мэта працы — распрацоўка праграмнага модуля, які рэалізуе аўтаматызаваны пошук інфармацыі ў рознародных крыніцах з выкарыстаннем метаду RAG і мадэлі LLaMA 3.1, а таксама параўнальны аналіз яго эфектыўнасці з традыцыйнымі і семантычнымі метадамі пошуку.

У працы вывучаны сучасныя метады здабывання і пошуку тэксту, уключаючы традыцыйныя (TF-IDF, BM25) і семантычныя (DPR) падыходы. Распрацаваны модуль уключае кампаненты для індэксацыі дадзеных, генерацыі эмбедзінгаў, апрацоўкі запытаў і тэставання. Праведзена тэставанне модуля на рэальных дадзеных з адзнакай якасці пошуку па метрыках Precision, Recall, EM.

Галоўныя вынікі:

• Павышэнне рэлевантнасці пошуку за кошт камбінацыі семантычнага здабывання і генерацыі кантэкстна-адаптаваных адказаў.

• Уніфікацыя апрацоўкі рознародных дадзеных (PDF, HTML, JSONL) у адзінай сістэме.

• Зніжэнне залежнасці ад ручной разметкі дадзеных.

Перспектывы развіцця ўключаюць пашырэнне падтрымкі фарматаў (HTML, JSONL), паляпшэнне апрацоўкі шматмоўных дадзеных і аптымізацыю параметраў RAG для павышэння дакладнасці і хуткасці працы.

## Abstract

Graduation thesis, 54 pages, 4 tables, 6 figures, 1 formula, 9 references.

Keywords: automated search, textual data, multi-format sources, Retrieval-Augmented Generation (RAG), LLaMA 3.1, semantic search, natural language processing.

Research object — methods of automated search for textual data in multi-format sources.

Research subject — application of the Retrieval-Augmented Generation (RAG) method based on the LLaMA 3.1 model to improve search relevance and accuracy in structured (JSONL, SQLite), semi-structured (HTML, DocBook), and unstructured (PDF, ODT) data.

Objective — development of a software module implementing automated information retrieval in heterogeneous sources using the RAG method and LLaMA 3.1 model, along with a comparative analysis of its effectiveness against traditional and semantic search methods.

The work examines modern text extraction and search methods, including traditional (TF-IDF, BM25) and semantic (DPR) approaches. The developed module includes components for data indexing, embedding generation, query processing, and testing. The module was tested on real-world data with search quality evaluated using Precision, Recall, and EM metrics.

Key results:

• Improved search relevance through the combination of semantic retrieval and context-adapted answer generation.

• Unified processing of heterogeneous data (PDF, HTML, JSONL) within a single system.

• Reduced dependency on manual data labeling.

Future development prospects include expanding format support (HTML, JSONL), enhancing multilingual data processing, and optimizing RAG parameters to improve accuracy and performance.