|  |  |  |
| --- | --- | --- |
|  |  |  |
| МИНОБРНАУКИ РОССИИ | | |
| Федеральное государственное бюджетное образовательное учреждение  высшего образования  **«МИРЭА – Российский технологический университет»**  **РТУ МИРЭА** | | |
| Институт искусственного интеллекта | | |
| Кафедра программного обеспечения систем радиоэлектронной аппаратуры | | |

|  |  |
| --- | --- |
| **ОТЧЕТ** | |
| **ПО НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ** | |
|  | |
| **Тема: «Разбиение текста книги по разделам»** | |
| Студент группы           КМБО-02-22 | *П.А. Аракчеева* |
| Руководитель практики | *А. В. Завьялов* |

|  |  |  |
| --- | --- | --- |
| «Отчет представлен к рассмотрению» | «\_\_\_» \_\_\_\_\_\_\_\_\_\_2024 г. | *(подпись студента)* |
|  |  |  |
| «Отчет утвержден.  Допущен к защите» | «\_\_\_» \_\_\_\_\_\_\_\_\_\_2024 г. | *(подпись руководителя)* |

Москва 2024

Оглавление

[Введение 3](#_Toc167288235)

[1 Постановка задачи 4](#_Toc167288236)

[1.1 Задание 4](#_Toc167288237)

[1.2 Исходные данные 4](#_Toc167288238)

[1.3 Планируемые результаты 4](#_Toc167288239)

[2 Описание работ и результатов 5](#_Toc167288240)

[2.1 Используемые средства разработки 5](#_Toc167288241)

[2.2 Содержание и результат выполненных работ 5](#_Toc167288242)

[2.3 Доступ к результатам 6](#_Toc167288243)

[3 Заключение 7](#_Toc167288244)

[Список использованных источников 8](#_Toc167288245)

# Введение

Для упрощения поиска информации в источниках для преподавателей была разработана система разбиения текста книги по разделам в соответствии с оглавлением.

# Постановка задачи

## Задание

Необходимо разработать приложение на языке С++, осуществляющее составление разметки, поиск информации в соответствии с разметкой. Результаты работы программы (текст книги, разбитый по разделам) необходимо сохранять в базу данных.

## Исходные данные

1. Пример текста книги
2. Таблица с информацией об оглавлении
3. Структура таблицы, в которую необходимо записать полученную информацию

## Планируемые результаты

Программа, выполняющая разметку текста через нечеткий поиск заданных разделов, позволяющая редактировать результаты поиска пользователем, обеспечивающая извлечения текста разделов в соответствии с разметкой. Также программа позволяет генерировать SQL-запросы и сохранять тем самым результаты разбора в БД PostgreSQL с согласованной структурой.

# 2 Описание работ и результатов

## 2.1 Используемые средства разработки

Программа написана с использованием языка программирования с++ не ниже 17 стандарта. Для написания исходного кода был использован текстовые редактор vscode. Также в ходе разработки проекта были использована библиотека libpqxx для соединения с базой данных postgresql. Компилятор GNU g++ (Ubuntu 11.4.0-1ubuntu1~22.04) 11.4.0. Система сборки cmake.

## 2.2 Содержание и результат выполненных работ

В соответствии с планом индивидуального задания были выполнены следующие пункты:

* Был изучен и использован алгоритм нахождения расстояния Левенштейна для нечеткого поиска заголовков в тексте
* Был изучен и использован алгоритм возрастающей подпоследовательности для составления соответствия между началом страницы и строкой в документе
* Изучен функционал библиотеки libpqxx для работы с базой данных postgresql. Пример результата запроса на рисунке 1.

На вход программе файлы подаются через аргументы командной строки

0 filename.txt output.txt - разметить filename.txt и записать размеченный файл в output.txt

1 filename.txt - занести в базу главы filename.txt (уже размеченного текста книги)

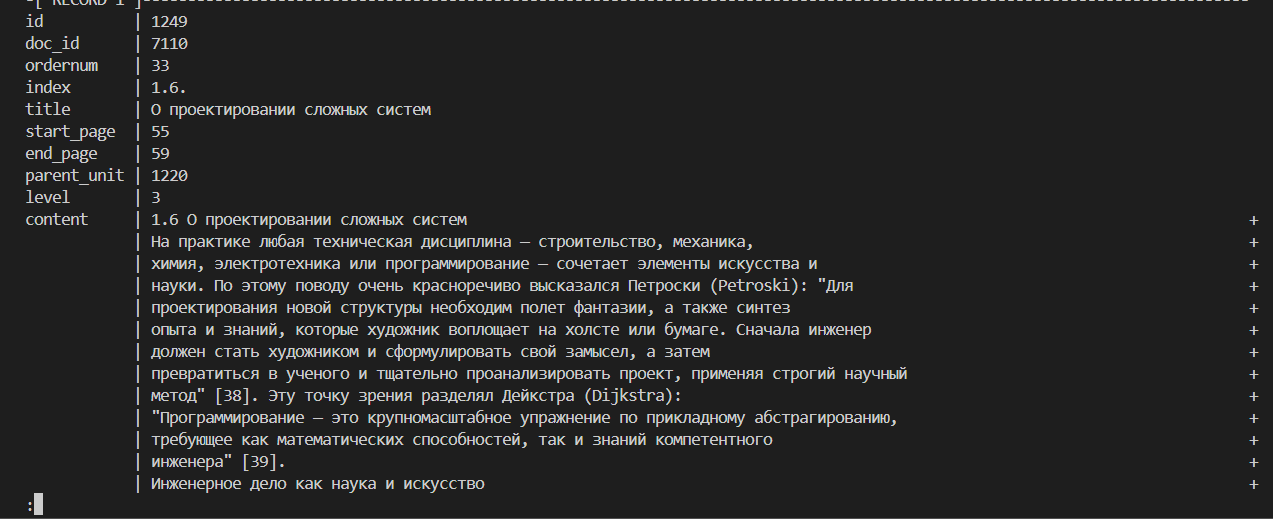


Рис. 1 – Результат запроса SELECT к заполненной таблице

Текст книги записывается в базу данных в соответствии с разметкой (рис.2). При желании после авторазметки текста (запуск с первым аргументом 0) пользователь может скорректировать положение тэгов в тексте.

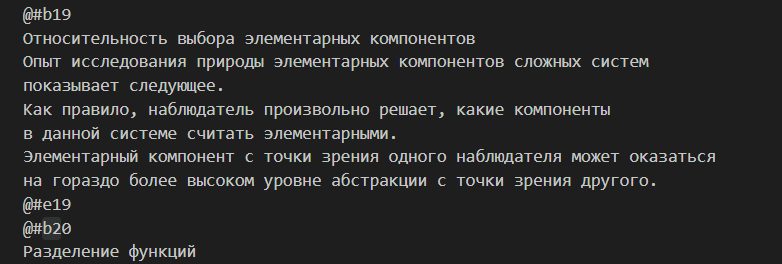


Рис. 2 – Пример разметки текста

## 2.3 Доступ к результатам

Полученные результаты практики расположены в репозитории Github и доступны по ссылке https://github.com/damplingo/pars\_book

# 3 Заключение

В процессе работы над программой была создана программа, которая разделяет текст книги на разделы.

Были изучены алгоритмы Левенштейна и возрастающей подпоследовательности. Также были рассмотрены особенности работы с русскими буквами в языке программирования C++: использование локалей и расширенной строки std::wstring.

Кроме того, были получены знания о современных стандартах языка C++ и о работе с PostgreSQL на этом языке.

# Список использованных источников

1. https://github.com/jtv/libpqxx
2. http://cppreference.com
3. https://habr.com/ru/articles/117063