

МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра математической кибернетики и компьютерных наук

КУРСОВАЯ РАБОТА

НА ТЕМУ: «РАЗРАБОТКА ПРОГРАММЫ ПОИСКА ПОВТОРОВ В ТЕКСТЕ»

Выполнила:

студентка 2 курса
251 группы
Яфарова Эльвина Эрнестовна

Научный руководитель:

д.т.н., проф.
Богомолов Алексей Сергеевич

Саратов, 2025

АКТУАЛЬНОСТЬ ТЕМЫ

- Обработка текстовых данных играет важную роль в научной, образовательной и издательской деятельности
- Выявление повторяющихся фрагментов позволяет оптимизировать тексты, устранять избыточность и улучшать их структуру
- Существующие решения не дают возможности автоматического поиска повторов в текстах



ЦЕЛЬ И ЗАДАЧИ КУРСОВОЙ РАБОТЫ

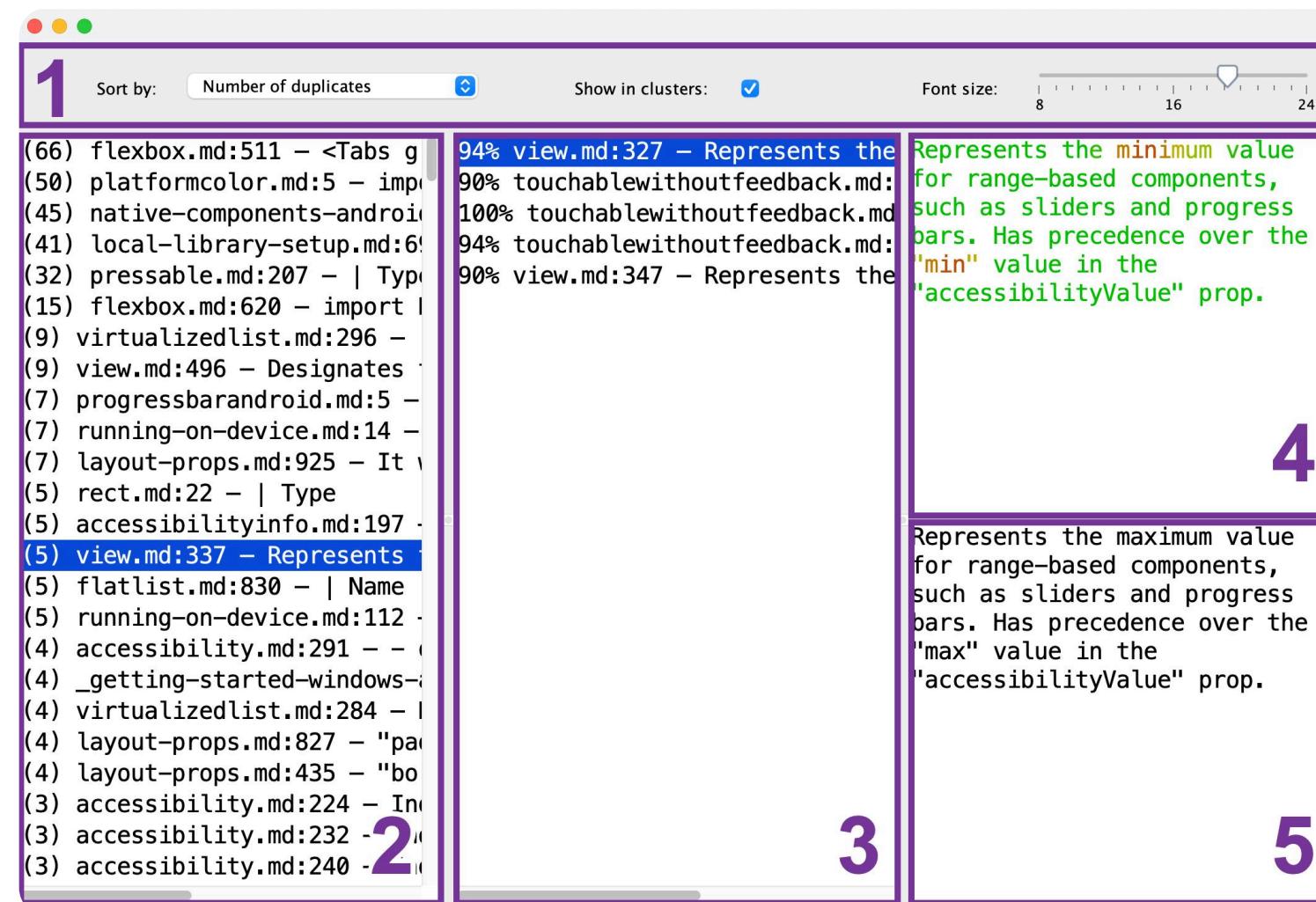
ЦЕЛЬ

Разработка утилиты для поиска повторяющихся фрагментов текста с заданными признаками, включая длину.

ЗАДАЧИ

- Анализ существующих решений
- Исследование требований и ожиданий пользователей
- Выбор технологий и инструментов
- Разработка алгоритма поиска повторяющихся фрагментов
- Создание пользовательского интерфейса
- Тестирование и анализ результатов

ОБЗОР КОНКУРЕНТОВ



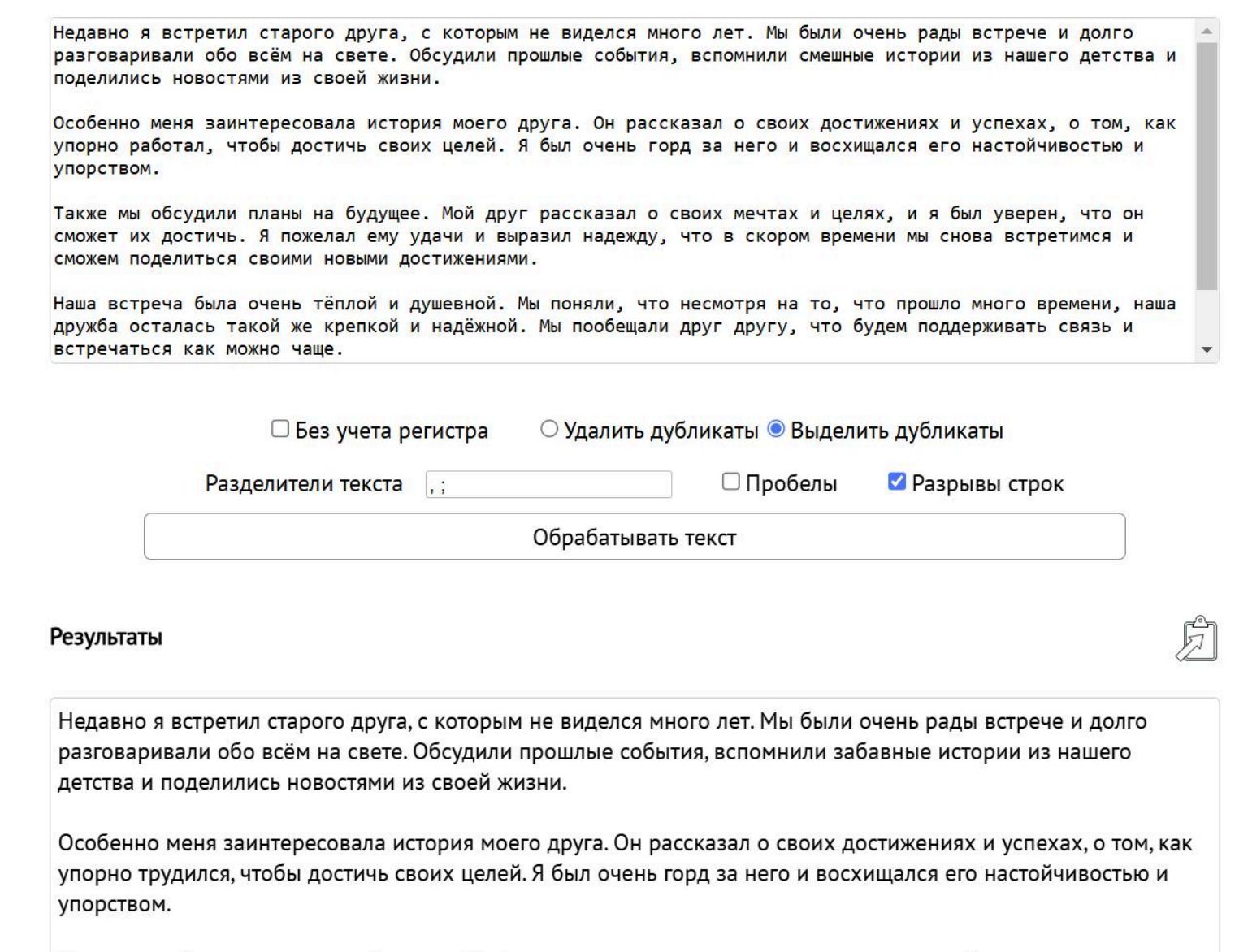
Duplicate Finder

Не находит повторы в тексте



Duplicate Word Finder

Ищет только отдельные слова



Андубликтатор

Удаление конкретных слов

ОБЗОР КОНКУРЕНТОВ

Подсчет уникальных слов

Текст

```
Вот дом,  
Который построил Джек.  
А это пшеница,  
Которая в темном чулане хранится  
В доме,  
Который построил Джек
```

Текст, содержащий повторы слов

Удалить повторы
Удаляет из текста повторяющиеся слова

Показывать количество
Показывает количество слов рядом со словом

Учитывать регистр

Исключить из подсчета
() % \$ + /

Слова, которые требуется исключить из подсчета

РАССЧИТАТЬ

Исходный текст для обработки:

```
Также мы обсудили планы на будущее. Мой друг рассказал о своих мечтах и целях, и я был уверен, что он сможет их достичь. Я пожелал ему удачи и выразил надежду, что в скором времени мы снова встретимся и сможем поделиться своими новыми достижениями.

Наша встреча была очень тёплой и душевной. Мы поняли, что несмотря на то, что прошло много времени, наша дружба осталась такой же крепкой и надёжной. Мы обещали друг другу, что будем поддерживать связь и встречаться как можно чаще.

Я благодарен судьбе за то, что она свела меня с таким замечательным человеком, как мой друг. Я уверен, что наша дружба будет длиться долгие годы и станет примером для других людей.

Недавно я встретил старого друга, с которым не виделся много лет. Мы были очень рады встрече и долго разговаривали обо всём на свете. Обсудили прошлые события, вспомнили смешные истории из нашего детства и поделились новостями из своей жизни.

Особенно меня заинтересовала история моего друга. Он рассказал о своих достижениях и успехах, о том, как упорно работал, чтобы достичь своих целей. Я был очень горд за него и восхищался его настойчивостью и упорством.

Также мы обсудили планы на будущее. Мой друг рассказал о своих мечтах и целях, и я был уверен, что он сможет их достичь. Я пожелал ему удачи и выразил надежду, что в скором времени мы снова встретимся и сможем поделиться своими новыми достижениями.

Наша встреча была очень тёплой и душевной. Мы поняли, что несмотря на то, что прошло много времени, наша дружба осталась такой же крепкой и надёжной. Мы обещали друг другу, что будем поддерживать связь и встречаться как можно чаще.

Я благодарен судьбе за то, что она свела меня с таким замечательным человеком, как мой друг. Я уверен, что наша дружба будет длиться долгие годы и станет примером для других людей.
```

Длина контекста
9 [2-30]

Порог срабатывания
600 [400-1000]

Коэффициент учета частотности слов
50 [0-100]

Обработать

Planetcalc

Выполняет базовый анализ, не ищет повторяющихся блоков

Свежий взгляд

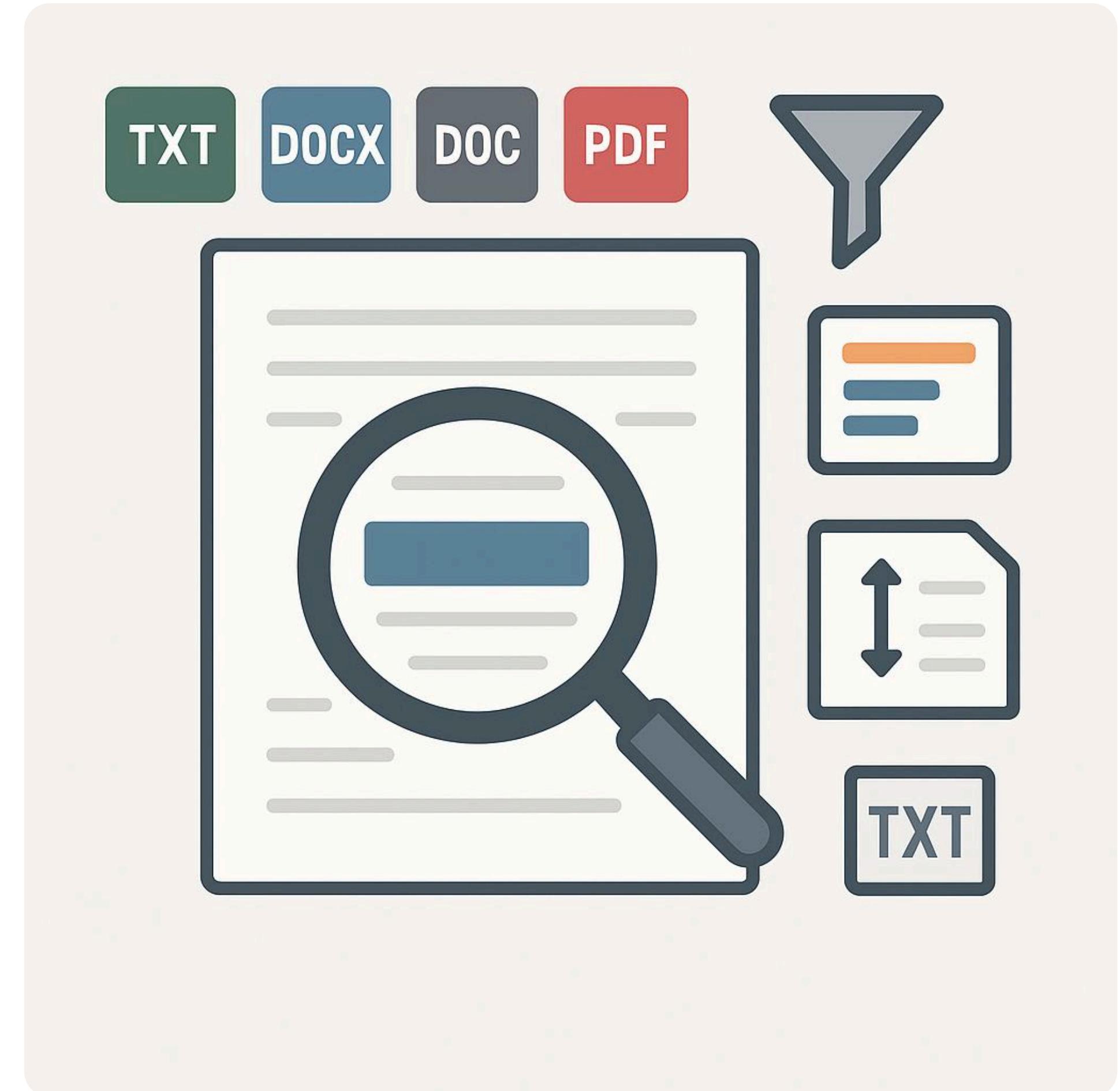
Не имеет гибких настроек

Вывод

Существующие решения не дают в достаточной степени автоматизированного инструмента по поиску повторяющихся фрагментов текста

ПОЛЬЗОВАТЕЛЬСКИЕ ТРЕБОВАНИЯ

- Обработка TXT, DOCX, DOC и PDF
- Настройка длины искомого фрагмента и фильтрация по начальному символу
- Сортировка результатов (по частоте повторений, алфавиту)
- Копирование и экспорт результатов (TXT)



СРЕДСТВА РЕШЕНИЯ ЗАДАЧИ

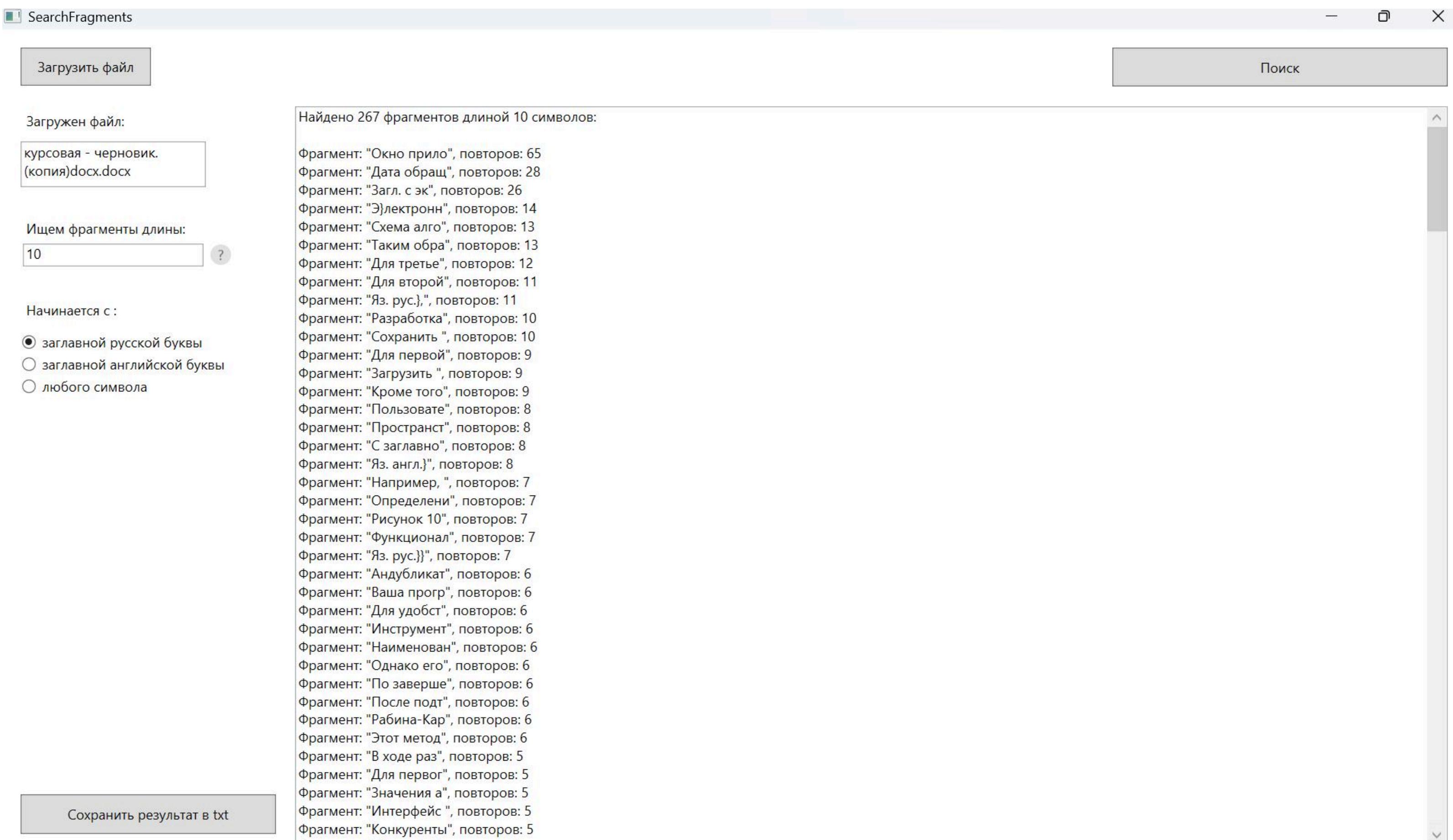
- Разработка программного продукта осуществляется на языке C# в рамках WPF-проекта;
- Для работы с файлами используются классы из пространства имен Microsoft.Win32, такие как OpenFileDialog и SaveFileDialog.
- Чтение текстовых данных из файлов формата TXT в проекте осуществляется с помощью метода File.ReadAllText из пространства имен System.IO.
- При работе с документами формата DOCX используется библиотека DocumentFormat.OpenXml.
- В случае работы с более старым форматом DOC, для корректной обработки текста используется библиотека Aspose.Words.
- Для обработки PDF-документов в проекте был выбран пакет iText7.
- Для выполнения операций по фильтрации, сортировке и обработке коллекций повторяющихся фрагментов применяется LINQ (System.Linq).
- Для корректного выделения текстовых фрагментов используются встроенные методы анализа символов Char.IsUpper и Char.IsWhiteSpace.

СРЕДСТВА РЕШЕНИЯ ЗАДАЧИ

Таблица сравнения алгоритмов поиска повторяющихся фрагментов

Алгоритм	Преимущества	Ограничения
Полный перебор (Brute Force)	Простая реализация	Низкая производительность на больших объемах данных
Хеш-функции (Рабин-Карп)	Быстрое нахождение совпадений, сокращает количество сравнений	Возможны коллизии, сложность работы с фрагментами произвольной длины
Суффиксный массив + LCP	Высокая эффективность, ускоренный поиск повторяющихся последовательностей	Требует больших вычислительных ресурсов и управления памятью
Гибридный метод (выбранный подход)	Оптимальный баланс между скоростью и гибкостью поиска	Сложнее в реализации по сравнению с Brute Force

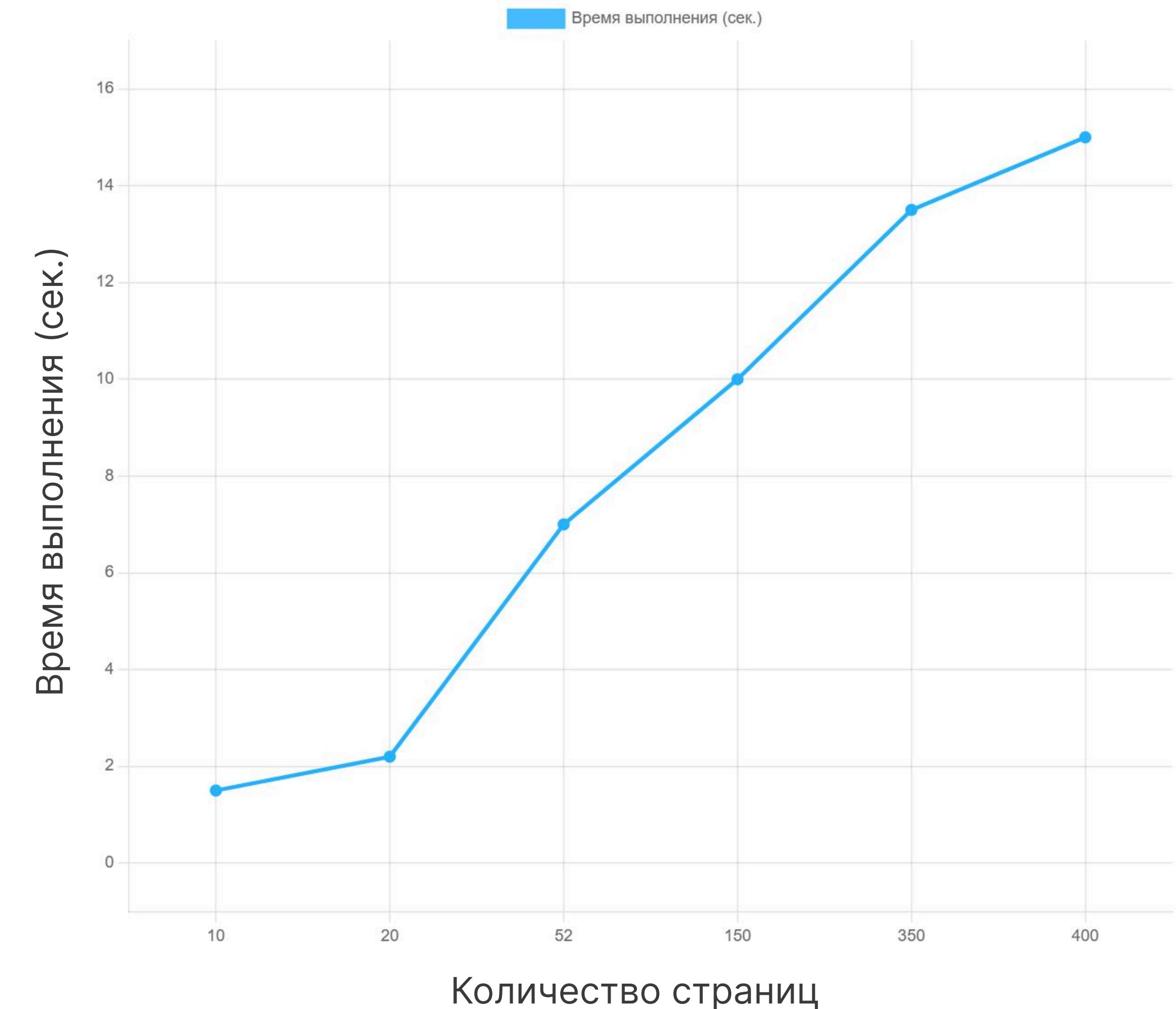
ИНТЕРФЕЙС ПРОГРАММЫ



В оконном приложении размещены следующие элементы управления: три кнопки **Button**, три метки **Label**, три текстовых поля **TextBox**, три переключателя **RadioButton**, одно текстовое поле с **ToolTip**. В окне отображения результата предусмотрены полосы прокрутки **VerticalScrollBarVisibility** и **HorizontalScrollBarVisibility**.

БЫСТРОДЕЙСТВИЕ ПРОГРАММЫ

Анализ результатов показывает, что при работе на стандартном офисном компьютере с большими объёмами текста, например, диссертацией более 350 страниц, время обработки составляет 13,5 секунды; для 40–70 страниц — 7 секунд, а для 10–20 страниц — 2,2 секунды.



ЗАКЛЮЧЕНИЕ

В ходе работы был разработан инструмент для поиска повторяющихся фрагментов текста с заданными параметрами. Реализованы загрузка файлов различных форматов, применение фильтров, сужающих поиск, и выгрузка результатов в удобном виде. Время обработки данных не превышает 14 секунд для книг обычных размеров (до 500 стр.) с распознаваемым текстовым слоем.

Практическая значимость разработанного ПО подтверждена его использованием при работе над совместной монографией сотрудников СГУ и ИПУ им. В.А. Трапезникова РАН (350 стр.).

Дальнейшее развитие: введение новых критериев поиска, поддержка семантического анализа текстов и реализация локальной проверки новых документов на повторы текста.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Маликов, А. В. Алгоритм обнаружения фактов дублирования информации в документированных результатах самостоятельной учебной деятельности студентов, устойчивый к незначительным изменениям текста / А. В. Маликов, А. С. Целиковский // Известия вузов. Северо-Кавказский регион. Серия: Технические науки. — 2011. — № 4. — С. 40–42.
2. Игнатов, Д. И. Разработка и апробация системы поиска дубликатов в текстах проектной документации / Д. И. Игнатов, С. О. Кузнецов, В. Б. Лопатникова, И. А. Селицкий // Бизнес-информатика. — 2008. — № 4. — С. 21–28.
3. Смит, Б. Методы и алгоритмы вычислений на строках / Б. Смит. — Москва: ООО “И.Д. Вильямс”, 2006. — С. 496.
4. Кудрина, Е. В. Основы алгоритмизации и программирования на языке C# / Е. В. Кудрина, М. В. Огнева. — Профессиональное образование. Москва: Издательство Юрайт, 2025. — С. 322.

ПРИЛОЖЕНИЕ

Включает следующие ключевые материалы:

- 1. Алгоритмы для решения**
- 2. Руководство пользователя**

АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ

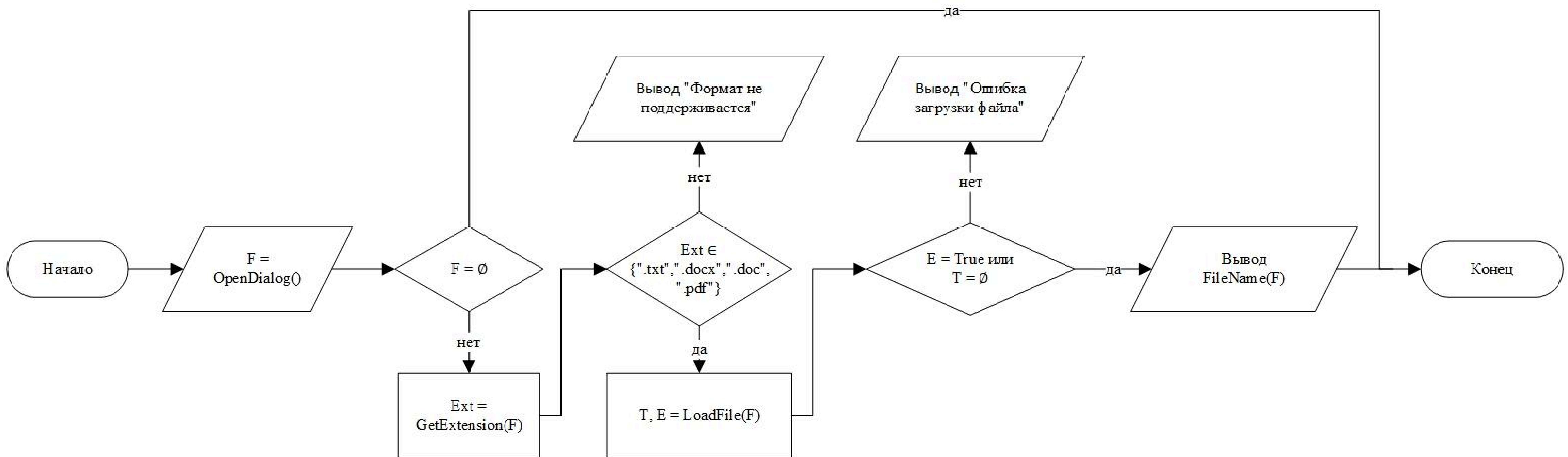


Схема алгоритма загрузки файла

АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ

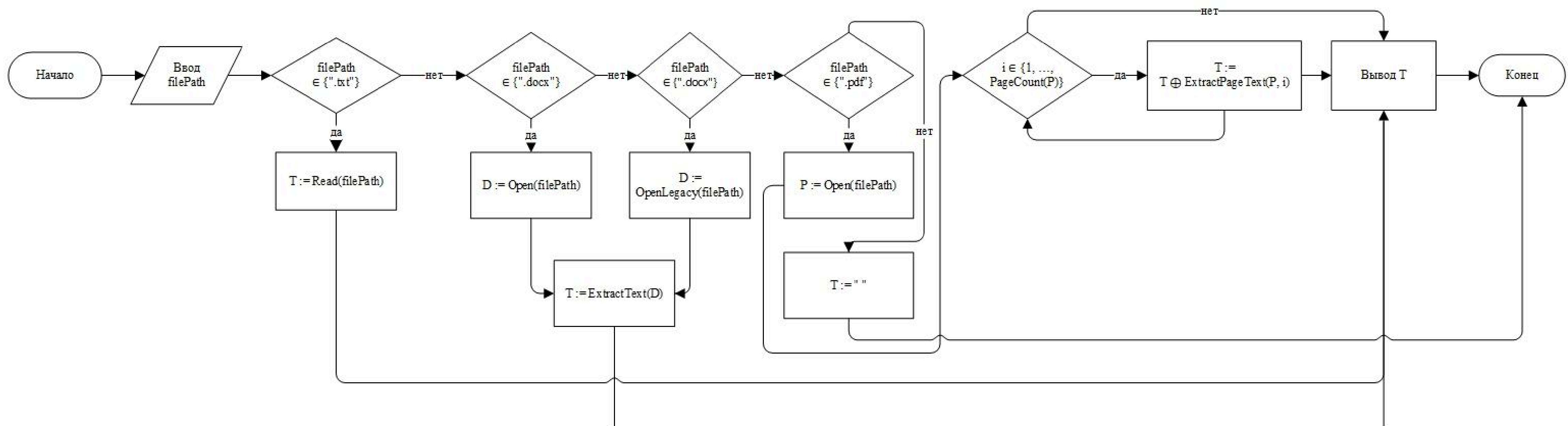


Схема функции загрузки файла

АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ

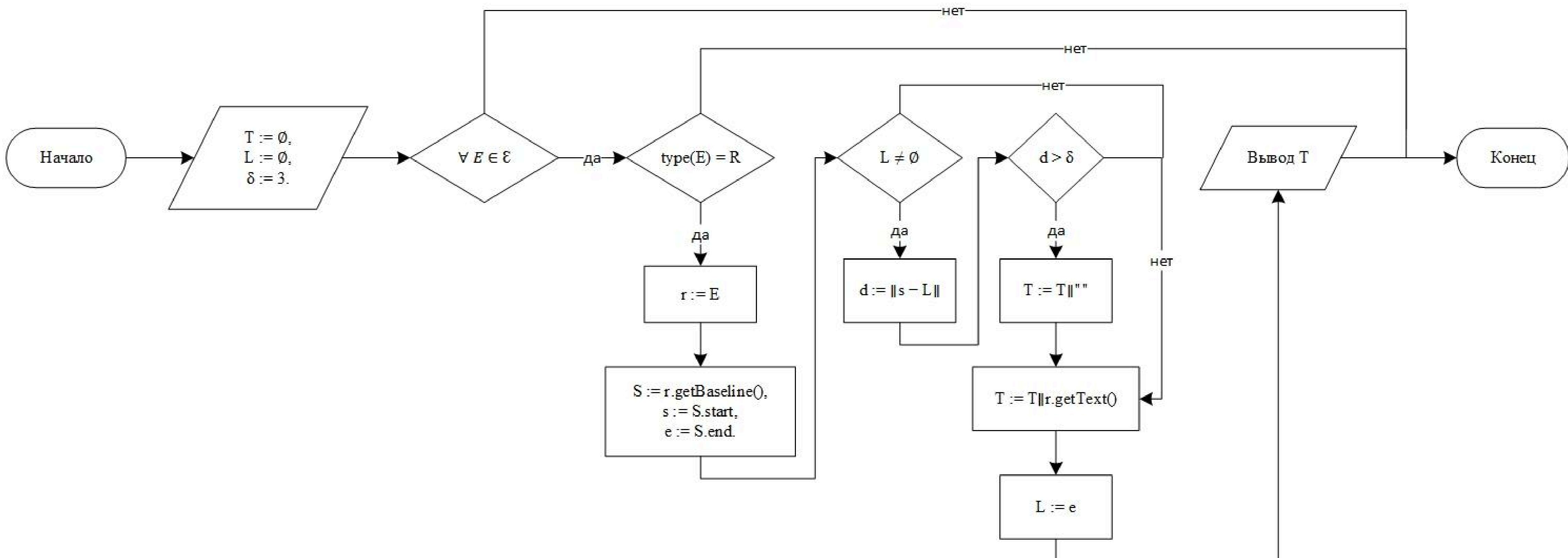


Схема алгоритма извлечения текста из PDF

АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ

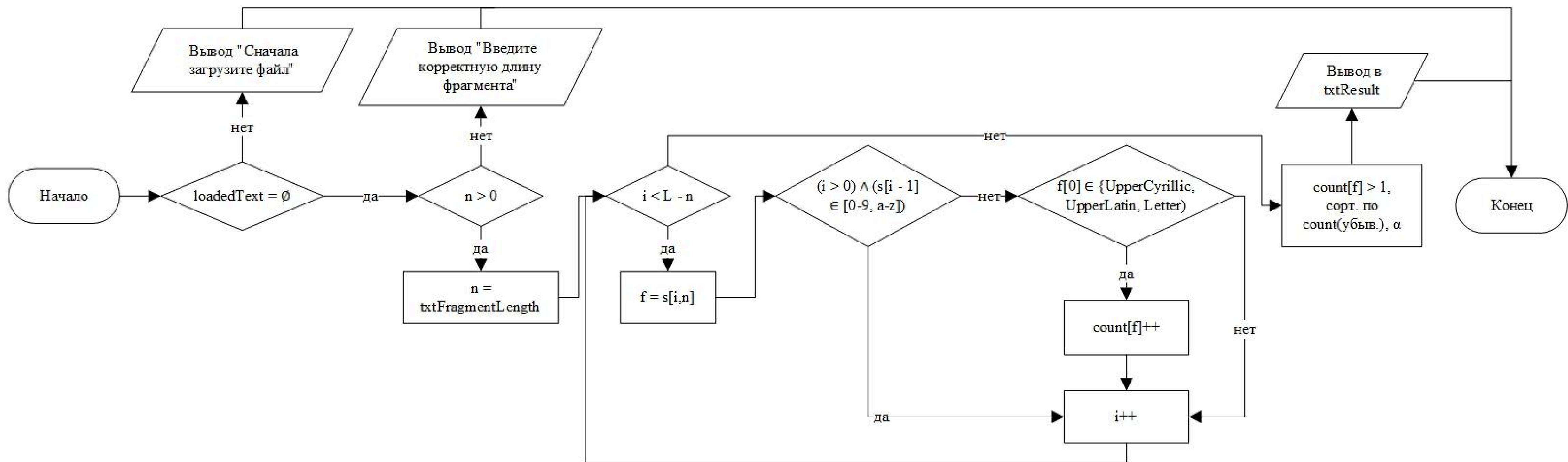


Схема алгоритма поиска повторяющихся фрагментов

АЛГОРИТМЫ ДЛЯ РЕШЕНИЯ

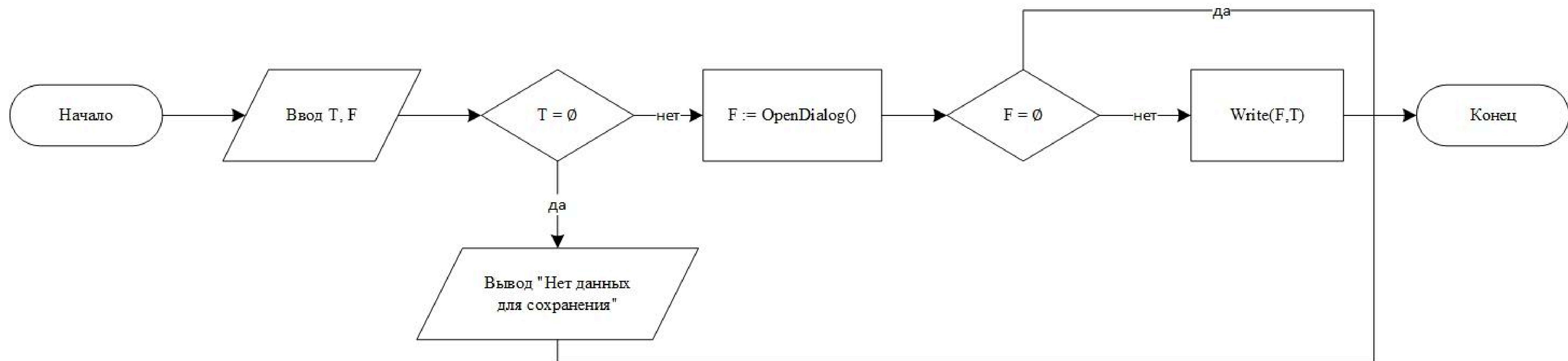


Схема алгоритма сохранения результатов

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

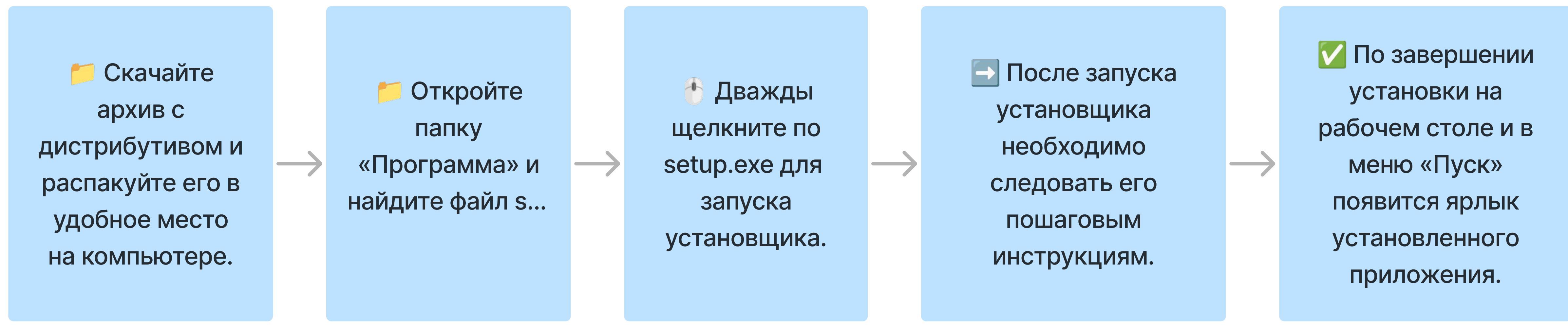


Схема установки программы

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

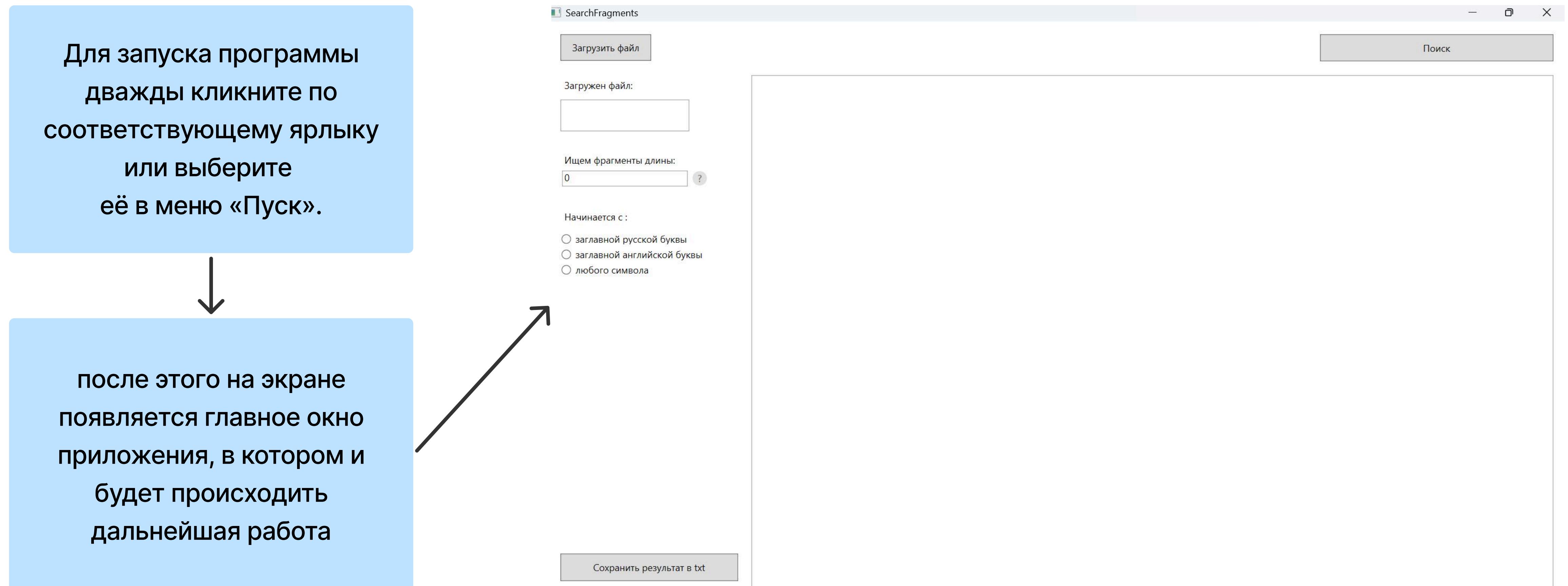


Рисунок 5 – Окно приложения «SearchFragments»: начальный запуск

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

При нажатии на кнопку
Загрузить файл
открывается диалоговое
окно с выбором файла

↓

При корректной загрузке
имя файла отобразится в
соответствующем поле ввода

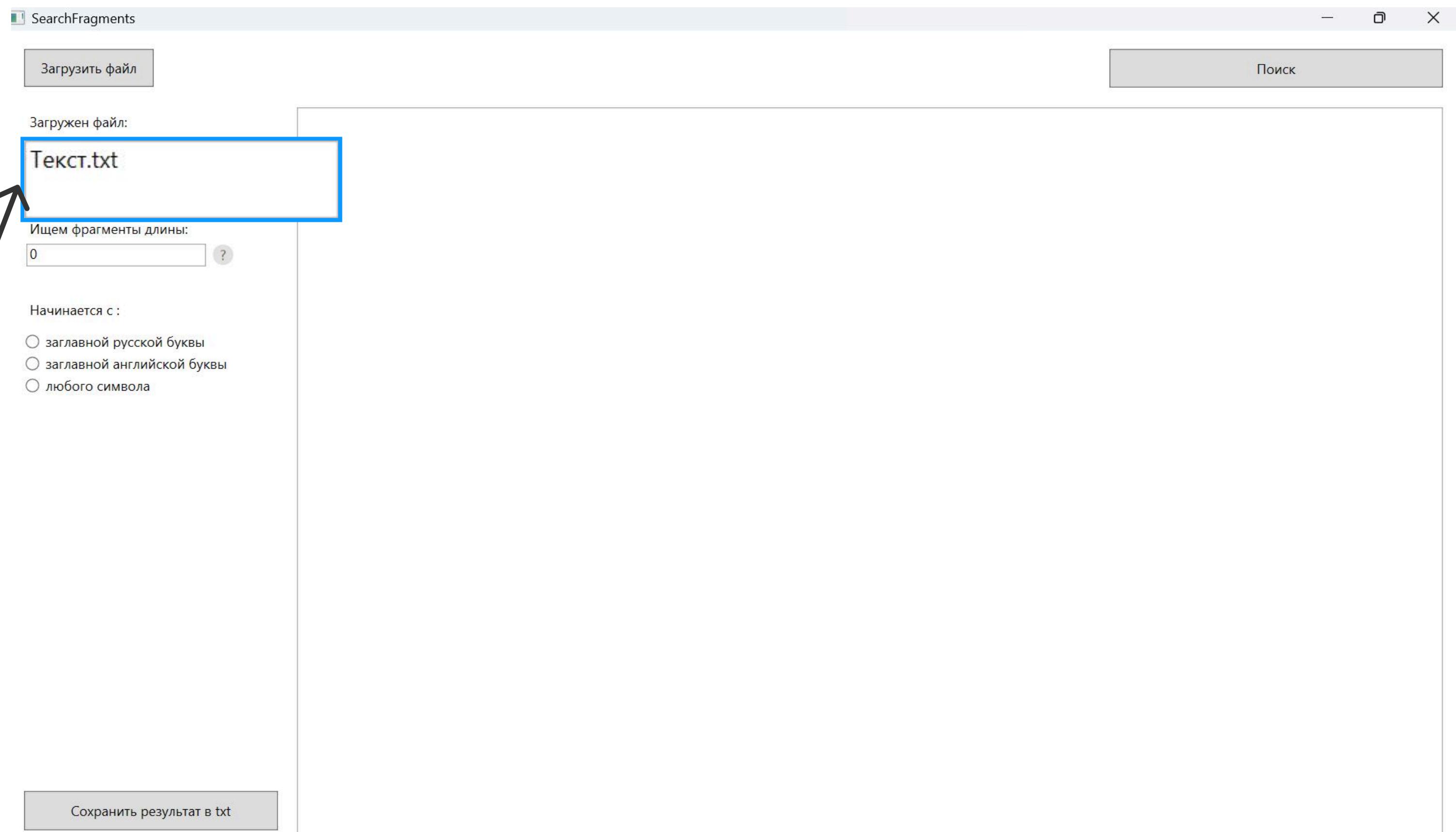


Рисунок 6 – Окно приложения «SearchFragments»: загрузка файла

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

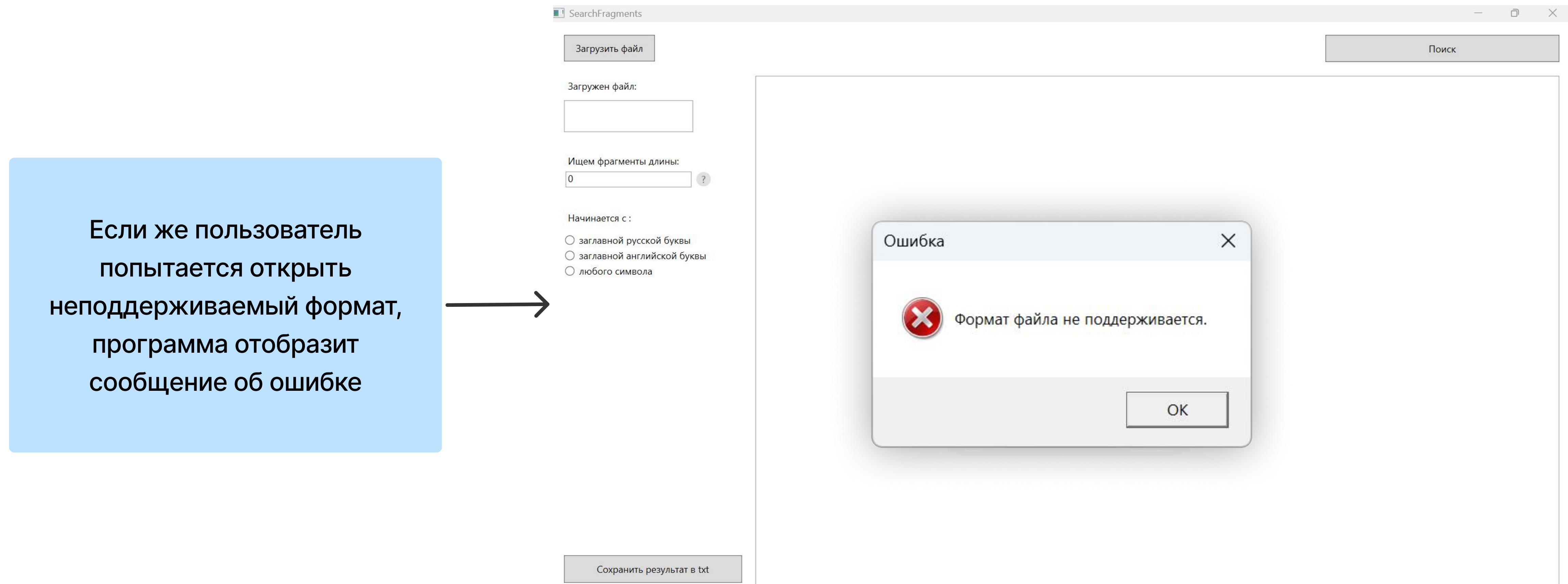


Рисунок 7 – Окно приложения «SearchFragments»: сообщение о неподдерживаемом формате загруженного файла

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

в поле «Длина фрагмента»
указывается требуемое
количество символов



Далее выбирается с какого
символа начинается фрагмент

- С заглавной русской буквы
- С заглавной английской буквы
- Любой символ

Затем по нажатию кнопки
Поиск приложение
анализирует содержимое файла
и отображает в поле результат

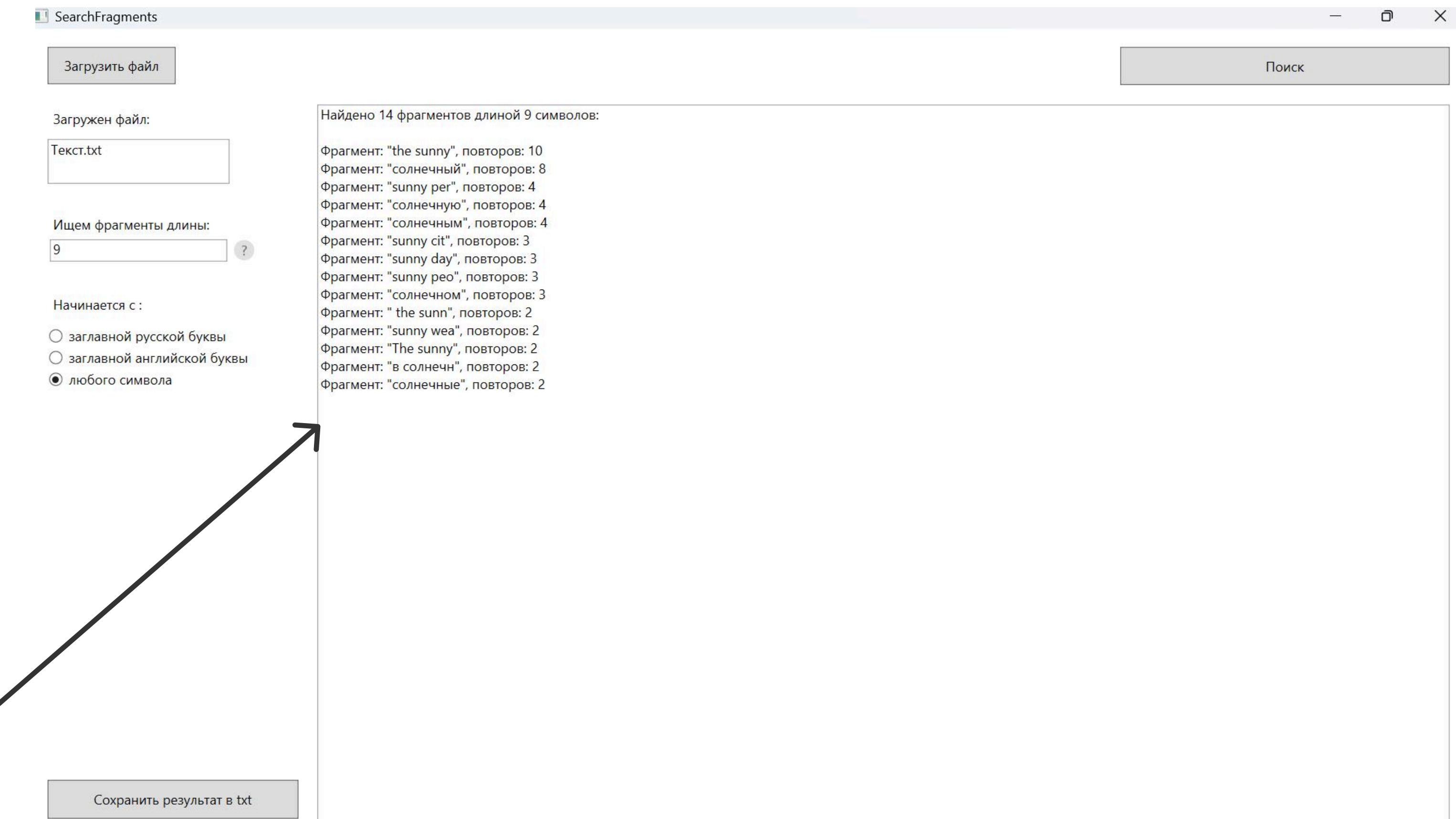


Рисунок 7 – Окно приложения «SearchFragments»: сообщение о неподдерживаемом формате загруженного файла

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

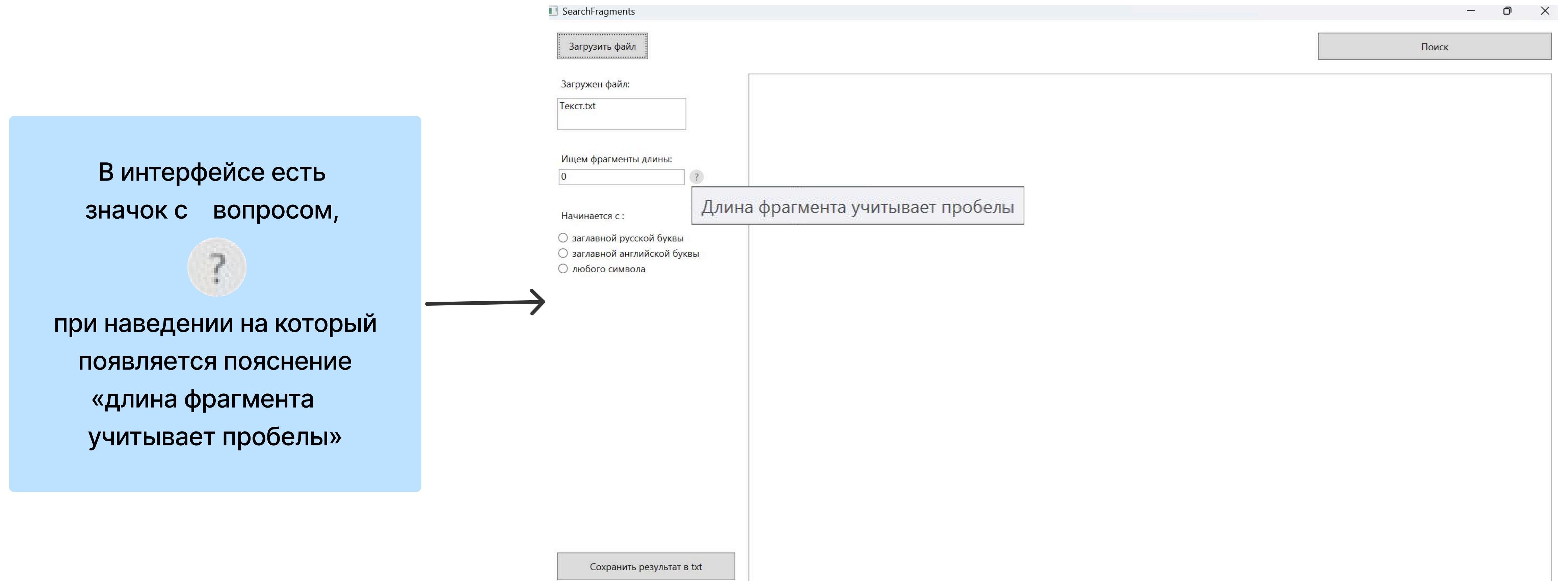


Рисунок 11 – Окно приложения «SearchFragments»: результат, когда повторяющихся фрагментов не найдено

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

Если в тексте нет повторений,
появляется уведомление
«Найдено 0 фрагментов N-ого
количество символов»

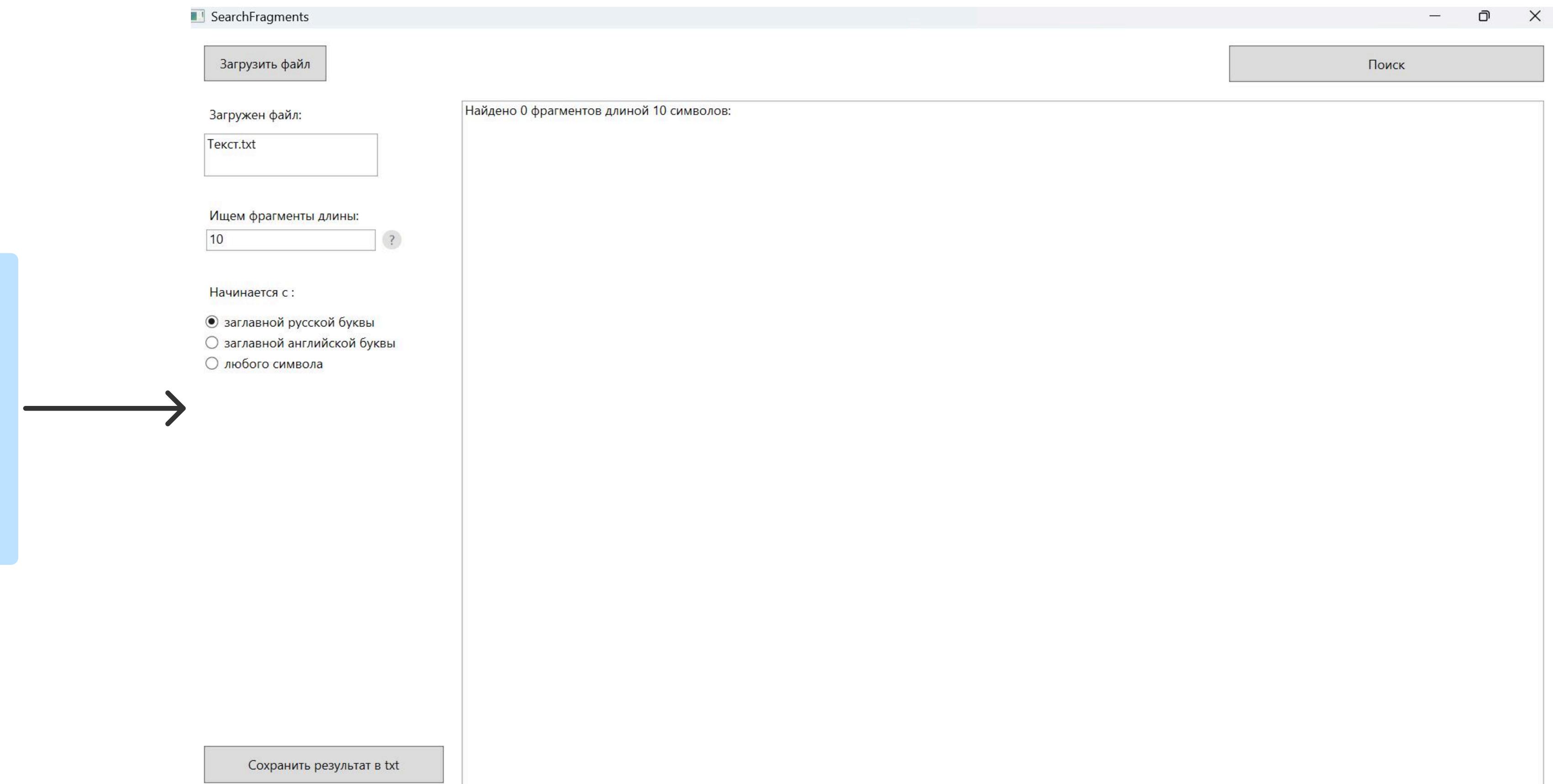


Рисунок 11 – Окно приложения «SearchFragments»: результат, когда повторяющихся фрагментов не найдено

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

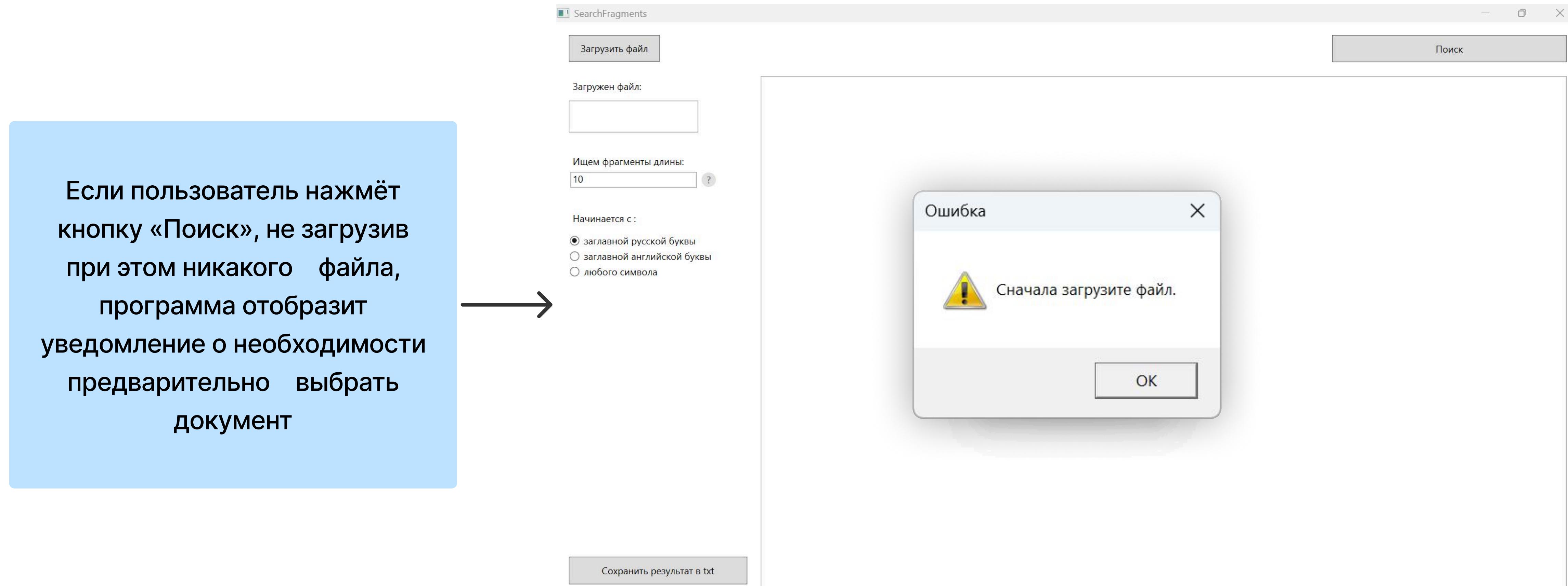


Рисунок 12 – Окно приложения «SearchFragments»: сообщение о просьбе пользователя перед работой загрузить файл

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

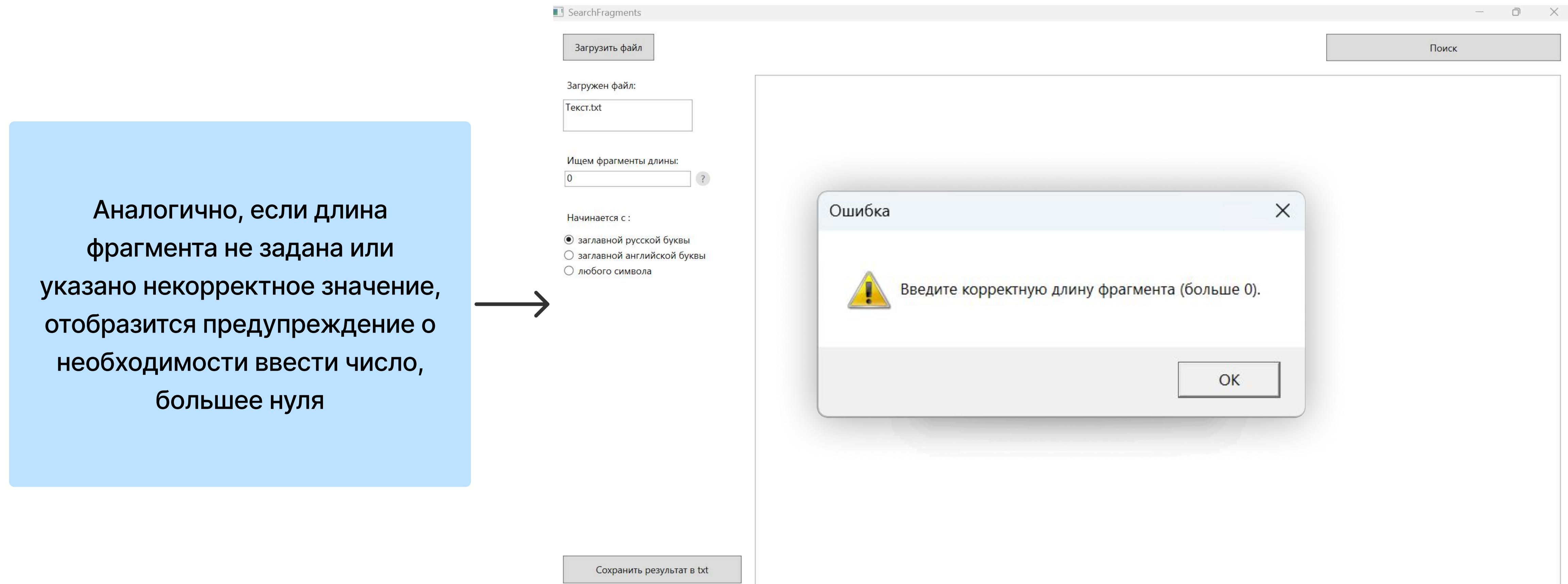


Рисунок 13 – Окно приложения «SearchFragments»: сообщение о просьбе пользователя перед работой ввести длину фрагмента больше нуля

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

В ситуациях, когда результат поиска оказывается слишком объёмным и не умещается в видимой области окна, предусмотрена полоса прокрутки ScrollViewer для просмотра всего списка

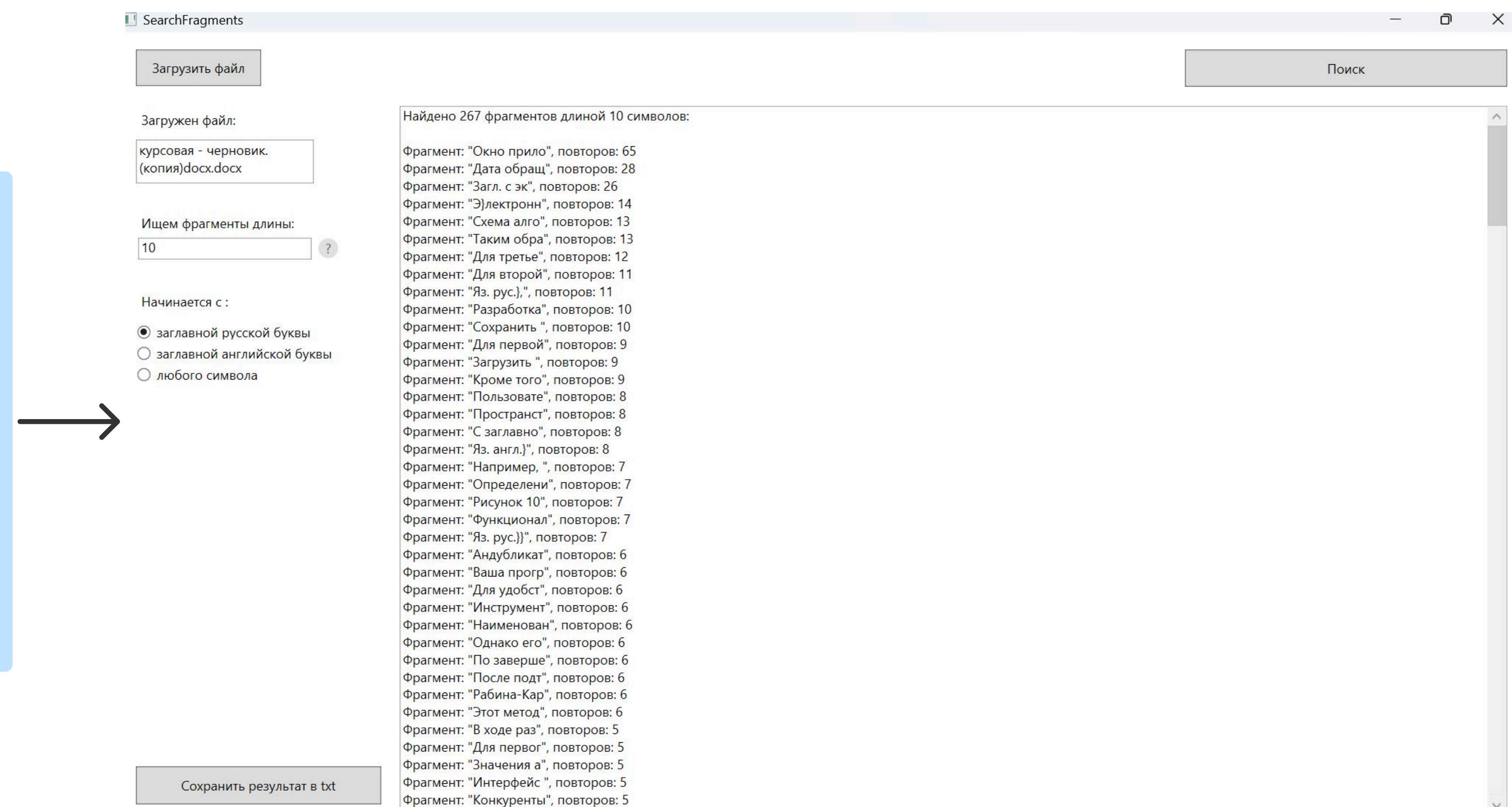


Рисунок 14 – Окно приложения «SearchFragments»: появление полосы прокрутки для просмотра всего списка

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

Завершив анализ, пользователь может
сохранить полученные данные,
нажав кнопку

Сохранить результат в txt



Появится диалоговое окно,
где следует указать имя файла
и выбрать место сохранения