

Инструкция по использованию программы `extract_pages`

Описание

`extract_pages` — это утилита для извлечения страниц из PDF документа на основе поиска текста. Программа ищет строки в PDF файле, извлекает страницы с найденным текстом и сохраняет их в новый PDF файл.

Требования

- `pdftgrep` — инструмент для поиска текста в PDF файлах.
- `pdftk` — инструмент для работы с PDF файлами (например, для извлечения страниц).

Подготовка

1. Убедитесь, что в той же папке, где находится PDF файл, существует текстовый файл с тем же именем, но с расширением `.txt`. Этот текстовый файл должен содержать строки, которые нужно искать в PDF файле.

Например, если ваш PDF файл называется `document.pdf`, то текстовый файл должен называться `document.txt`.

2. В текстовом файле каждая строка должна содержать одну строку текста, которую нужно искать в PDF документе.

Использование

1. **Базовое использование:** Чтобы извлечь страницы, на которых содержатся искомые строки, выполните следующую команду:

```
bash
Копировать код
./extract_pages.sh document.pdf
```

В результате будет создан файл `extracted_pages.pdf`, содержащий все страницы, на которых найдены искомые строки.

2. **Использование с опцией `-back`:** Если вам нужно извлечь не только страницы с найденным текстом, но и следующие за ними страницы, используйте опцию `-back`:

```
bash
Копировать код
./extract_pages.sh -back document.pdf
```

Этот вариант извлечет страницы с найденным текстом и следующую за ними страницу и сохранит их в файл `extracted_pages.pdf`.

Логирование

Все операции программы, включая найденные страницы и возможные ошибки, записываются в файл `script.log`, который создается или обновляется в текущей директории.

Примеры

1. Извлечение страниц с найденным текстом:

Например, если `document.txt` содержит строки:

```
Копировать код
Первая строка
Вторая строка
```

И вы запускаете команду:

```
bash
Копировать код
./extract_pages.sh document.pdf
```

Программа найдет и извлечет страницы, на которых есть "Первая строка" и "Вторая строка".

2. Извлечение страниц с найденным текстом и следующей за ними страницей:

Если вы используете:

```
bash
Копировать код
./extract_pages.sh -back document.pdf
```

Программа извлечет не только страницы с "Первой строкой" и "Второй строкой", но и страницы, которые следуют за ними.

Устранение неполадок

- Если программа не может найти строки в PDF файле, убедитесь, что текст в `document.txt` совпадает с текстом в PDF файле, включая регистр букв.
- Если программа не находит `document.txt`, убедитесь, что файл существует и находится в той же директории, что и PDF файл.

Завершение работы

После завершения работы программы все найденные страницы будут объединены в файл `extracted_pages.pdf`, который будет находиться в той же директории, что и исходный PDF файл.