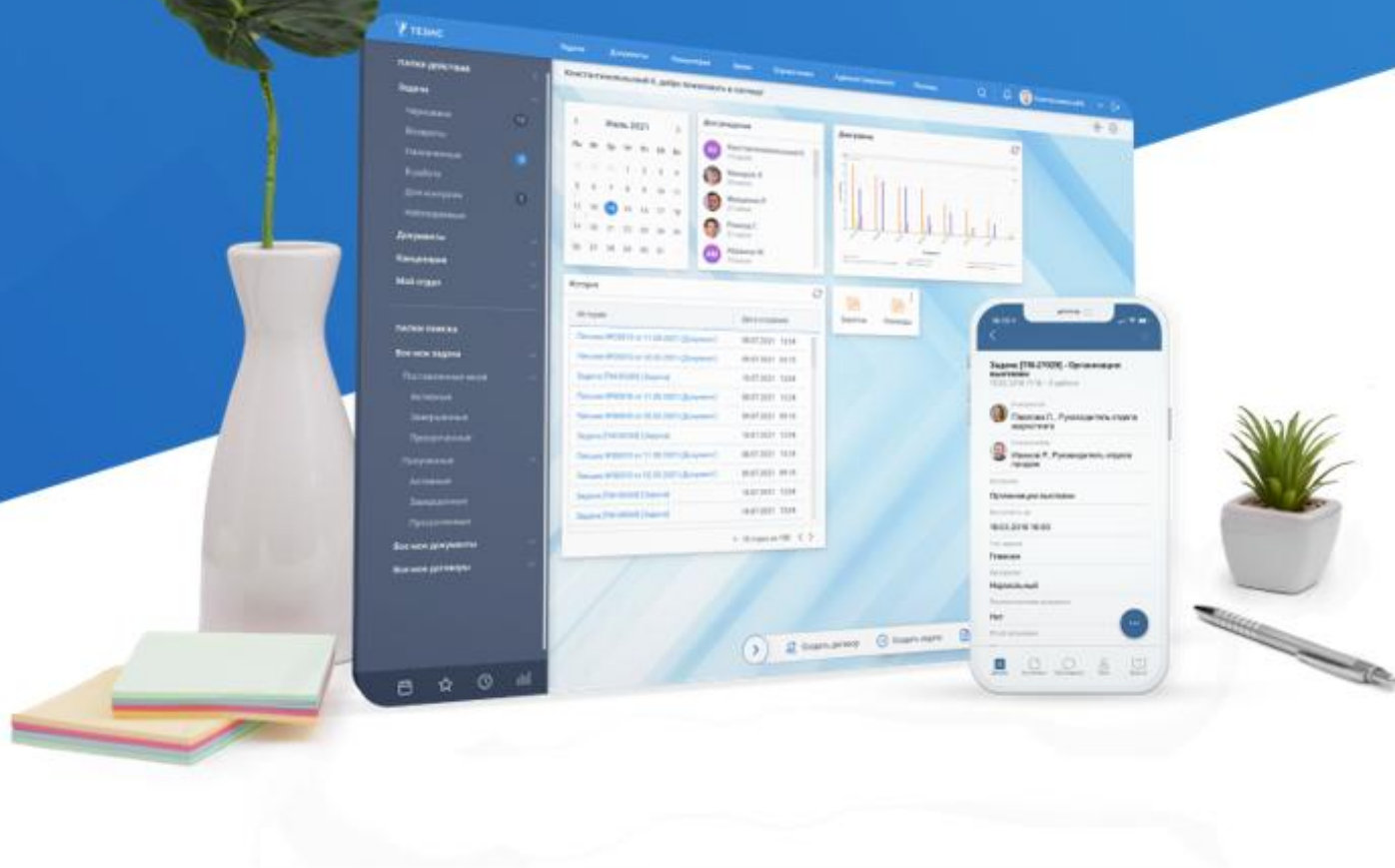




Установка и настройка Tesseract OCR

Памятка Администратора
Версия 5.4

г. Самара, 2025

Оглавление

1. УСТАНОВКА.....	2
1.1. Установка на Ubuntu (ОС Linux).....	2
1.2. Установка на ОС Windows	2
1.3. Настройка конфигурационных файлов	5

1. Установка

Tesseract OCR – это бесплатная система оптического распознавания текста (OCR), поддерживающая множество языков, включая русский. Результаты распознавания автоматически индексируются в СЭД ТЕЗИС, что позволяет выполнять полнотекстовый поиск по вложениям. Распознанный контент также может использоваться чат-ботом, ИИ-ассистентом и другими компонентами системы.

1.1. Установка на Ubuntu (OS Linux)

Приложение хранится в стандартных репозиториях ОС Linux.

Для его установки откройте терминал и выполните следующие команды:

```
sudo apt update  
sudo apt install tesseract-ocr
```

Для проверки установки выполните команду:

```
tesseract -version
```

Если выводится версия – установка прошла успешно.

Для установки языковых пакетов на русском языке выполните команду:

```
sudo apt install tesseract-ocr-rus
```

Для установки пакетов на других языках, например, на английском, немецком и французском выполните соответственно команды для каждого из языков:

```
sudo apt install tesseract-ocr-eng  
sudo apt install tesseract-ocr-deu  
sudo apt install tesseract-ocr-fra
```

Для проверки распознавания выполните команду:

```
tesseract test.png output -l rus
```

Результат будет в файле output.txt.

1.2. Установка на ОС Windows

Для установки на ОС Windows выполните следующие действия:

1. Скачайте файл установки.

Рекомендуем использовать версию 5.5.0, скачать её можно по [ссылке](#).

2. Запустите скачанный файл в формате *.exe и следуйте мастеру установки.

Выберите необходимый языковой пакет, например, русский и завершите установку.

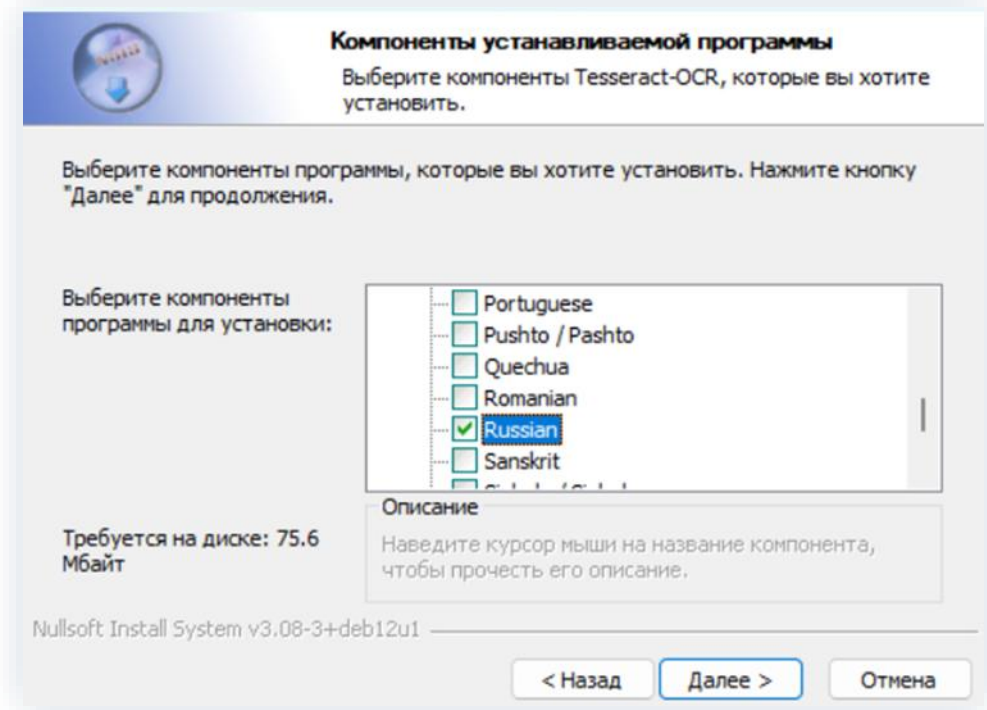


Рисунок 1. Мастер установки Tesseract-OCR

3. После установки Tesseract добавьте в системную переменную Path путь до файла tesseract.exe.

По умолчанию путь «C:\Program Files\Tesseract-OCR».

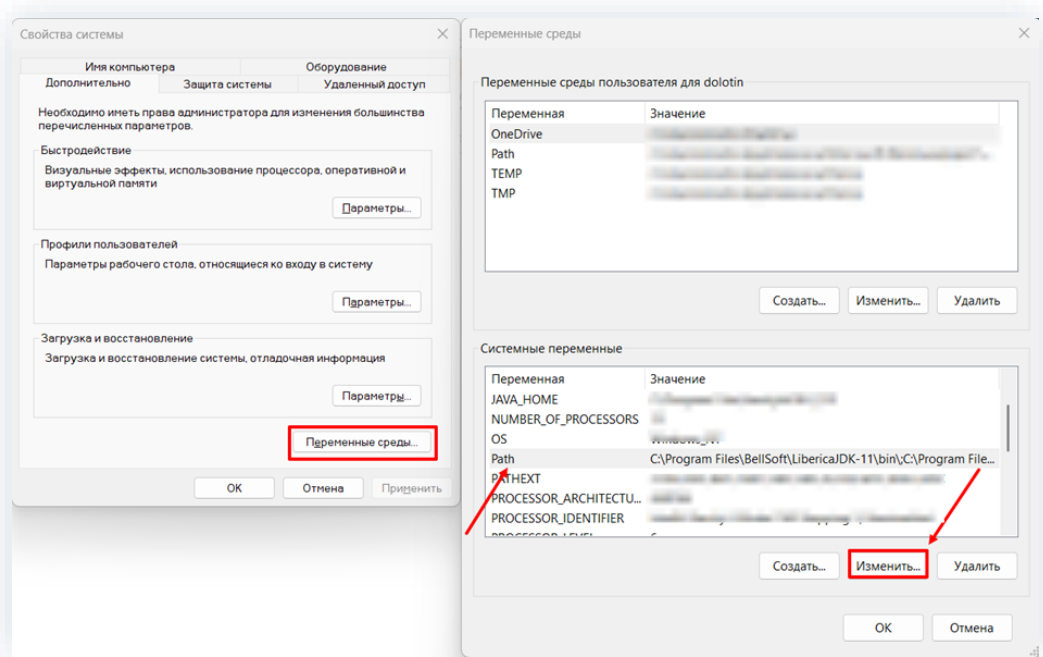


Рисунок 2. Переменные среды

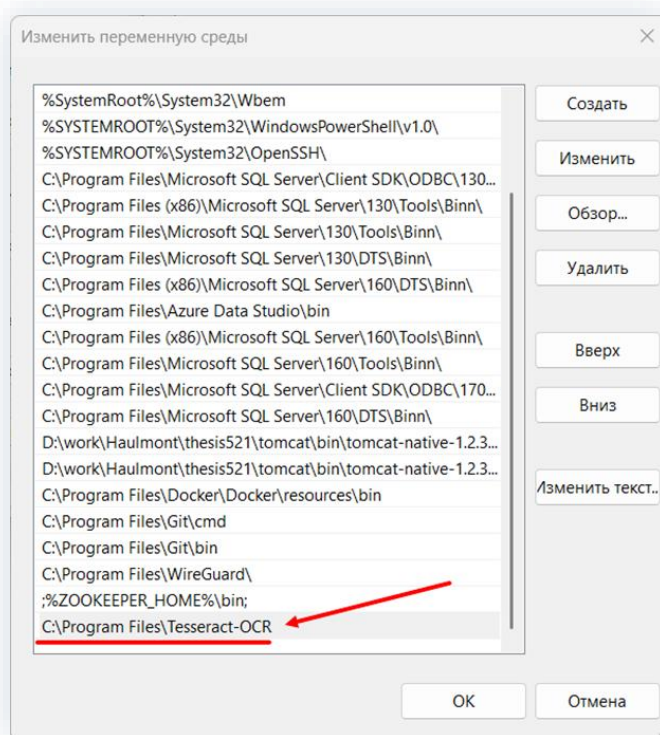
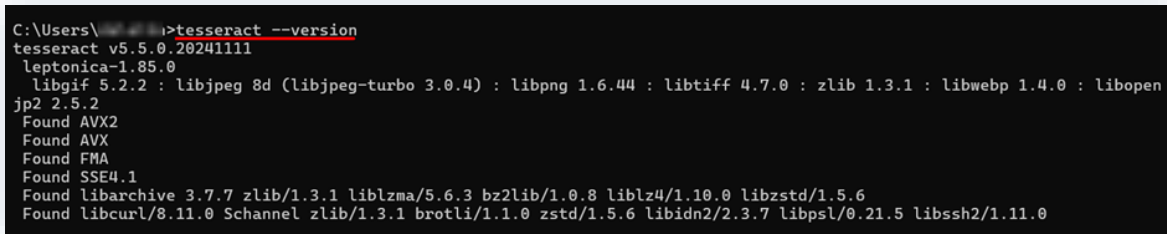


Рисунок 3. Изменение переменной среды

4. Перезапустите сервер для применения изменений из п.3.

- После перезагрузки в командной строке проверьте установку с помощью команды:

```
tesseract -version
```



```
C:\Users\...>tesseract --version
tesseract v5.5.0.20241111
leptonica-1.85.0
libgif 5.2.2 : libjpeg 8d (libjpeg-turbo 3.0.4) : libpng 1.6.44 : libtiff 4.7.0 : zlib 1.3.1 : libwebp 1.4.0 : libopen
jp2 2.5.2
Found AVX2
Found AVX
Found FMA
Found SSE4.1
Found libarchive 3.7.7 zlib/1.3.1 liblzma/5.6.3 b2lib/1.0.8 liblz4/1.10.0 libzstd/1.5.6
Found libcurl/8.11.0 Schannel zlib/1.3.1 brotli/1.1.0 zstd/1.5.6 libidn2/2.3.7 libpsl/0.21.5 libssh2/1.11.0
```

Рисунок 4. Окно командной строки

1.3. Настройка конфигурационных файлов

Для настройки конфигурационных файлов необходимо выполнить следующие действия:

- Установите значение в файле «tomcat\conf\app-core\local.app.properties» для следующих параметров:

```
thesis.fts.tesseract.enabled=true
```

- Сохраните изменения в файле.
- Перезапустите tomcat для применения изменений.