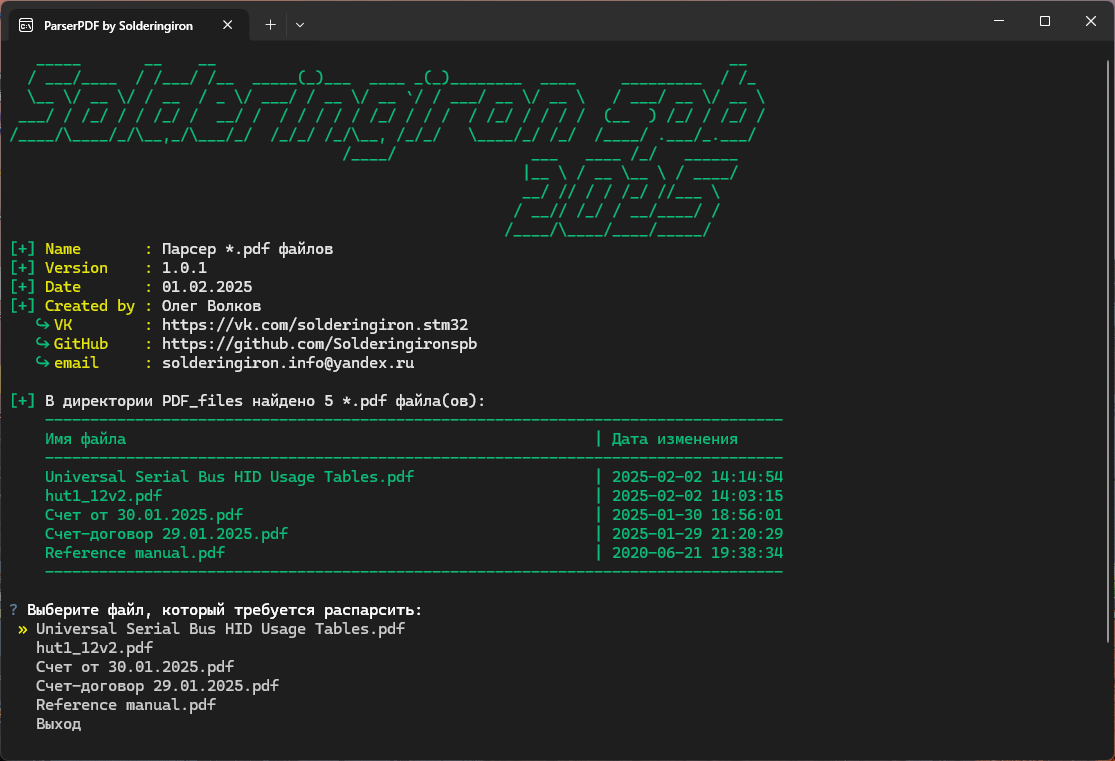
ИНСТРУКЦИЯ по работе с программой **ParserPDF by Solderingiron**



**ParserPDF by Solderingiron** является бесплатным ПО.

Программа предназначена для парсинга pdf файлов, а именно таблиц в них.

Скомпилирована только под Windows x64 системы.

***Работает только с pdf файлами, где данные не являются сканом.***

Для корректной работы программы требуются MS Office, ghostscript и java.

Если с MS Office все понятно, то для установки ghostscript и java напишу небольшую инструкцию:

1. **Установка ghostscript**

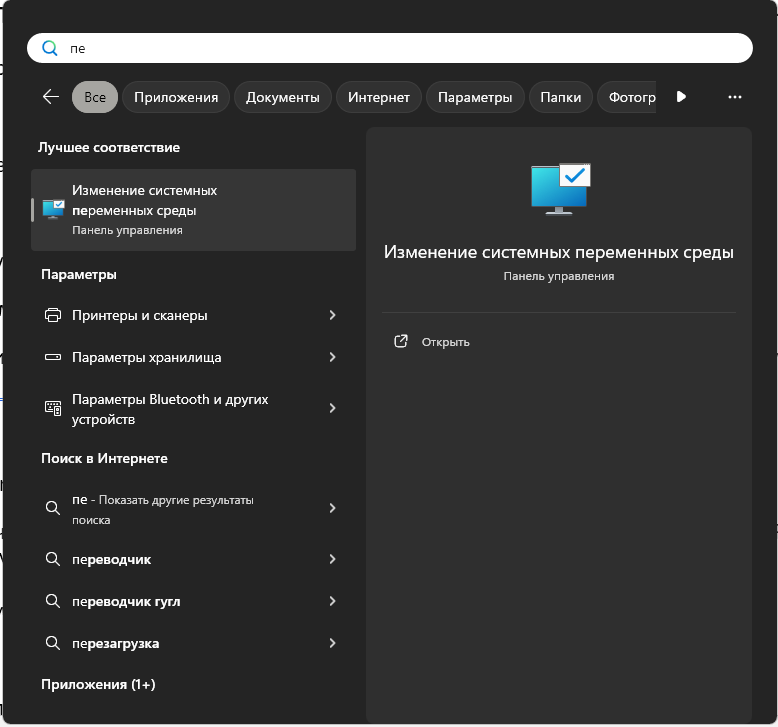
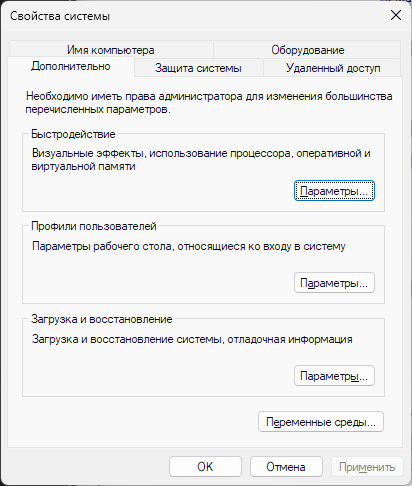
Установить ghostscript: <https://ghostscript.com/releases/gsdnld.html>

Вам нужен Ghostscript 10.04.0 for Windows (64 bit) -> [Ghostscript AGPL Release](https://github.com/ArtifexSoftware/ghostpdl-downloads/releases/download/gs10040/gs10040w64.exe)

1. **Добавление пути в PATH в Свойствах системы**

После установки нужно добавить ghostscript в Path:

Заходим в Изменение системных переменных среды->Переменные среды…



1. Вы увидите список переменных среды пользователя (вверху) и системных переменных (внизу). PATH присутствует в обоих расположениях.

Изменим переменные среды текущего пользователя:

нужная переменная Path->изменить

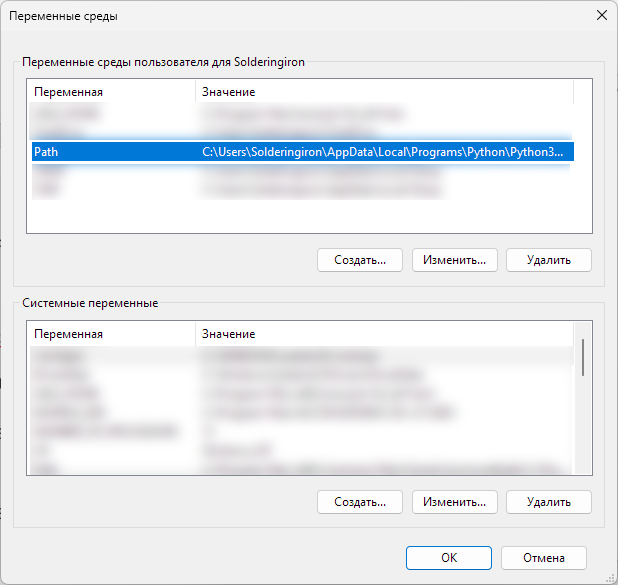
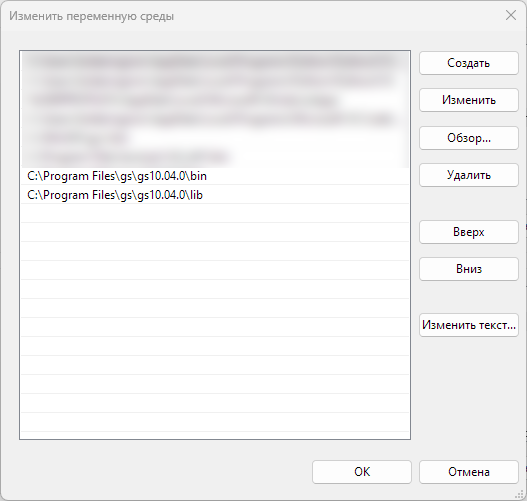
Для добавления нового пути нажмите «Создать», а затем впишите новый путь, который требуется добавить в переменную PATH в новой строке. Вместо нажатия «Создать» можно дважды кликнуть по новой строке для ввода нового пути.

Добавим:

C:\Program Files\gs\gs10.04.0\bin

C:\Program Files\gs\gs10.04.0\lib

**gs10.04.0** – это версия установленного ghostscript. В вашем случае может отличаться.

Далее чтоб избежать проблем с правильной работой ghostscript, переходим в директорию с ним:

C:\Program Files\gs\gs10.04.0\bin и переименуем gswin64c.exe в gs.exe

1. **Установка java**

<https://www.java.com/ru/download/manual.jsp>

Вам нужна [Windows Автономный режим (64-разрядная)](https://javadl.oracle.com/webapps/download/AutoDL?BundleId=251656_7ed26d28139143f38c58992680c214a5)

1. **Как и в прошлый раз, после установки, Вам нужно добавить пути в PATH системы(см. 2 и 3 пункты).**

Добавим:

C:\Program Files\Java\jre1.8.0\_441\bin

C:\Program Files\Java\jre1.8.0\_441\lib

jre1.8.0\_441 – это версия установленной java. В вашем случае может отличаться.

**Как пользоваться программой:**

**PDF\_files** - это папка, куда нужно складывать \*.pdf файлы для парсинга таблиц

**Excel\_files** - это папка, куда будут записываться распарсенные \*.pdf файлы и сохраняться в \*.xlsx

*По мере парсинга, все, что программа примет за таблицу - будет записываться на новую страницу*

*Excel документа: part\_1, part\_2 и т.д.*

В программу интегрировано 2 библиотеки парсинга: **Camelot** и **Tabula**.

Они используются для извлечения таблиц из PDF-документов, но имеют различия в подходах,

функциональности и качестве результатов.

### 1. **Подход к извлечению данных**

* **Camelot**:
  + Использует методы, основанные на анализе структуры PDF (например, распознавание линий и текста).
  + Поддерживает два основных метода:
    - **Lattice** (решетка): для таблиц с четкими границами (линиями).
    - **Stream** (поток): для таблиц без четких границ, где данные разделены пробелами.
  + Лучше справляется с таблицами, имеющими сложную структуру.
* **Tabula**:
  + Использует метод извлечения данных на основе текстового содержимого PDF.
  + Работает лучше с таблицами, где данные разделены пробелами или имеют простую структуру.
  + Может испытывать трудности с таблицами, где нет четкого разделения между ячейками.

### 2. **Точность и качество извлечения**

* **Camelot**:
  + Обычно дает более точные результаты, особенно для таблиц с границами.
  + Может корректно обрабатывать сложные таблицы, включая объединенные ячейки.
  + Лучше сохраняет структуру таблицы.
* **Tabula**:
  + Может терять структуру таблицы, если границы ячеек нечеткие.

### 3. **Производительность**

* **Camelot**:
  + Может быть медленнее при обработке больших PDF-файлов, особенно с использованием метода Lattice.
  + Требует больше ресурсов для анализа сложных таблиц.
* **Tabula**:
  + Обычно работает быстрее, но может уступать в точности для сложных таблиц.

### Когда использовать Camelot:

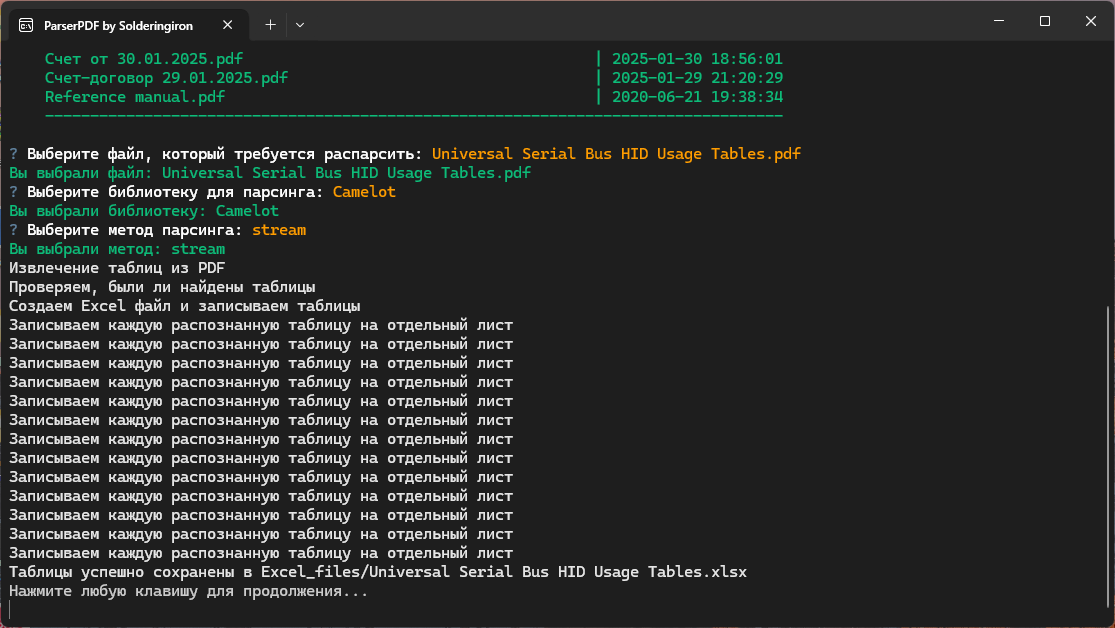
* Если таблицы в PDF имеют сложную структуру (например, объединенные ячейки или четкие границы).
* Если требуется высокая точность извлечения данных.

### Когда использовать Tabula:

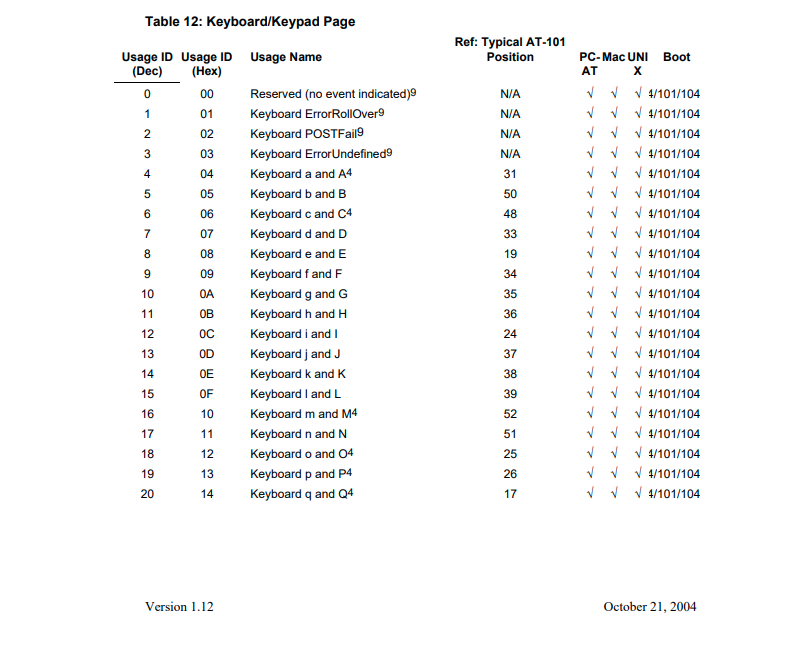
* Если таблицы простые и разделены пробелами.

*Если во время парсинга, при использовании одной библиотеки, Вы не получили результат или результат Вас не устроил – попробуйте использовать другую библиотеку.*

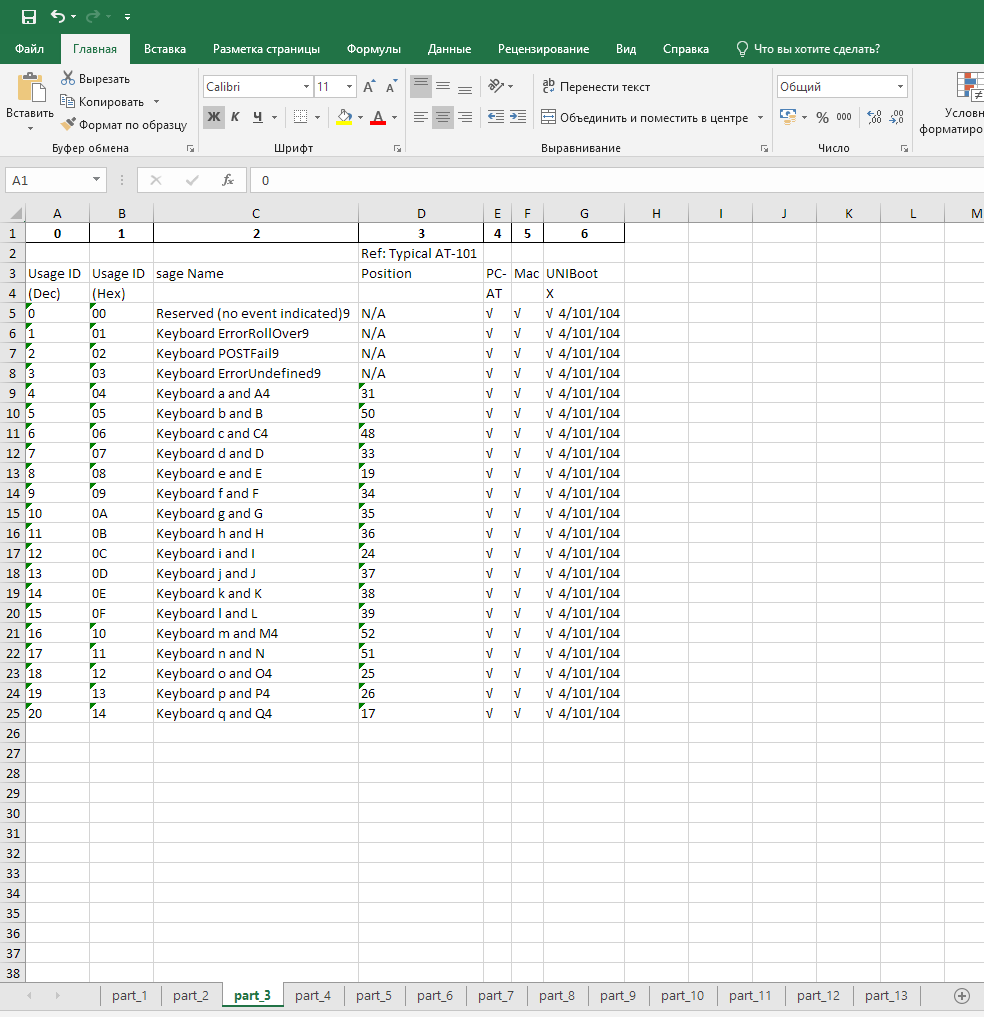
**Скриншоты:**

****

**Скриншот 1 - Окно программы. Парсим файл, предварительно выбрав библиотеку Camelot и метод stream**

****

**Скриншот 2 - Таблица в pdf файле, которую хотели бы перенести в Excel**

****

**Скриншот 3 – Результат парсинга таблиц из pdf файла**