

Хабр



КАК СТАТЬ АВТОРОМ



Зарплаты IT-специалистов



Войти



empenoso

10 мая 2023 в 05:42

## Как убрать пустые оборотные страницы из PDF после двухстороннего сканирования

Средний

6 мин

9.9K

Open source\*, PDF, Софт, Лайфхаки для гиков

Кейс

Около двух месяцев назад я написал статью [как сканировать многостраничные двухсторонние документы, если под рукой только обычный сканер с автоподачей](#), в которой затронул проблему того, что МФУ часто имеют дуплексную двустороннюю печать, но односторонний сканер.

Однако после решения проблемы быстрого сканирования больших двухсторонних документов, была обнаружена ещё одна проблема — некоторое количество страниц могут оказаться односторонними. И это означает, что PDF будет иметь белые страницы, например, со сканами перфораций или отверстий под кольца.

Конечно, можно удалить несколько страниц из PDF вручную, но что если таких файлов сотни, а сами документы имеют несколько десятков или даже сотен страниц как на фотографии?

РЕКЛАМА

**Получи грант за код**

Конкурс open source проектов



Большой многостраничный документ

[▶ TL;DR](#)

## Вариант удаления пустых страниц из pdf при помощи локальной программы

Перед тем как начать писать свой скрипт я честно пытался разобраться как удалить пустые страницы из пдф при помощи штатных средств какой-нибудь программы:

1. Пытался сделать это при помощи бесплатной открытой [PDFsam Basic](#), которая доступна под Linux и Windows, и MacOS, потому что в интернете нашёл инструкции, но они оказались устаревшими.
2. Пытался сделать это при помощи Adobe Acrobat Pro, но у меня не получилось. Делал по инструкции:

**Получи грант за код**

Конкурс open source проектов

2. Нажмите на вкладку «Инструменты» в верхней строке меню.
3. Выберите «Страницы» из списка инструментов справа.
4. Нажмите «Обрезать» в меню инструментов «Страницы».
5. В диалоговом окне «Обрезка страниц» выберите параметры «Удалить белые поля» и «Удалить белые поля для всех страниц».
6. Нажмите «ОК», чтобы применить изменения.

Эти действия должны были автоматически удалить все пустые страницы из файла PDF, но у меня этого не произошло.

*Adobe Acrobat Pro и удаление пустых страниц*



**Получи грант за код**

Конкурс open source проектов

1. Загрузите файл PDF: выберите «Файл» > «Открыть» или нажмите Ctrl + O на клавиатуре, затем найдите и выберите файл PDF, из которого вы хотите удалить пустые страницы.
2. После загрузки PDF-файла щелкните вкладку «Организация» на верхней панели инструментов.
3. Выбрав все страницы, нажмите кнопку «Удалить пустые страницы».

Прогресс пробежал, но пустые страницы оставались на месте для любых из трех вариантов.

PDF-XChange Editor

Использование локальной программы, конечно, было бы лучшим вариантом, потому что это гарантировало, что PDF-файлы останутся на компьютере, обеспечивая конфиденциальность и безопасность по сравнению с использованием онлайн-инструментов.



**Получи грант за код**

Конкурс open source проектов

Но раз с локальными инструментами у меня не пошло, решил попробовать онлайн сервисы.

Я смог найти несколько доступных онлайн-инструментов, которые могли бы помочь автоматически удалить пустые страницы из PDF-файла:

1. Sejda (<https://www.sejda.com/delete-pdf-pages>)
2. Smallpdf (<https://smallpdf.com/delete-pages-from-pdf>)
3. DeftPDF (<https://deftpdf.com/delete-pdf-pages>)

Ни в одном из них я не смог найти опцию автоматического распознавания пустых страниц, хотя в поисковике попадались ссылки на несуществующие сейчас страницы (pdf remove blank pages) этих сервисов.

Ну и конечно использование онлайн-инструментов может поставить под угрозу конфиденциальность и безопасность ваших документов.

## Вариант удаления пустых страниц из pdf при помощи локального bash скрипта и консольной программы PDFtk

После постигшей неудачи решил написать свой собственный скрипт который удалит пустые страницы из всех pdf файлов в текущем каталоге.

При изучении вопроса наткнулся на [большую дискуссию](#), где обсуждался вопрос как лучше [удалить пустые страницы из pdf при помощи командной строки](#). Предлагались разные методы, но у меня были все документы сканированные и это значит, что даже на пустом листе какая-то информация всё равно была — сканы отверстий под перешивку или просто грязь со сканера.

Решил что будет следующий алгоритм:

1. Разделяю PDF документ на отдельные файлы.
2. Страницы меньше определенного размера удаляю.
3. Склеиваю оставшиеся страницы обратно.



**Получи грант за код**

Конкурс open source проектов

После нехитрых манипуляций получился файл `blank_page_remover.sh` :

```
# Подробнее в статье Как убрать пустые оборотные страницы из PDF после двухстороннего с
# https://habr.com/ru/articles/733754/
# Михаил Шардин https://shardin.name/

#!/bin/bash
datetime=$(date +"%Y-%m-%d_%H-%M-%S")
# Создаём единый лог файл для всех действий и папку куда перемещаем вырезанные страницы
log_file="blank_page_remover_${datetime}.log"
touch $log_file
mkdir removed
# Перебираем все PDF файлы в текущем каталоге
for file in *.pdf; do
    echo "Работаем с $file..." >> "$log_file"
    # Разделяем PDF файл на отдельные страницы
    echo "Разделяем $file на отдельные страницы..." >> "$log_file"
    pdftk "$file" burst output "${file%.*}_pg_%04d.pdf" >> "$log_file" 2>&1
    # Удаляем файлы страниц, размер которых меньше чем XX килобайт
    echo "Удаляем файлы страниц, размер которых меньше чем 35 килобайт..." >> "$log_file"
    for page in "${file%.*}_pg_*.pdf; do
        size=$(wc -c < "$page")
        if [[ $size -lt 35000 ]]; then
            echo "Удаляем $page (размер: $size байт)..." >> "$log_file"
            mv "$page" "removed/"
            #rm "$page"
        fi
    done
    # Склеиваем оставшиеся страницы в новый файл
    echo "Склеиваем оставшиеся страницы в новый файл..." >> "$log_file"
    pdftk "${file%.*}_pg_*.pdf cat output "${file%.*}_без пустых.pdf" compress >> "$log_
    # Удаляем временные файлы
    echo -e "Удаляем временные файлы...\n" >> "$log_file"
    rm "${file%.*}_pg_*.pdf
done
```

Для работы скрипта понадобится Pdftk (сокращение от PDF Toolkit) — это инструмент



**Получи грант за код**

Конкурс open source проектов

## Как воспользоваться скриптом удаления пустых страниц из PDF документа

Чтобы выполнить сценарий bash на компьютере, выполните следующие действия в зависимости от операционной системы:

### Для Linux и macOS:

1. Откройте Терминал: нажмите `Ctrl + Alt + T` в Linux или откройте **Терминал** из папки **Приложения > Утилиты** в macOS.
2. Перейдите в каталог, где находится скрипт: используйте команду `cd`, за которой следует путь к каталогу. Например:  
`cd /путь/к/скрипту`
3. Сделайте скрипт исполняемым:  
`chmod +x blank_page_remover.sh`
4. Выполните этот сценарий. Запустите сценарий, введя `./`, а затем имя сценария:  
`./blank_page_remover.sh`
5. PROFIT!  
Скрипт создаст новые pdf файлы без пустых страниц и подробный лог действий.

Терминал в Ubuntu и результат выполнения скрипта `blank_page_remover.sh`

### Для Windows (используя GitBash или WSL):

1. Установите GitBash или WSL: если вы еще этого не сделали, установите [GitBash](#) или



**Получи грант за код**

Конкурс open source проектов

2. Откройте Git Bash или WSL: щелкните правой кнопкой мыши папку, содержащую скрипт, и выберите `GitBash здесь` или `Открыть в WSL`.
3. Сделайте скрипт исполняемым:  

```
chmod +x blank_page_remover.sh
```
4. Выполните этот сценарий. Запустите сценарий, введя `./`, а затем имя сценария:  

```
./blank_page_remover.sh
```
5. PROFIT!  
Скрипт создаст новые pdf файлы без пустых страниц и подробный лог действий.

Актуальная версия скрипта [всегда доступна на гитхабе](#).

## Заключение

Удаление пустых страниц из PDF-файлов после двустороннего сканирования может оказаться непростой задачей, особенно при работе с большими объемами документов. Тем не менее, эта статья предоставила вам решение в виде использования автоматического локального сценария bash с консольной программой PDFtk.

Следуя подробным инструкциям вы сможете эффективно избавиться от пустых страниц и поддерживать чистый профессиональный вид отсканированных PDF-документов.

Независимо от объема или сложности ваших файлов, это решение упростит ваш рабочий процесс и сэкономит ваше время и усилия.

**Автор:** Михаил Шардин

 [Моя онлайн-визитка](#)

 [Telegram «Умный Дом Инвестора»](#)

10 мая 2023 г.

**Теги:** [bash](#), [pdftk](#), [сканирование](#), [документы](#)

**Хабы:** [Open source](#), [PDF](#), [Софт](#), [Лайфхаки для гиков](#)



**Получи грант за код**

Конкурс open source проектов





Оставляя свою почту, я принимаю [Политику конфиденциальности](#) и даю согласие на получение рассылок

**212**

Карма

**63.3**

Рейтинг

**Михаил Шардин** @empenoso

Автоматизация / Данные / Финансы / Умные дома

[Подписаться](#)[Сайт](#) [Сайт](#) [GitHub](#)**Получи грант за код**

Конкурс open source проектов

 Комментарии 10

## Публикации

ЛУЧШИЕ ЗА СУТКИ

ПОХОЖИЕ



Tirarex

14 часов назад

### Как я делал сеть на 2,5 гигабита с минимальным бюджетом — апгрейд, доступный каждому



Простой



9 мин



11K

Тutorial



+41



83



34



Erwinmal

18 часов назад

### Сэндвич, сэр? История британских бутербродов от аристократических салонов до вокзальных буфетов



Простой



13 мин



3K

Ретроспектива



+41



20



2



oneastok

19 часов назад

### Умное зеркало на Raspberry Pi: пошаговое руководство



Простой



4 мин



4.5K

Обзор

Перевод



+22



56



15



iLushkersky

14 часов назад



Получи грант за код

Конкурс open source проектов

 +17 5 10**TrexSelectel**

16 часов назад

## Nintendo Virtual Boy: неожиданное возрождение виртуальной реальности из 90-х

 5 мин 1.2K +14 3 3**mio\_anni**

19 часов назад

## От мини-ЭВМ и перфокарт к IDE и фреймворкам. Как поменялось программирование за 50 лет — взгляд изнутри

 12 мин 1.9K +12 15 35**RED\_OS\_M**

19 часов назад

## Станислав Петров: «Ключевые отличия РЕД ОС М от Android – вовсе не в интерфейсе»

 Средний 8 мин 6.6K[Интервью](#) +10 10 43**Albert\_Wesker**

19 часов назад

## Миф о быстром и медленном пути выполнения программы

 Средний 11 мин 1.5K[Обзор](#)[Перевод](#)**Получи грант за код**

Конкурс open source проектов

**kilokanat**

5 часов назад

## Механическая клавиатура LARKeyboard

**Простой**

5 мин



541

Тutorial



+8



4



2

**beeline\_cloud**

10 часов назад

## Научный «дипфейк»? Как галлюцинации нейросетей — и другие проблемы — просачиваются в академические статьи

**Простой**

8 мин



905

Аналитика



+8



11



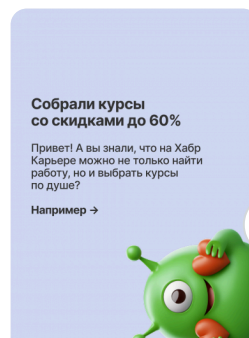
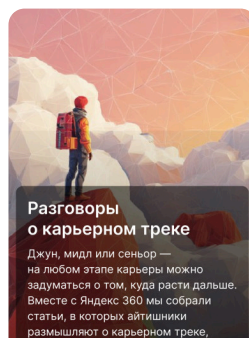
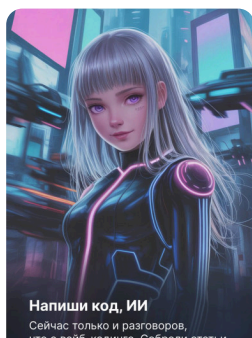
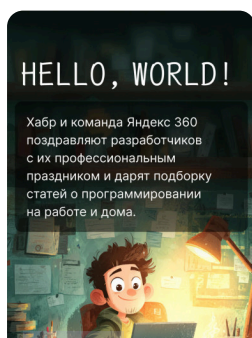
2

## «Раньше играл в Танки, сейчас веду кибертурниры» — истории современных студентов

Турбо

Показать еще

### ИСТОРИИ

**Получи грант за код**

Конкурс open source проектов

Чай, торт и код: с  
Днём программиста!

Made in AI

Чего хотят лиды в  
бигтехе?

Как расти в ИТ:  
советы, гайды и  
опыт сеньоров

Курсы со скидками  
до 60%

## ВАКАНСИИ

### SRE/DevOPS инженер

до 500 000 ₽ · Карма8 · Можно удаленно

### Linux администратор HPC стека

от 200 000 до 350 000 ₽ · Сбер · Москва

### Linux администратор HPC стека

от 200 000 до 400 000 ₽ · Сбер · Москва

### DevOps-инженер (Junior)

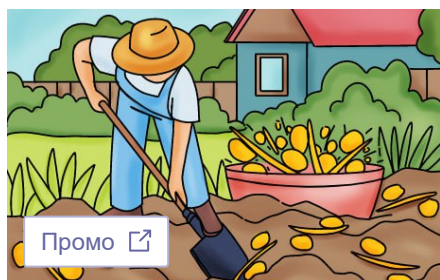
от 40 000 до 62 000 ₽ · DBI · Можно удаленно

### Программист C++

от 110 000 ₽ · Ридан · Москва

Больше вакансий на [Хабр Карьере](#)

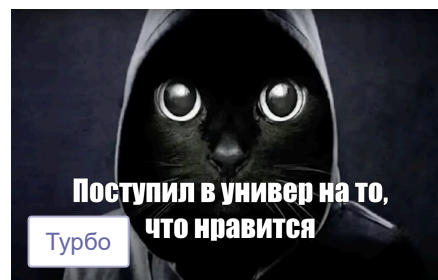
## МИНУТОЧКУ ВНИМАНИЯ



Не нужно копать картошку —  
хватит скидок в Промокодусе



Прямая трансляция IT Elements:  
ключевые доклады



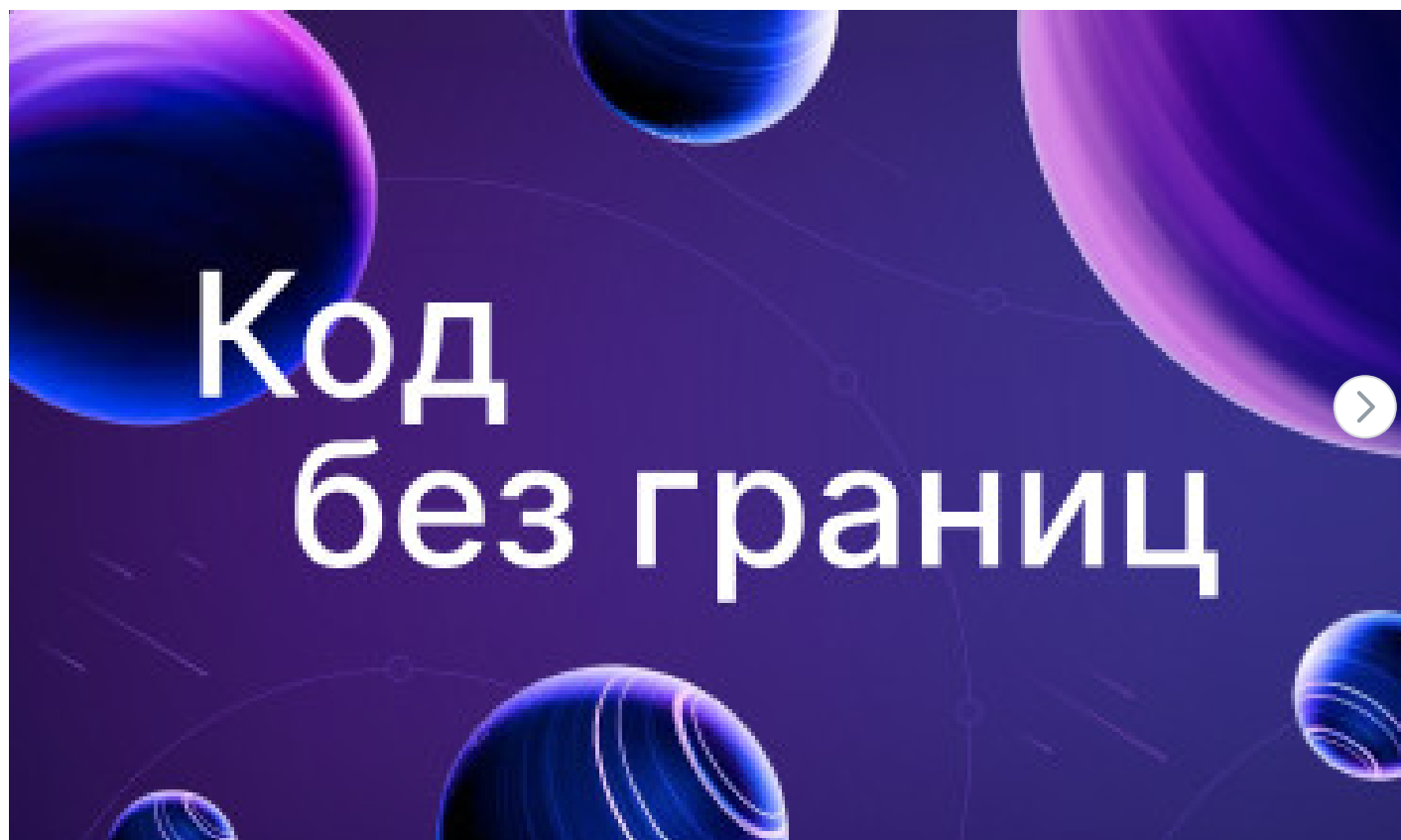
Где обучают вести  
кибертурниры и разрабатывать  
игры

## БЛИЖАЙШИЕ СОБЫТИЯ



### Получи грант за код

Конкурс open source проектов



3 сентября – 31 октября

## Программа грантов для развития open source проектов «Код без границ»

Онлайн

Разработка

Больше событий в календаре

Хабр



Получи грант за код

Конкурс open source проектов

Техническая поддержка

© 2006–2025, Habr



**Получи грант за код**

Конкурс open source проектов