2024-10-08

Лабораторная работа № 2

Написать программу, которая принимает на вход верхнеуровневый URL (строку), а на выходе у нее — ZIP-архив со всеми PDF-файлами и извлеченным из них текстом (просто в `ИМЯ_ОРИГИНАЛЬНОГО_ФАЙЛА.txt`), которые можно получить по этому URL и его дочерним адресам (т.е. программа должна уметь ходить по ссылкам). В архив можно упаковать через модуль `zipfile` стандартной библиотеки. В архиве должна сохраниться структура подразделов исходного сайта.

Пример

На странице по адресу https://abcd.local есть ссылки на файлы `A1.pdf`, `A2.pdf` и т.д. Также на странице https://abcd.local есть ссылка на страницу https://abcd.local/boring, где есть ссылки на файлы `B1.pdf`, `B2.pdf`, `B3.pdf` и т.п. Все эти файлы нужно сохранить и извлечь из них текст.
Мы кладем `A1.txt` и `A2.txt` в корень ZIP-архива, а файлы `B1.txt`, `B2.txt`, `B3.txt` в папку `boring`, которая находится в корне архива.

Для работы с PDF рекомендуется использовать `pypdf` pypdf · PyPI

```
pip install pypdf
```

```
from pypdf import PdfReader
```

```
reader = PdfReader("example.pdf")
number_of_pages = len(reader.pages)
page = reader.pages[0]
text = page.extract_text()
```

Если текст сохранен в PDF-файле в виде изображения, нужно этот текст распознать посредством OCR.

Вариант 1 — Tesseract

Одна из самых известных OCR-библиотек — GitHub - tesseract-ocr/tesseract: Tesseract Open Source OCR Engine (main repository)

Есть биндинг для Питона — pytesseract · PyPI

Вариант 2 — OCRMyPDF

ocrmypdf · PyPI (тоже использует Тессеракт под капотом) Доки: OCRmyPDF documentation — ocrmypdf 15.1.1.dev8+g2b0e149 documentation

Как вспомогательный инструмент может пригодиться pdf2image · PyPI. HO! Сложно завести под Windows, т.к. требует для работы Poppler. См. также документацию по установке Installation — pdf2image latest documentation

Плюс утилита общего назначения для работы с изображениями — `PIL` (Python Imaging Library) Pillow · PyPI

Если совсем все печально, то можно прибегнуть к OpenCV (opencv-python · PyPI), но не рекомендуется в рамках этой лабораторной.

> ⚠ **Warning**

Каждый PDF-файл должен быть сохранен ровно 1 раз на самом верхнем уровне.
Если у нас файл `A1.pdf` есть и на странице https://abcd.local, и на странице https://abcd.local/boring, сохраняете только со страницы https://abcd.local

Какие можно взять источники?

1. Образование на оф. сайте Политеха
2. Федеральные государственные образовательные стандарты на сайте Политеха — примеры с текстом в виде изображений