07.04.22

**Парсинг PDF-файлов в формат \*.JSON с помощью Python**

Цель: Для чего требуется парсинг документов?

В последнее время в работе с данными чаще всего используется формат данных JSON, основанный на JavaScript. За счёт своей лаконичности по сравнению с XML формат JSON может быть более подходящим для сериализации сложных структур. Применяется в веб-приложениях как для обмена данными между браузером и сервером (AJAX), так и между серверами (программные HTTP-сопряжения).

Поскольку формат JSON является подмножеством синтаксиса языка JavaScript, то он может быть быстро десериализован встроенной функцией eval().

Задача: создать программу для конвертации (парсинга) PDF-файлов в файлы JSON. Для перевода файла требуются следующие составляющие:

* Соединение с интернетом,
* API-ключ,
* Библиотеки «os» и «requests».

Теория:

1. Создание PDF-файла

Источником данных послужила [таблица маршрутов](https://www.aeroflot.ru/ru-ru/online_services/flights/ways_map/ways_map_table) на сайте Аэрофлота. Координаты мест были получены на [этом сайте](https://3planeta.com/gps-tools/ru/index.html). Разрешенную высота была описана в нормативно-правовых документах на [сайте](http://www.consultant.ru/document/cons_doc_LAW_91259/2bace5cdaae2198a7f91960512ad02251771530d/) Консультант Плюс. После этого была создана таблица со следующими полями (рис. 1):

* Маршрут,
* Точка маршрута,
* Координаты,
* Разрешенная высота (м),
* Направления движения,
* Расстояние (км),
* Среднее время по расписанию (час).

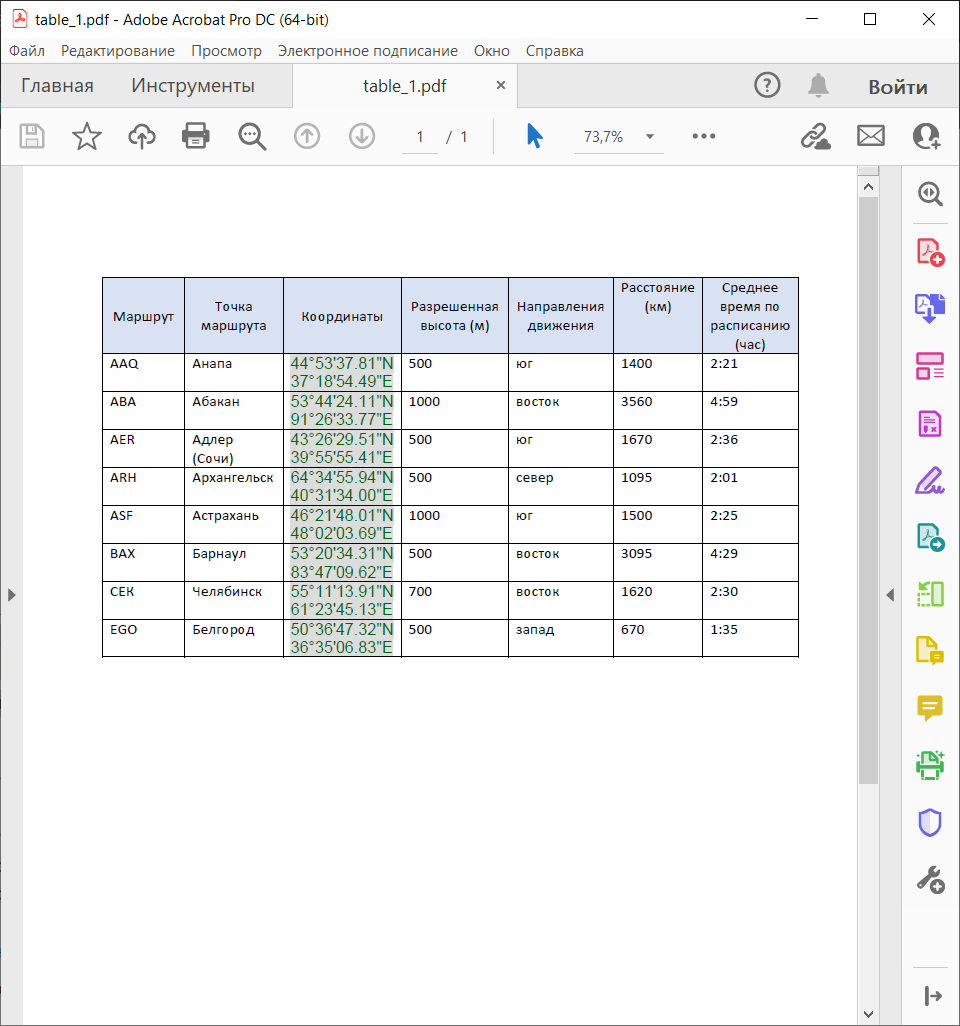


Рисунок 1 – таблица маршрутов в формате \*.pdf.

1. Получение API-ключа:

Для работы с программой необходимо получить специальный API-ключ. Это требуется для авторизации пользователя и открытия доступа программе к необходимым для нее ресурсам.

Для этого необходимо зарегистрироваться в системе по [ссылке](https://app.pdf.co/login). После создания аккаунта будет доступен API-ключ на вкладке «View your API key» (рис. 2).

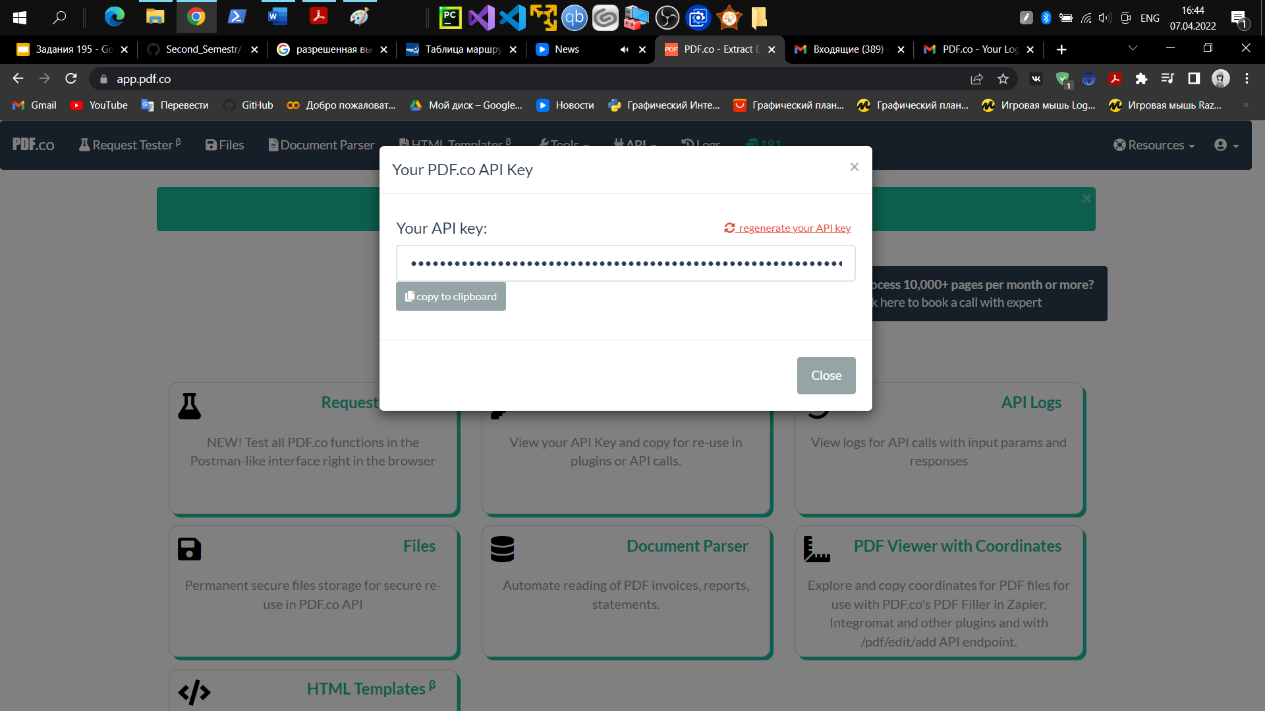


Рисунок 2 – API-ключ в зашифрованном виде.

Полученный ключ нужно вставить в переменную API\_KEY, находящуюся в исходном коде программы, потом запустить ее. Позже появится файл в формате \*.JSON с данными таблицы из файла PDF (рис. 3).

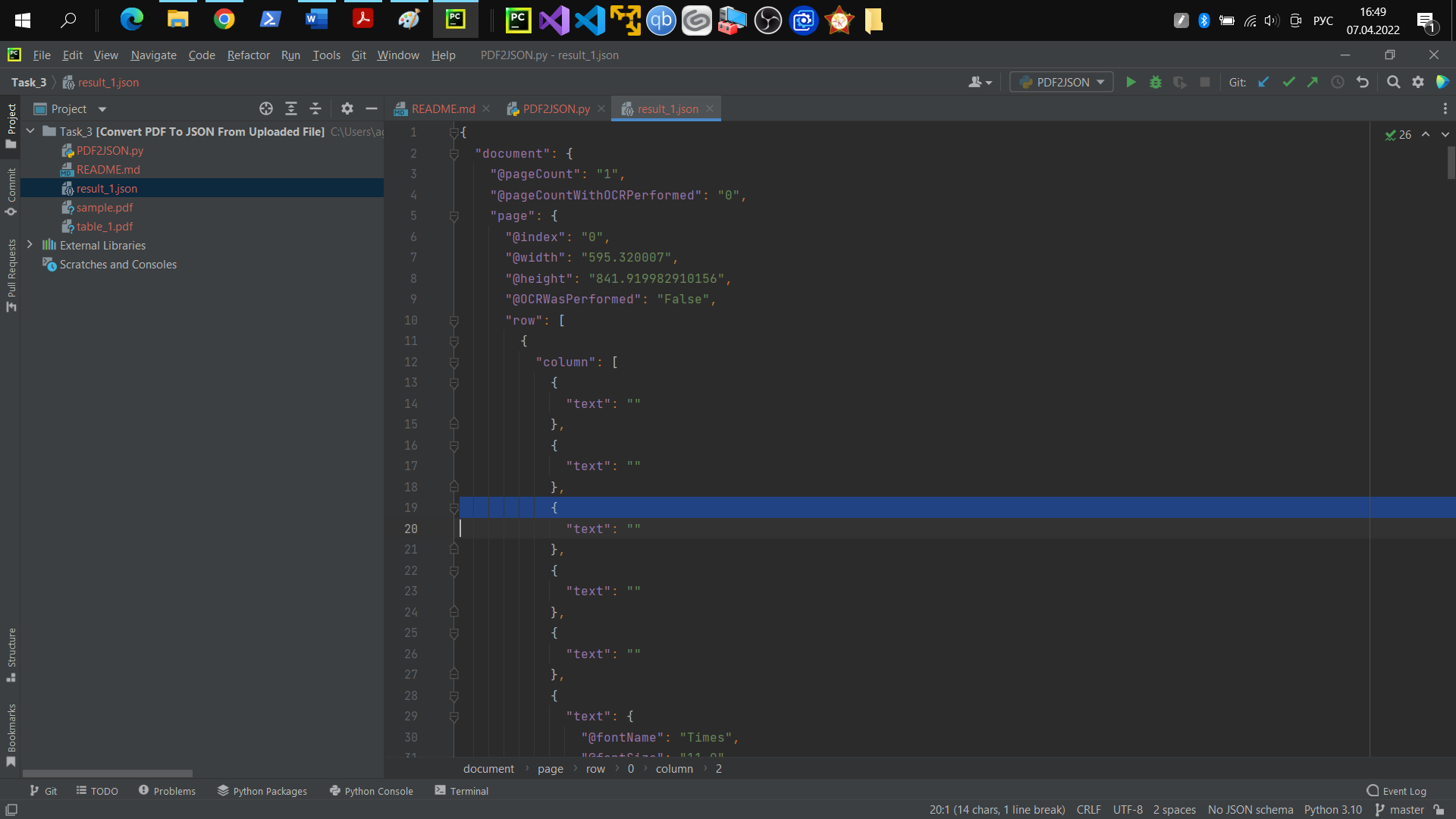


Рисунок 3 – результат работы программы.

1. Описание

Архив со всеми данными доступен на GitHub