Описание программы, разработанной для парсинга PDF

В ходе обработки результатов лабораторных исследований грунтов потребовалось оперативно изучить содержание PDF-файла, представлявшего несколько десятков однотипных паспортов испытаний песков. Поскольку требовалось совершить одну и ту же операцию – копирование конкретных значений с каждой страницы документа в TXT-файл, решение задачи ручным способом заняло бы достаточно много времени. В связи с этим было решено разработать parser на языке Python для ускорения работы и исключения ошибок.

При решении задачи были использована библиотека PyPDF2, а также стандартные методы работы с файлами. Представленный код содержит определение функции **parse(path)** и её вызов. Функция получает от пользователя путь к PDF-файлу, содержащему паспорта испытаний грунтов и проводит постраничный анализ содержимого с помощью цикла **for**. Методами **str.find** и **str.rfind** находятся первые и последние индексы строки-страницы, ведущие к информации о значениях **одометрического (k_od)** и **компрессионного (k_k)** модулей деформации грунта, а также глубине (**depth**) отбора образца грунта, который испытывался в лаборатории. После записи значений индексов в переменные, в открытый TXT-файл функцией **print** по значениям индексов осуществляется запись **срезов** строки-страницы, содержащих полную информацию, необходимую для анализа. В дальнейшем информация полученного **TXT**-файла была использована для расчёта проектных свойств грунтов, что позволило достичь поставленных целей.

В рассматриваемом случае применение разработанной программы позволило сократить время работы сотрудника решения задачи на один рабочий день и позволило использовать наработки на других проектах для оптимизации процесса обработки и интерпретации геологических данных.