Техническое задание

Создание и заполнение справочника на основе PDF документа

Описание задачи:

Написать программу на языке Python, которая будет осуществлять парсинг текстового содержимого из PDF документа.

Требования к функциональности:

Программа должна принимать входной PDF файл в качестве аргумента или через интерфейс пользователя, название документа "SAE J1939-71".

Программа должна извлекать текстовую информацию из PDF документа и сохранять извлеченный текст в указанный формат базы данных.

Образец pdf файла.

Технические требования:

БД - postgresql или sqlite

Требования к распознаванию текста:

Программа должна использовать одну из библиотек для PDF-парсинга, таких как PyPDF2, pdfminer.six, PDFMiner или на Ваше усмотрение.

Структура БД:

Таблица с данными

Таблица № 1

Наименование столбика	Описание	Поля с документа	Примечание
ID	Идентификатор пакета CAN шины	ID	
Data_length	Длинна	Data_length	
Length	Длина занимаемая значением	Length	
Name	Наименование	Parameter Name	
RusName	Наименование на русском		
Scaling	Коэффициент	Slot Scaling	
Range	Диапазон	Slot Range	
SPN		SPN	

Описание процесса получения информации с PDF:

- 1. Для заполнения таблицы в начале нас интересует раздел 5.3... документа. Разбегаясь по разделу 5.3... Из данного раздела забираем в таблицу № 1 значения согласно таблице.
- Далее в разделе 5.2... ищем подробную расшифровку каждого параметра. Поиск производим по 3 признакам. По PGN, названию (Parameter Name в желтой зоне) и Параграфу документа (выделено зеленым). с раздела 5.2 нас интересуют параметры выделенные синим (Рис № 2)

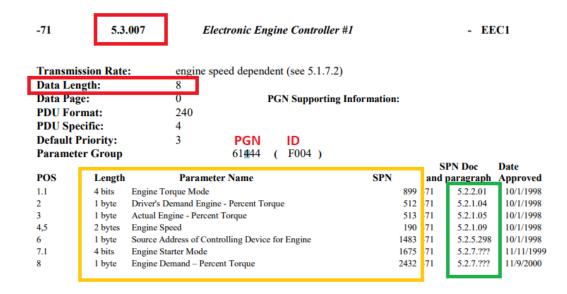


рис № 1 (пример из раздела 5.3)

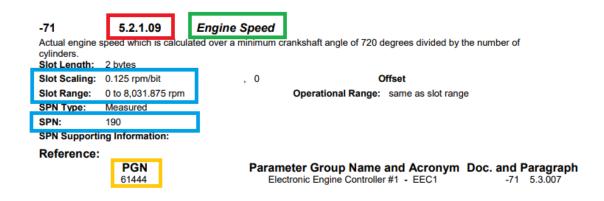


рис № 2 (пример из раздел 5.2)

Требования к документации:

Написать документацию, описывающую функциональность программы, ее использование и возможные настройки.

Ограничения и примечания:

Программа не должна использовать сторонние коммерческие библиотеки или сервисы без предварительного согласования.

Решение:

Ссылка на GitHub с полным описанием README и инструкцией по запуску.