**PDF Invoice Parsing**

**Задача:** разработать решение на python для извлечения структурированной информации с инвойсов, которые хранятся в PDF документах.

Примеры входных документов: <https://drive.google.com/drive/folders/1JNmtoOWKAEvlbz7vCivnBpOiQ25mTvwN?usp=sharing>

Пример выходного документа:

<https://drive.google.com/file/d/1YUw3ghGBlWdvJD8duBv1JMFMPra6gKfr/view?usp=sharing>

На первом этапе выполнения задания необходимо отделить текст документа от таблицы, после чего разобрать таблицу по колонкам.

Для этой задачи нет ограничений в используемых технологиях (т.е. это может быть, например, и java).

Как один из возможных вариантов решения можно попробовать воспользоваться этой библиотекой <https://jonathanlink.ch/PDFLayoutTextStripper.html> , преобразовать документ в текст, затем с помощью эвристических правил отделить таблицу от контента.

В результате необходимо разработать изолированный компонент на python (внутри себя он может использовать что-угодно), на вход которому будет подаваться документ, а на выходе разобранная таблица(ы) из документа.