**Алгоритм решения задачи**

В общем алгоритм решения задачи делится на два главных этапа: 1) разбиение таблицы на ячейки и 2) распознавание содержимого каждой. И первый пункт является самым объёмным, из-за отсутствия в какой-либо библиотеки Python одной соответствующей функции.

Но первым делом, конечно, производится считывание файла с фото протокола в сером цвете. Затем фото переводится в бинарные данные с избавлением от всех шумов и сведением всего к двум цветам: чёрному и белому (фон и надписи соответственно) -> reading\_photo()

1. Разбиение таблицы на ячейки
   1. Распознание вертикальных линий, формирующих таблицу -> vertical\_lines\_detection(…)
   2. Распознание горизонтальных линий, формирующих таблицу -> horizontal\_lines\_detection(...)
   3. Объединение горизонталей и вертикалей, для формирования цельной таблицы без содержимого -> combining\_horizontals\_and\_verticals(...)
   4. Находим контуры таблицы -> cv2.findContours(...)
   5. Сортируем контуры таблицы по порядку их расположения в таблице (сверху вниз) -> sort\_contours(...)
   6. Формируем из контуров список ячеек с рамками -> form\_cells(...)
   7. Разбиваем список рамок на ячейки по строкам -> rows(...)
   8. Сортируем строки в каждом столбце (слева направо) -> rows\_sort(...)
2. Распознание значений в ячейках -> recognition(...)
   1. Рассматриваем циклично содержимое каждой ячейки
   2. Подгоняем содержимое для лучшего распознавания расширением и размыванием -> cv2.dilate(...) & cv2.erode(...)
   3. Анализируем содержимое с помощью ML технологий и формируем список значений на соответствующих позициях -> pytesseract.image\_to\_string(...)

Последним пунктом, конечно же, является запись матрицы значений в excel файл -> data\_to\_excel(...)

**Софт**

Программа написана на языке программирования Python с использованием различных его модулей:

OpenCV (cv2) – библиотека для всевозможных действий с фотографиями - используется для сведения фото таблицы к упорядоченному набору её ячеек (1 глобальный этап), а также для улучшения распознаваемого участка фото, ячеек, на 2 этапе

PyTesseract — это пакет python для разработки OCR (Optical character recognition), где движком OCR будет Tesseract от google. Модуль применяется на 2 глобальном этапе, чтобы преобразовать отсканированный документ в редактируемый формат – excel таблицу

Matplotlib.pyplot – модуль для визуализации данных: выводим распознанные горизонтали и вертикали, саму таблицу, ячейки с рамками

NumPy – необходим для упрощения работа со списками данных

Pandas – библиотека применяется для записи таблицы распознанных значений в excel документ