Lad_Academy Task 2

Scaning pdf with OCR and text searching

В данной директории расположен проект решения второй задачи вступительного испытания.

About

- Сравнив несколько **OCR**-библиотек было принято решение использовать **EasyOCR** для получения печатного текста из pdf. Реализация библиотеки позволяет дообучить ее для получения более чистых данных с различными шрифтами и рукописными текстами.
- Для получения текста из русских рукописных документов была найдена другая модель **shiftlab_ocr**, показывающая в этой области показатели лучше, чем модель для печатного текста.
- Реализован алгоритм, позволяющий работать с pdf и представлять их в виде набора страниц в формате png для сканирования.
- Реализовано условие, позволяющее улучшить и скомбинировать применение двух моделей, но с необходимостью определенного именования файла пользователем.
- Реализована очистка данных для уменьшения объема и ускорения поиска нужного текста.
- В коде присутствуют комментарии для дополнительного пояснения.

Usage

- Загрузить необходимые pdf в директорию проекта.
- Если есть pdf рукописного текста, следует переименовать файл с добавлением `hand_` к началу.

  ```
  document.pdf -> hand_document.pdf
  ```

- Получение результата от поиска текста предполагается запуском команды с помощью командной строки:

  ```
  python main.py <подстрока>
  e.g.  python main.py декабрь
  ```

Features

- Поддержка поиска текста в нескольких pdf сразу.
- Алгоритм реализован таким образом, чтобы экономить ресурсы и не запускать *OCR* для повторного поиска другого слова (подстроки) в сканированных файлах: при повторном запуске поиск текста будет по результирующему json'у и не будет требовать повторного сканирования самих pdf.
- Использование различных *OCR*-моделей для рукописного и печатного текста позволяет снизить потери качества выходных данных для различных pdf.