

Lad_Academy Tasks

Вступительные испытания на стажировку в 2023 году по направлению

Аналитика данных и Data-Science

В данном репозитории в папках приведены решения задач (с пояснениями):

- Написан алгоритм парсинга данных по вакансиям для IT-специалистов с ресурса: nn.hh.ru (Нижегородский регион). Алгоритм позволяет получить информацию о количестве вакансий по направлениям "Аналитика данных" и "Data Science" в разрезе уровней (Junior, Middle, Senior, Lead). Результат представлен в табличном и графическом виде.
- Написан алгоритм распознавания и поиска текста из pdf файлов. Примеры файлов: <https://www.archive-nnov.ru/?id=24998> (печатный документ), <https://www.archive-nnov.ru/?id=37046> (рукописный документ). Результат представлен в виде ссылки на страницу документа, где заявленное слово(подстрока) была распознана.

Features

- Запуск алгоритмов предполагается полностью из командной строки среды разработки с установленными зависимостями, записанными в файле requirements.txt . Установить соответствующие зависимости можно командой из директории с данным файлом:

```
pip install -r requirements.txt
```

Tech

Frameworks & libraries (for Python 3.8.18):

- [Scrapy](#) - An open source and collaborative framework for extracting the data you need from websites. In a fast, simple, yet extensible way.
- [EasyOCR](#) - Ready-to-use OCR Python module with 80+ supported languages and all popular writing scripts including: Latin, Chinese, Arabic, Devanagari, Cyrillic, etc.
- [shiftnlab ocr](#) - SHIFT OCR is a library for russian handwriting text segmentation and character recognition.
- [PyMuPDF](#) - PyMuPDF is a high performance Python library for data extraction, analysis, conversion & manipulation of PDF (and other) documents.
- NumPy, Pandas - base data management python libs

About dev

Козырев Дмитрий Вадимович, 21, бакалавр ПМИ ННГУ им. Лобачевского, обучаюсь на ФИИТ в магистратуре (1 курс). Интересует развитие в области DA и DS. Если Вас устроят мои компетенции, был бы рад работать в компании Lad. Contact: dmitry.v.kozyrev@gmail.com