

# Lad\_Academy Task 1

## Parsing with Scrapy and data analysis

В данной директории расположен проект решения первой задачи вступительного испытания.

### About

- Было принято решение считывать название каждой вакансии и требуемый опыт работы для определения уровня рекрутируемого исходя из усредненного расчета опыта работы в области:
  - опыт не требуется - Junior
  - 1-3 года работы - Middle
  - 3-6 лет работы - Senior
  - 6+ лет работы - Lead
- Данный подход определен из-за следующих условий:
  - возможности градации сайта hh.ru
  - внутри каждой компании названия уровней могут быть уникальными
  - некоторые вакансии не имеют четкого обозначения уровня рекрутируемого
- Реализован алгоритм парсинга данных по вакансиям с ресурса: nn.hh.ru (Нижегородский регион) с помощью паука `scrapy.Spider`. Данные о каждой вакансии выводятся в текстовый файл `vacancies.json`.
- Алгоритм аналитики реализован в классе и позволяет получить информацию о количестве вакансий по направлениям "Аналитика данных" и "Data Science" в разрезе уровней (Junior, Middle, Senior, Lead). При необходимости можно выводить и таблицы вакансий как переменных класса.
- Результат выводится в табличном (и сохраняется в `result.json`) и графическом виде (сохраняется в `result.png`).
- В коде присутствуют комментарии для дополнительного пояснения.

### Usage

Получение результата предполагается в два последовательных этапа с помощью командной строки:

- Запускает скрипт парсинга с помощью Scrapy в асинхронном режиме.

```
python crawl.py
```

- Запускает аналитику собранных данных.

```
python main.py
```

### Features

- Можно дополнить лист `start_urls` (reference) ссылками на списки вакансий для разных запросов, чтобы обходить большее количество вакансий с помощью паука в

файле `hh_spider.py` .

- Реализован поиск ключевых слов в названиях вакансий, чтобы отбрасывать вакансии непредметных направлений и сортировать в пределах предметного (некоторое подобие *cleaning data*, которое можно расширять в рамках проекта).
- Вывод в табличку и в виде круговой диаграммы.