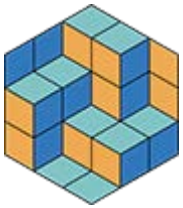


Извлечение атрибутов из объявлений на hh.ru

Олег Мельников, 3 курс МКН

Научный руководитель:
Сергей Резник, Яндекс

Проблема

- В теории, у абстрактного объявления ожидается наличие строгих характеристик и неформального текста
- Фактически, в ведущих сервисах объявлений о работе строгая часть не очень проработана - в отличие от недвижимости, авто, гостиниц и т.п.
- hh.ru: поиск по строгим атрибутам не очень мощный, поиск по тексту тоже не достаточен (можно искать по слову Scala, но в выдачу попадет и вакансия, где это ключевое требование, и та, где это nice-to-have)
- Для аналитики тенденций на рынке труда хочется вытаскивать полезную информацию из неформальных описаний вакансий

Как это выглядит

- https://spb.hh.ru/vacancy/38487280

```
<p><strong>Страховой Дом ВСК</strong> осуществляет страховую деятельность с 11 февраля 1992 года Компания входит в ТОП
    -10 компаний лидеров на рынке страхового бизнеса. Общий штат более 7000 сотрудников по всей России. В ИТ
    подразделении уже более 500 сотрудников.</p>
<p>Наши Сотрудники – это единая команда, поэтому взаимодействие внутри коллектива строится на принципах взаимного
    уважения и профессионализма.</p>
<p>Сейчас мы находимся в поиске <strong>ведущего администратора Linux </strong>сразу на несколько проектов, один из них:
</p>
<p><strong>Стратегическая программа «Прорыв ДМС» -</strong> это выход на новый уровень сервиса для клиентов компании по
    добровольному медицинскому страхованию. Работа в кроссфункциональных командах по agile-принципам. Решение
    амбициозных задач по разработке высокотехнологичных инструментов для взаимодействия с клиентами и партнерами.</p>
<p><strong>Ключевые задачи:</strong></p>
<ul>
    <li>Построение непрерывной интеграции и доставки (CI/CD);</li>
    <li>Администрирование и настройка Linux-систем (RHEL, CentOS, Debian);</li>
    <li>Администрирования HTTP-серверов (Nginx, Apache HTTP Server);</li>
    <li>Написание и работа с SQL-запросами (PostgreSQL, MySQL);</li>
    <li>Администрирование серверов приложений (PHP-FPM, PHP, Node.Js, Redis, RabbitMQ);</li>
    <li>Опыт работы с системами контроля версий git;</li>
    <li>Мониторинг доступности, производительности и создание проверок для системы мониторинга (Zabbix);</li>
    <li>Автоматизация задач на скриптовых языках (Bash, sh);</li>
</ul>
<p><strong>Для нас важно:</strong></p>
<ul>
    <li>Знание Linux-систем на уровне администратора (RHEL, CentOS);</li>
    <li>Опыт построения непрерывной интеграции и доставки (CI/CD);</li>
```

Задача

По набору текстовых объявлений с hh.ru (для простоты ограничиться IT-сегментом) извлекать:

- Перечень требуемых компетенций
- Перечень желательных компетенций
- Есть ли гибкое расписание (если в тексте сказано с “работа в офисе с 10 до 18”, установить этот атрибут в False)
- Подобласть (Разработка, Тестирование, DevOps, Data Science)

Pipeline - 1

- Скачиваем много объявлений разных IT вакансий (Python, C++, Java, ...)
- Парсим сырое тело объявления на разделы

```
32  <p><strong>Плюсом будет:</strong></p>
33  <ul>
34      <li>Знание Agile подходов.</li>
35      <li>Опыт написания скриптов на языке Python;</li>
36      <li>Администрирование и опыт работы с Apache ActiveMQ, Tomcat;</li>
37      <li>Умение читать код.</li>
38  </ul>
```

Pipeline - 2

- Очищаем текст (tokenization, stopwords, lemmatization)
- Обучаем модель распознавания разделов

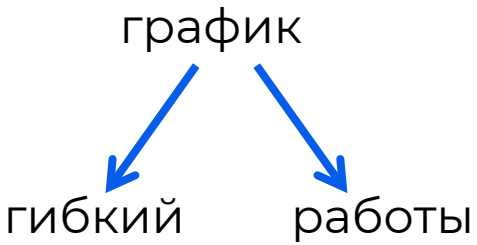
1. Чем предстоит заниматься/ключевые задачи
2. Требования/мы ждем от вас
3. Nice-to-have
4. Что мы предлагаем/почему мы/общее описание команды/плюшки от компании
5. Условия работы
6. Текущий стек

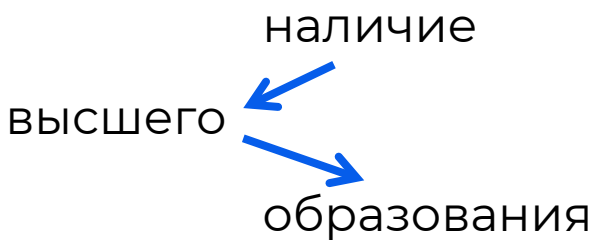
Pipeline - 3

- Извлекаем требования (NER, POS-tagging)

Python – upos: PROPN, ner: S-MISC

- Извлекаем информацию про график работы (Dependency parsing)





- Извлекаем информацию о подобласти (эвристики)

Результат работы

Вакансия

{..} result.json > ...

```
{
    "Обязательные компетенции": [
        "Windows",
        "CI",
        "CD",
        "Node.Js",
        "RHEL",
        "PHP",
        "Redis",
        "Linux",
        "MySQL",
        "Bash",
        "Zabbix",
        "CentOS",
        "Администрирование серверов Linux",
        "PostgreSQL",
        "Nginx"
    ],
    "Желательные компетенции": [
        "Tomcat",
        "Agile",
        "ActiveMQ",
        "Python",
        "Apache"
    ],
    "Гибкий график работы": false,
    "Подобласть": "DevOps"
}
```

Оценка качества

- Качество классификации разделов

CV accuracy = 78%, baseline = 38%

- Качество ключевого алгоритма методом сэмплирования

Presicion – 92%, Recall – 82%

Используемые технологии

- Для выкачки данных и парсинга:

- Для обработки текста:

Ссылки

- Мой репозиторий - https://github.com/melnikoff-oleg/hh-load
- Stanza - https://stanfordnlp.github.io/stanza/index.html

Спасибо за внимание

Контакты

- telegram: @melnikoff_oleg
- mail: oleghools@gmail.com