3. МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

А.О. Абрамов [1], К.М. Филатов [1], А.М. Перегримов [1], Ю.В. Боганюк [1,2]

[1] Тюменский государственный университет, г.Тюмень

[2] Научно-технический университет «Сириус», г.Сочи

УДК 004.912

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ АНАЛИЗА АКТУАЛЬНЫХ ТРЕБОВАНИЙ РЫНКА ТРУДА НА ОСНОВЕ ТЕКСТОВ IT ВАКАНСИЙ

Аннотация. В статье представлен метод анализа требований рынка труда на основе текстов вакансий, позволяющий определить требуемые навыки и уровень их знаний для вакансий с интернет ресурса по поиску работы hh.ru.

Ключевые слова: анализ текста, рынок труда, text mining

Введение

Ежегодно вузы выпускают большое количество специалистов разного характера, которым требуется найти подходящую под их уровень знаний работу. Возникает проблема получения информации об актуальных требуемых навыках и уровнях владения ими по специализациям ИТ-сферы. Автоматизация исследования рынка труда, значительно сократит время получения информации о рынке труда. Такое решение позволит новым специалистам трудоустроиться в кратчайшие сроки.

Цель работы - разработать приложение для оценки уровня требуемых навыков на рынке труда в IT сфере, на основе текстов вакансий.

Необходимо решить задачи:

- Разработать выгрузчик данных о вакансиях;
- Разработать алгоритм обработки текстов вакансий;
- Разработать базу данных;
- Разработать приложение, предоставляющее пользователю доступ к базе.

Анализ текстов вакансий

На рисунке 1 представлен пример текста вакансии.

Мы являемся немецкой компанией с головным офисом в г. Висбаден и филиалом в России, г. Краснодар.

Наша специализация заключается в предоставлении индивидуальных IT-решений для ресторанного бизнеса, с которыми мы являемся ведущими в Германии. Кроме того, мы также действуем в Австрии, Швейцарии и Польше. Помимо ресторанов мы работаем с производителями кассовых систем для ресторанов и порталами для заказа еды

Сейчас мы находимся в поиске **опытного PHP разработчика** для работы в офисе в г. Краснодар. Также возможен вариант удаленного сотрудничества.

Что мы предлагаем:
- Гибкое начало рабочего дня;
- Работа в международной команде;
- Интересные и разнообразные проекты, возможность переключения между ними;
- Возможность релокации или командировок в центральную часть Германии;
- Компенсация затрат на обучение и развитие;
- Оплачиваемые отпускные и больничные дни;
- Возможность удаленной работы после 6 мес работы в компании.

Что мы ожидаем:
- Опыт разработки на PHP более 3 лет;
- Навыки реализации REST API на PHP;
- Знание протокола, HTTP-глаголов и того, как они используются для взаимодействия между скриптами на веб-серверах и клиентах;
- Опыт работы с MySQL базами данных;
- Опыт работы с Git и Git Flow.

Рис.1. Пример текста вакансии

Для анализа содержания вакансии, требуется 2 компонента: словарь характеризующих слов и коллекция оцениваемых навыков. Словарь характеризующих слов состоит из выражений, с помощью которых

работодатели оценивают навыки, например «Уверенное знание» или «Опыт работы». Для получения такого словаря, используется TF-IDF мера.

TF-IDF — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов [1; 105-108].

Мера высчитывается по формуле $TF\text{-}IDF = TF_{x,y} * IDF_{x,Y}$, где $TF_{x,y}$ - это частота появления слова x в документе y, $IDF_{x,Y}$ - это логарифм отношения общего числа документов в коллекции Y к числу документов, в которых встречается слово x.

С помощью TF-IDF меры, для каждого слова определяется значение, которое говорит о важности слова в коллекции документов, где максимум – это наиболее важные слова и минимум – это наименее важные слова. То есть чем больше у слова значение TF-IDF метрики, тем больше оно смысла несет относительно документа.

Также слова проходят дополнительную обработку стеммером, который находит основу слова. Это нужно для исключения однокоренных слов, несущих одинаковый смысл, но имеющих различия в написании, например «Знание» и «Знания».

По итогам обработки данных были получены оценки наиболее и наименее важных слов (см. Табл.1, Табл. 2).

Табл.2. Наименее важные слова

Оценочное слово	TF-IDF мера
«работ»	0.07365
«разработк»	0.05154
«опыт»	0.04948
«знан»	0.04524

«умен»	0.02416

Табл.1. Наиболее важные слова

Оценочное слово	TF-IDF мера
«больничн»	0.006
«настройк»	0,00888
«механизм»	0.00532

Так как компьютер не может определить, является ли слово подходящим для оценки, после выявления всех важных слов среди большого количества вакансий был произведен ручной отбор оценочных слов.

Затем необходимо распределить оценочные слова по 3 уровням знаний: Junior, Middle, Senior. Для этого проводится поиск вакансий с указанным в названии уровнем и сохраняются все вхождения оценочных слов в текст вакансии.

Примеры словарей оценочных слов:

- **Junior**: Junior, Начальное знание, знание, понимание, умение;
- **Middle**: Middle, уверенное знание, хорошее владение, опыт работы, хорошее знание, владение, опыт;
- **Senior**: Senior, уверенное знание, большой опыт работы, опыт разработки.

Для составления коллекции навыков, требуется извлечь все технологии, указанные работодателями в описании вакансий. Для этого перед обработкой, для всех вакансий, опубликованных за последний день, навыки, указанные работодателями как ключевые, выгружаются и добавляются в общую базу.

После предварительного сбора данных, выполняется основной алгоритм обработки. Текст вакансии разбивается на отдельные

предложения, в которых производится обход в обратном порядке для поиска навыков. При выявлении навыка, он помещается во временный массив. Процесс поиска навыков производится, пока не найдется любое из оценочных слов или не кончится предложение. В случае нахождения оценочного слова, из всех навыков во временном массиве составляются наборы «Навык-оценочное слово», например «SQL – Опыт Работы» или «C++ - Знание». По завершению алгоритма, все полученные наборы преобразуются в наборы «Навык-Уровень», например «C# - Junior», «Python – Senior».

Результаты

С помощью словарей, можно проанализировать требуемый уровень знаний навыков на рынке труда. На рисунке 2 можно увидеть количество упоминаний уровней знаний навыков в текстах вакансий.

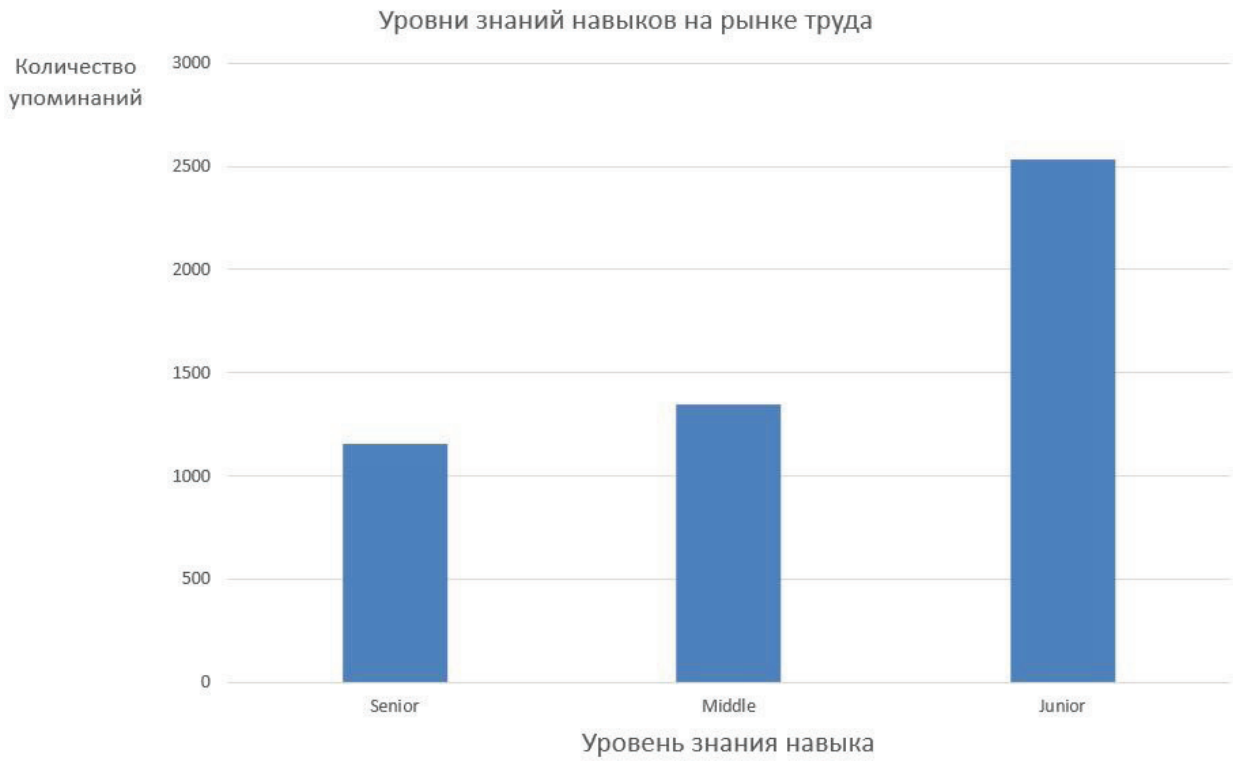


Рис.2. Количество упоминаний уровней знаний навыков в текстах вакансий.

Можно заметить, что в основном работодатели оценивают навыки на уровень Junior, по сравнению с Senior и Middle.

Некоторые работодатели заранее указывают в названии вакансии требуемый уровень знаний. На рисунке 3 можно увидеть количество вакансий с заранее указанным уровнем знаний.

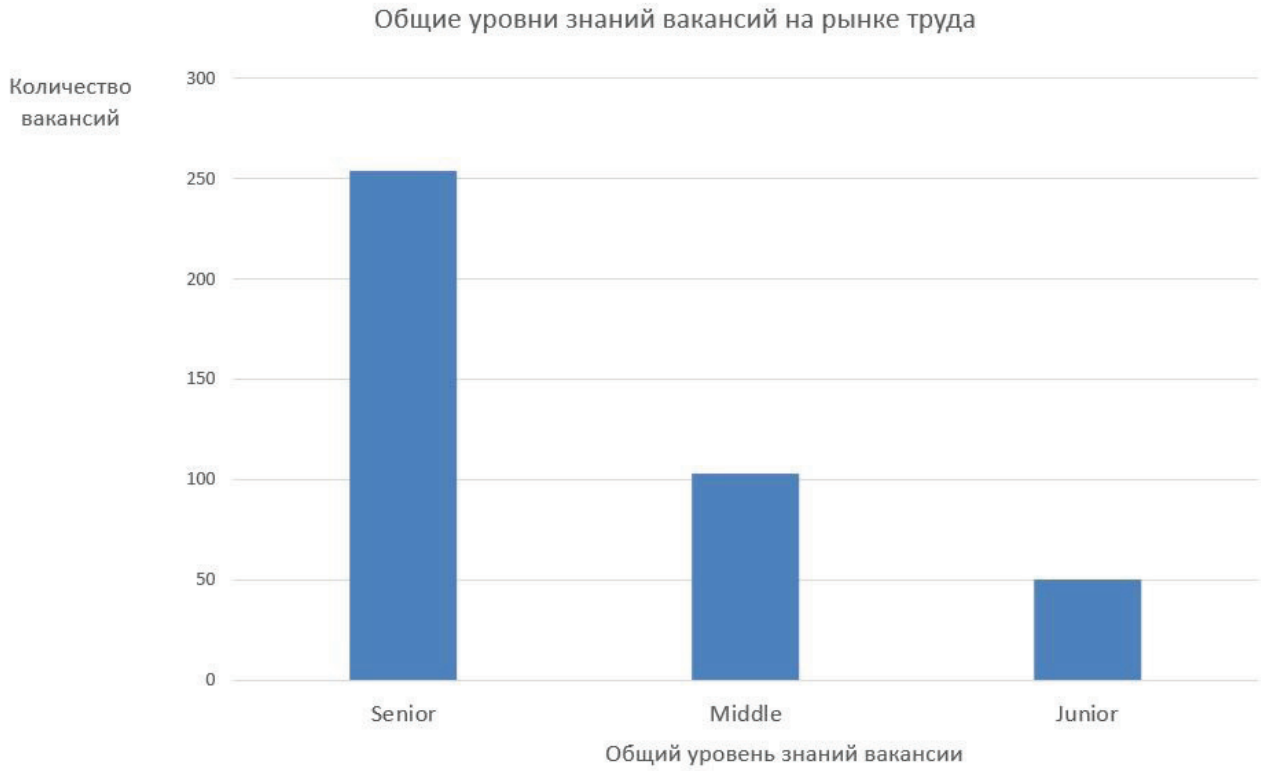


Рис.3. Количество вакансий с заранее указанным уровнем знаний.

Можно заметить, что результат исследования обратный. Это можно объяснить тем, что помимо основных навыков, которые устанавливают общий уровень вакансии, работодатели указывают дополнительные навыки, которые как правило достаточно знать на уровне Junior.

Структура проекта

Приложение состоит из 4 модулей: выгрузчик, обработчик текстов вакансий, база данных и пользовательское приложение. С помощью API hh.ru выгрузчик загружает вакансии в обработчик [2; 87 – 92]. После обработки вакансии добавляются в общую базу.

Выгрузка данных и обработчик данных были реализованы на языке Python с использование библиотек: Scikit-learn, Requests, PyStemmer, NumPy. В качестве СУБД была выбрана Microsoft SQL server.

На рисунке 4 представлена архитектура проекта

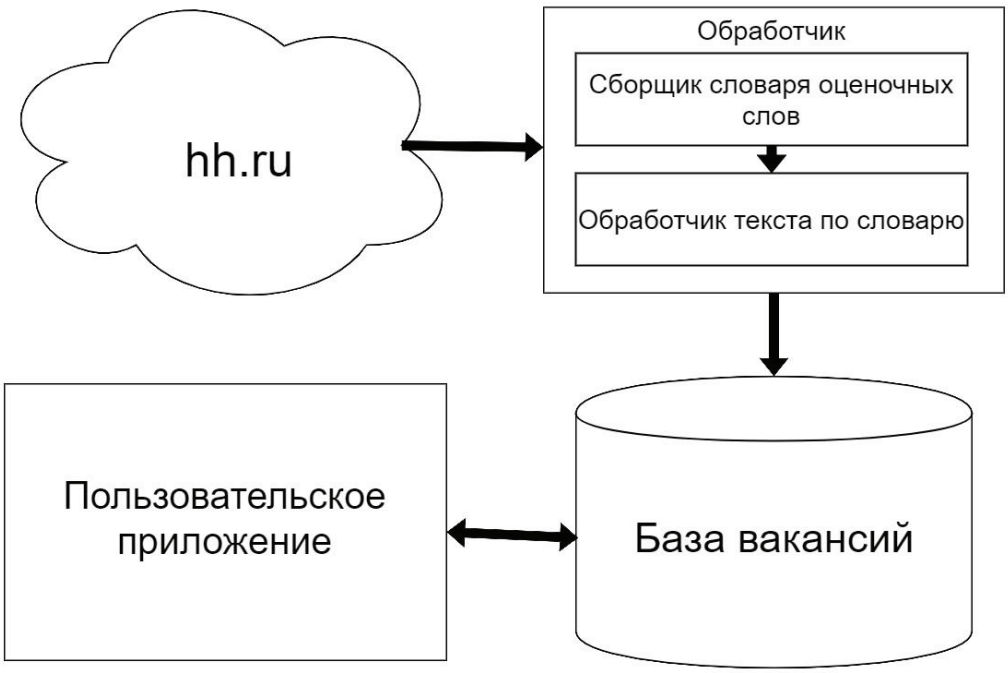
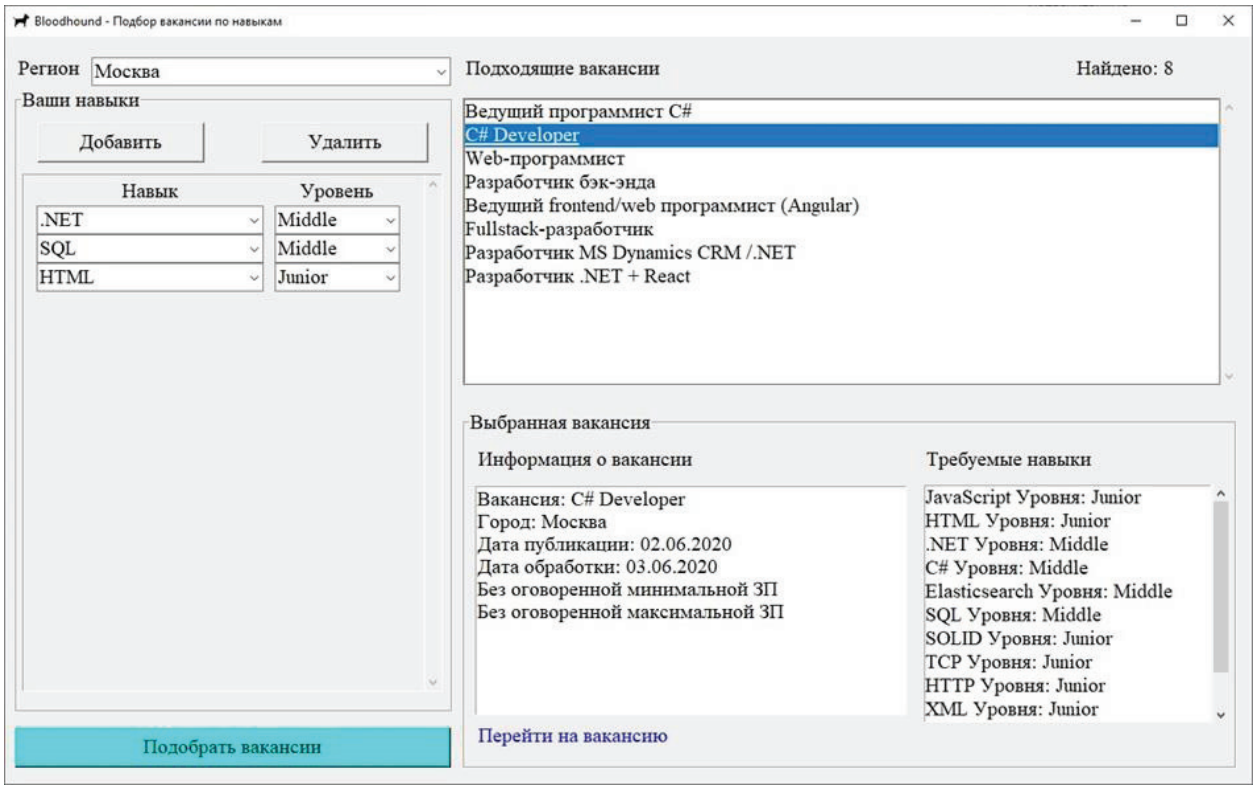


Рис.4. Архитектура проекта

База данных хранит в себе список вакансий, список уникальных навыков и список регионов. Для каждой связи между вакансией и навыком в отдельном списке хранятся наборы «навык-уровень».

Пользовательское приложение позволяет по заданным навыкам и уровням знаний навыков определить подходящие вакансии. На рисунке 5 представлен пример работы приложения.



PPuc. 5. Пример работы приложения

Заключение

В результате работы, с помощью API hh.ru, был разработан выгрузчик данных о вакансиях. Разработан анализатор коллекции текстов вакансий для нахождения оценочных слов с помощью TF-IDF меры. Разработан анализатор текста вакансии, с помощью собственного алгоритма обхода предложения и поиска наборов «Навык-Оценочное слово». Разработана база данных для хранения информации о вакансиях с помощью MySQL server. Был проведен анализ указанных навыков и общего уровня вакансий. Разработано пользовательское приложение, предоставляющее пользователю доступ к базе.

В дальнейшем планируется добавления классификатора, который сможет определять общий уровень вакансии по набору навыков и их уровней знаний. Это позволит подбирать вакансии не только по уровням отдельно взятых навыков, но и по общему уровню знаний пользователя.

Благодарности

Статья подготовлена в рамках разработки образовательного кейса для НТУ Сириус при финансовой поддержке РФФИ в рамках научного проекта № 19-37-51028.

СПИСОК ЛИТЕРАТУРЫ

1. Jurafsky D. Speech and Language Processing / D.Jurafsky // -Stanford: Stanford University 2019. -621p.

2. Митчелл Р. Скрапинг веб-сайтов с помощью Python / Р.Митчелл // -Москва: ДМК Пресс, 2016. -280с.

3. Лутц М. Программирование на Python / М.Лутц // -Санкт-Петербург: Символ-Плюс, 2011. 1 том - 992с.