В.Е. Филиппов [1]***, Ю.В. Боганюк*** [1,2]***, М.С. Воробьева*** [1,2]

[1] *Тюменский государственный университет, г. Тюмень*

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

УДК 004.912

РАЗРАБОТКА ПРИЛОЖЕНИЯ ДЛЯ СОСТАВЛЕНИЯ ТРЕБОВАНИЙ К ПРОФЕССИОНАЛЬНЫМ НАВЫКАМ НА ОСНОВЕ МЕТОДОВ АНАЛИЗА ТЕКСТОВ

Аннотация. В статье описываются результаты разработки веб-приложения для составления требований к профессиональным навыкам на основе методов анализа текстов, основываясь на исследовании требований рынка труда в сфере IT, с опорой на данные, полученные с агрегаторов вакансий. Поставленная задача решалась с помощью технологий сбора статистики, анализа текстов вакансий, технологий машинного обучения и построения ассоциативных правил.

Ключевые слова: актуальные требования рынка труда; обработка естественного языка; машинное обучение; Apriori; обучение ассоциативных правил; Associations rules learning; извлечение информации; анализ данных; Web Mining; Text Mining.

Введение

Содержание профессиональных компетенций, критерии и целевые показатели развития как специалиста в наше время очень динамичны. Поэтому тем, кто хочет освоить востребованные профессии, необходимо постоянно мониторить и анализировать изменяющиеся требования рынка труда [1], а в рамках современной концепции индивидуализации образования выбирать перспективные образовательные программы и в соответствии с запросами изучать учебные курсы [2].

Для решения задачи определения актуальных требований рынка труда, то есть требований работодателей к соискателям, был применен анализ корпуса текстов вакансий с помощью средств машинного обучения, что позволило извлечь и проанализировать информацию, используя технологии обработки естественного языка [3].

Необходимо спроектировать и разработать инструмент, который предоставит возможность решить описанные проблемы, опираясь на актуальные данные.

Используемые методы для извлечения данных

В процессе разработки приложения и анализа текстов вакансий потребуется решить задачу извлечения данных из неструктурированного текста, в котором в свободной форме описаны технологии и навыки, владения которыми работодатель ожидает от соискателей. Эта задача известна как задача извлечения данных из текста (Text mining), которая также является подзадачей обработки естественного языка (NLP).

В качестве одного из методов предобработки текста для семантического анализа потребуется использовать токенизацию – подход для разбиения текста на токены в виде отдельных слов, которые в дальнейшем передаются на вход другим алгоритмам обработки естественного языка [4].

Для того чтобы корректно обрабатывать слова, требуется лемматизация – процесс преобразования слова в его значимую базовую форму, не просто удаляя их окончания, а извлекая лемму, основываясь на контексте, в котором находится данное слово [5].

В качестве инструмента для морфологического анализа слов, приведения к нормальной форме, извлечения грамматической информации о слове использовалась библиотека pymorphy2 языка программирования Python [6].

Подход к решению задачи извлечения из текстов вакансий упоминаний о технологиях, владения которыми требует работодатель от соискателей, основан на задании правил извлечения фактов на основе анализа текстов предметной области с дальнейшим составлением правил для извлечения и словарей ключевых слов. В качестве конкретного инструмента для семантического анализа русскоязычных текстов был выбран Yargy – инструмент для извлечения структурированной информации из текстов на естественном языке [7].

Для определения квалификации по списку технологий требовалось выполнить разбиение на классы *«Junior», «Middle», «Senior»,* основываясь на наборе данных, состоящем из извлеченных с помощью семантического анализа текстов вакансий списков технологий.

Для решения задач определения квалификации и прогнозирования диапазона заработной платы была выбрана модель машинного обучения Случайный лес [8].

Для решения задачи формирования рекомендаций о том, какие навыки целесообразно изучить на основе стека технологий, которыми пользователь уже владеет, требовалось проанализировать информацию о том, какие технологии часто встречаются вместе в текстах вакансий. Для этого было решено применить алгоритм обучения ассоциативных правил, заключающемся в том, чтобы на основе имеющегося набора данных, надо найти закономерности между наборами объектов [9]. В дальнейшем по построенным правилам будет производиться определение рекомендуемых к изучению технологий. Для решения задачи поиска ассоциативных правил использовался алгоритм Apriory [10].

Реализация функционала приложения

В качестве источника данных был выбран портал HeadHunter, агрегирующий вакансии по различным видам профессий. Данный сервис предоставляет доступ к данным по API. На запросы, отправляемые к

интерфейсу сервиса в формате JSON, были получены 27073 вакансии, связанные с IT-областью(дата выборки 21.05.2020).

Перед тем как происходит извлечение списка технологий из текста вакансии, производится его предобработка: очищается от html-тегов, производится разбиение на предложения с помощью библиотеки Razdel, происходит токенизация предложений, после чего поток токенов подается на вход синтаксическому анализатору, который выдает конечный список технологий из текста данной вакансии.

Для выборки из текстового описания вакансии списка технологий и их важности использовался синтаксический анализатор, реализованный в библиотеке Yargy, для которого предварительно были составлены правила для извлечения информации для данной предметной области.

В целях улучшения качества моделей машинного обучения в наборе данных были оставлены только часто встречающиеся технологии. С учетом удаления редких технологий и отбрасывания вакансий, из текстов которых не удалось извлечь информацию о списке требований, набор данных составил 4333 записи. На сформированном наборе данных обучаются модели для прогнозирования диапазона заработной платы и уровня квалификации, а также строятся ассоциативные правила.

При работе с приложением, веб-интерфейс которого был разработан с применением фреймворка Dash, пользователь вводит на странице свои данные (список технологий, регион проживания, уровень владения технологиями (см. рис.1)) и получает результат обработки его данных обученными моделями, которые предсказывают диапазон заработной платы и уровень квалификации для совокупности введенных им технологий.

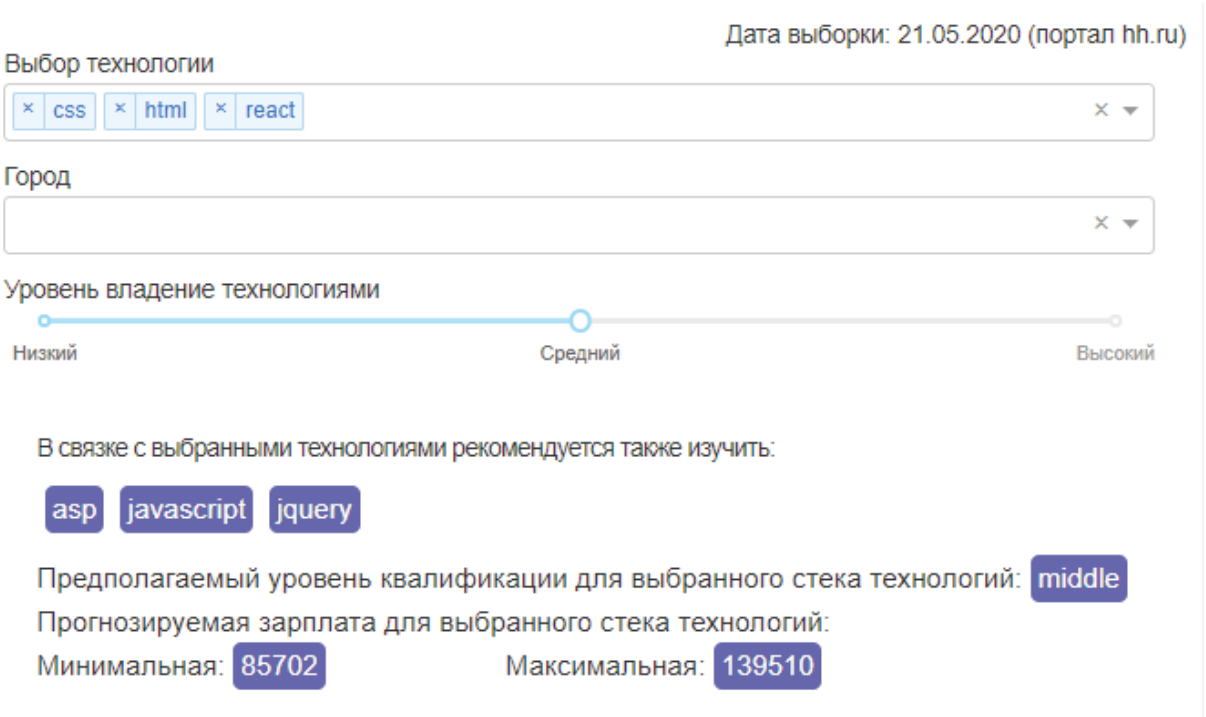
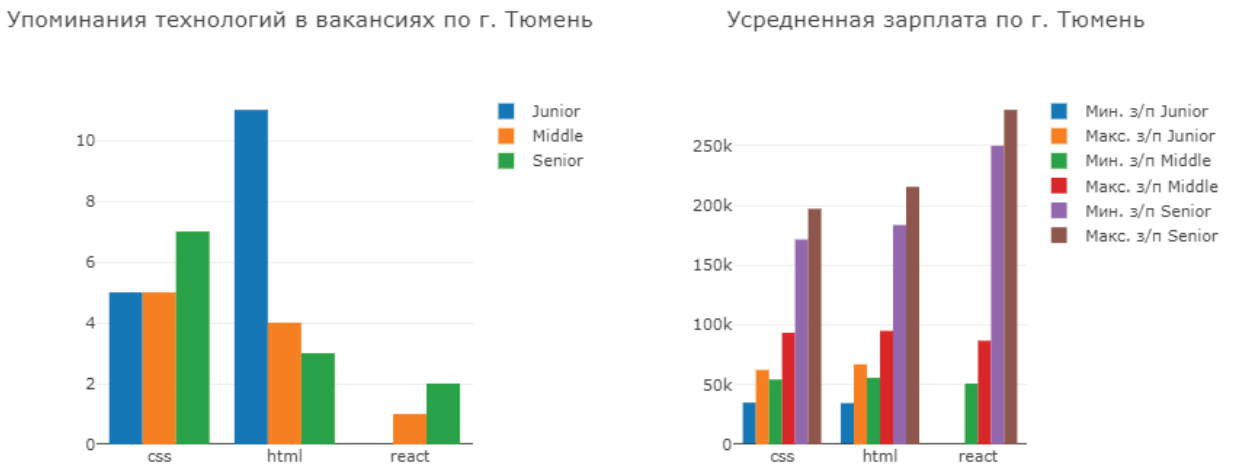


Рис. 1. Часть интерфейса веб-приложения

По введенными технологиям *css*, *html*, *react* и выбранному среднему уровню владения система по выборке на 21 мая 2020 года выдала прогноз возможного диапазона заработной платы от 85702 до 139510, предполагаемый уровень квалификации *middle*, а также порекомендовала изучить вместе с этими технологиями *asp*, *javascript* и *jquery*.

В дополнение к полученной информации для каждой указанной технологии отображается усредненный уровень заработной платы специалистов с различной квалификацией как по выбранному пользователем городу (см. рис.2), так и по всей России (см. рис.3).

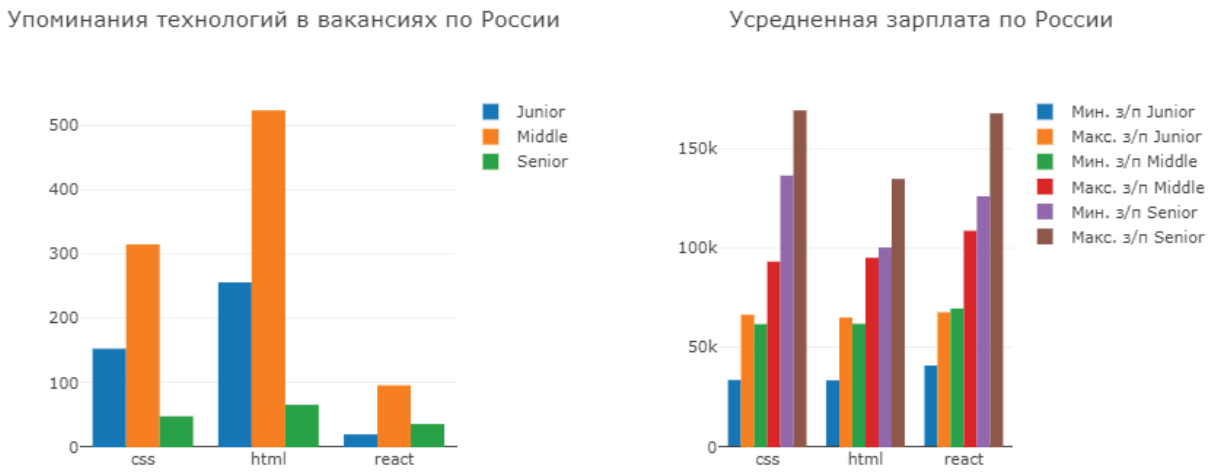


Технология	Всего	Junior	Middle	Senior	Мин. з/п Junior	Макс. з/п Junior	Мин. з/п Middle	Макс. з/п Middle	Мин. з/п Senior	Макс. з/п Senior
css	17	5	5	7	35000	62364	54356	93456	171555	197208
html	18	11	4	3	34545	67025	55668	95115	183630	215606
react	3	0	1	2			50891	86820	250000	280000

Рис. 2. Часть интерфейса веб-приложения с отображением данных по выбранному городу

По выбранным технологиям *css*, *html*, *react* можно сказать, что в Тюмени не наблюдается высокая востребованность в специалистах, владеющих *react*, однако имеется много вакансий для тех, кто разбирается в верстке на *html* и *css*, причем начинающие специалисты в этой области не могут претендовать на высокий уровень заработной платы.

Из общероссийской статистики видно, что владение технологиями для верстки сайтов на уровне *Middle* востребовано. Более сложная для освоения технология *react* имеет меньшую популярность среди работодателей, эти специалисты получают на 10-20 тысяч больше, чем работающие с *css* и *html*.



Технология	Всего	Junior	Middle	Senior	Мин. з/п Junior	Макс. з/п Junior	Мин. з/п Middle	Макс. з/п Middle	Мин. з/п Senior	Макс. з/п Senior
css	516	153	315	48	33765	66422	61723	93099	136262	168924
html	845	256	523	66	33576	65037	61897	95008	100188	134529
react	152	20	96	36	40953	67664	69563	108567	125870	167441

Рис. 3. Часть интерфейса веб-приложения с отображением данных по России

Результаты

В процессе решения задачи анализа актуальных требований рынка труда были обучены модели для определения минимального и максимального уровня заработной платы, а также предполагаемого уровня квалификации по списку технологий и уровню владения ими.

Модели обучались на данных 4333 вакансий с размером тестовой выборки 20%. В результате для модели, прогнозирующей минимальный уровень заработной платы, был получен результат, характеризуемый медианной абсолютной ошибкой 5472,67 и средней абсолютной ошибкой 9343,06.

Для модели, определяющей максимальный уровень зарплаты, был получен результат, характеризуемый медианной абсолютной ошибкой 10394,29 и средней абсолютной ошибкой 16537,28.

Для модели, определяющей уровень *(«Junior», «Middle», «Senior»)*, был получен результат, характеризуемый 65% верных ответов, точностью и полнотой модели 0,61 и 0.46 соответственно.

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. Захарова И.Г. Методы машинного обучения для информационного обеспечения управления профессиональным развитием студентов // Образование и наука. 2018. Т. 20. № 9 – С. 91-114.
2. Я.В. Ланг, М.С. Воробьева. Моделирование процесса построения электронных учебных курсов на основе учебных объектов. // Вестник Тюменского государственного университета. 2009. № 6. С. 230-234.
3. М.С. Воробьева, А.А. Дубаков. Разработка приложения для определения тематики текста с использованием алгоритмов кластеризации. // Математическое и информационное моделирование: сборник научных трудов. Вып. 17. – Тюмень: Издательство Тюменского государственного университета, 2018. – С. 85-95.
4. Jonathan J. Webster, Chunyu Kit. Tokenisation is the initial phase of NLP. // Proc of COLING-92, Nantes, Aug. 23-28, 1992. С. 1106-1110.
5. Grzegorz Chrupała. Simple Data-Driven Context-Sensitive Lemmatization. [Электронный ресурс]. URL:

https://pdfs.semanticscholar.org/0139/bed746c44262ed6e12eb025143538d86bda2.pdf?_ga=2.62015084.1387677874.1589640610-544942190.1587184978 (дата обращения: 12.05.2020).

6. Репозиторий и документация к системе извлечения фактов Yargy. [Электронный ресурс]. URL: https://github.com/natasha/yargy (дата обращения: 12.05.2020).

7. Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages. [Электронный ресурс] URL: https://arxiv.org/pdf/1503.07283v1.pdf (дата обращения: 07.05.2020).

8. Amit Y., Geman D. Shape quantization and recognition with randomized trees. Neural Computation 9, 1997 – C. 1545-1588

9. Зайко Т.А., Олейник А.А., Субботин С.А. Ассоциативные правила в интеллектуальном анализе данных. // Вестник Национального технического университета Харьковский политехнический институт. Серия: Информатика и моделирование 2013.

10. Agrawal, Rakesh, and Ramakrishnan Srikant. Fast algorithms for mining association rules. Proc. 20th int. conf. very large data bases, VLDB. Vol. 1215. 1994.