3. МЕТОДЫ И ТЕХНОЛОГИИ АНАЛИЗА ДАННЫХ

А.О. Абрамов[1], К.М. Филатов[1], А.М. Перегримов[1], Ю.В. Боганюк[12]

[1] Тюменский государственный университет, г. Тюмень

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

УДК 004.912

РАЗРАБОТКА СЕРВИСА ДЛЯ ОПРЕДЕЛЕНИЯ АКТУАЛЬНЫХ ГРУПП НАВЫКОВ СПЕЦИАЛИСТА НА ОСНОВЕ ТЕКСТОВ ВАКАНСИИ

Аннотация. В статье представлен метод анализа требований рынка труда на основе текстов вакансий, позволяющий определить какие персональные качества ценятся работодателями и какие группы навыков специалиста наиболее востребованы для вакансий с интернет ресурса по поиску работы hh.ru.

Ключевые слова: анализ текста, рынок труда, кластеризация, 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 требованиях рынка труда по специализациям ИТ сферы. Автоматизация исследования текстов вакансий значительно сократит время получения информации о рынке труда. Такое решение позволит новым специалистам оценить свои перспективы и трудоустроиться в кратчайшие сроки.

Цель работы – разработать приложение для определения актуальных групп навыков специалиста на рынке труда в ИТ сфере на основе текстов вакансий

Необходимо решить задачи:

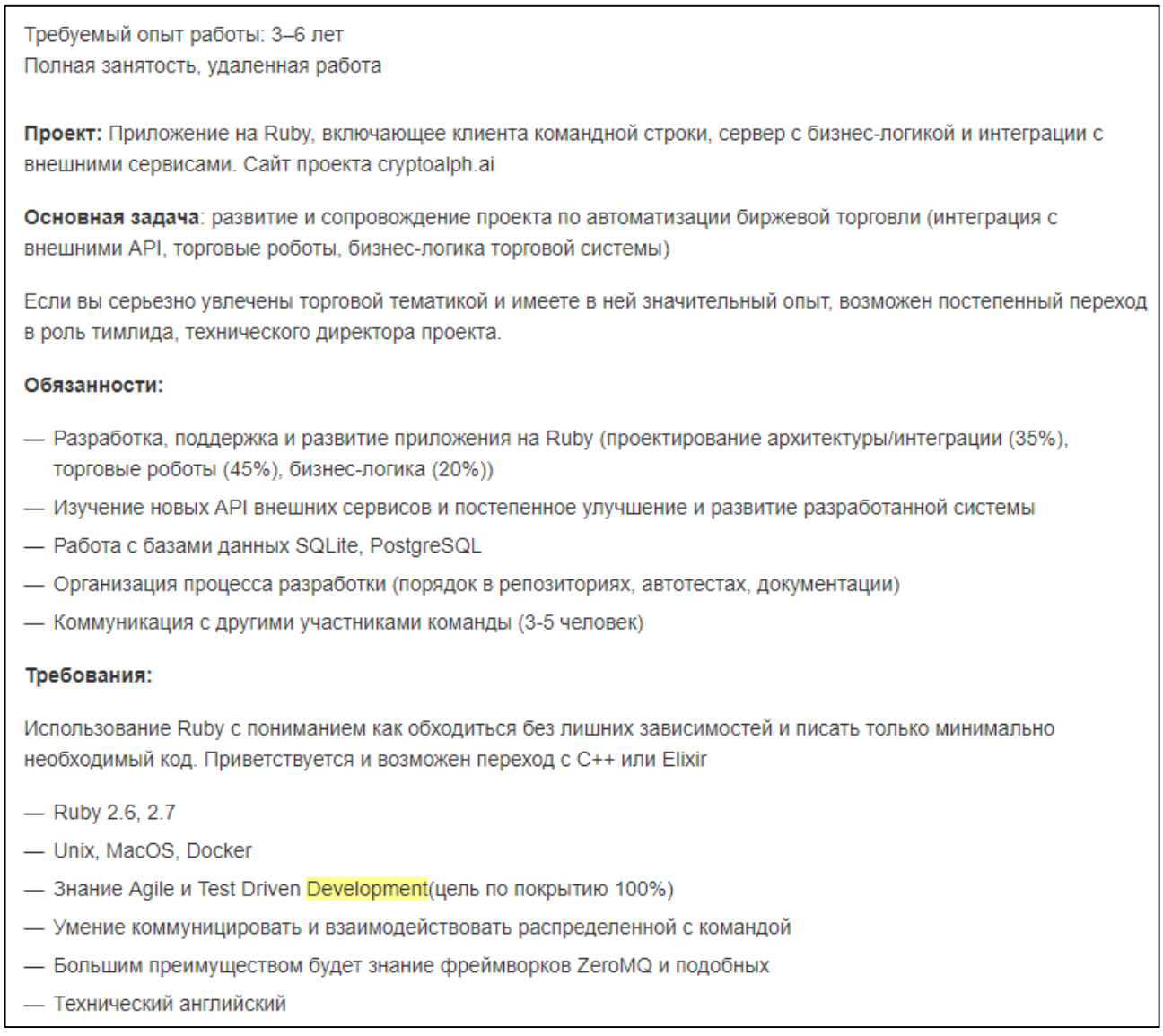
- Определить структуру базы данных;
- Автоматизировать выгрузку вакансий и обновление уже имеющихся;
- Разработать метод извлечения не только технологий, но и других навыков, которые требуются работодателям;
- Разработать метод формирования навыков ИТ в группы по специализациям исходя из вакансий;

- Разработать приложение, предоставляющее пользователю доступ к результатам анализа вакансий;
- Интегрировать решение в систему "ЦСС".

Анализ текстов вакансий

Ежедневно публикуется до 2000 вакансий, и важно разработать общий алгоритм извлечения навыков из текстов вакансий, который будет работать вне зависимости от структуры текста.

На рисунке 1 представлен пример текста вакансии.



Требуемый опыт работы: 3–6 лет
Полная занятость, удаленная работа

Проект: Приложение на Ruby, включающее клиента командной строки, сервер с бизнес-логикой и интеграции с внешними сервисами. Сайт проекта cryptoalph.ai

Основная задача: развитие и сопровождение проекта по автоматизации биржевой торговли (интеграция с внешними API, торговые роботы, бизнес-логика торговой системы)

Если вы серьезно увлечены торговой тематикой и имеете в ней значительный опыт, возможен постепенный переход в роль тимлида, технического директора проекта.

Обязанности:

— Разработка, поддержка и развитие приложения на Ruby (проектирование архитектуры/интеграции (35%), торговые роботы (45%), бизнес-логика (20%))
— Изучение новых API внешних сервисов и постепенное улучшение и развитие разработанной системы
— Работа с базами данных SQLite, PostgreSQL
— Организация процесса разработки (порядок в репозиториях, автотестах, документации)
— Коммуникация с другими участниками команды (3-5 человек)

Требования:

Использование Ruby с пониманием как обходиться без лишних зависимостей и писать только минимально необходимый код. Приветствуется и возможен переход с C++ или Elixir

— Ruby 2.6, 2.7
— Unix, MacOS, Docker
— Знание Agile и Test Driven Development(цель по покрытию 100%)
— Умение коммуницировать и взаимодействовать распределенной с командой
— Большим преимуществом будет знание фреймворков ZeroMQ и подобных
— Технический английский

Рис. 1. Пример текста вакансии

Для анализа содержания вакансии, помимо словарей характеризующих слов и оцениваемых навыков, требуется коллекция персональных качеств работника. Коллекция персональных качеств состоит из навыков, которые

требуются работодателям вне зависимости от профессии, например «Ответственность» или «Пунктуальность». Для получения такого словаря нужно анализировать не просто отдельные слова, а целые сущности. С такой целью справляется анализ биграмм.

Биграмма – это последовательность из двух элементов, проще говоря словосочетание, за исключением того, что часть речи в таком сочетании не имеет значения. С определением наиболее важных биграмм в тексте поможет TF-IDF – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов [2; 169-174].

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

Важно учесть, что все слова в тексте проходят отдельную обработку лемматизацией, которая приводит слово к его нормальной форме. Это требуется для исключения слов, несущих одинаковый смысл, но имеющих различия в написании, например «Знание» и «Знания». Лемматизация более аккуратно обрабатывает слова в отличие от стеммера, но за это приходится платить большим временем обработки.

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

Таблица 1. Наиболее важные биграммы

Биграмма	TF-IDF мера
опыт работа	0.04673
английский язык	0.03542
работа команда	0.03275
опыт разработка	0.03103
умение разбираться	0.02779
уверенный знание	0.02326

Таблица 2. Наименее важные биграммы

Биграмма	TF-IDF мера
оплата труд	0.003
разработка внедрение	0.0015
дружный команда	0.0013
работа база	0.0009
мозговой штурм	0.0009
международный компания	0.0008

После выявления всех важных биграмм среди большого количества вакансий был произведен ручной отбор персональных качеств.

Пример словаря персональных качеств: работать в команде, английский язык, совместная работа, коммуникабельность, ответственность.

После предварительного сбора данных выполняется основной алгоритм обработки. Для вакансии составляются наборы «Навык-Уровень знаний» в том числе и для личных качеств, за исключением того, что уровень знаний не определяется благодаря оценочному слову, а выбирается заранее и имеет название – “Soft”.

Кластеризация

После обработки каждая вакансия имеет свой набор требуемых технологий, и стоит задача разгруппировать вакансии по специализациям. Эту задачу решает кластеризация.

Кластеризация – это объединение объектов или наблюдений в группы, называемые кластерами, на основе близости значений их признаков. Обучение модели в таком случае ведется только относительно входных данных, без вмешательства в ход обучения «учителя», что является преимуществом кластеризации [1; 151-168].

Перед тем, как приступить к разбиению на кластеры нужно подготовить входные данные. Для этого каждая вакансия представляется в виде вектора из

нулей и единиц, где единицы характеризуют наличие определенного навыка в вакансии, а нули его отсутствие.

Затем набор из векторов новых вакансий подается в обученную на старых данных модель и позволяет ей доучиться на новых данных и провести группировку.

На рисунках 2 и 3 представлены облака слов разных групп навыков.

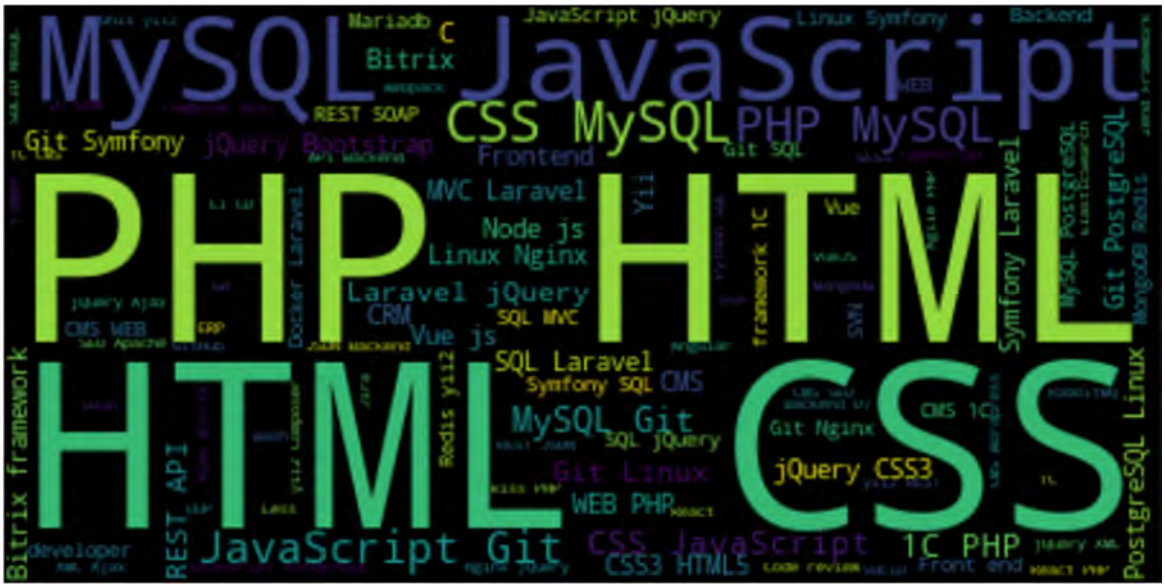


Рис. 2. Облако слов для группы мобильная разработка

Рис. 3. Облако слов для группы веб-разработки

Важно отметить, что имена группам выдаются автоматически на основе наиболее востребованных навыков данной группы.

Результаты

С помощью обработанных текстов вакансий можно проанализировать требуемые персональные навыки на рынке труда. На рисунке 4 можно увидеть статистику упоминаний персональных навыков в текстах вакансий.

Рис. 4. Количество вакансий для персональных навыков

Можно заметить, что очень часто работодатели требуют знать английский язык, уметь работать в команде и быть коммуникабельным. Это объясняется тем, что на нынешнем рынке труда все больше набирают популярность командные проекты, и для успешного выполнения заданий от программиста требуется взаимодействовать с коллегами и работать как часть большой структуры. Большинство полезной для программиста литературы сейчас на английском языке, и его знание просто необходимо для работы с актуальными технологиями.

С помощью полученных групп вакансий можно проанализировать актуальные группы навыков на рынке труда. На рисунке 5 можно увидеть статистику количества вакансий по разным группам навыков.

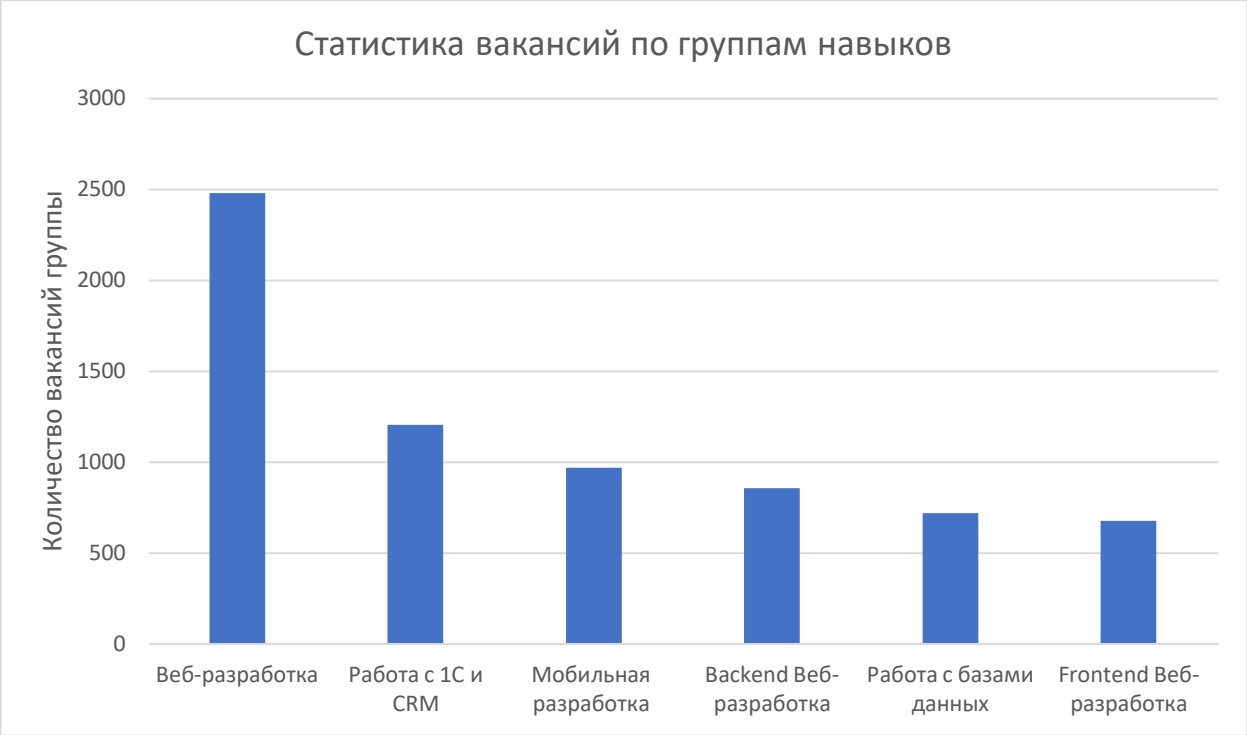


Рис. 5. Количество вакансий по группам навыков

Можно заметить, что лидирующую позицию на рынке труда занимает веб-разработка. Это объясняется тем, что интернет стал неотъемлимой частью жизни современного человека. С помощью создания сайта можно распространять информацию и доносить ее до огромного количества пользователей сети. Важно учесть, что для реализации веб-приложений очень часто разделяют обязанности программистов на Backend и Frontend разработку, что подвержает востребованность умения работать в команде. Так же очень популярна разработка мобильных приложений, что объясняется большой популярностью смартфонов.

Структура проекта

Система «ЦСС» представляет из себя сервер, который состоит из двух частей: Frontend и Backend. Backend часть включает в себя выгрузчик, обработчик текста, кластеризацию, интерфейс для работы с базой данных и интерфейс для отправки данных в Frontend. Frontend часть включает в себя пользовательский интерфейс и интерфейс для получения данных из Backend части.

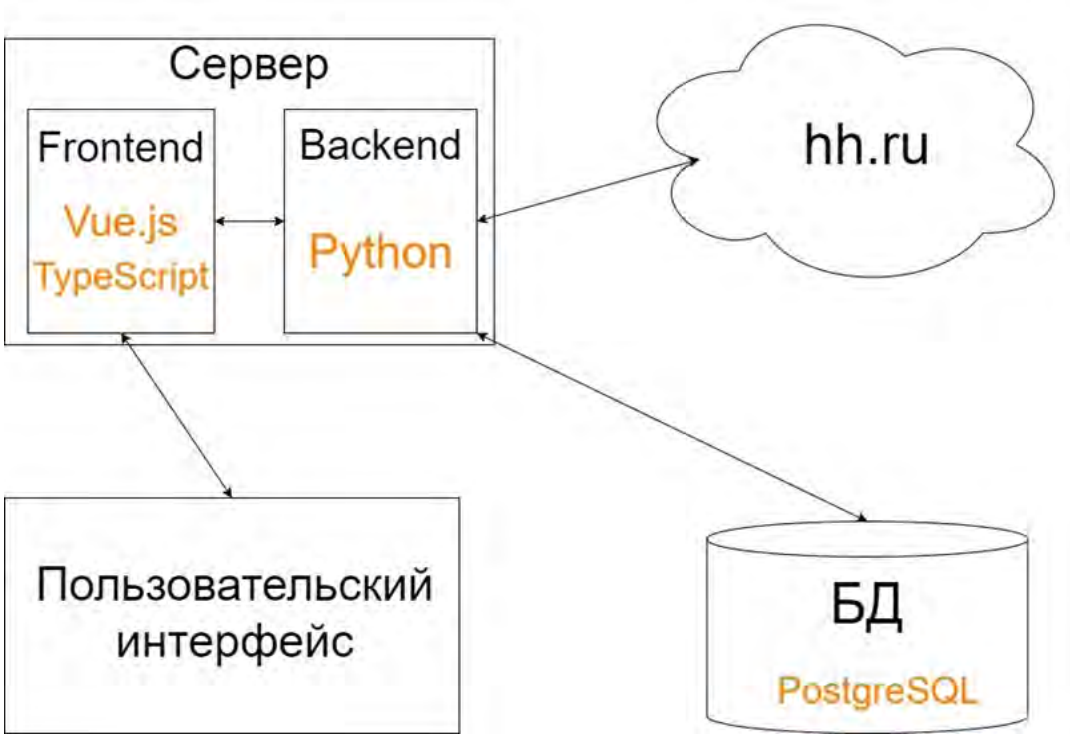


Рис. 6. Схема проекта

С помощью API hh.ru выгрузчик загружает вакансии. После обработки вакансии добавляются в общую базу. Раз в неделю из базы выгружаются данные для кластеризации, после чего информация о группе вакансий обновляется.

Выгрузка, обработчик данных и кластеризация были реализованы на языке Python с использованием библиотек: Scikit-learn, Requests, PyStemmer, NumPy, NLTK. В качестве СУБД в проекте используется PostgreSQL.

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

Пользовательское приложение предоставляет пользователю статистику по выбранной группе навыков. На рисунках 7-9 представлен пример работы приложения.

Рис. 7. Анализ технологий выбранной группы

После выбора группы пользователь видит облако слов выбранной группы и может получить статистику по одному из пяти наиболее востребованных навыка выбранной группы.

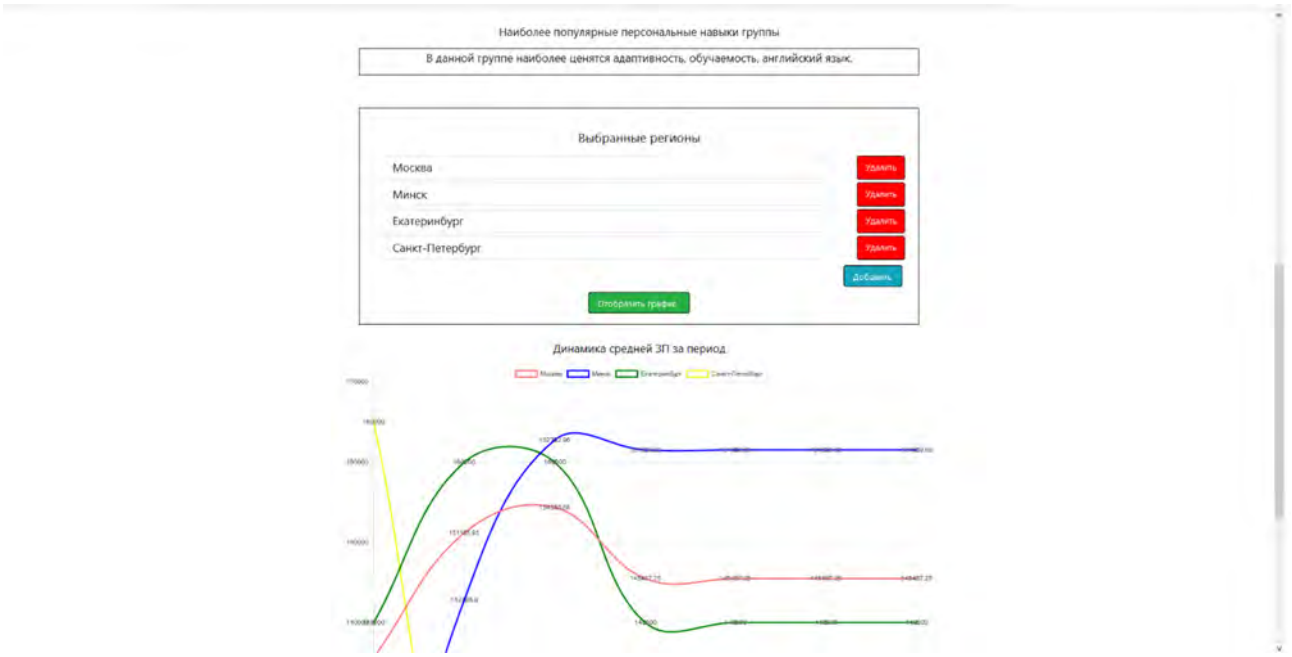


Рис. 8. Анализ популярных персональных навыков
и выбор регионов для анализа

После анализа технологий пользователь видит наиболее популярные персональные навыки по выбранной группе и может пронаблюдать изменения средней заработной платы по регионам, которые он сам выберет.

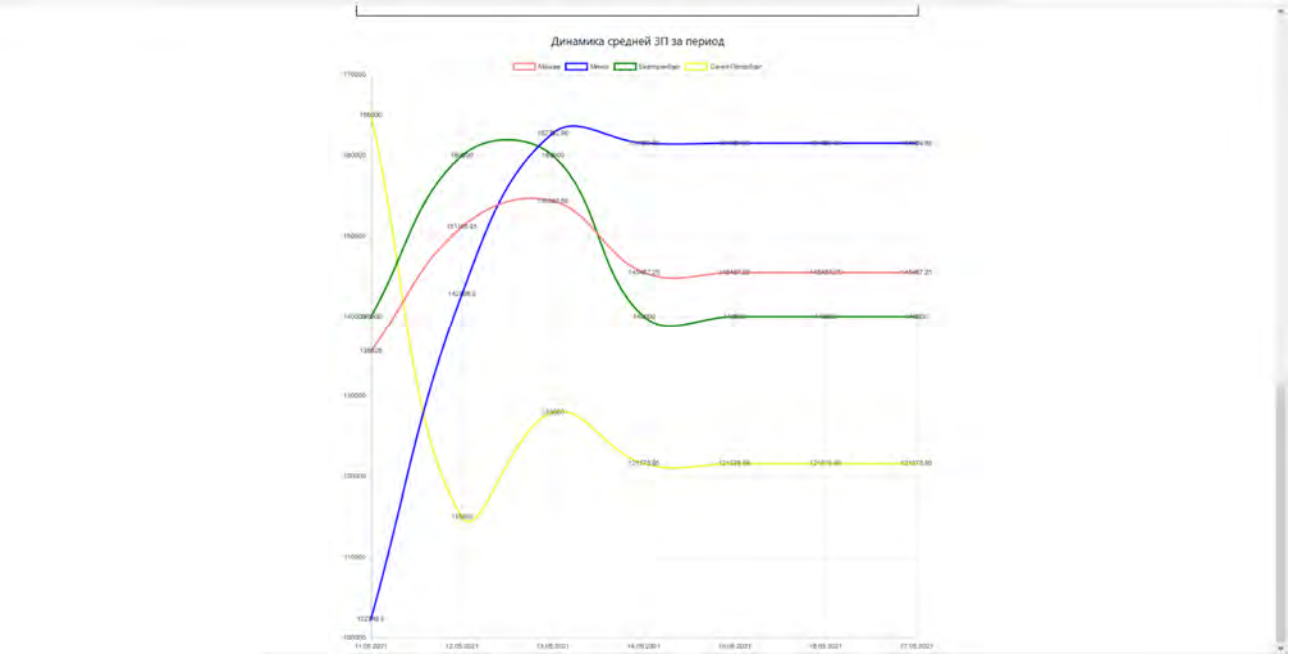


Рис. 9. График изменения средней заработной платы по выбранным регионам

Заключение

В результате работы был разработан анализатор коллекции текстов вакансий для нахождения персональных навыков специалиста с помощью TF-IDF меры. Разработан анализатор текста вакансии, с помощью собственного алгоритма обхода предложения и поиска наборов «Навык-Оценочное слово». Определена структура базы данных для хранения информации о вакансиях с помощью PostgreSQL. Разработан метод группировки вакансий по специализациям. Разработано пользовательское приложение, предоставляющее пользователю доступ к результатам анализа вакансий.

В дальнейшем планируется совершенствование группировки путем разработки более сложного метода кластеризации. Это позволит более точно определять группы и уменьшит влияние «выбросов» на группы. Также планируется усовершенствовать метод извлечения персональных навыков для ускорения работы обработчика текста.

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. Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP / А.А. Барсегян, С.М. Куприянов, В.В. Степаненко, И.И. Холод. – СПб.: «БХВ-Петербург», 2007. – 384 с.

2. Бенгфорт Б. Прикладной анализ текстовых данных на Python / Б. Бенгфорт, Р. Билбро, Т. Охеда. – СПб.: Питер, 2020. – 368 с.