TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа информационных технологий и робототехники
Направление подготовки 09.03.02 «Информационные системы и технологии»
Отделение школы (НОЦ) информационных технологий

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Разработка и рефакторинг программного сервиса для идентификации пользователей-экспертов в социальной сети в заданной предметной области

УДК 004.5.056.523:316.472.4

Студент

Группа	ФИО	Подпись	Дата
8И5А	Кондратьева Анна Александровна		

Руководитель

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Лунёва Е. Е.	к. т. н.		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ОСГН ШБИП	Шулинина Ю. И.	–		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ООД ШБИП	Немцова О. А.	–		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Цапко И. В.	к. т. н.		

Томск – 2019

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код	Результат обучения
Профессиональные и общепрофессиональные компетенции	
Р1	Применять базовые и специальные естественнонаучные и математические знания для комплексной инженерной деятельности по созданию, внедрению и эксплуатации геоинформационных систем и технологий, а также информационных систем и технологий в бизнесе
Р2	Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач
Р3	Ставить и решать задачи комплексного анализа, связанные с созданием геоинформационных систем и технологий, информационных систем в бизнесе, с использованием базовых и специальных знаний, современных аналитических методов и моделей
Р4	Выполнять комплексные инженерные проекты по созданию информационных систем и технологий, а также средств их реализации (информационных, методических, математических, алгоритмических, технических и программных)
Р5	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания геоинформационных систем и технологий, а также информационных систем и технологий в бизнесе
Р6	Внедрять, эксплуатировать и обслуживать современные геоинформационные системы и технологии, информационные системы и технологии в бизнесе, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды
Универсальные (общекультурные) компетенции	
Р7	Использовать базовые и специальные знания в области проектного менеджмента для ведения комплексной инженерной деятельности
Р8	Осуществлять коммуникации в профессиональной среде и в обществе в целом. Владеть иностранным языком (углублённый английский язык), позволяющим работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты комплексной инженерной деятельности
Р9	Эффективно работать индивидуально и в качестве члена команды, состоящей из специалистов различных направлений и квалификаций
Р10	Демонстрировать личную ответственность за результаты работы и готовность следовать профессиональной этике и нормам ведения комплексной инженерной деятельности
Р11	Демонстрировать знания правовых, социальных, экологических и культурных аспектов комплексной инженерной деятельности, а также готовность к достижению должного уровня физической подготовленности для обеспечения полноценной социальной и профессиональной деятельности

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа информационных технологий и робототехники
Направление подготовки 09.03.02 «Информационные системы и технологии»
Отделение школы (НОЦ) информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Цапко И. В.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

бакалаврской работы

(бакалаврской работы, дипломного проекта/работы, магистерской диссертации)

Студенту:

Группа	**ФИО**
8И5А	Кондратьевой Анне Александровне

Тема работы:

Разработка и рефакторинг программного сервиса для идентификации пользователей-экспертов в социальной сети в заданной предметной области

Утверждена приказом директора (дата, номер)	3654/с от 13.05.2019

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Цель работы – разработка и рефакторинг элементов существующего программного сервиса для идентификации пользователей-экспертов в социальной сети в заданной предметной области. Объектом исследования является существующий программный сервис для идентификации пользователей-экспертов в социальной сети в заданной предметной области, а также методы

	поиска семантически близких слов в социальных сетях. Подробно сервис описан в работе, получить доступ к которой можно по ссылке – http://earchive.tpu.ru/handle/11683/48410. Для задач семантического анализа используется коллекция векторных представлений слов, извлечённых из социальной сети Twitter – https://nlp.stanford.edu/projects/glove/.
Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	• Анализ предметной области в задаче поиска пользователей-экспертов в социальных сетях; • изучение теории по семантическому анализу; • формирование требований к разработке модулей сервиса; • проектирование и разработка модулей сервиса; • рефакторинг; • разработка программной документации.
Перечень графического материала *(с точным указанием обязательных чертежей)*	Презентация в формате *.pptx на 17 слайдах.

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Социальная ответственность	Немцова Ольга Александровна, ассистент ООД ШБИП
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Шулинина Юлия Игоревна, ассистент ОСГН ШБИП

Названия разделов, которые должны быть написаны на русском и иностранном языках:

Заключение

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Лунёва Е. Е.	к. т. н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5А	Кондратьева Анна Александровна		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Министерство науки и высшего образования Российской Федерации
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский Томский политехнический университет» (ТПУ)

Школа информационных технологий и робототехники
Направление подготовки 09.03.02 «Информационные системы и технологии»
Уровень образования бакалавриат
Отделение школы (НОЦ) информационных технологий
Период выполнения весенний семестр 2018/2019 учебного года

Форма представления работы:

бакалаврская работа

(бакалаврская работа, дипломный проект/работа, магистерская диссертация)

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	

Дата контроля	Название раздела (модуля)/вид работы (исследования)	Максимальный балл раздела (модуля)
01.03.2019	*Анализ предметной области*	10
15.03.2019	*Проектирование модулей сервиса*	10
15.05.2019	*Разработка и рефакторинг модулей сервиса*	45
15.04.2019	*Финансовый менеджмент, ресурсоэффективность и ресурсосбережение*	10
17.05.2019	*Социальная ответственность*	10
31.05.2019	*Оформление пояснительной записки*	15

Составил преподаватель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Лунёва Е. Е.	к. т. н.		

СОГЛАСОВАНО:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Цапко И. В.	к. т. н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	**ФИО**
8И5А	Кондратьевой Анне Александровне

Школа	ИШИТР	**Отделение (НОЦ)**	ОИТ
Уровень образования	бакалавриат	**Направление/специальность**	09.03.02 Информационные системы и технологии

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	Оклад инженера – 21760 руб. Оклад руководителя – 33664 руб.
2. *Нормы и нормативы расходования ресурсов*	Премиальный коэффициент 30%; Коэффициент доплат и надбавок 20%; Районный коэффициент 30%; Коэффициент дополнительной заработной платы 12%; Накладные расходы 16%.
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	Коэффициент отчислений на уплату во внебюджетные фонды 30%.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения*	– Анализ конкурентных технических решений.
2. *Планирование и формирование бюджета научных исследований*	Формирование плана и графика разработки: – определение структуры работ; – определение трудоемкости работ; – разработка графика Гантта. Формирование бюджета затрат на научное исследование: – материальные затраты; – затраты на специальное оборудование; – заработная плата (основная и дополнительная); – отчисления на социальные цели; – накладные расходы.
3. *Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования*	– Определение потенциального эффекта исследования.
Перечень графического материала *(с точным указанием обязательных чертежей)***:**	
1. *Оценочная карта конкурентных технических решений* 2. *Матрица SWOT* 3. *График Гантта* 4. *Расчет бюджета затрат*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Ассистент ОСГН ШБИП	Шулинина Ю. И.	—		

Задание принял к исполнению студент:

Группа	**ФИО**	**Подпись**	**Дата**
8И5А	Кондратьева Анна Александровна		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8И5А	Кондратьевой Анне Александровне

Школа	ИШИТР	**Отделение (НОЦ)**	Информационных технологий
Уровень образования	бакалавриат	**Направление/специальность**	09.03.02 Информационные системы и технологии

Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Данная работа представляет собой комплекс улучшений существующего сервиса, направленного на идентификацию пользователей-экспертов в социальной сети в заданной предметной области. Объектом исследования является существующий программный сервис для идентификации пользователей-экспертов в социальной сети в заданной предметной области, а также методы поиска семантически близких слов в социальных сетях. Рабочее место для использования сервиса представляет собой помещение, оборудованное персональным компьютером.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	В разделе описаны правовые и организационные мероприятия по обеспечению рабочего места разработчика сервиса и пользователя сервиса. Организация рабочего места, оборудованного персональным компьютером, осуществляется в соответствии со следующими нормативными документами: ГОСТ 12.2.033-78 ССБТ, ГОСТ 12.2.032-78 ССБТ, СанПиН 2.2.2/2.4.1340-03.
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	В данном разделе представлен анализ опасных и вредных факторов, которые могут возникать при разработке или эксплуатации сервиса. Вредные факторы: • электромагнитное излучение; • повышенный уровень шума; • недостаточная освещённость рабочей зоны; • статические физические нагрузки; • перенапряжение зрительных анализаторов. Опасные факторы: • опасность поражения электрическим током; • опасность возникновения короткого замыкания; • повышенный уровень статического электричества.
3. Экологическая безопасность	Негативное влияние на окружающую среду связано с утилизацией люминесцентных ламп, используемых для освещения рабочего помещения, а также с эксплуатацией персонального компьютера.
4. Безопасность в чрезвычайных ситуациях	В данном разделе описана наиболее вероятная чрезвычайная ситуации при эксплуатации разработанного сервиса – возникновение пожара.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ассистент ООД ШБИП	Немцова О. А.	—		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И5А	Кондратьева Анна Александровна		

РЕФЕРАТ

Выпускная квалификационная работа содержит 90 страниц, 25 рисунков, 8 таблиц, 31 источник, 3 приложения.

Ключевые слова: Twitter, социальные сети, пользователи-эксперты, социальные графы, задача поиска ключевых игроков.

Объектом исследования является существующий программный сервис для идентификации пользователей-экспертов в социальной сети в заданной предметной области, а также методы поиска семантически близких слов в социальных сетях.

Цель работы – разработка и рефакторинг элементов существующего программного сервиса для идентификации пользователей-экспертов в социальной сети в заданной предметной области.

В процессе исследования проводилось:

- изучение переданного в разработку программного сервиса для идентификации пользователей-экспертов в социальной сети Twitter;
- теоретический обзор материалов по семантическому анализу текстов, отображению слов в векторное пространство, получению меры схожести векторов.

В результате исследования был разработан модуль рекомендаций хештегов по заданному пользователем хештегу для решения задачи повышения качества и увеличения полноты выборки исходных для анализа данных.

Степень внедрения: последняя версия сервиса доступна в интернете по адресу socgraph.tpu.ru.

В будущем планируется наращивать функционал приложения, добавлять новые модули, внедрять новые алгоритмы для анализа выгружаемых из социальной сети данных.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ГОСТ – межгосударственный стандарт.

Рефакторинг – перепроектирование кода, переработка кода, равносильное преобразование алгоритмов, то есть процесс изменения внутренней структуры программы, не затрагивающий её внешнего поведения и имеющий целью облегчить понимание её работы.

СанПиН – санитарные нормы и правила.

ASP.NET – технология создания веб-приложений и веб-сервисов от компании Microsoft.

CSS – формальный язык описания внешнего вида документа, написанного с использованием языка разметки.

HTML – стандартизированный язык разметки документов в интернете.

JavaScript – прототипно-ориентированный скриптовый язык программирования.

JSON (JavaScript Object Notation) – текстовый формат обмена данными.

KPP-POS – Key Player Problem/ Positive. Задача поиска влиятельных пользователей социальной сети.

LCS (longest common subsequence) – наибольшая общая подпоследовательность.

MVC (model-view-controller) – схема использования нескольких шаблонов проектирования.

SVD (singular-value decomposition) – разложение матрицы по сингулярным значениям.

SWOT-анализ (Strengths, Weaknesses, Opportunities, Threats) – метод стратегического планирования, заключающийся в выявлении факторов внутренней и внешней среды организации.

Twitter – это социальная сеть, сервис микроблогов, в которых можно публиковать короткие текстовые сообщения, называемые «твиты» (от англ. tweet). Размер одного твита ограничен 140 символами, включая пробелы.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Социальные сети играют важную роль в коммуникации современных людей. Каждый год число их пользователей увеличивается, при этом один человек способен быть участником нескольких социальных сетей. Высокий уровень вовлеченности человека в процесс виртуального общения в совокупности с привычкой получать и передавать информацию с помощью социальных сетей вызывает интерес у маркетологов, психологов, аналитиков, социологов, рекламодателей и государственных структур [1].

Данные, полученные из социальных сетей, могут быть подвергнуты анализу: от прогнозирования спроса на тот или иной товар до мониторинга общественного мнения и предсказания результатов выборов [2].

Для повышения эффективности анализа выгружаемых из социальных сетей данных целесообразно определять подмножества пользователей-экспертов, которые являются лидерами общественного мнения в заданной предметной области или по заданной тематике. В этом состоит актуальность разработки сервиса, позволяющего автоматизировать поиск пользователей-экспертов в социальных сетях.

В качестве источника данных для анализа была использована социальная сеть Twitter. Выбор обусловлен тем, что концепция Twitter заключается в публикации записей с небольшим объёмом текста, что значительно упрощает его семантический анализ и ускоряет выборку исходных данных.

Целью выпускной квалификационной работы является разработка и рефакторинг элементов существующего программного сервиса для идентификации пользователей-экспертов в социальной сети в заданной предметной области. Одной из основных задач стала разработка модуля рекомендаций хештегов по указанному хештегу, определяющему некоторую предметную область. Этот программный модуль позволяет пользователям получать список семантически близких хештегов по интересующей их теме, тем самым расширяя область поиска пользователей-экспертов.

1 Анализ предметной области в задаче поиска пользователей-экспертов в социальных сетях

До перехода к этапам проектирования, рефакторинга и разработки компонентов программного сервиса требуется провести анализ предметной области, изучить теоретические основы идентификации пользователей-экспертов в социальных графах, а также выяснить, какие задачи необходимо решить для достижения поставленной цели.

1.1 Задача идентификации пользователей-экспертов в заранее заданной предметной области

В настоящее время социальные сети используются людьми не только в развлекательных целях, но также являются эффективными поставщиками информации, на основании которой специалисты различных профилей могут узнавать о проблемах, волнующих общество, или, проводя анализ полученных данных, прогнозировать какие-либо явления или тенденции.

Социальные сети являются благоприятной средой для свободного выражения собственного мнения пользователей относительно конкретных тем, событий, продуктов или услуг. Замечено, что некоторые пользователи имеют способность влиять на других пользователей социальных сетей. Они являются своеобразными «лидерами мнений» [3]: записи на их страницах в социальных сетях получают много комментариев, к их мыслям и мнению прислушиваются, иногда у них просят советов или помощи. Зачастую описанное явление происходит в рамках одной определённой предметной области.

Таким образом, главным аспектом задачи поиска пользователей-экспертов является выявление пользователей, которые распространяют информацию максимально быстро, эффективно и наибольшему количеству других пользователей социальной сети. Публикации таких пользователей находят позитивный отклик, их авторы рассматриваются как специалисты в данной теме. Другими словами, необходимо найти пользователей, максимально связанных с другими пользователями и способных оказывать влияние их мнение.

Следовательно, пользователя-эксперта также можно называть влиятельным пользователем.

Задача поиска влиятельных пользователей социальной сети получила название KPP-POS (Key Player Problem/Positive) [4].

Данная задача может быть решена дистанционными способами: в них задействуются пути между узлами социальных графов [5]. Среди этих способов наиболее известными и эффективными считаются методы, основанные на расчётах показателей коммуникационной эффективности [6], информационной энтропии [7] или показателя Боргатти [4]. Данные методы применяются к социальным графам, построенным по принципу, описанному в работе [8].

В работе [9] показано, что нельзя выделить один лучший метод поиска пользователей-экспертов. Эффективность и валидация методов решения задачи KPP-POS представлена на подробно описанных в литературе социальных графах, в частности, таких как krebs [4], mexican [7, 8] и некоторых других, которые представляют собой невзвешенные и неориентированные графы с заранее известными ключевыми игроками, где большинство известных способов дают сходные, но не идентичные результаты. Это объясняется вариацией характеристик социальных графов, построенных по выборке исходных данных. Для повышения достоверности результатов можно воспользоваться преимуществами наиболее эффективных методов идентификации влиятельных пользователей и применить их в совокупности, интерпретируя сводные результаты с использованием методов машинного обучения, как показано в работах [8, 9]. Следует также отметить, что в 2018 г. В ТПУ был разработан программный сервис, последняя версия которого доступна по адресу http://socgraph.tpu.ru [10]. Данная работа посвящена улучшению сервиса и расширению его функционала.

Однако эффективность решения задачи идентификации пользователей-экспертов зависит от того, насколько полны и точны исходные данные, по которым идентифицируется пользователь-эксперт. Данная задача требует верного определения набора ключевых слов, а также анализа текстов сообщений

на предмет принадлежности к заданной предметной области. Кроме того, рекомендации по группам пользователей-экспертов могут быть пересмотрены с учетом дополнительной информации, например, географического положения пользователя-эксперта. Интересной может быть информация о том, какие пользователи составляют целевую аудиторию, на которую пользователи-эксперты оказывают влияние.

Решение таких задач вместе с увеличением информативности представления результатов может повысить эффективность идентификации-пользователей экспертов.

Таким образом, дальнейшую работу над проектом необходимо разделить на две части: анализ методики сбора исходных данных из социальной сети по заданной предметной области и представление результатов сбора и анализа исходных данных.

1.2 Задача сбора исходных данных из социальной сети по заданной предметной области

Задача идентификации пользователей решается на основе коллекции исходных данных, которая извлекается из социальной сети Twitter по заранее заданной предметной области. В связи с этим значительно снижается вероятность того, что результат анализа будет недостоверным, определённые алгоритмом пользователи-эксперты точно будут иметь авторитет в указанной сфере. От полноты и разнообразия коллекции данных зависит результат решения поставленной задачи. Одним из методов, помогающих улучшить качество выборки данных, является семантический анализ. Он основывается на рассмотрении смысловой структуры слов, нахождении связей между ними, даже если на её наличие указывают только косвенные признаки. Оценка связей между словами производится с помощью определённых метрик, выбранных как наиболее подходящих для задач, решаемых настоящей работой.

1.2.1 Семантический анализ

Алгоритмы автоматического понимания, распознавания, выделения смысловой составляющей текстов могут быть реализованы на основе рассмотрения отношений между единицами текста. Это возможно благодаря методам семантического анализа языка. В настоящей работе рассматриваются латентно-семантический анализ, word2vec и GloVe.

1.2.1.1 Латентно-семантический анализ

Это метод обработки информации на естественном языке, построенный на анализе взаимосвязи библиотеки документов и терминов, в них встречающихся, и выявлении характерных факторов (тематик), присущих всем документам и терминам.

В основе метода латентно-семантического анализа (ЛСА) лежат принципы факторного анализа, в частности, выявление латентных связей изучаемых явлений или объектов [11]. При классификации или кластеризации документов этот метод используется для извлечения контекстно-зависимых значений лексических единиц при помощи статистической обработки больших корпусов текстов.

Иными словами, идея латентно-семантического анализа состоит в следующем: если в исходном вероятностном пространстве, состоящем из векторов (вектором может быть слово, предложение, абзац, документ и т.д.) между двумя любыми векторами не наблюдается никакой зависимости, то после некоторого алгебраического преобразования данного векторного пространства зависимость может появиться, причем её величина будет определять силу ассоциативно-семантической связи между этими двумя словами (предложениями, абзацами) [12].

Исходной информацией для ЛСА является матрица термы-на-документы, которая описывает набор обучающих данных. Далее эта матрица раскладывается во множество ортогональных матриц (применяется разложение матрицы по сингулярным значениям (Singular Value Decomposition, SVD)). Далее

выполняются некоторые математические операции с матрицами, в результате чего термы (слова) и документы представляются в виде векторов в общем пространстве гипотез. Близость между этими векторами определяется как их скалярное произведение.

Метод ЛСА применяется для решения трёх основных задач:

- сравнение двух термов (слов) между собой;
- сравнение двух документов между собой;
- сравнение терма и документа.

Достоинствами латентно-семантического анализа можно считать следующее:

- качественное выявление скрытых зависимостей внутри множества документов;
- применение метода как с обучением, так и без него;
- использование матрицы близости, основанной на частотных характеристиках документов и лексических единиц;
- решение проблем полисемии и омонимии.

Также этот метод имеет и недостатки:

- снижение скорости вычисления при увеличении объема входных данных;
- вероятностная модель метода не соответствуют реальности: предполагается, что слова и документы распределены нормально, хотя в действительности их распределение ближе к распределению Пуассона.

1.2.1.2 Word2vec

В 2013 году исследователи Google под руководством Томаша Миколова разработали группу моделей word2vec [13]. Принцип их работы следующий. Согласно предположению, слова, находящиеся в похожих контекстах, должны описывать близкие по смыслу термины. Исходя из этого, находятся контекстные связи между словами. В основу определения «близкие» заложена контекстная близость слов.

Эти модели обучаются без учителя на большом корпусе текстов, после чего порождают векторное пространство слов. Размерность пространства, порожденного word2vec гораздо меньше, чем размерность пространства, полученного при унитарном кодировании, при котором она равна размеру словаря. К тому же, пространство погружения word2vec плотнее пространства погружения, полученного унитарным кодированием.

У word2vec выделено две основные архитектуры:

- непрерывный мешок слов (Continuous Bag Of Words, CBOW);
- skip-граммы.

Модель с архитектурой CBOW предсказывает слово, если известно окно окружающих его слов. В архитектуре skip-грамм модель предсказывает окружающие слова по заранее известному центральному слову. При этом порядок контекстных слов не влияет на предсказание. Оба варианта word2vec являются мелкими нейронными сетями.

Получаемые на выходе нейронной сети векторные представления слов позволяют находить семантическое расстояние (или контекстную близость) между словами. Основываясь на этом значении, word2vec предсказывает слова. Точность предсказаний может быть повышена большим объёмом обучающих корпусов.

Word2vec применяется в различных задачах обработки естественного языка. Например, для:

- кластеризации слов по принципу их семантической близости;
- выявления семантической близости слов;
- анализа настроений [14].

1.2.1.3 GloVe

Модель GloVe (Global Vectors) была описана в 2013 году тремя учеными (Джеффри Пеннингтон, Ричард Сокер, Кристофер Мэннинг) из Стэнфордского университета в их совместной работе «Global Vectors for Word Representation» [15]. Этот алгоритм описывается как обучаемый без учителя, его целью является

получение векторных представлений слов. Обучение производится на агрегированной глобальной статистике совместной встречаемости слов из корпуса, а получившиеся представления затем располагаются в векторном пространстве слов.

В отличие от word2vec, GloVe – основанная на счётчиках модель, а не прогностическая. На первом этапе строится матрица совместной встречаемости пар (слово, контекст) в обучающем корпусе. Элементы матрицы описывают частоту встречаемости слова, представленного этой строкой в контексте, представленном этим столбцом (рисунок 1).

Рисунок 1 – Преобразование матрицы совместной встречаемости

Затем происходит преобразование матрицы в пару матриц (слово, признак) и (признак, контекст). Это называется факторизацией матрицы, и выполняется оно итеративно методом стохастического градиентного спуска. Уравнение выглядит следующим образом (формула 1):

$$R = P \cdot Q \approx R', \tag{1}$$

где R – исходная матрица совместной встречаемости.

Сначала P и Q инициализируются случайными значениями, а R' воссоздается их перемножением. Разница между реконструированной матрицей R' и исходной R показывает, как нужно изменить P и Q, чтобы R' стала более похожа на R, то есть уменьшилась ошибка реконструкции. Это повторяется, пока алгоритм градиентного спуска не сойдется, и ошибка реконструкции не станет меньше установленного значения. Полученная в этом случае матрица (слово, признак) и становится векторным пространством в смысле GloVe.

1.2.2 Метрики

Методы семантического анализа языка довольно часто пользуются различными метриками для оценки «расстояния» между рассматриваемыми единицами текста. Прежде, чем приступать к оценке расстояния, необходимо отобразить тексты на векторное пространство с помощью алгоритмов погружения слов. Два из таких алгоритмов – word2vec и GloVe, – были описаны ранее. После выполнения погружения слов можно переходить к этапу анализа этого пространства.

Следует ввести математическое понятие «метрики». В множестве точек, называемом пространством, метрикой называется функция d(x,y), которая в качестве аргументов принимает две точки, а возвращает вещественное число [16]. При этом функция удовлетворяет четырём аксиомам:

1. $d(x, y) \geq 0$ (неотрицательность расстояния).
2. $d(x, y) = 0$, когда $x = y$ (все расстояния положительны, кроме расстояния от точки до нее самой).
3. $d(x, y) = d(y, x)$ (симметричность).
4. $d(x, y) \leq d(x, z) + d(z, y)$ (неравенство треугольника).

Далее рассматриваются метрики, применение которых имеет смысл в контексте определения семантической близости различных единиц текста.

1.2.2.1 Евклидовы метрики

Точками n-мерного евклидова пространства являются векторы из n вещественных чисел. Традиционно в этом пространстве используется метрика, называемая L_2-нормой и определенная следующим образом (формула 2):

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \quad (2)$$

Вычисляется сумма квадратов расстояний по каждому измерению, а затем из нее извлекается квадратный корень.

Для евклидовых пространств есть и другие метрики. Для любой константы r можно определить L_r-норму d по формуле 3:

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \left(\sum_{i=1}^{n} |(x_i - y_i)|^r \right)^{1/r}. \quad (3)$$

При $r=2$ получается обычная L_2-норма. Часто используется L_1-норма, называемая манхэттенским расстоянием. В этом случае расстояние между двумя точками вычисляется как сумма абсолютных величин разностей координат по каждому измерению.

1.2.2.2 Расстояние Жаккара

Коэффициент Жаккара является мерой близости, но не настоящей метрикой. Чем ближе множества, тем больше их коэффицинт Жаккара. Метрикой Жаккара называют выражение вида (формула 4):

$$d(x, y) = 1 - SIM(x, y), \quad (4)$$

где $SIM(x, y)$ – отношение размера пересечения множеств к размеру их объединения.

1.2.2.3 Косинусное расстояние

Косинусное расстояние имеет смысл в пространствах, в которых определены измерения, в частности, в евклидовых пространствах. Косинусное расстояние между двумя точками определяется как угол между соответствующими им векторами. Этот угол изменяется в пределах от 0 до 180 градусов.

Для получения косинусного расстояния сначала необходимо вычислить косинус угла, а затем взять арккосинус. Косинус угла вычисляется по формуле 5:

$$\cos(\widehat{\vec{x}, \vec{y}}) = \frac{x \cdot y}{L_2(x) * L_2(y)}. \quad (5)$$

Косинус угла между векторами x и y равен скалярному произведению этих векторов, поделённому на произведение L_2-норм x и y (их евклидовых расстояний от начала координат).

Интерпретация этой метрики геометрическая: чем ближе значение арккосинуса к единице, тем меньше угол между векторами, тем сильнее они «похожи» друг на друга.

1.2.2.4 Редакционное расстояние

Это расстояние имеет смысл использовать в случае, когда точками представлены строки.

В одном случае расстоянием между строками x и y называется наименьшее количество операций вставки и удаления одного символа, в результате которых x превращается в y.

Второй способ определить редакционное расстояние $d(x, y)$ – вычислить наибольшую общую подпоследовательность (LCS) x и y. LCS строк x и y – это строка, образованная удалением элементов из x и y, а ее длина не меньше длины любой строки, построенной таким способом. Тогда редакционное расстояние может быть вычислено как разность суммы длин x и y и удвоенной длины их LCS.

1.2.2.5 Расстояние Хэмминга

В векторном пространстве расстояние Хэмминга между двумя векторами определяется как количество позиций, в которых они отличаются. Как правило, оно применяется для булевых векторов, все элементы которых равны 0 или 1 (например, 101011), но допустимо выбирать элементы из любого множества.

1.2.2.6 Обоснование решения о применении метрики

В ходе изучения способов оценки расстояния между словами, отображёнными на векторное пространство, были рассмотрены различные аргументы, доказывающие, что метрика косинусное расстояние является

наилучшим способом для оценки контекстной близости слов среди описанных в подразделах 1.2.2.1 – 1.2.2.5.

Наиболее важным показателем рассматриваемой метрики стало то, что разработчики модели GloVe, выходные данные который были применены в данной работе, использовали косинусное расстояние для оценки близости полученных векторов [17].

Эффективность применения косинусной меры совместно с предварительно обученными моделями, какими являются GloVe и word2vec, описана в сравнении с другими методами оценки схожести текстов в источнике [18].

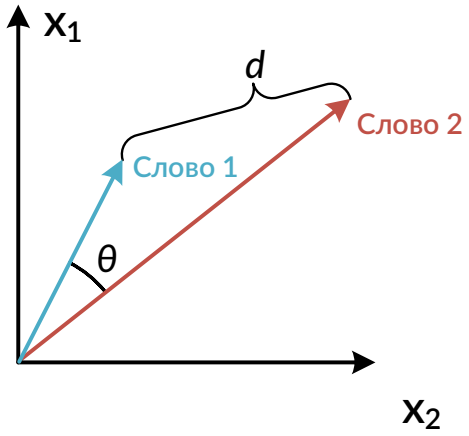

Рисунок 2 – Косинусное (θ) и евклидово (d) расстояния

Таким образом, косинусную меру или её вариации, в отличие от евклидовой метрики, более эффективно использовать, когда идет речь о семантическом сходстве. Величины в векторном представлении некоторого слова зависят от того, насколько часто это слово встречается в документах. Это влияет на модуль векторного представления. Иногда модуль вектора одного слова может быть значительно больше модуля вектора другого слова, что делает евклидово расстояние большим для двух слов, встречающихся в одном тексте. Это обусловлено тем, что одно из слов может быть более популярным, чем второе. Однако не исключена высокая степень семантического сходства между ними. Это происходит потому, что евклидово расстояние зависит от длины

вектора, тогда как косинусное сходство зависит от угла между векторами (рисунок 2).

В источниках [18, 19, 20, 21] отмечается, что угловая мера более устойчива к вариациям количества встречаемости между терминами, которые семантически похожи, тогда как на величину векторов влияют количество встречаемости и неоднородность соседства слов. Слова-соседи используются при обучении нейронной сети для получения векторного представления слов.

Аргументом в пользу косинусного расстояния стало также и то, что отображенные моделью в векторное пространство слова (векторы слов) имеют достаточно большое количество измерений: в векторах выходных данных модели GloVe – от 25 координат. Поэтому для вычисления, например, евклидова расстояния потребовалось бы предварительно нормировать значения координат векторов, а затем приступать к вычислению метрики. В случае использования косинусной меры такой необходимости не возникает.

Таким образом, для решения задачи настоящей работы была использована метрика косинусное расстояние.

1.3 Задача представления результатов сбора и анализа исходных данных

Одним из этапов анализа данных является визуализация – представление этих данных конечному пользователю в виде, который повышает эффективность работы человека, занятого их изучением.

Модули, отвечающие за визуализацию данных, являются важными составными частями программных систем, выполняющих интеллектуальный анализ данных, особенно тех, которые ориентированы на обработку больших объёмов информации со сложной внутренней структурой. В области бизнес-аналитики визуализация применяется широко и на всех этапах работы с данными: от выборки исходных данных до окончательных результатов.

Применительно к теме анализа данных социальных сетей обычно применяются следующие виды визуализации:

- визуализация исходных данных;
- визуализация выборки, подготовленной для обработки;
- визуализация окончательных результатов.

Современные средства представления данных обеспечивают простое отображение большого числа разнотипных данных, демонстрируют кластеры данных, их размеры, схожесть и отличие, позволяют рассматривать данные в контексте, анализируя связи между ними, помогают выбрать подходы к поиску нужной информации. Такими средствами могут являться, например, подключаемые к проектам библиотеки или предоставляемые по API возможности различных сервисов: Vis.js, D3, Cytoscape.js, Polymaps, JointJs, Google, Яндекс, OpenStreetMap и другие.

Актуальность решения данной задачи состоит в том, что с помощью визуального анализа можно выявить различные закономерности, существующие между данными, а также получить информацию о различных показателях выборки: географическом или математическом распределении элементов, неочевидных в числовом выражении тенденциях, относительности и близости соседних элементов и так далее.

1.4 Анализ переданного в разработку сервиса

В рамках выпускной квалификационной работы был выполнен комплекс улучшений существующего программного сервиса. Этот сервис был разработан в Томском политехническом университете [22]. Основной задачей сервиса является поиск пользователей-экспертов в социальной сети Twitter. Несмотря на то, что сервис на момент получения задания работал на production-сервере и был доступен для пользователей, требовалось выполнить рефакторинг и разработку одного из его модулей.

Основными функциями веб-приложения являются:

- Выборка данных для заданной предметной области из социальной сети Twitter;

- Анализ данных и предоставление возможности скачать результаты анализа для исследования во внешних приложениях;
- Визуализация результатов анализа.

Главная страница сервиса изображена на рисунке 3. Она содержит информацию о приложении и блок навигации.

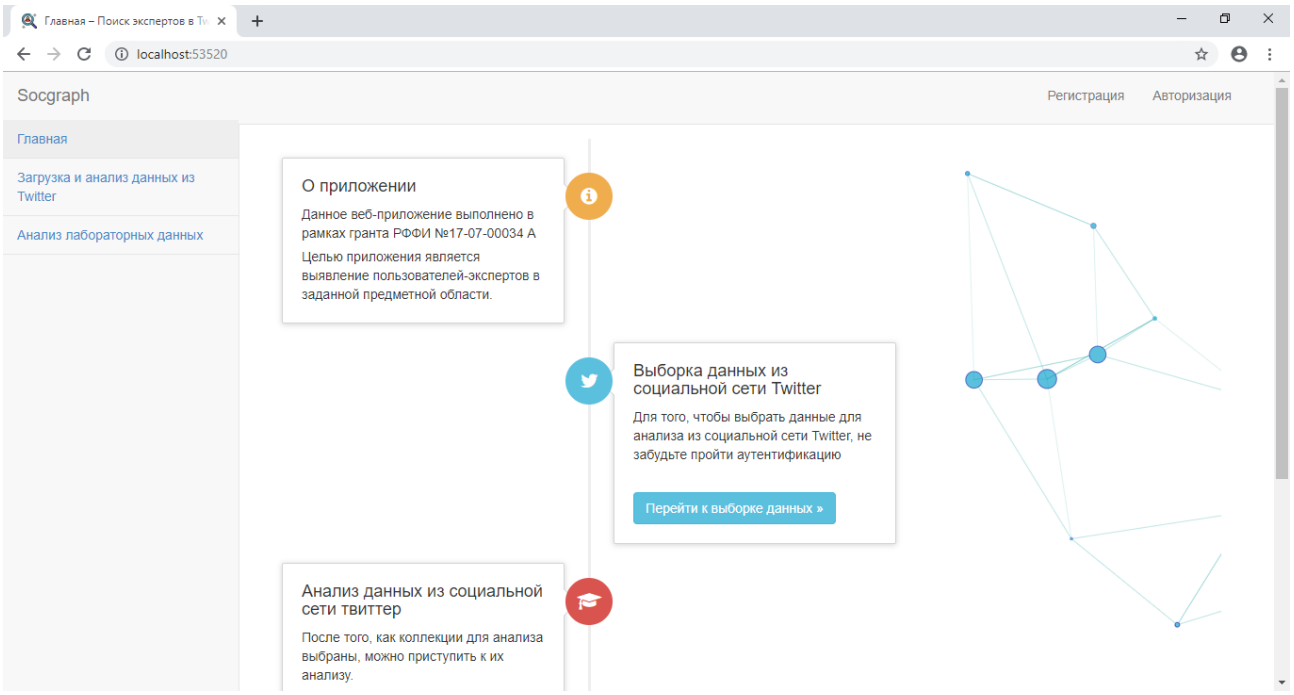

Рисунок 3 – Главная страница сервиса

1.4.1 Предоставляемые сервисом функции

Концепция разработанного программного приложения предполагает взаимодействие сервиса и пользователя. Пользователь имеет возможность выполнять операции, определенные функциональными требованиями к сервису.

В соответствии с требованиями, приложение предоставляет пользователю следующие возможности:

- Позволяет указывать ключевое слово (хештег) для задания интересующей предметной области;
- позволяет устанавливать количество выгружаемых из социальной сети Twitter записей;
- позволяет устанавливать количество искомых пользователей-экспертов;

- осуществляет поиск в социальной сети Twitter публикаций, содержащих указанное ключевое слово;
- осуществляет обработку полученных из Twitter данных;
- идентифицирует пользователей-экспертов на основе обработанных данных;
- визуализирует результат поиска пользователей-экспертов, в том числе посредством обращения к API Google chart и с помощью библиотеки для визуализации данных D3;
- позволяет хранить и сохранять загруженные коллекции данных для дальнейшей обработки.

1.4.2 Варианты использования

Варианты использования приложения определяются исходя из функциональных требований к программному сервису. Описываемый программный сервис предполагает взаимодействие с пользователем, которому позволено производить определённый набор действий с приложением.

Для визуальной интерпретации функциональных возможностей используются диаграммы вариантов использования. Этот вид диаграмм применяется для описания набора действий, которые заявлены как особенности системы.

Варианты использования сервиса представлены на рисунке 4.

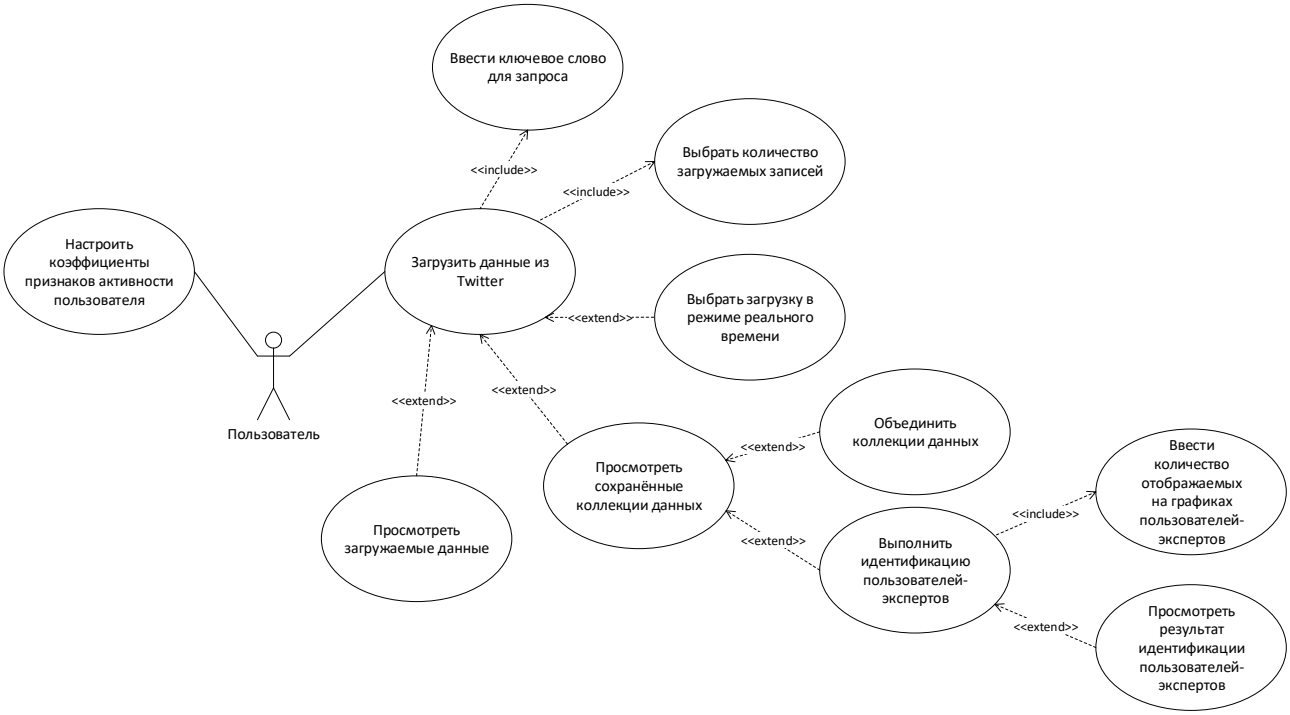

Рисунок 4 – Диаграмма вариантов использования программного сервиса

Варианты использования, изображённые на рисунке 4, были подвергнуты детальному анализу, в результате которого сформировался план работ по улучшению программного сервиса.

1.4.3 Архитектура программного сервиса

Архитектура приложения – это совокупность решений об организации проектируемой системы. Описываемый сервис построен по принципу клиент-серверной архитектуры. Все вычислительные процессы, связанные с загрузкой данных из Twitter, их обработкой и анализом, протекают на серверной стороне, где также хранятся коллекции загруженных ранее данных. На стороне клиента происходит получение информации от пользователя: задание параметров коллекций, визуализация данных, задание числа экспертов для отображения графиков, ввод ключевого слова и т.д. Схема архитектуры программного сервиса представлена на рисунке 5. Архитектура компонента-сервера изображена на рисунке 6.

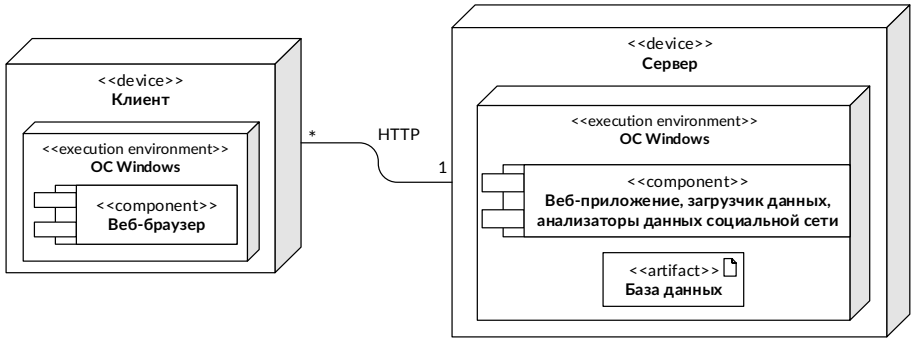

Рисунок 5 – Диаграмма развертывания программного сервиса

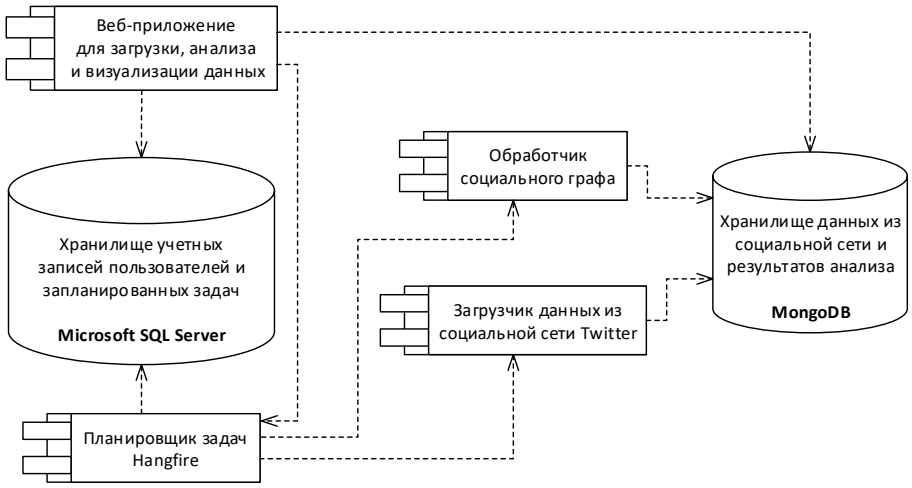

Рисунок 6 – Компонентная архитектура сервера

Серверная часть программного сервиса отвечает за загрузку и анализ данных. Загрузка данных из сети Twitter обеспечивается библиотекой Tweetinvi. Хранение данных аккаунтов пользователей сервиса обеспечивается системой управления базами данных MS SQL. При этом хранение коллекций данных обеспечивается системой управления базами данных MongoDB. Этот выбор сделан ввиду сложной и объёмной структуры данных, получаемых при помощи библиотеки Tweetinvi. Программный сервис некоторые задачи выполняет отдельно от основного потока. Такая функция обеспечивается планировщиком задач Hangfire. Принцип работы Hangfire изображён на рисунке 7 [23].

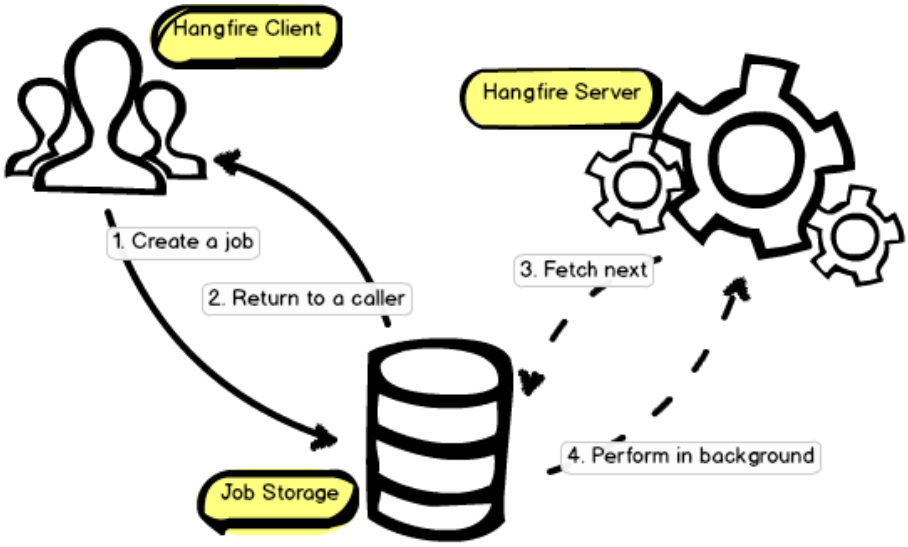

Рисунок 7 – Описание работы планировщика задач Hangfire

Работа с планировщиком происходит следующим образом: процесс-клиент добавляет задачу в базу данных, а процесс-сервер периодически опрашивает базу данных и выполняет задачи из очереди. После окончания выполнения задачи процесс-сервер сохраняет результат её выполнения в базу данных, которая возвращает этот результат процессу-клиенту.

1.4.4 Разработка предложений по улучшению сервиса

В ходе ознакомления с программным сервисом были выделены группы предлагаемых изменений, которые позволили бы повысить эффективность работы с приложением, а также улучшить восприятие предоставляемой информации. Улучшения, которые было решено внедрить в проект, отображены на схеме (рисунок 8).

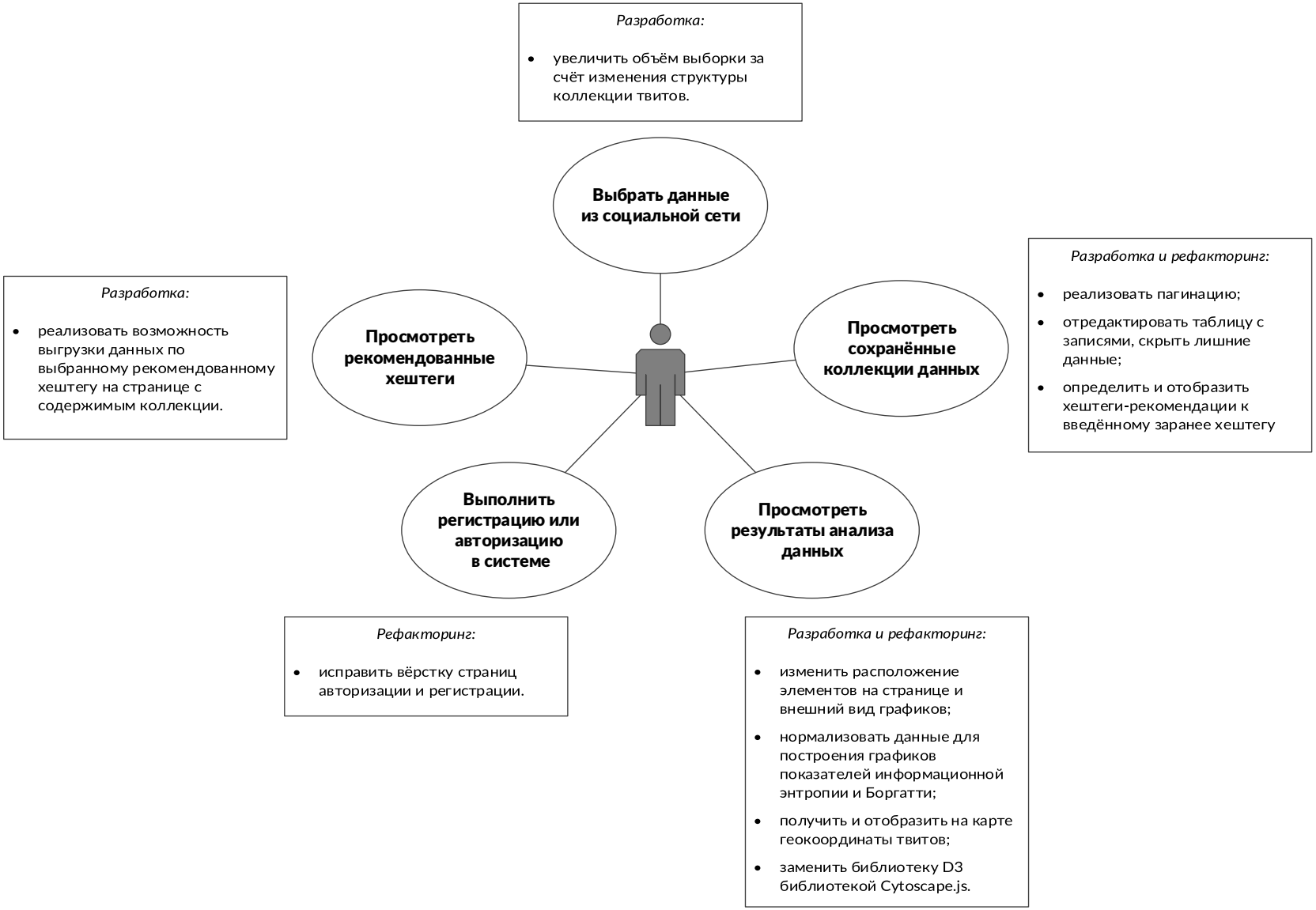

Рисунок 8 – Схема предлагаемых улучшений

1.5 Цели и задачи выпускной квалификационной работы

Исходя из анализа предметной области, приведённого в подразделах 1.1 – 1.4, целью выпускной квалификационной работы является разработка и рефакторинг элементов существующего программного сервиса для идентификации пользователей-экспертов в социальной сети в заданной предметной области.

Для достижения поставленной цели необходимо решить следующие задачи:

- ознакомиться с веб-сервисом и изучить теорию;
- разработать предложения по улучшению сервиса;
- провести визуальный рефакторинг страниц приложения;
- разработать модуль рекомендаций хештегов по заранее заданному хештегу;
- внедрить в проект библиотеку для анализа и визуализации социальных графов Cytoscape.js;
- исправить вёрстку страниц авторизации и регистрации пользователей;
- внедрить виджет карты для отображения геометок, полученных из твитов загруженных коллекций.

2 Проектирование и разработка модулей сервиса

Данный раздел содержит описание этапов создания модулей программного приложения – проектирования и реализации. На этапе проектирования были сформулированы требования к выполнению поставленных задач.

2.1 Описание применяемых технологий

В качестве основного языка программирования использован C# – язык, реализующий принципы объектно-ориентированного программирования. Средой разработки программного сервиса стал продукт от компании Microsoft – Visual Studio 2017 Enterprise. Проект разработан с применением технологии ASP.NET, реализующей шаблон проектирования Model-View-Controller (MVC), концепция которого предусматривает разделение приложения на три компонента: модель, представление, контроллер. Данные компоненты выполняют следующие функции:

- класс представления отвечает за визуализацию компонентов и обеспечение взаимодействия пользователя и приложения;
- класс модели описывает структуру используемых данных;
- класс контроллера обеспечивает корректное взаимодействие модели, представления и базы данных.

Ввиду того, что разработка связана с получением данных из социальной сети Twitter, для легкого и надёжного доступа к API Twitter применялась .NET-библиотека Tweetinvi.

Отдельные фрагменты кода разработаны на скриптовом языке программирования JavaScript, использованы библиотеки Cytoscape.js для визуализации социального и jQuery для отправки запросов на сервер без необходимости полностью перезагружать страницы приложения.

2.2 Функциональные требования к разрабатываемым модулям

Действия, выполняемые приложением, определяются функциональными требованиями к этому приложению. От предоставляемого программой функционала зависит будущее продукта на рынке: спрос, конкурентоспособность, целевая аудитория. Также функциональные требования позволяют определить компоненты архитектуры будущего приложения.

Таким образом, разработанный модуль, обеспечивающий сбор расширенных коллекций исходных данных из социальной сети по заданной предметной области должен предоставлять следующие функции:

- просмотр рекомендованных хештегов для заранее определённой предметной области;
- возможность выборки исходных данных в предметных областях, характеризуемых рекомендованными хештегами.

Компоненты, отвечающие за визуализацию исходных данных и результатов их анализа, должны предоставлять возможности:

- сбора и визуализации информации о географическом положении пользователей твитов, хранящихся в коллекциях данных;
- просмотра крупных и информативных графиков, отображающих результаты анализа данных;
- удобного просмотра исходных данных на страницах, не содержащих лишней информации;
- комфортного взаимодействия пользователя и приложения в результате правильного расположения элементов на страницах.

2.3 Разработка вариантов использования, соответствующих функциональным требованиям к модулям программного сервиса

На основе приведённых в подразделе 2.2 функциональных требований были определены соответствующие варианты использования. Некоторые уже реализуемые сервисом варианты использования в ходе разработки были

изменены. На рисунке 9 представлена диаграмма вариантов использования сервиса, актуальная на текущий момент.

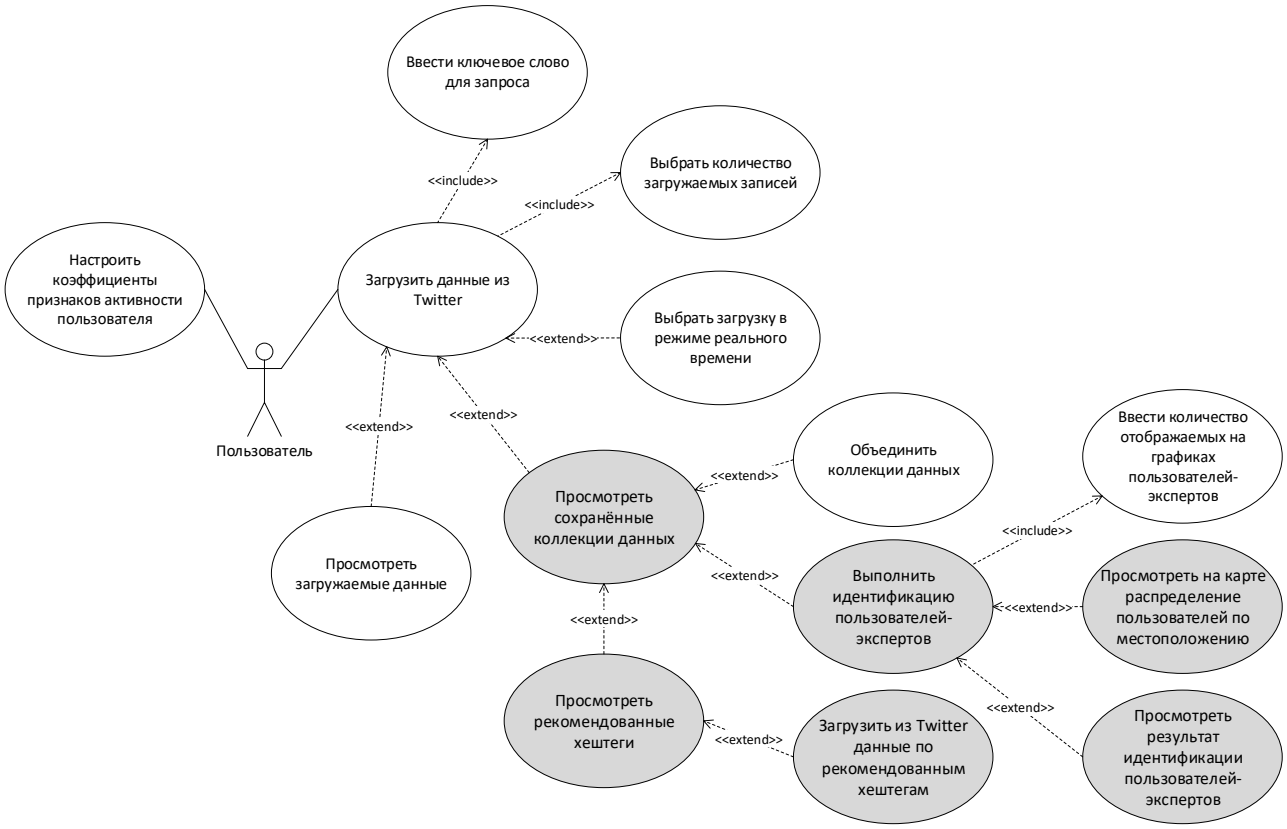

Рисунок 9 – Диаграмма вариантов использования сервиса после решения поставленных задач

Варианты использования, которые были затронуты или введены в рамках настоящей работы, отображены на диаграмме в виде элементов с заливкой серого цвета. Исходная диаграмма представлена на рисунке 4 подраздела 1.4.2.

2.4 Архитектура программных модулей

Исходная архитектура программного сервиса для идентификации пользователей-экспертов в социальной сети представлена на рисунке 6 подраздела 1.4.3. На рисунке 10 представлена схема компонентов сервера, актуальная на момент завершения разработки в рамках настоящей работы.

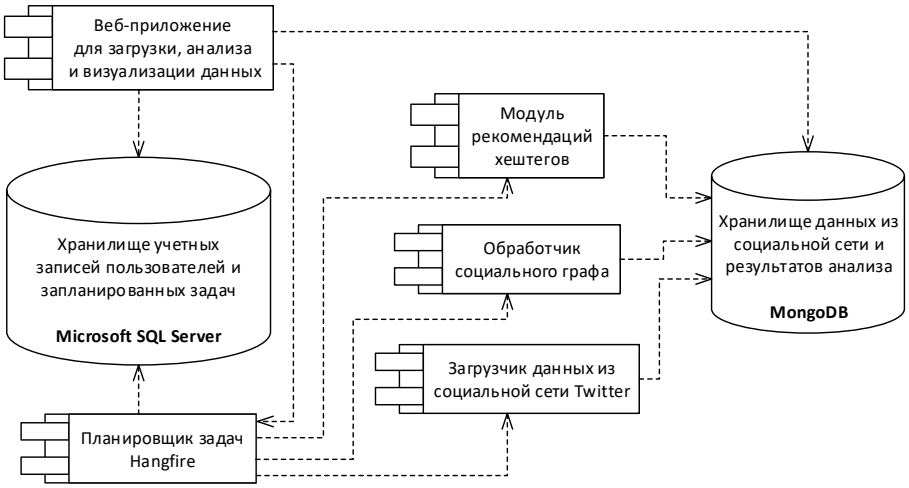

Рисунок 10 – Компонентная архитектура сервера после решения поставленных задач

Модуль рекомендаций хештегов был создан в ходе решения задачи сбора исходных данных из социальной сети по заданной предметной области. На рисунке 11 изображена микроархитектура программного компонента.

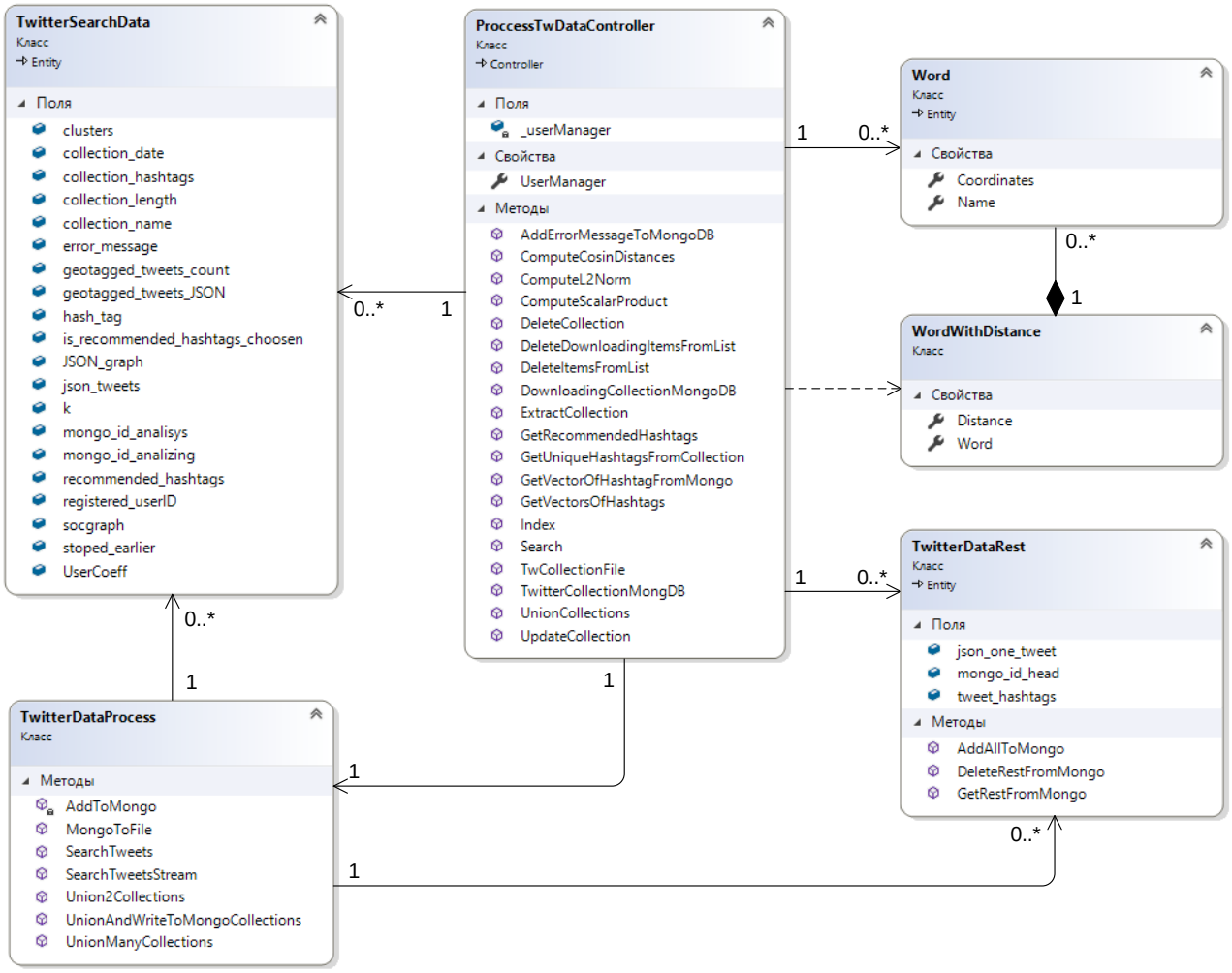

Рисунок 11 – Диаграмма классов модуля рекомендаций хештегов

При разработке этого модуля были модифицированы уже существующие в приложении классы, а также добавлены новые, необходимые для решения текущей задачи. Представленные на рисунке 11 классы выполняют следующие функции в приложении:

- ProcessTwDataController реализует методы контроллера и отвечает за управление коллекциями и обработку полученных коллекций твитов для определения рекомендаций;
- TwitterSearchData представляет собой модель коллекции данных;
- TwitterDataProcess реализует методы взаимодействия с базой данных, в которой хранятся коллекции твитов и результаты анализа этих коллекций;
- TwitterDataRest отвечает за загрузку данных из сети Twitter по заданному ключевому слову, также реализует методы доступа к хранилищу данных;
- Word представляет собой модель для хранения ключевого слова – хештега;
- WordWithDistance представляет собой модель, в которой хештегу сопоставлено его векторное представление.

Архитектура программных модулей, представленных на рисунке 10, была изменена в соответствии с предъявляемым функциональными требованиями. Полный список изменений с указанием модуля, к которому они относятся, представлен в таблице А.1 (приложение А).

2.5 Разработка модуля рекомендаций хештегов по заданной предметной области

Сбор рекомендаций – задача семантического анализа хештегов, которые пользователи упоминают в своих постах в Twitter. Для ее решения необходимо определить некоторое количество хештегов, наиболее близких по смыслу к введенному пользователем хештегу. Сравнить хештеги можно при помощи метрики. На основании аргументов, содержащихся в подразделе 1.2.2.6, была

выбрана метрика косинусное расстояние. Так как она предполагает работу с векторами, возникла необходимость отображения хештегов в векторное пространство.

Векторные представления слов извлекались из созданной на базе Стэнфордского университета коллекции, содержащей 1,2 миллиона слов, употребляемых в социальной сети Twitter и представленных в виде векторов с 25, 50, 100, или 200 координатами, которыми являются вещественные числа [24]. Данная коллекция была получена в результате обработки слов, извлечённых из социальной сети Twitter, моделью GloVe, описанной в подразделе 1.2.1.3. Для того, чтобы сравнивать векторные представления слов непосредственно в приложении, коллекция была загружена в базу данных под управлением MongoDB. Это было выполнено с помощью разработанного вспомогательного приложения, этапами работы которого являются:

1) установка соединения с сервером MongoDB;

2) построчное извлечение слов и координат соответствующих им векторов из файла с коллекцией данных Twitter;

3) запись слов и координат соответствующих им векторов в сущность для последующей работы с MongoDB;

4) добавление слов в базу данных.

Общий алгоритм подбора рекомендаций по введённому пользователем хештегу представлен диаграммой в нотации DFD на рисунке 12.

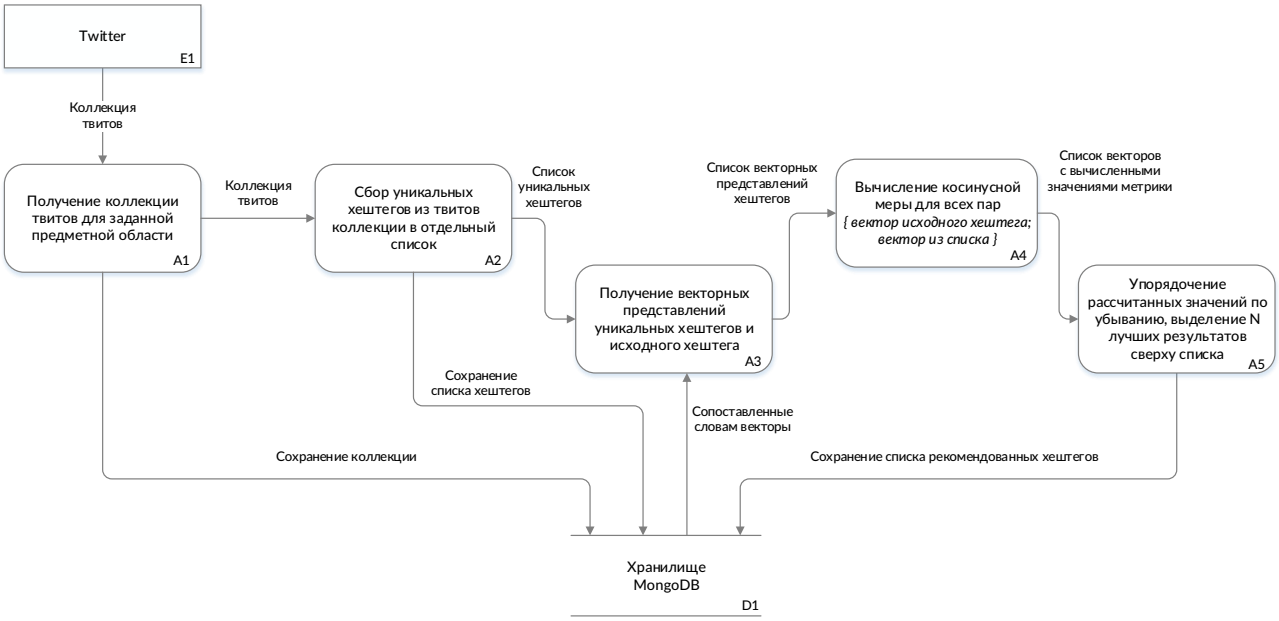

Рисунок 12 – Диаграмма потоков данных для процесса подбора рекомендаций

Представление результатов пользователю происходит при обращении к соответствующей странице веб-приложения, отображающей состав коллекции твитов и рекомендуемые хештеги.

2.6 Разработка модуля визуализации информации о местоположении пользователей

Данные о местоположении пользователей социальной сети могут помочь в выявлении зависимостей между предметными областями и регионами, в которых эти пользователи находятся. А также, получая и анализируя информацию о геопозиции, можно понять, какие темы волнуют пользователей различных стран или континентов, где аудитория более активная, а где менее активная, какие темы важны для пользователей различных государств.

В рамках настоящей работы было выполнено внедрение виджета карты в представление с результатами анализа коллекции. Для достижения результата потребовалось провести анализ возможных вариантов реализации. В качестве встраиваемых модулей карт рассматривались решения Google Maps, Яндекс.Карты и OpenStreetMap. Отказ от использования Google Maps был обусловлен тем, что предоставляемый сервисом бесплатный тарифный план не удовлетворял потребностям разрабатываемого приложения. OpenStreetMap

обладает плохо поддерживаемым API, поэтому выбор был сделан в пользу Яндекс.Карт.

Сущность выгруженной из Twitter записи может содержать в качестве информации о местоположении два объекта: Place и Coordinates. В них соответственно указывается название места, откуда была отправлена запись и геокоординаты этого места. В отдельных случаях в сущности было указано только место. В этом случае использовался Геокодер Яндекс.Карт – сервис, позволяющий определить координаты объекта по его адресу или определить адрес по координатам. На рисунке 13 представлена структурная схема работы с API Яндекс.Карт.

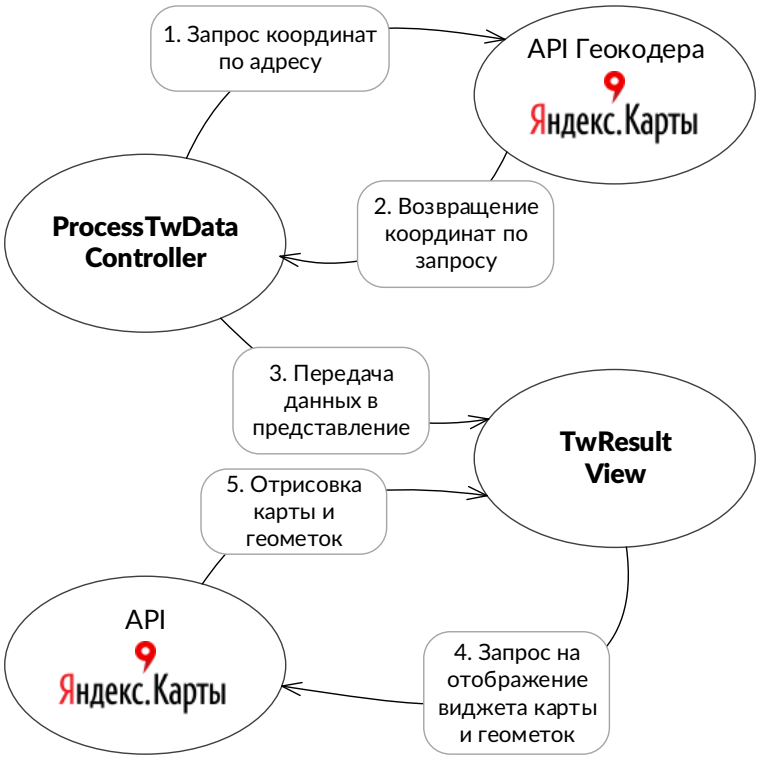

Рисунок 13 – Структурная схема алгоритма работы с API Яндекс.Карт

Для взаимодействия описываемого приложения с Геокодером Яндекс.Карт был разработан вспомогательный модуль на стороне клиента, главными задачами которого являются отправление запросов на получение географических координат по указанному адресу и обработка ответов от сервера.

2.7 Визуальный рефакторинг страниц приложения

Визуальный рефакторинг связан с исправлением недостатков пользовательского интерфейса. В проект вместо библиотеки D3 (Data-Driven Documents) была внедрена библиотека Cytoscape.js, предназначенная специально для работы с графами и имеющая в арсенале встроенные функции для расчёта различных коэффициентов и показателей центральности. Такое решение способно удовлетворить вероятные потребности пользователя в средствах анализа графа.

Для комфортного просмотра содержимого выгруженных из социальной сети Twitter коллекций была добавлена пагинация, предоставляющая возможность выбора количества отображаемых твитов на каждой из страниц.

Другие изменения, коснувшиеся визуальных компонентов программы, представлены в таблице 1.

Таблица 1 – Изменения, связанные с отображением веб-страниц

Название компонента-представления	Характер изменений
Login.cshtml	Внесены изменения, связанные с исправлением вёрстки страницы.
Register.cshtml	
ExtractCollection .cshtml	1. В коде представления появились изменения, связанные с передачей хештега между двумя представлениями посредством записи выбранного хештега в URL-параметр загружаемой страницы. 2. Исправлена вёрстка страницы, удалены лишние элементы. 3. Добавлена автоматическая перезагрузка страницы, пока не будут отображены рекомендованные хештеги.

Продолжение таблицы 1

Название компонента-представления	Характер изменений
ProccessTwData /Index.cshtml	В коде представления появились изменения, связанные с автозаполнением поля для ключевого слова, если это представление было открыто в результате нажатия на какой-либо рекомендованных хештег в представлении с содержимым коллекции твитов.
TwResult.cshtml	1. Изменены элементы интерфейса, структура страницы, изменено отображение графиков с результатами идентификации экспертов. 2. Исправлена вёрстка. 3. Внесены изменения в код, отвечающий за отрисовку диаграммы, на которой показаны результаты выделения кластеров.
DrawGraph.cshtml	Изменён размер элемента, в котором происходит отрисовка визуализированного социального графа.
Borgatti.cshtml	Проведён подбор параметров отображения диаграммы с учётом наилучшего отображения данных.
Clusters.cshtml	
DrawGraph.cshtml	

2.8 Результаты работы

В результате выполнения выпускной квалификационной работы был разработан модуль рекомендаций хештегов и проведён рефакторинг функциональных возможностей приложения.

На рисунке 14 представлена страница для задания параметров выборки твитов.

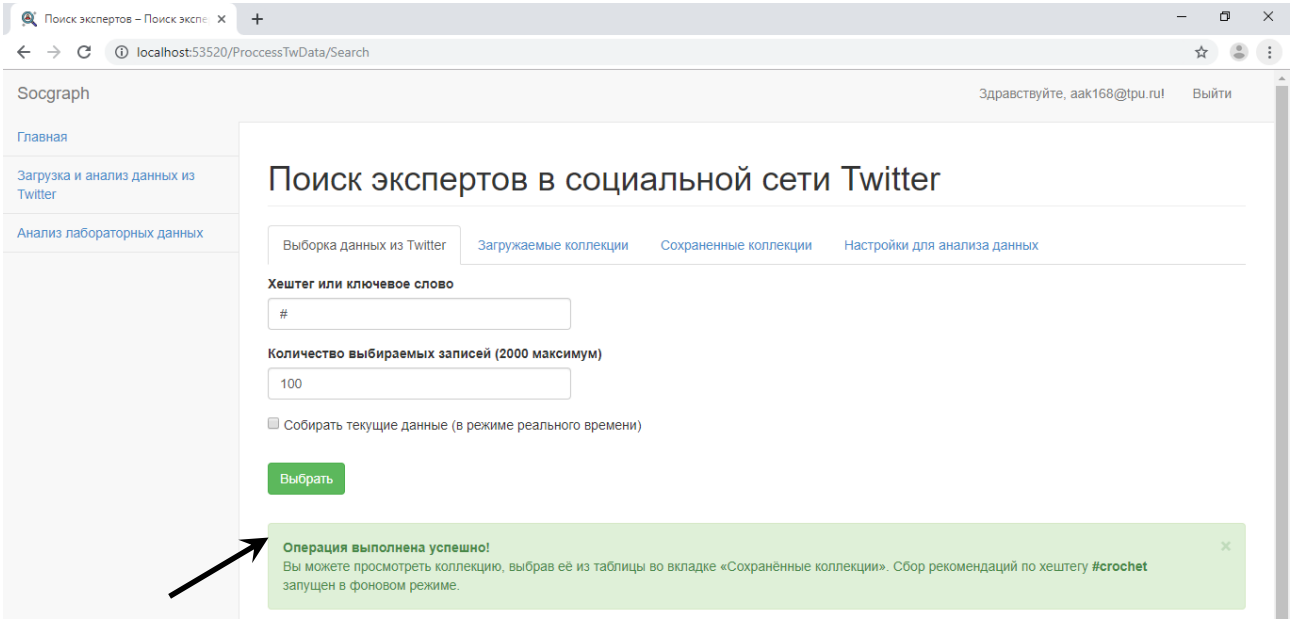

Рисунок 14 – Страница для задания параметров выборки твитов (после отправки запроса)

После отправки запроса на экране отображается информационное сообщение о том, что работа по сбору рекомендаций выполняется в фоновом режиме (показано стрелкой на рисунке 14). Результат выполнения этой работы можно увидеть на рисунках 15-17: на странице, содержащей информацию о твитах в выбранной коллекции, отображается панель, в которой присутствуют 15 хештегов, наиболее похожих по смыслу на заданный пользователем исходный хештег.

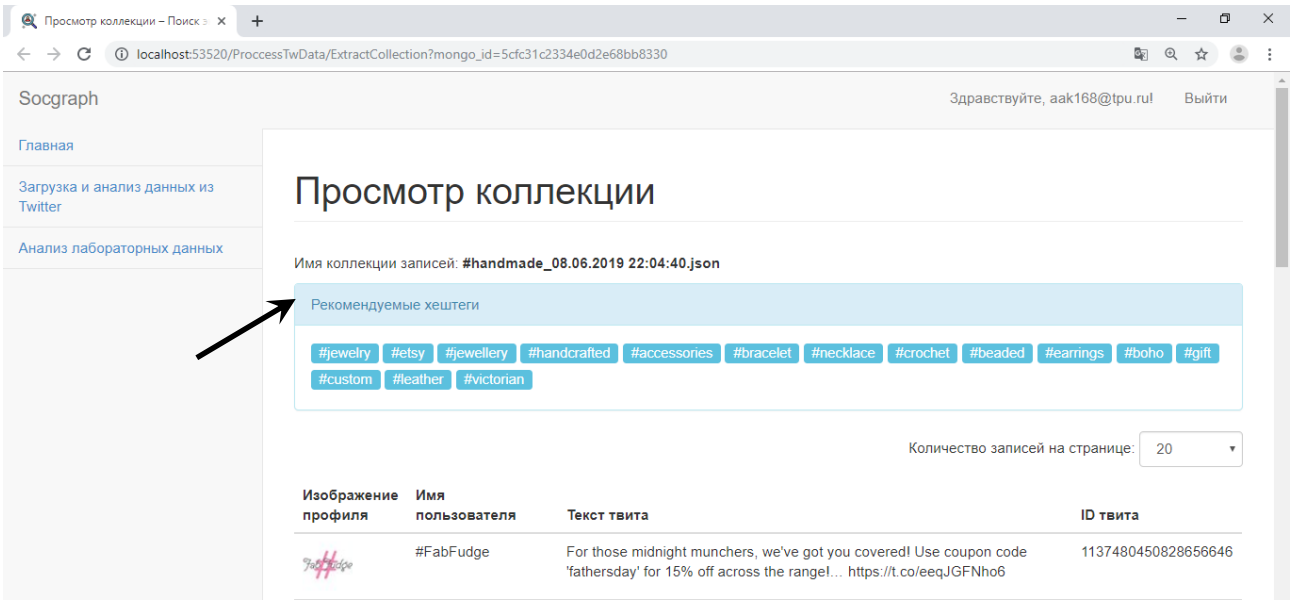

Рисунок 15 – Отображение панели с хештегами, рекомендованными для хештега “handmade”

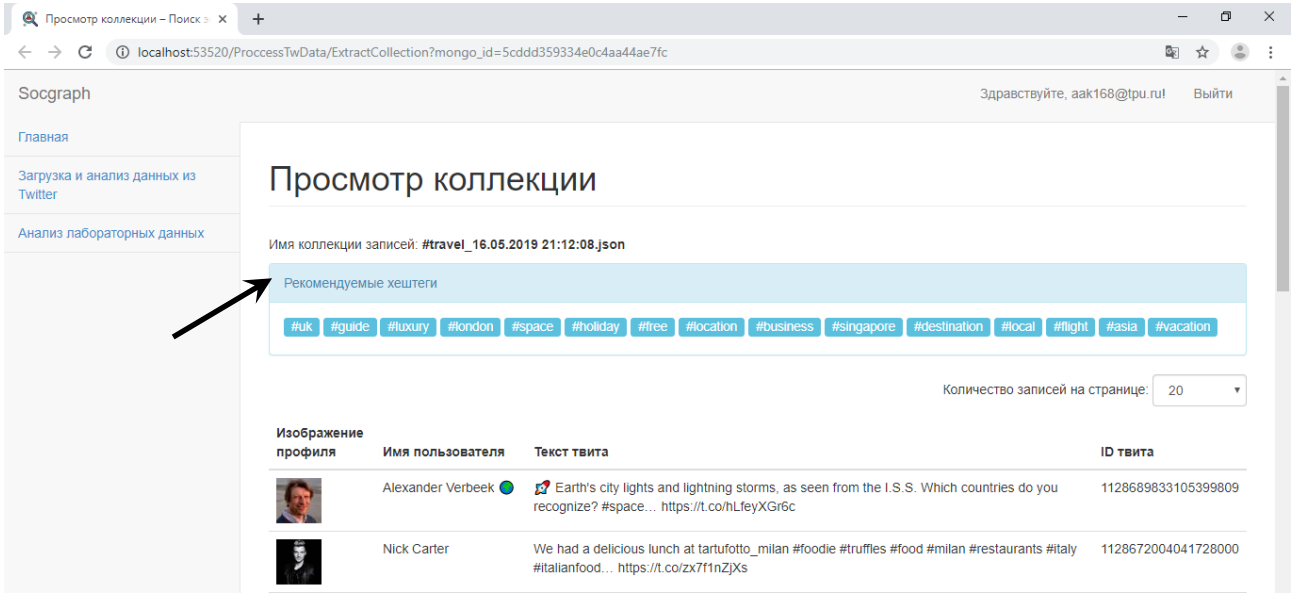

Рисунок 16 – Отображение панели с хештегами, рекомендованными для хештега “travel”

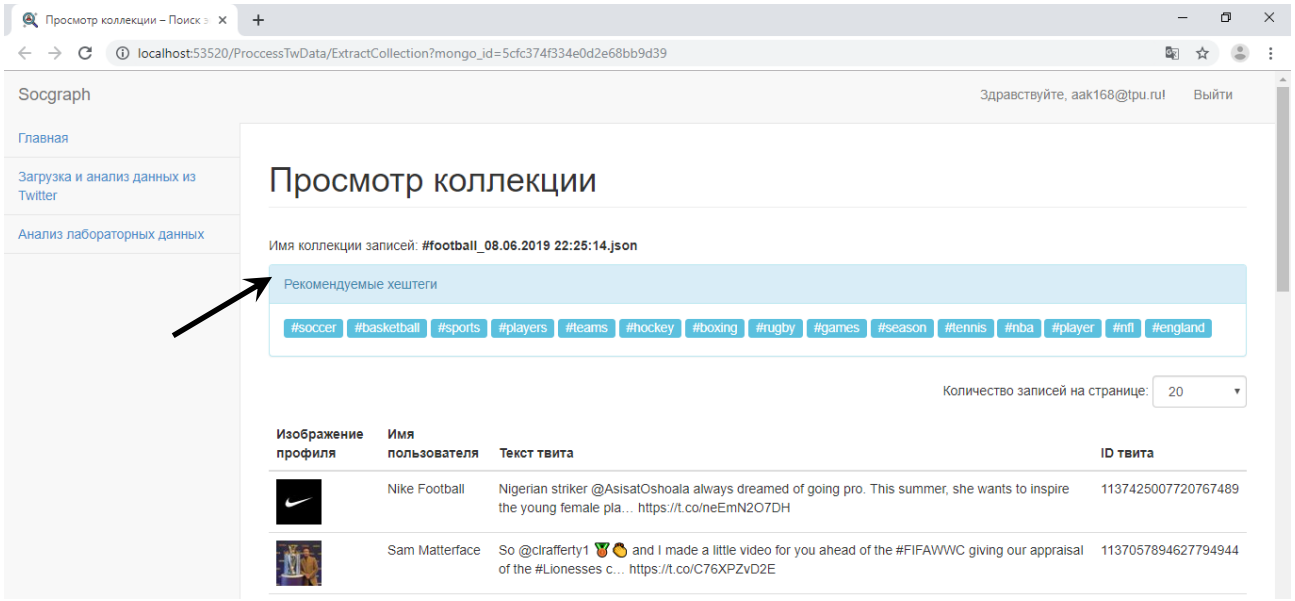

Рисунок 17 – Отображение панели с хештегами, рекомендованными для хештега “football”

Рисунки 15-17 демонстрируют результат работы модуля рекомендаций для трёх различных предметных областей: handmade, travel, football.

Качество решения задачи отображения на карте геометок, полученных из записей пользователь Twitter, напрямую зависит от отношения числа твитов с отмеченным местоположением к общему числу твитов. На рисунках 18-20 представлены результаты отображения геометок на карте. Процентное

соотношение твитов, отмеченных на карте ко всем твитам – 3%, 5% и 4% для выборок объёмом 100, 500, 1000 твитов соответственно.

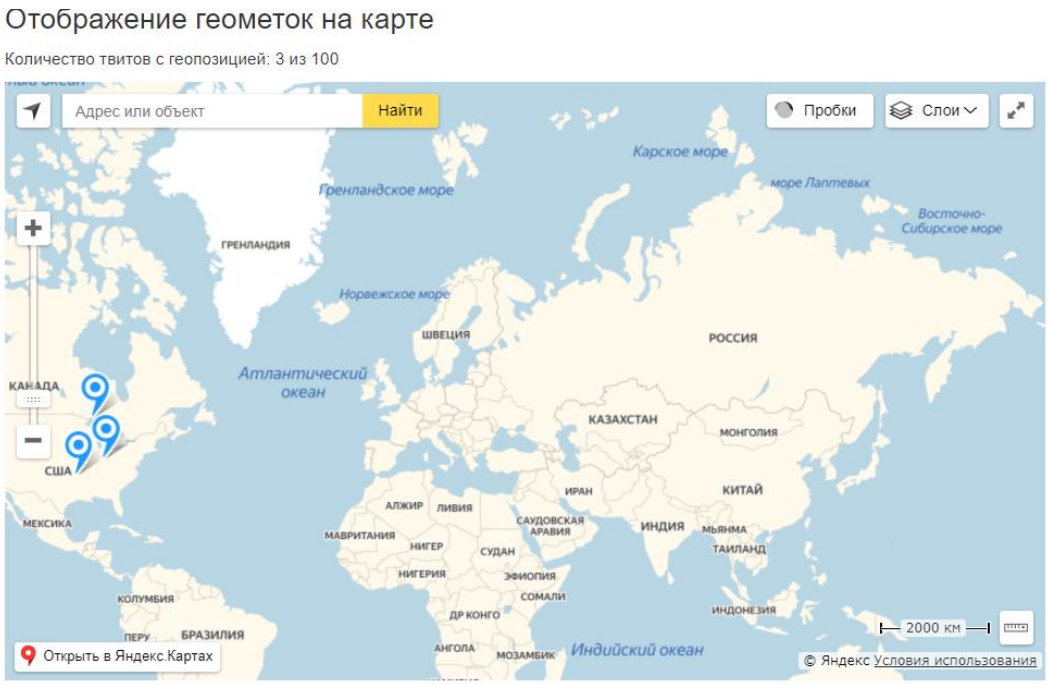

Рисунок 18 – Отображение геометок, полученных из записей в Twitter, на карте (3% геометок в выборке)

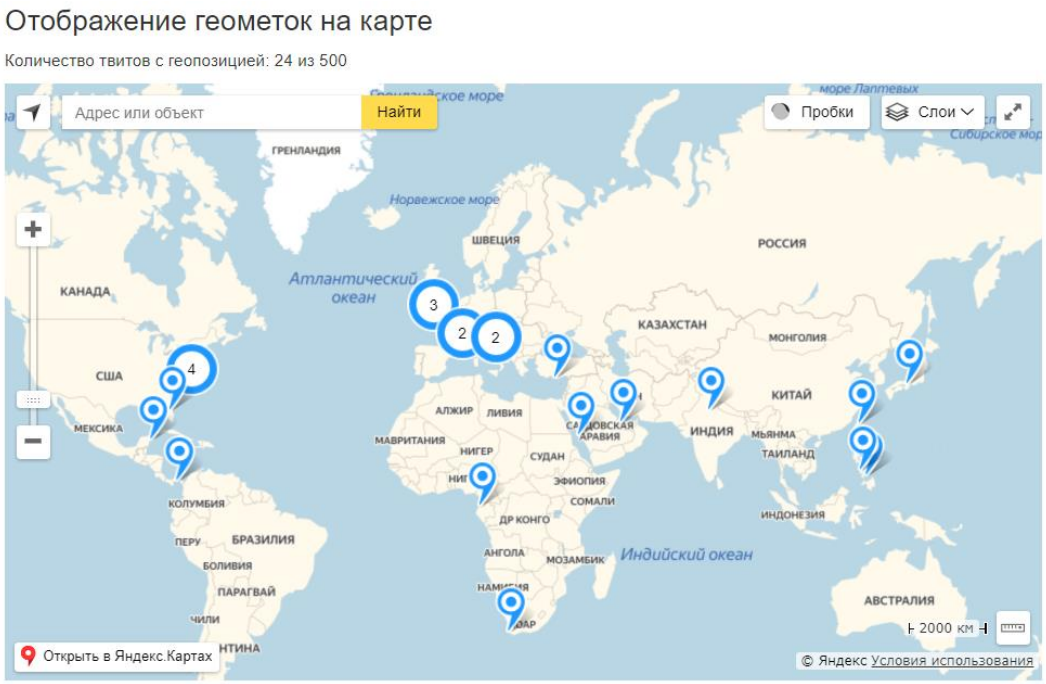

Рисунок 19 – Отображение геометок, полученных из записей в Twitter, на карте (5% геометок в выборке)

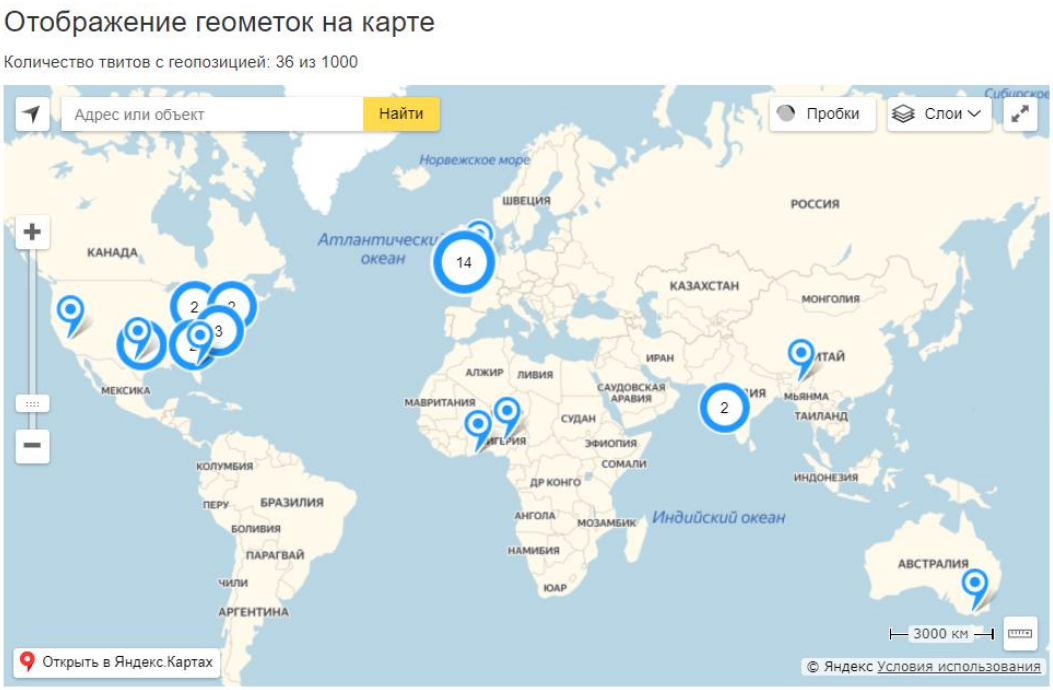

Рисунок 20 – Отображение геометок, полученных из записей в Twitter, на карте (4% геометок в выборке)

При выполнении задач в контексте настоящей работы были исправлены недостатки визуализации исходных данных и результатов анализа. Веб-страница для просмотра коллекций изъятых из Twitter записей была дополнена постраничной навигацией (рисунок 21), которая отсутствовала ранее, из-за чего просмотр полной коллекции данных был невозможен по причине невозможности вывода в представление всего содержимого коллекции.

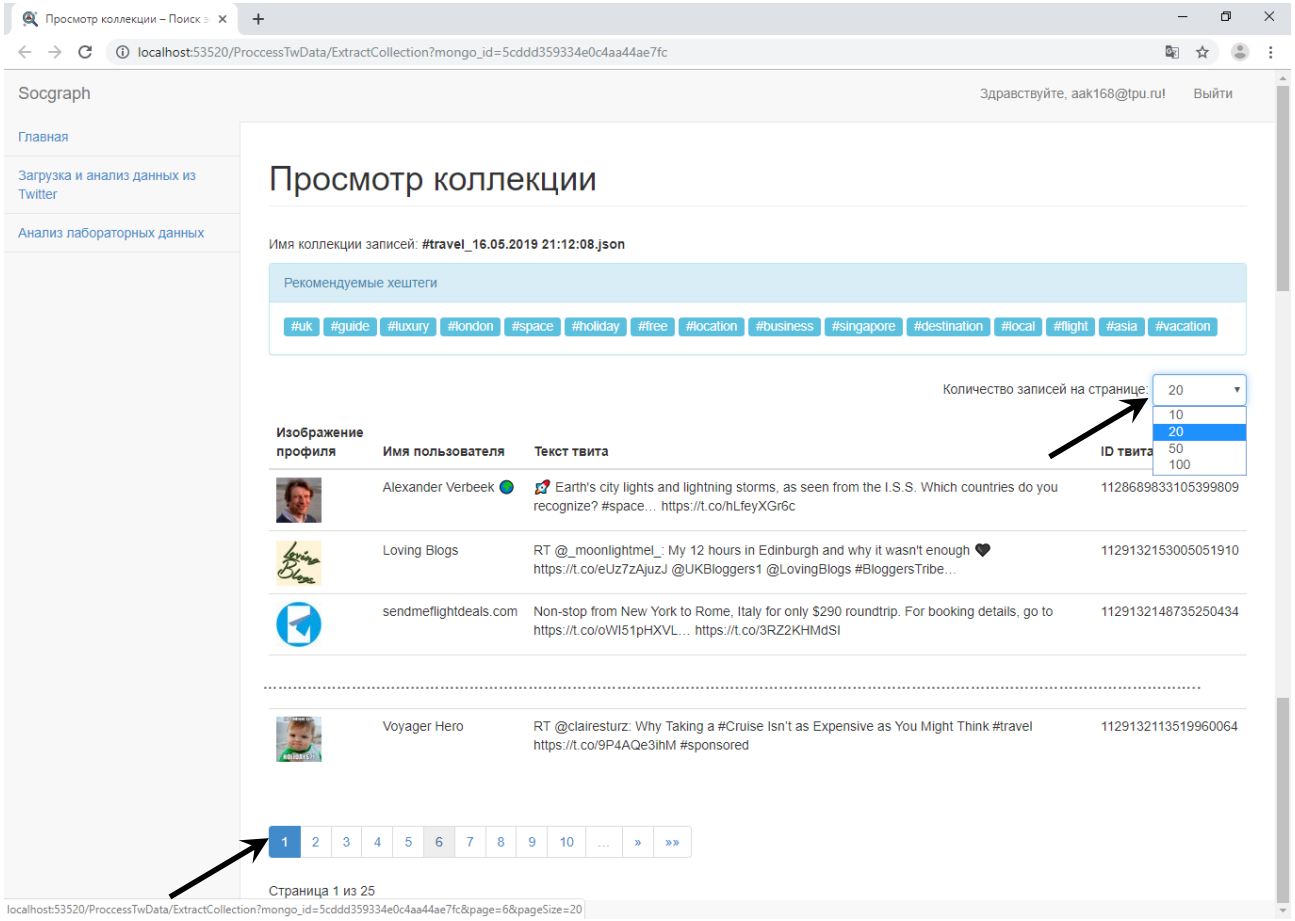

Рисунок 21 – Постраничная навигация на странице просмотра коллекции твитов

На рисунке Б.1 (приложение Б) приведён скриншот веб-страницы, изображённой на рисунке 21 до внесения изменений в её вёрстку.

Страница, представляющая результаты анализа коллекций, также получила изменения. Вместо библиотеки для визуализации данных D3 внедрена специализированная библиотека Cytoscape.js (рисунок 22).

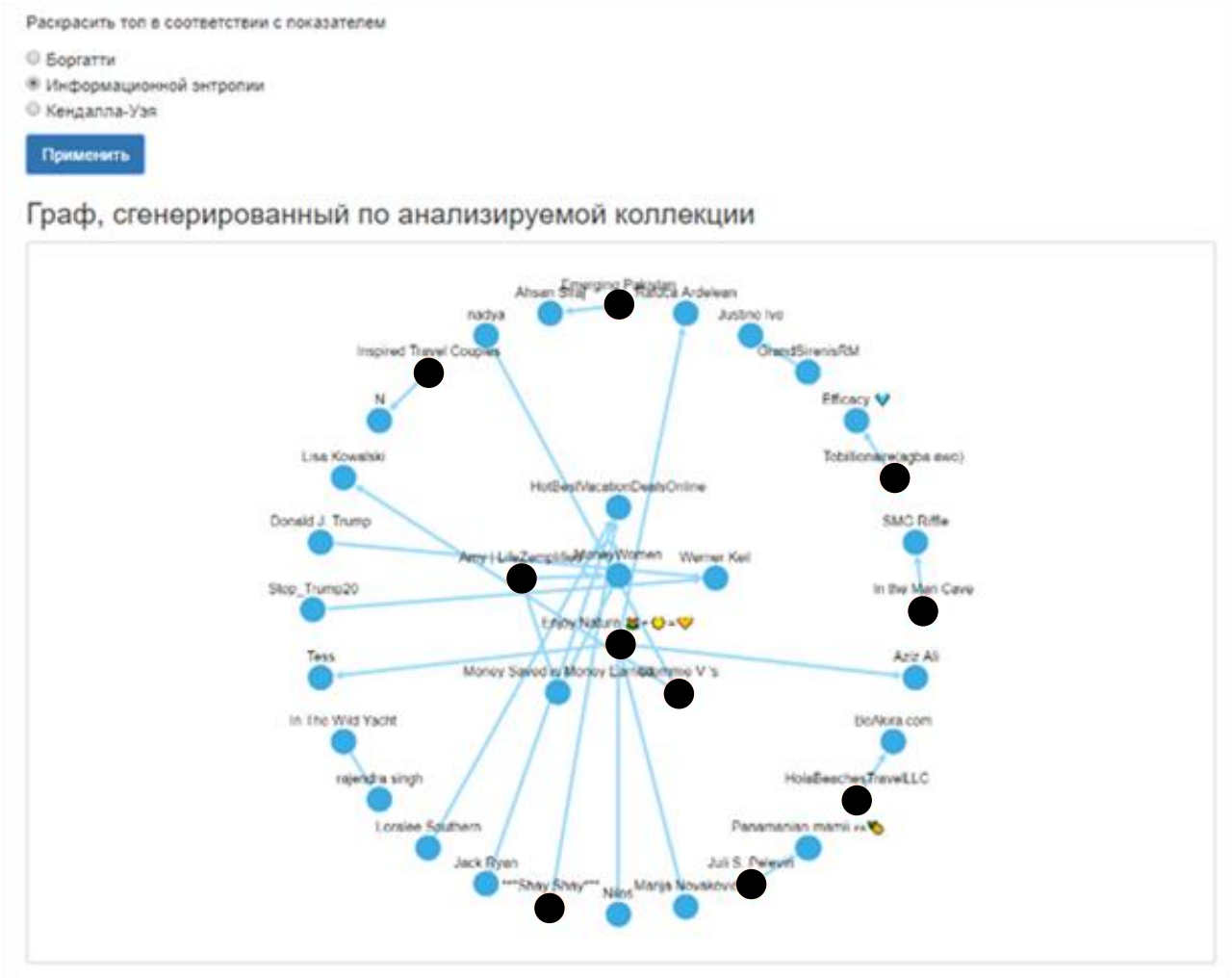

Рисунок 22 – Визуализация социального графа средствами Cytoscape.js

На графе введённое пользователем количество пользователей-экспертов выделено другим цветом (чёрные круги на рисунке 22). Нормализованы значения для построения диаграмм, отображающих результаты идентификации пользователей-экспертов, изменена шкала для представления кластеров пользователей-экспертов (рисунок 23).

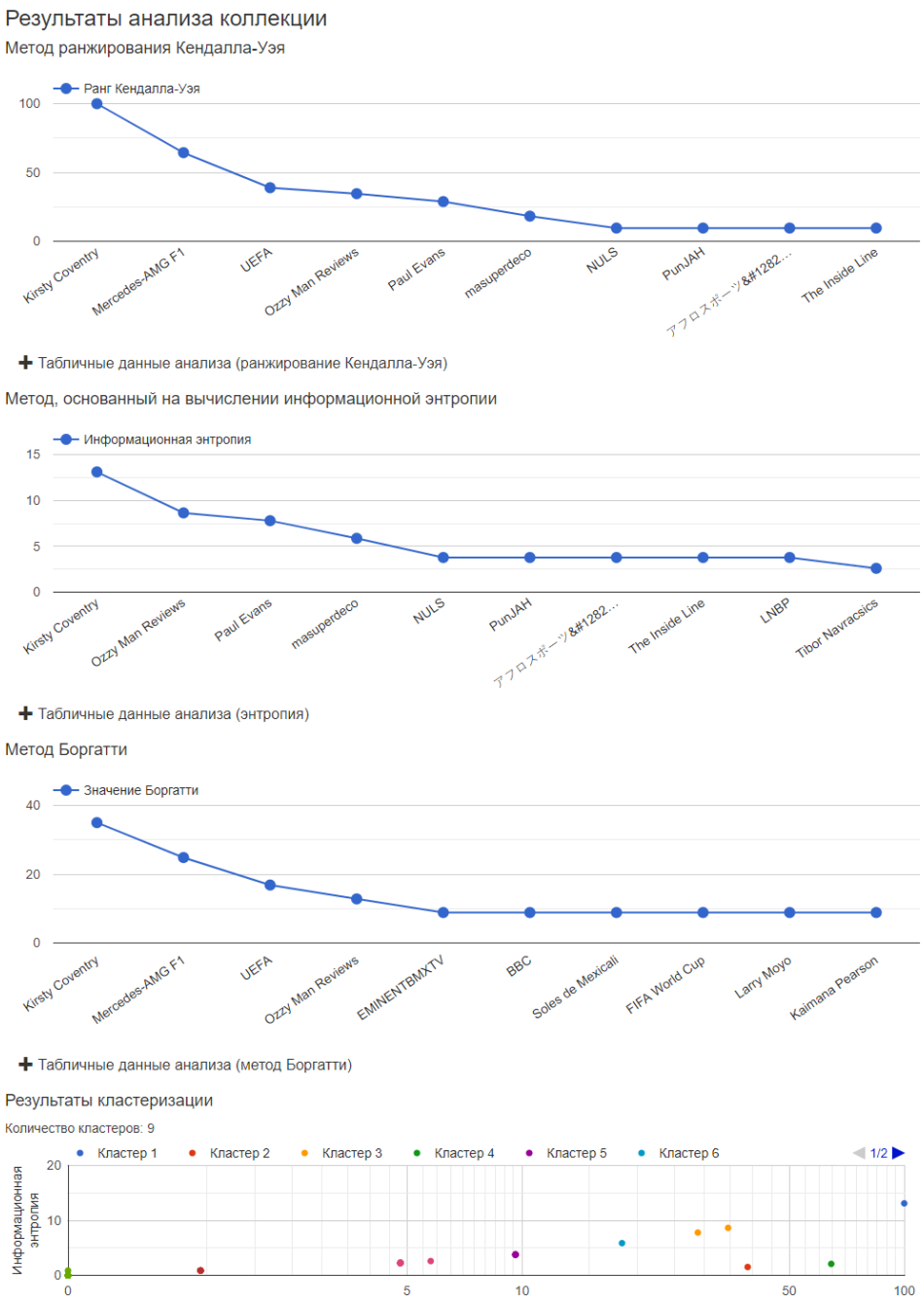

Рисунок 23 – Визуализация результатов идентификации пользователей после внесения изменений

Изменённый внешний вид диаграмм упрощает визуальную оценку результатов анализа. Разреженная около начала системы координат логарифмическая шкала на диаграмме «Результаты кластеризации» решает проблему скопления точек графика около нуля.

На рисунках Б.2 и Б.3 (приложение Б) приведены скриншоты веб-страниц до внесения в них изменений.

3 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

3.1 Оценка коммерческого потенциала и перспективности проведения научных исследований с позиции ресурсоэффективности и ресурсосбережения

3.1.1 Потенциальные потребители результатов исследования

Социальные сети – важная часть современной коммуникации. С ростом числа пользователей растет и влияние социальных сетей на реальную жизнь: всё больше интереса вызывают Twitter, Facebook, Instagram у специалистов разного профиля. Маркетологи, социологи, психологи, аналитики, специалисты по подбору персонала считают информацию о людях в соцсетях важной, подходящей для интерпретации. Поэтому они используют ее в своих целях: прогнозируют какие-либо события или отношение к ним, изучают общественное мнение или спрос на товары, проверяют, какова будет реакция пользователей на ту или иную тему.

Разработанный программный сервис позволяет получить информацию о пользователях-экспертах в заданной предметной области. Он может быть использован как физическими лицами, так и различными организациями.

Однако наибольшей заинтересованностью обладают коммерческие и некоммерческие организации, поэтому в качестве основных потребителей продукта будут рассматриваться именно они.

Концепция продукта отвечает одному из критериев сегментации рынка – сфере деятельности организаций. Поиск авторитетных пользователей производится в контексте одной предметной области, которая, вероятно, совпадает с областью деятельности организации, использующей данный сервис.

3.1.2 Анализ конкурентных технических решений

Анализ конкурентных технических решений помогает понять, конкурентоспособен ли разрабатываемый продукт. При анализе должны быть рассмотрены достоинства и недостатки решений-конкурентов.

Основными конкурентными решениями являются сервисы «Brandwatch Audience», «BuzzSumo», «Empire.Kred». Стоит отметить, что все сервисы предоставляют пользователям только платные тарифные планы.

Разработанный сервис функционально несколько отличен от «BuzzSumo» и «Empire.Kred», но схож с «Brandwatch Audience»: оба сервиса предоставляют пользователю данные о влиятельных пользователях-экспертах, а не выделяют целевую аудиторию по запрошенному ключевому слову. Существенным недостатком «Brandwatch Audience» является высокая стоимость ежемесячной подписки на использование.

Основными факторами конкурентоспособности были выбраны следующие критерии:

- повышение производительности труда пользователя;
- удобство эксплуатации;
- предоставляемые возможности;
- стоимость использования;
- проникновение на рынок.

Детальный анализ конкурентоспособности был выполнен в процессе составления оценочной карты (таблица 2).

Таблица 2 – Оценочная карта для сравнения конкурентных технических разработок

№ п/п	Конкуренты	Факторы конкурентоспособности					Итоговая оценка
		Повышение производительности труда пользователя	Удобство эксплуатации	Предоставляемые возможности	Стоимость использования	Проникновение на рынок	
1	«Brandwatch Audience»	9 1,638	10 2,27	7 1,589	1 0,227	8 1,088	6,812
2	«BuzzSumo»	7 1,274	9 2,043	6 1,362	4 0,908	5 0,68	6,267
3	«Empire.Kred»	8 1,456	8 1,816	5 1,135	5 1,135	3 0,408	5,95
4	Разрабатываемое техническое решение	8 1,456	9 2,043	10 2,27	10 2,27	1 0,136	8,175
	b_j	4	5	5	5	3	22
	w_j	0,182	0,227	0,227	0,227	0,136	-

Исходя из результатов, полученных при составлении оценочной карты конкурентных технических решений, можно заметить, что разрабатываемое приложение имеет более высокую конкурентоспособность по выбранным критериям: его оценка составила 8,175. Это обусловлено тем, что эксплуатация рассмотренных сервисов имеет высокую стоимость, а также они не обладают тем уникальным функционалом, который предлагает пользователю разрабатываемое решение.

По оценкам факторов конкурентоспособности был построен многоугольник конкурентоспособности, представленный на рисунке 24.

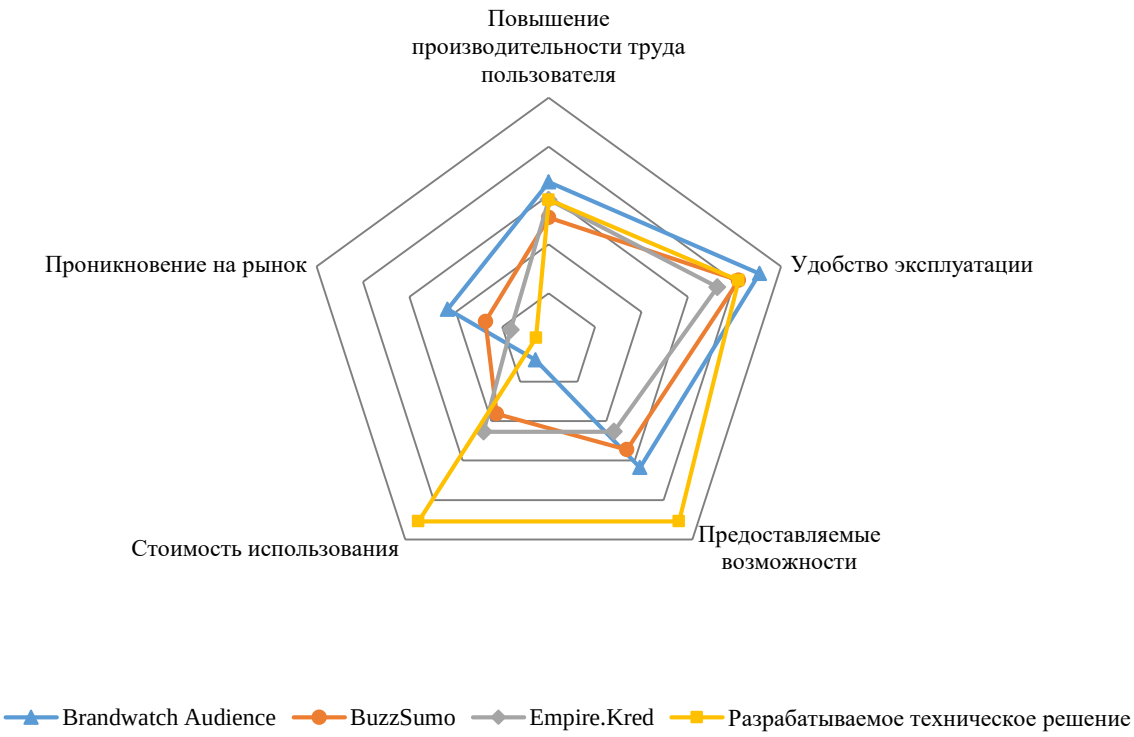

Рисунок 24 – Многоугольник конкурентоспособности

Многоугольник конкурентоспособности показывает, что разработанный программный продукт имеет преимущества в виде стоимости использования и предоставляемых возможностей.

Таким образом, сервис для выявления пользователей-экспертов в социальной сети обладает достаточно высокой конкурентоспособностью среди аналогов.

3.1.3 SWOT-анализ

На первом этапе SWOT-анализа определяются сильные и слабые стороны разработанного решения, а также возможности для развития и угрозы, которые могут оказать внешнее воздействие на продукт.

Второй этап анализа заключается в выявлении соответствий сильных и слабых сторон разработанного решения внешним условиям окружающей среды. Эти соответствия способны помочь в принятии решения о необходимости стратегических изменений.

Результат SWOT-анализа представлен в таблице 3.

Таблица 3 – SWOT-анализ проекта

		Внутренние факторы	
		Сильные стороны проекта: 1. Актуальность разработки. 2. Предоставление сервиса в пользование бесплатно. 3. Простой пользовательский интерфейс. 4. Конкурентоспособность продукта. 5. Широкая целевая аудитория.	**Слабые стороны проекта:** 1. Низкий уровень проникновения на рынок. 2. Потребность в финансировании и рекламе для привлечения клиентов. 3. Ограниченный объём выгружаемой для анализа информации.
Внешние факторы	**Возможности:** 1. Рост клиентской базы разработанного сервиса. 2. Увеличение тарифов у конкурентов. 3. Получение инвестиций. 4. Наращивание функционала.	**Направления развития:** 1. Расширение функционала с учётом пользовательских интересов на основе полученных инвестиций. 2. Использование современных тенденций в области пользовательских интерфейсов для обеспечения комфортного использования сервиса.	**Сдерживающие факторы:** 1. Низкий уровень проникновения на рынок осложняет привлечение инвестиций. 2. Низкий уровень проникновения на рынок замедляет рост популярности приложения.

Продолжение таблицы 3

Внешние факторы	**Угрозы:** 1. Появление новых конкурентоспособных продуктов. 2. Невостребованность сервиса у клиентов. 3. Появление новых способов определения авторитетности пользователей. 4. Отсутствие кроссплатформенной поддержки.	**Угрозы развития:** 1. Незаинтересованность пользователей в сервисе снижает конкурентоспособность продукта. 2. Появление новых методов поиска пользователей-экспертов может привести к созданию приложений, работающих быстрее и эффективнее данной разработки.	**Уязвимости:** 1. Отсутствие интереса пользователей к продукту и его слабое проникновение на рынок могут привести к завершению его использования и обслуживания. 2. Отсутствие поддержки различных операционных систем может сократить клиентскую базу.

SWOT-анализ показал, что одной из основных угроз проекта является появление новых, более конкурентоспособных продуктов, а также непопулярность сервиса среди потенциальных пользователей. Чтобы этого избежать, необходимо вовремя предусмотреть возможные неудачи и принять меры по их профилактике.

3.2 Планирование научно-исследовательских работ

3.2.1 Структура работ в рамках научного исследования

Данный раздел содержит перечень этапов и работ, осуществлённых в рамках проведения научного исследования. Распределение исполнителей по видам работ представлено в таблице 4.

Таблица 4 – Перечень работ и распределение исполнителей

№ работы	Наименование работы	Исполнители работы
1	Выбор научного руководителя бакалаврской работы	Кондратьева А. А.
2	Составление и утверждение темы бакалаврской работы	Лунёва Е. Е., Кондратьева А. А.
3	Составление календарного плана-графика выполнения бакалаврской работы	Лунёва Е. Е.
4	Подбор и изучение литературы по теме бакалаврской работы	Кондратьева А. А.
5	Изучение возможностей предоставленного для разработки сервиса	Кондратьева А. А.
6	Предложение изменений программного сервиса	Лунёва Е. Е., Кондратьева А. А.
7	Рефакторинг элементов интерфейса сервиса	Кондратьева А. А.
8	Разработка модуля рекомендаций	Кондратьева А. А.
9	Усовершенствование визуализации результатов анализа выбранных из соцсети данных	Кондратьева А. А.
10	Согласование выполненной работы с научным руководителем	Лунёва Е. Е., Кондратьева А. А.
11	Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Кондратьева А. А.
12	Подведение итогов, оформление работы	Кондратьева А. А.

3.2.2 Определение трудоёмкости выполнения работ

Обычно при разработке приложений основную часть её стоимости составляют трудовые затраты. В связи с этим важно определить трудоёмкость работы каждого из участников научного исследования. При определении трудоёмкости были использованы показатели трудоёмкости и длительности работ, измеряемые в человеко-днях и днях соответственно.

Для построения календарного плана-графика необходимо произвести расчёты временных показателей проведения научного исследования (таблица 5).

Таблица 5 – Временные показатели проведения научного исследования

Наименование работы	Исполнители работы	Трудоёмкость работ, человеко-дни			Длительность работ, дни	
		tmin	tmax	tож	Тр	Тк
Выбор научного руководителя бакалаврской работы	Кондратьева А. А.	1	3	1,8	2	2
Составление и утверждение темы бакалаврской работы	Лунёва Е. Е.	1	3	1,8	1	1
	Кондратьева А. А.	0,25	0,5	0,35	1	1
Составление календарного плана-графика выполнения бакалаврской работы	Лунёва Е. Е.	3	7	4,6	5	6
Подбор и изучение литературы по теме бакалаврской работы	Кондратьева А. А.	5	7	5,8	6	7
Изучение возможностей предоставленного для разработки сервиса	Кондратьева А. А.	5	7	5,8	6	7
Предложение изменений программного сервиса	Лунёва Е. Е.	3	5	3,8	2	2
	Кондратьева А. А.	3	5	3,8	2	2
Рефакторинг элементов интерфейса сервиса	Кондратьева А. А.	10	14	11,6	12	14
Разработка модуля рекомендаций	Кондратьева А. А.	12	15	13,2	13	16
Усовершенствование визуализации результатов анализа выбранных из соцсети данных	Кондратьева А. А.	10	15	12	12	15

Продолжение таблицы 5

Наименование работы	Исполнители работы	Трудоемкость работ, чел-дни			Длительность работ, дни	
		tmin	tmax	tож	Тр	Тк
Согласование выполненной работы с научным руководителем	Лунёва Е. Е.	2	5	3,2	2	2
	Кондратьева А. А.	0,5	5	2,3	1	1
Выполнение других частей работы (финансовый менеджмент, социальная ответственность)	Кондратьева А. А.	7	14	9,8	10	12
Подведение итогов, оформление работы	Кондратьева А. А.	7	14	9,8	10	12
ИТОГО	Лунёва Е. Е.	6	14	9,2	10	12
	Кондратьева А. А.	60,75	99,5	76,25	75	92

Расчёты показали, что общее количество календарных дней, проведённых в работе над проектом составило 12 и 92 дня для научного руководителя и студента соответственно.

3.2.3 Разработка графика проведения научного исследования

График проведения научных работ представляется в форме диаграммы Гантта – ленточного горизонтального графика, на котором работы по теме представляются протяжёнными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

Для удобства построения графика, длительность каждого из этапов работ из рабочих дней следует перевести в календарные дни.

Согласно производственному календарю (для 6-дневной рабочей недели), в 2019 году 365 календарных дней, 299 рабочих дней, 66 выходных/праздничных дней.

Коэффициент календарности определяется по следующей формуле:

$$Ткал = \frac{Ткал}{Ткал - Твых - Тпр} \quad (1)$$

Таким образом, согласно формуле 1 был определен коэффициент календарности для 2019 года:

$$Ткал = \frac{Ткал}{Ткал - Твых - Тпр} = \frac{365}{365 - 66} = 1{,}22.$$

Построенная диаграмма Гантта приведена на рисунке 25.

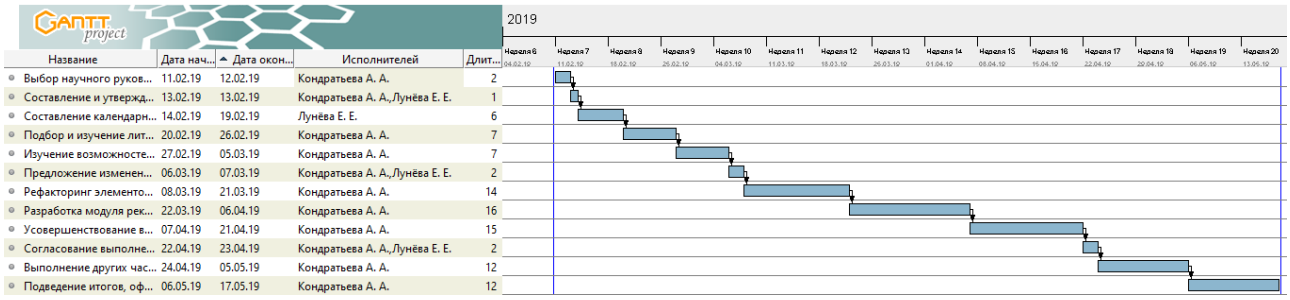

Рисунок 25 – Диаграмма Гантта

Дата начала работы над проектом – 11 февраля 2019 года, дата окончания работы – 17 мая 2019 года. Общая длительность работы над проектом составила 96 дней, учитывая параллельное выполнение задач исполнителями.

3.2.4 Бюджет научно-технического исследования

3.2.4.1 Расчет материальных затрат научно-технического исследования

Данная статья включает стоимость всех материалов, используемых при разработке проекта: сырья, материалов, комплектующих. В эту статью включаются и транспортные расходы, равные 15% от общей стоимости материальных затрат.

В стоимость материальных затрат могут включаться затраты на канцелярские принадлежности.

Сумма материальных затрат, необходимых для проведения данного исследования, состоит только из затрат на канцелярские принадлежности: она равна 1500 рублей.

3.2.4.2 Расчёт затрат на специальное оборудование для научных (экспериментальных) целей

Для написания ВКР не было совершено покупок дорогостоящего оборудования. Поэтому в качестве затрат на оборудование была рассчитана амортизация.

При написания ВКР был использован ноутбук первоначальной стоимостью 42 000 рублей. Срок полезного использования компьютера составляет от 2 до 3 лет. При расчётах выбран максимальный срок – 3 года. Длительность написания ВКР – 5 месяцев. Амортизация рассчитывается линейным способом по приведённым ниже формулам 2-5.

Норма амортизации рассчитывается по формуле:

$$\text{Ан} = \frac{1}{n} * 100\%, \quad (2)$$

где n – срок полезного использования.

Годовые амортизационные отчисления рассчитываются по формуле:

$$\text{Аг} = \text{С} * \text{Ан}, \quad (3)$$

где С – стоимость оборудования.

Ежемесячные амортизационные отчисления рассчитываются по формуле:

$$\text{Ам} = \frac{\text{Аг}}{12} \quad (4)$$

Итоговая сумма амортизации основных средств рассчитывается по формуле:

$$\text{А} = \text{Ам} * m, \quad (5)$$

где m – планируемая длительность написания ВКР.

По формулам 2-5 рассчитывается амортизация ноутбука:

- норма амортизации:

$$\text{Ан} = \frac{1}{n} * 100\% = \frac{1}{3} * 100\% = 33{,}33\%$$

- годовые амортизационные отчисления:

$$Аг = 42\,000 * 0{,}33 = 13\,860 \text{ рублей.}$$

- ежемесячные амортизационные отчисления:

$$Ам = \frac{13\,860}{12} = 1\,155 \text{ рублей.}$$

- итоговая сумма амортизации основных средств:

$$А = 1\,155 * 5 = 5\,775 \text{ рублей.}$$

Таким образом, итоговые затраты на амортизацию составили 5 775 рублей.

3.2.4.3 Основная заработная плата исполнителей темы

В данном разделе приведены расчёты затрат на заработную плату участникам разработки.

Для расчёта основной заработной платы студента был взят оклад, равный окладу ассистента без степени – 21 760 рублей. Оклад научного руководителя составляет 33 664 рубля.

Баланс рабочего времени, данные которого будут использованы при расчётах далее, приведён в таблице 6.

Таблица 6 – Баланс рабочего времени

Показатели рабочего времени	Дни
Календарные дни	365
Нерабочие дни (праздники/выходные)	66
Потери рабочего времени (отпуск/невыходы по болезни)	56
Действительный годовой фонд рабочего времени	243

Основная заработная плата рассчитывается по формуле:

$$Зосн = Здн * Тр * (1 + Кпр + Кд) * Кр, \quad (6)$$

где Здн – среднедневная заработная плата, рублей;

Кпр – премиальный коэффициент (0,3);

Кд – коэффициент доплат и надбавок (0,2);

Кр – районный коэффициент (1,3);

Тр – продолжительность работ, выполняемых работником, рабочих дней.

Средняя заработная плата рассчитывается по следующей формуле:

$$Здн = \frac{Зм * М}{F_д}, \quad (7)$$

где Зм – месячный оклад работника, рублей.

М – количество месяцев работы без отпуска в течение года (для 6-дневной рабочей недели $М = 10{,}4$ месяца);

$F_д$ – действительный годовой фонд рабочего времени персонала, рабочих дней.

Расчёты основной заработной платы студента и научного руководителя проекта, представленные ниже, были произведены по формулам 6-7.

Среднедневная заработная плата для студента составила:

$$Здн = \frac{Зм * М}{F_д} = \frac{21760 * 10{,}4}{243} = 931{,}29 \text{ рублей.}$$

Основная заработная плата для студента составила:

$$Зосн = Здн * Тр * (1 * Кпр * Кд) * Кр = 931{,}29 * 75 * (1 + 0{,}3 + 0{,}2) * 1{,}3$$

$$\approx 136\,201 \text{ рубль}$$

Среднедневная заработная плата для доцента составила:

$$Здн = \frac{Зм * М}{F_д} = \frac{33664 * 10{,}4}{243} = 1440{,}76 \text{ рублей.}$$

Основная заработная плата для доцента составила:

$$Зосн = Здн * Тр * (1 + Кпр + Кд) * Кр = 1440{,}76 * 10 * (1 + 0{,}3 + 0{,}2) * 1{,}3$$

$$\approx 28\,095 \text{ рублей.}$$

В таблице 7 представлен расчёт основной заработной платы студента и научного руководителя: общая сумма составила 164 296 рублей.

Таблица 7 – Расчёт основной заработной платы

Исполнители	Здн, руб.	Кпр	Кд	Кр	Тр	Зосн, руб.
Студент	931,29	0,3	0,2	1,3	75	136 201
Научный руководитель	1440,76	0,3	0,2	1,3	10	28 095
					Итого:	164 296

3.2.4.4 Дополнительная заработная плата исполнителей темы

В данном разделе рассчитаны дополнительные заработные платы студента и научного руководителя проекта. Эта статья учитывает доплаты за отклонение от нормальных условий труда, а также выплаты, связанные с обеспечением гарантий и компенсаций (при совмещении работы с обучением, при предоставлении ежегодного оплачиваемого отпуска и т. д.).

Дополнительная заработная плата составляет 12–15 % от основной заработной платы. В расчётах использовалось значение 12%.

Дополнительная заработная плата рассчитывается по формуле:

$$Здоп = Зосн * 0{,}12 \quad (8)$$

Дополнительная заработная плата студента, рассчитанная по формуле 8:

$$Здоп = Зосн * 0{,}12 = 136\,201 * 0{,}12 \approx 16\,344 \text{ рублей.}$$

Дополнительная заработная плата научного руководителя, рассчитанная по формуле 8:

$$Здоп = Зосн * 0{,}12 = \ 28\,095 * 0{,}12 \approx 3\,371 \text{ рублей.}$$

Суммарная дополнительная заработная плата составила 19 715 рублей.

3.2.4.5 Отчисления во внебюджетные фонды (страховые отчисления)

Страховые отчисления – это обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования, пенсионного фонда и медицинского страхования от затрат на оплату труда работников.

Страховые отчисления составляют 30% от суммы основной и дополнительной заработной планы и рассчитываются по формуле:

$$Рст = (Зосн + Здоп) * 0{,}3 \quad (9)$$

Страховые отчисления студента, рассчитанные по формуле 9:

$$Рст = (Зосн + Здоп) * 0{,}3 = (136\,201 + 16\,344) * 0{,}3 \approx 45\,764 \text{ рубля.}$$

Страховые отчисления научного руководителя, рассчитанные по формуле 9:

$$Рст = (Зосн + Здоп) * 0{,}3 = (28\,095 + 3\,371) * 0{,}3 \approx 9\,440 \text{ рублей.}$$

Общая сумма страховых отчислений составила 55 204 рубля.

3.2.4.6 Накладные расходы

Накладные расходы включают в себя различные прочие затраты организации, которые не были включены ни в одну из ранее описанных статей расходов: печать и ксерокопирование материалов исследования, оплата услуг связи, электроэнергии, почтовые расходы и т.д.

Накладные расходы вычисляются как 16% от суммы следующих статей затрат: материальных затрат, затрат на специальное оборудование, затрат на основную и дополнительную заработную плату и страховых взносов.

Формула для расчёта накладных расходов:

$$Рн = (Рм + Рсп + Зосн + Здоп + Рст) * 0{,}16, \quad (10)$$

где Рм – материальные затраты;

Рсп – затраты на специальное оборудование;

Зосн – затраты на основную заработную плату;

Здоп – затраты на дополнительную заработную плату;

Рст – страховые взносы.

Накладные расходы, рассчитанные по формуле 10:

$$Рн = (1\,500 + 5\,775 + 164\,296 + 19\,175 + 55\,204) * 0{,}16 \approx 39\,352 \text{ рубля.}$$

Накладные расходы проекта составили 39 352 рубля.

3.2.4.7 Формирование бюджета затрат научно-исследовательского проекта

Рассчитанные в предыдущих разделах статьи затрат формируют бюджет затрат научно-исследовательского проекта (таблица 8).

Таблица 8 – Расчёт бюджета затрат научно-технического проекта

Наименование	Сумма, руб.	Удельный вес, %
Материальные затраты	1500	0,53
Затраты на специальное оборудование	5 775	2,03
Затраты на основную заработную плату	164 296	57,67
Затраты на дополнительную заработную плату	19 175	6,73
Страховые взносы	55 204	19,38
Накладные расходы	38 956	13,67
Общий бюджет	284 906	100%

Данные таблицы 8 свидетельствуют о том, что бюджет проекта составил 284 906 рублей, при этом больше половины бюджета – это затраты на основную и дополнительную заработную плату.

3.3 Определение потенциального эффекта исследования

Потенциальными потребителями программного сервиса для выявления пользователей-экспертов в социальной сети можно считать организации, которые нуждаются в поиске людей, являющихся авторитетными в заданной предметной области.

Конкурентный анализ показал, что разрабатываемое решение является конкурентоспособным с точки зрения функционала и стоимости использования.

Длительность исследования для студента составила 75 рабочих дней (или 92 календарных дня). Длительность исследования для научного руководителя составила 10 рабочих дней (или 12 календарных дней).

Потенциальная стоимость исследования оценивается в 284 906 рублей.

Положительный эффект исследования заключается в том, что оно упрощает задачу поиска экспертов среди пользователей соцсетей, автоматизируя этот процесс.

4 Социальная ответственность

Целью выпускной квалификационной работы является рефакторинг клиентской и серверной частей существующего программного сервиса, направленного на идентификацию пользователей-экспертов в заданной предметной области. Сервис представляет собой веб-приложение, следовательно, разработка и эксплуатация этого сервиса предполагают использование помещения, оборудованного персональным компьютером.

Данный раздел включает описание экологической и производственной безопасности при разработке и использовании сервиса, также в нём рассмотрены вероятные ситуации опасного и чрезвычайного характеров, предложены способы их профилактики.

Понятие «социальная ответственность» включает в себя вопросы соблюдения прав персонала на труд, выполнения требований к обеспечению безопасности труда, охране окружающей среды и снижению вредных воздействий на неё.

Целью данного раздела является принятие проектных решений, исключающих несчастные случаи в производстве, а также снижение вредных воздействий на окружающую среду.

4.1 Правовые и организационные вопросы обеспечения безопасности

4.1.1 Специальные правовые нормы трудового законодательства

Нормы трудового законодательства регулируют отношения, связанные с использованием личного труда. Организация рабочего процесса обязательно должна осуществляться в соответствии с Трудовым кодексом Российской Федерации [25].

Согласно Трудовому кодексу Российской Федерации, а также инструкции по охране труда [26], организация труда должна происходить согласно следующим пунктам:

- рабочая смена должна составлять не более 8 часов;
- продолжительность непрерывной работы с компьютером без регламентированного перерыва не должна превышать 2-х часов (устанавливаются перерывы продолжительностью 15 минут);
- обеденный перерыв должен составлять не менее 30 минут и не более 1 часа.

Перед приёмом сотрудника на работу работодатель обязан провести инструктаж по технике безопасности, а также обеспечить периодическое проведение инструктажа в дальнейшем. Одной из обязанностей работодателя также является обеспечение сотрудников соответствующим нормам рабочим местом:

- рабочее место сотрудника должно быть организовано с учетом эргономических требований согласно ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования» [27];
- конструкция рабочей мебели должна предусматривать возможность индивидуальной регулировки (т.е. адаптироваться под рост сотрудника).

Работодатель обязан предпринимать меры, необходимые для профилактики производственного травматизма, профессиональных и других видов заболеваний работников.

4.1.2 Организационные мероприятия по компоновке рабочей зоны

Рабочее место сотрудника, использующего персональный компьютер, должно быть организовано в соответствии с требованиями, указанными в ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования» [27] и СанПиН 2.2.2/2.4.1340-03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы» [28]. Данные документы описывают нормы и требования к организации рабочей зоны и помещения, в котором располагается эта зона.

Общие требования к организации рабочего места сотрудника, работа которого связана с использованием персонального компьютера:

- экран монитора должен находиться от глаз пользователя на расстоянии 600 – 700 мм, но не ближе 500 мм с учетом размеров алфавитно-цифровых знаков и символов;
- деятельность с персональным компьютером должна чередоваться с перерывами, в которые пользователь не будет использовать персональный компьютер во избежание утомления пользователя;
- конструкция рабочего стула (кресла) должна обеспечивать поддержание рациональной рабочей позы при работе с компьютером. Рабочий стул (кресло) должен быть подъемно-поворотным, регулируемым по высоте и углам наклона сиденья и спинки.
- поверхность сиденья, спинки и других элементов стула (кресла) должна быть полумягкой, с нескользящим, слабо электризующимся и воздухопроницаемым покрытием, обеспечивающим легкую очистку от загрязнений.

4.2 Производственная безопасность

Разработка сервиса происходила в помещении с рабочим местом, оборудованным персональным компьютером с ЖК-дисплеем.

В данном разделе рассмотрены вредные и опасные факторы, приведены рекомендации по их устранению.

4.2.1 Вредные факторы

4.2.1.1 Электромагнитное излучение

Длительное воздействие электромагнитных излучений, источником которых является ПК, может стать причиной снижения иммунитета, возникновения мигреней, ухудшения памяти, а также стать катализатором развития серьезных заболеваний.

Большое количество компьютерной техники в помещении, к примеру, в компьютерном классе, приводит к возникновению повышенного уровня электромагнитного излучения.

Допустимый уровень электромагнитного излучения устанавливается в СанПиН 2.2.2/2.4.1340-03 [28].

Защита от электромагнитного излучения может быть осуществлена установлением предельно допустимого уровня напряженности, который, согласно СанПиН 2.2.2/2.4.1340-03, составляет не более 8 кА/м, при этом уровень магнитной индукции составляет 10 мТл [28].

4.2.1.2 Повышенный уровень шума

Источниками шума на рабочем месте могут являться системный блок персонального компьютера, устройства вентиляции, кондиционирования помещения, оргтехника и люди, находящиеся в помещении.

Ненормированные показатели шума на рабочем месте оказывают психологическое влияние на состояние сотрудника, вследствие чего падает уровень концентрации и сосредоточенности, понижается эффективность выполнения задач. Уровень шума также влияет на уровень стресса сотрудника и его утомляемость. Длительное воздействие шума может привести к снижению слуха.

Во избежание перечисленных выше последствий воздействия шума, требуется соблюдать установленные в СанПиН 2.2.2/2.4.1340–03 «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы» требования. Согласно нормам, представленным в этом документе, уровень шума не должен превышать 50 дБА [28].

Для снижения уровня шума могут быть использованы следующие средства:

- устройства вентиляции и кондиционирования со сниженным уровнем шума;
- звукопоглощающий корпус ПК;
- вентиляторы охлаждения корпуса ПК со сниженным уровнем шума;
- охлаждающие подставки для ноутбуков.

4.2.1.3 Недостаточная освещённость рабочей зоны

Освещение рабочего места имеет большое значение и оказывает влияние на работу сотрудника, а также на его физическое состояние. Значение имеет как естественное, так и искусственное освещение. Недостаток освещения ведет к ухудшению зрения работника.

Согласно СанПиН 2.2.2/2.4.1340-03 [28], освещенность на поверхности рабочего стола пользователя персонального компьютера должна быть 300 – 500 лк. Освещение не должно создавать бликов на поверхности экрана. Освещенность поверхности экрана не должна быть более 300 лк.

Помимо этого, существуют некоторые общие рекомендации и требования к организации освещения на рабочем месте, например:

- рабочие столы следует размещать таким образом, чтобы мониторы были ориентированы боковой стороной к световым проемам, а естественный свет падал преимущественно слева;
- искусственное освещение в помещениях для эксплуатации компьютера должно осуществляться системой общего равномерного освещения.

Соблюдение вышеуказанных мер позволит избежать негативного влияния на зрение работника.

4.2.1.4 Статические физические нагрузки

Статическое мышечное напряжение, возникающее при работе с персональным компьютером возникает из-за сидячего положения пользователя. Концентрация внимания на действиях, происходящих на мониторе компьютера, приводят к перенапряжению мышц шейного отдела позвоночника, вследствие чего увеличивается риск возникновения сколиоза, остеохондроза и ухудшения мозгового кровообращения.

Возникновению подобных проблем способствует оснащение рабочего места нерегулируемой мебелью (отсутствует возможность регулировать высоту

стула и рабочей области), а также отсутствием подставки для ног, подлокотников и т.п.

Нормы организации рабочего пространства пользователя ПК регулируются в соответствии с СанПиН 2.2.2/2.4.1340-03 [28].

Для того, чтобы минимизировать влияние статических нагрузок при работе с компьютером, требуется оснастить рабочее место правильно подобранной мебелью, соблюдать режимы работы и отдыха, а также выполнять ряд специальных физических упражнений.

4.2.1.5 Перенапряжение зрительных анализаторов

Наблюдение за действиями, происходящими на экране компьютера, приводит к повышению нагрузки на зрительные анализаторы пользователя. Ввод текста также является причиной перенапряжения зрительных анализаторов, так как пользователю приходится переносить свой взгляд с клавиатуры на монитор и обратно.

Длительная работа с ПК приводит к ежедневному утомлению зрительных анализаторов и к ухудшению зрения в длительной перспективе.

Минимизация влияния указанного фактора может быть произведена через чередование режимов работы и отдыха. При этом перерывы в работе должны производиться без использования ПК. Согласно СанПиН 2.2.2/2.4.1340-03 [28], суммарное время перерывов должно составлять не менее 90 минут при 8-ми часовой рабочей смене.

Для профилактики перенапряжения может проводиться зрительная гимнастика, способствующая расслаблению и восстановлению зрительных анализаторов.

4.2.2 Опасные факторы

4.2.2.1 Опасность поражения электрическим током

Пользователь персонального компьютера взаимодействует непосредственно с электрооборудованием, поэтому всегда есть вероятность

поражения электрическим током. В результате поражения током человек может получить механические повреждения, возникающие из-за сокращения мышц под действием тока и ожоги.

Для того, чтобы избежать получения травм, пользователю персонального компьютера следует проверить свое рабочее место перед началом работы: убедиться, что розетки закреплены надежно и отсутствуют оголенные провода или видимые признаки неисправности оборудования. В случае обнаружения неисправностей, нельзя начинать работу с оборудованием или пытаться самостоятельно исправить неполадки. Неисправности должны быть устранены квалифицированным специалистом.

4.2.2.2 Опасность возникновения короткого замыкания

Короткое замыкание возникает при высоком уровне напряжения в сети и может спровоцировать пожар. Опасность представляет возможное поражение электрическим током.

Мерой предосторожности является использование стабилизатора напряжения или сетевого фильтра, которые позволяют защитить оборудование от перепадов напряжения. Необходимо также, чтобы электропроводка была скрытой.

4.2.2.3 Повышенный уровень статического электричества

Опасность повышенного уровня статического электричества состоит в том, что оно может вызывать головные боли, чрезмерную раздражительность и эмоциональность у работника, а также нарушение сна. Наибольшая опасность статического электричества состоит в возможности возникновения быстрого искрового разряда между частями оборудования. Такой заряд может привести к выходу оборудования из строя, электрическим травмам у работника или же к возникновению пожара.

Значение показателей уровня напряженности электростатических полей на рабочем месте регулируется ГОСТ 12.1.045–84 ССБТ «Электростатические

поля. Допустимые уровни на рабочих местах и требования к проведению контроля». В этом документе предельно допустимый уровень напряженности электростатических полей составляет 60 кВ/м в течение 1 часа [29].

Для минимизации количества статического электричества следует заземлять оборудование и коммуникации, а также повышать уровень влажности воздуха в помещениях.

4.3 Экологическая безопасность

Разрабатываемый проект представляет собой программу, поэтому негативное влияние на экологию окружающей среды при взаимодействии пользователя с продуктом связано с эксплуатацией персонального компьютера, а также с его утилизацией.

Негативное воздействие на окружающую среду может оказать неправильная утилизация батареи компьютера-ноутбука, которая содержит в себе тяжелые металлы, кислоты и щелочи, способные стать источником загрязнения литосферы или гидросферы. Для того, чтобы избежать возможного ущерба окружающей среде, при утилизации аккумуляторных батарей следует обращаться в специализированные учреждения, занимающиеся утилизацией и переработкой аккумуляторных батарей [30].

Другим опасным фактором воздействия на литосферу является неправильная утилизация люминесцентных ламп, которые используются для искусственного освещения на рабочих местах. Одно из опасных веществ, которое может загрязнить атмосферу, гидросферу и литосферу, а также стать причиной отравления человека и других живых существ – ртуть, содержащаяся в люминесцентных лампах в объеме от 10 до 70 мг. По истечении срока службы ламп (пять лет) их требуется сдавать на утилизацию в специализированные учреждения. Утилизация и транспортировка ламп должна производиться в соответствии с ГОСТ 6825-91 [31].

4.4 Безопасность в чрезвычайных ситуациях

Работа с электроприборами, в частности, с персональным компьютером, связана с риском возникновения пожара. Пожар – это неконтролируемый процесс горения, который происходит вне специального очага и наносит материальный ущерб, вред здоровью и жизни людей, интересам общества и государства.

Пожар может случиться из-за обрывания проводов, замыкания электропроводки оборудования, неисправности розеток или выключателей, при несоблюдении правил пожарной безопасности и т.д.

Существует основной ряд мер по предотвращению возникновения пожара:

- помещение требуется содержать в чистоте;
- должны быть свободны и ничем не загромождены проходы, лестничные клетки, коридоры и двери эвакуационных выходов;
- мебель и другие предметы не должны своим расположением препятствовать эвакуации в случае пожара;
- нельзя допускать использования неисправных электроприборов, розеток, рубильников, вилок и прочего оборудования;
- нельзя допускать курения в помещении и применения открытого огня;
- нельзя допускать использования электроприборов с поврежденной изоляцией.

В качестве мер предосторожности нужно ознакомиться с расположением огнетушителей и других средств пожаротушения, с планом эвакуации, а также убедиться в доступности представленных в плане эвакуационных выходов.

4.5 Итоги по разделу

В данном разделе были рассмотрены основные вредные и опасные факторы, которые могут возникнуть при работе с персональным компьютером.

Также были предложены меры, которые позволят предотвратить негативное воздействие компьютера на человека. В качестве вероятной чрезвычайной ситуации был представлен пожар. Меры противопожарной безопасности, а также требования к организации рабочего места и помещения в были также рассмотрены в данной работе.

Помимо этого, раздел содержит основные нормы и требования как к рабочему месту, так и к порядку организации труда.

Рабочее место, использовавшееся при разработке программного сервиса, соответствует всем описанным нормам организации рабочего места.

Нормы организации труда регулируются, прежде всего, Трудовым кодексом Российской Федерации. Следует также отметить, что ни один внутренний регламент организации не должен противоречить Трудовому кодексу.

Исследование, проведённое при выполнении данного раздела, показало, что нет необходимости дополнительно модернизировать рабочее место пользователя или внедрять дополнительные меры безопасности при разработке и эксплуатации сервиса, достаточно выполнять приведённые выше рекомендации.

ЗАКЛЮЧЕНИЕ

Социальные сети – инструмент быстрого и эффективного распространения информации в обществе. Как и в любых сообществах, в социальных сетях есть персоны, имеющие авторитет – пользователи-эксперты, поиск которых является целью описанного в работе программного сервиса. Выявление таких пользователей необходимо для разных целей, областей и сфер деятельности.

Актуальность работы состоит в подготовке полной и разнообразной коллекции данных для анализа, так как от качества исходной выборки в значительной степени зависит результат определения авторитетных пользователей. Результаты работы сервиса способны заменить или дополнить маркетинговые и социологические исследования, помочь в изучении общественной реакции.

В результате выполнения выпускной квалификационной работы была достигнута её цель – разработан модуль рекомендаций хештегов по заданному хештегу, характеризующему конкретную предметную область, а также выполнен рефакторинг элементов сервиса в соответствии с разработанными требованиями. В проект была внедрена библиотека Cytoscape.js, добавлен новый вид визуализации результатов – выгружаемые из Twitter данные с информацией о местоположении авторов отображаются на карте. Внесено множество исправлений в структуру страниц веб-приложения.

Один из важных результатов работы – получение свидетельства о государственной регистрации программы для ЭВМ, представленного на рисунке В.1 (приложение В).

Заключительным этапом разработки стало размещение актуальной версии проекта на production-сервере [10].

В дальнейшем планируется расширять функционал приложения, добавлять новые модули, внедрять новые алгоритмы для анализа выгружаемых из социальной сети данных.

CONCLUSION

Social networks are a tool for the rapid and effective dissemination of information in society. As in any communities, there are persons with authority in social networks – users-experts, the search for which is the main purpose of the software service described in this work. The identification of such users is necessary for different purposes, areas and scopes of activity.

The relevance of the work is to prepare a complete and diverse collection of data for analysis, since the quality of the initial sample largely depends on the result of determining authoritative users. The results of the service can replace or enlarge marketing and sociological research, help in the study of social reaction.

As a result of the final qualifying work the goal was achieved – a module of hashtag recommendations for a given hashtag characterizing a specific subject area was developed, and the service elements were refactored in accordance with the software requirements. Cytoscape.js library was implemented in the project. New visualization of results was added: data uploaded from Twitter with information about the location of the authors are displayed on the map. Many corrections have been made to the page structure of the web application.

One of the important results of the work is obtaining a certificate of state registration of the computer program, presented in Annex B.

The final stage of the development was the placement of the current version of the project on the production server [10].

In the future it is planned to increase the functionality of the application, add new modules, introduce new algorithms for the analysis of data uploaded from the social network.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Камаев Р.А., Григорьева В.В. Коммуникации государственных структур и населения в социальных медиа: специфика и направления развития // Научный журнал «Вестник Московского финансово-юридического университета МФЮА». – 2018. – №3. – С. 191-203.
2. Budiharto W., Meiliana M. Prediction and analysis of Indonesia Presidential election from Twitter using sentiment analysis // Journal of Big Data. – 2018. – №1.
3. Лазуткина Е. В. Лидеры мнений в информационном пространстве блогосферы рунета // Вестник НГУ. Серия: История, филология. – 2016. – Т. 15, №6. – С. 51-59.
4. Borgatti S.P. Identifying sets of key players in a social network // Computational & Mathematical Organization Theory. – 2006. Vol. 12, № 1. – P. 21-34.
5. Veremyev A., Prokopyev O.A., Pasiliao E.L. Critical nodes for distance-based connectivity and related problems in graphs // Networks. – 2015. Vol. 66, № 3. – P. 170-195.
6. Ortiz-Arroyo D. Discovering Sets of Key Players in Social Networks // Computational Social Networks Analysis. – 2010. – P. 27-47.
7. Shetty J., Adibi J. Discovering important nodes through graph entropy the case of enron email database // Third International Workshop on Link Discovery. – 2015. – P. 74-81.
8. Luneva E.E., Zamyatina V.S., Banokin P.I., Yefremov A.A. Estimation of social network user's influence in a given area of expertise // Journal of Physics: Conference Series. – 2017. –Vol. 803, № 1. – P. 1-6.
9. Лунева Е.Е., Ефремов А.А., Банокин П.И. Способ идентификации пользователей-экспертов в социальных сетях // Программные системы и вычислительные методы. – 2018. – № 4. – С.86-101.
10. Поиск экспертов в Twitter. URL: http://socgraph.tpu.ru/ (дата обращения: 31.05.2019).

11. Латентно-семантический анализ // Википедия. URL: https://ru.wikipedia.org/wiki/Латентно-семантический_анализ (дата обращения: 31.05.2019).

12. Латентно-семантический анализ и искусственный интеллект (ЛСА и ИИ) // Habr. URL: https://habr.com/ru/post/230075/ (дата обращения: 31.05.2019).

13. Word2vec // Википедия. URL: https://ru.wikipedia.org/wiki/Word2vec (дата обращения: 31.05.2019).

14. Немного про word2vec: полезная теория // NLPx. Tales of Data Science. URL: http://nlpx.net/archives/179 (дата обращения: 31.05.2019).

15. GloVe: Global Vectors for Word Representation // The Stanford NLP Group. URL: https://nlp.stanford.edu/projects/glove/, (дата обращения: 31.05.2019).

16. Лесковец, Юре. Анализ больших наборов данных / Юре Лесковец, Ананд Раджараман, Джеффри Д. Ульман - М.: ДМК Пресс, 2016. - 498 с.

17. Jeffrey Pennington, Richard Socher, Christopher D. Manning GloVe: Global Vectors for Word Representation // roceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). - Doha, Qatar: Association for Computational Linguistics , 2014. - С. 1532–1543.

18. Text Similarities : Estimate the degree of similarity between two texts // Medium. URL: https://medium.com/@adriensieg/text-similarities-da019229c894 (дата обращения: 31.05.2019).

19. Векторная модель // Википедия. URL: https://medium.com/@adriensieg/text-similarities-da019229c894 (дата обращения: 31.05.2019).

20. How does word2vec work? Can someone walk through a specific example? // Quora. URL: https://www.quora.com/How-does-word2vec-work-Can-someone-walk-through-a-specific-example/answer/Ajit-Rajasekharan (дата обращения: 31.05.2019).

21. Why do we use cosine similarity on Word2Vec (instead of Euclidean distance)? // Quora. URL: https://www.quora.com/Why-do-we-use-cosine-similarity-on-Word2Vec-instead-of-Euclidean-distance (дата обращения: 31.05.2019).

22. Замятина В. С. Способ и программный сервис для идентификации пользователей-экспертов в социальных сетях: магистерская диссертация / В. С. Замятина; Национальный исследовательский Томский политехнический университет (ТПУ), Инженерная школа информационных технологий и робототехники (ИШИТР), Отделение информационных технологий (ОИТ); науч. рук. Е. Е. Лунёва. – Томск, 2018.

23. Overview // Hangfire Documentation URL: https://docs.hangfire.io/en/latest/ (дата обращения: 31.05.2019).

24. GloVe: Global Vectors for Word Representation // The Stanford NLP Group. URL: https://nlp.stanford.edu/projects/glove/ (дата обращения: 31.05.2019).

25. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 01.04.2019) // КонсультантПлюс. URL: http://www.consultant.ru/document/cons_doc_LAW_34683/ (дата обращения: 31.05.2019).

26. Типовая инструкция по охране труда при работе на персональном компьютере (утв. Приказом Минсвязи РФ от 02.07.2001 N 162) // КонсультантПлюс. URL: http://www.consultant.ru/document/cons_doc_LAW_79762/ (дата обращения: 31.05.2019).

27. Рабочее место при выполнении работ сидя. Общие эргономические требования. // Интернет и Право. URL: https://internet-law.ru/gosts/gost/31970/ (дата обращения: 31.05.2019).

28. Постановление Главного государственного санитарного врача РФ от 03.06.2003 N 118 (ред. от 21.06.2016) "О введении в действие санитарно-эпидемиологических правил и нормативов СанПиН 2.2.2/2.4.1340-03" (вместе с "СанПиН 2.2.2/2.4.1340-03. 2.2.2. Гигиена труда, технологические процессы, сырье, материалы, оборудование, рабочий инструмент. 2.4. Гигиена детей и подростков. Гигиенические требования к персональным электронно-вычислительным машинам и организации работы. Санитарно-эпидемиологические правила и нормативы", утв. Главным государственным

санитарным врачом РФ 30.05.2003) (Зарегистрировано в Минюсте России 10.06.2003 N 4673) // КонсультантПлюс URL: http://www.consultant.ru/document/cons_doc_LAW_42836/ (дата обращения: 31.05.2019).

29. ГОСТ 12.1.045-84 ССБТ. Электростатические поля. Допустимые уровни на рабочих местах и требования к проведению контроля. // Интернет и Право. URL: https://www.internet-law.ru/gosts/gost/2729/ (дата обращения: 31.05.2019).

30. Постановление Главного государственного санитарного врача РФ от 30.04.2003 N 80 "О введении в действие Санитарно-эпидемиологических правил и нормативов СанПиН 2.1.7.1322-03" (вместе с "СанПиН 2.1.7.1322-03. 2.1.7. Почва. Очистка населенных мест, отходы производства и потребления, санитарная охрана почвы. Гигиенические требования к размещению и обезвреживанию отходов производства и потребления. Санитарно-эпидемиологические правила и нормативы", утв. Главным государственным санитарным врачом РФ 30.04.2003) (Зарегистрировано в Минюсте РФ 12.05.2003 N 4526) // КонсультантПлюс. URL: http://www.consultant.ru/document/cons_doc_LAW_42228/ (дата обращения: 31.05.2019).

31. ГОСТ 6825-91. Лампы люминесцентные трубчатые для общего освещения. // Интернет и Право. URL: https://www.internet-law.ru/gosts/gost/28169/ (дата обращения: 31.05.2019).

Приложение А
(обязательное)

Описание изменений в кодовой базе с указанием модулей приложения

Таблица А.1 – Описание изменений в кодовой базе проекта

Модуль приложения	Название компонента	Тип компонента	Характер изменений
Обработчик социального графа	ProccessTwData Controller.cs	Класс контроллера	1. Добавлены фильтры авторизации. 2. Заданы начальные параметры для пагинации, добавленной в соответствующее представление. 3. Добавлены модели для хранения слов и их векторных представлений в MongoDB (Word и WordWithDistances). 4. Разработаны методы, составляющие модуль поиска хештегов-рекомендаций к заданному пользователем хештегу. 5. К именам коллекций добавлена метка времени в текущем часовом поясе. 6. Написан код для выбора твитов, имеющих сведения о геокоординатах или местоположении автора. 7. Написан код для получения геокоординат твитов, имеющих информацию только в виде адреса. 8. Реализовано добавление коллекций твитов с геокоординатами в основные коллекции твитов в MongoDB.

Продолжение таблицы А.1

Модуль приложения	Название компонента	Тип компонента	Характер изменений
Обработчик социального графа	TWData Analizer Controller.cs	Класс контроллера	1. Добавлены фильтры авторизации. 2. Запись в объект ViewBag следующих данных: • количество твитов с геокоординатами в выбранной коллекции; • общее количество твитов в коллекции; • информация о твитах с геокоординатами в формате JSON.
	AnalizeTW Data.cs	Класс библиотеки	1. Написан код, нормализующий значения показателей информационной энтропии и Боргатти. 2. Написан код метода, возвращающего JSON-строку формата, принимаемого Cytoscape.js для отображения социального графа.
	TwitterData Process.cs	Класс библиотеки	3. Модель коллекции твитов TwitterSearchData расширена новыми свойствами, которые нужны для записи рекомендованных и отмеченных геометкой хештегов в MongoDB. 4. Добавлен метод, извлекающий список уникальных хештегов из всех твитов коллекции. 5. Изменены параметры поиска твитов: в качестве языка установлен английский, в качестве категории выбираемых твитов выбраны все типы. 6. Добавлен код, запускающий задачу выборки рекомендованных хештегов в отдельном потоке.

Продолжение таблицы А.1

Модуль приложения	Название компонента	Тип компонента	Характер изменений
Обработчик социального графа	PageInfo.cs	Модель	Описана модель сущности «пагинация», хранящая в себе свойства для формирования постраничного отображения коллекций данных.
	Pagination View Model.cs	Модель представления	Описана модель представления, содержащая свойства, в которых хранятся коллекция данных и модель PageInfo с начальными значениями для отображения информации в постраничном виде.
Веб-приложение для загрузки, анализа и визуализации данных	Login .cshtml	Представление	Внесены изменения, связанные с исправлением вёрстки страницы авторизации.
	Register .cshtml	Представление	Внесены изменения, связанные с исправлением вёрстки страницы регистрации нового аккаунта.
	Extract Collection .cshtml	Представление	1. Заданы параметры отображения количества страниц для пагинации. 2. В коде представления появились изменения, связанные с передачей хештега между двумя представлениями посредством записи выбранного хештега в URL-параметр загружаемой страницы. 3. Исправлена вёрстка страницы, удалены лишние элементы. 4. Добавлена автоматическая перезагрузка страницы, пока не будут отображены рекомендованные хештеги.
	ProccessTwData/ Index.cshtml	Представление	В коде представления появились изменения, связанные с автозаполнением поля для ключевого слова, если это представление было открыто в результате нажатия на какой-либо рекомендованных хештегов в представлении с содержимым коллекции твитов.

Продолжение таблицы А.1

Модуль приложения	Название компонента	Тип компонента	Характер изменений
Веб-приложение для загрузки, анализа и визуализации данных	TwResult .cshtml	Представление	1. Изменены элементы интерфейса, структура страницы, изменено отображение графиков с результатами идентификации экспертов. 2. Исправлена вёрстка. 3. Были внесены изменения, связанные с отображением социального графа с помощью библиотеки Cytoscape.js. 4. Добавлен код для отображения геометок твитов на карте. 5. Внесены изменения в код, отвечающий за отрисовку диаграммы, на которой показаны результаты выделения кластеров.
	DrawGraph .cshtml	Представление	Изменён размер элемента, в котором происходит отрисовка визуализированного социального графа.
	Borgatti .cshtml	Представление	Проведён подбор параметров отображения диаграммы с учётом наилучшего отображения данных.
	Clusters .cshtml	Представление	Проведён подбор параметров отображения диаграммы с учётом наилучшего отображения данных.
	DrawGraph .cshtml	Представление	Проведён подбор параметров отображения диаграммы с учётом наилучшего отображения данных.
	Yandex Geocoder	Вспомогательный модуль	Данный модуль внедрён в проект для облегчения работы с API Геокодера Яндекс.Карт. Его использование необходимо при определении геокоординат автора твита по адресу.

Приложение Б
(справочное)

Скриншоты веб-страниц до внесения в них визуальных исправлений

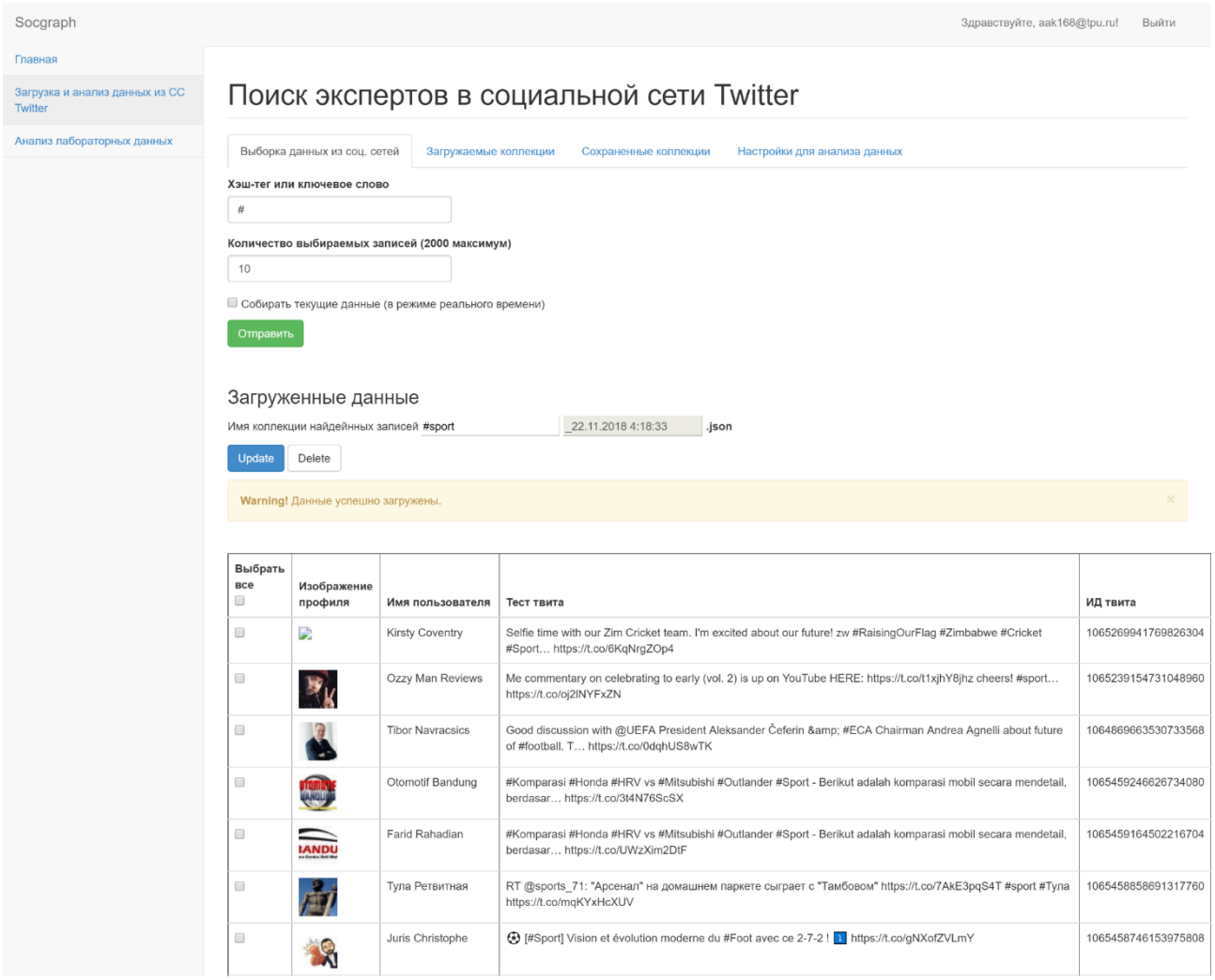

Рисунок Б.1 – Просмотр коллекции данных в исходной версии проекта

Раскрасить топ в соответствии с показателем

- Боргатти
- Информационной энтропии
- Кендалла-Уэя

Применить

Граф, сгенерированный по анализируемой коллекции

Рисунок Б.2 – Визуализация социального графа с помощью библиотеки D3

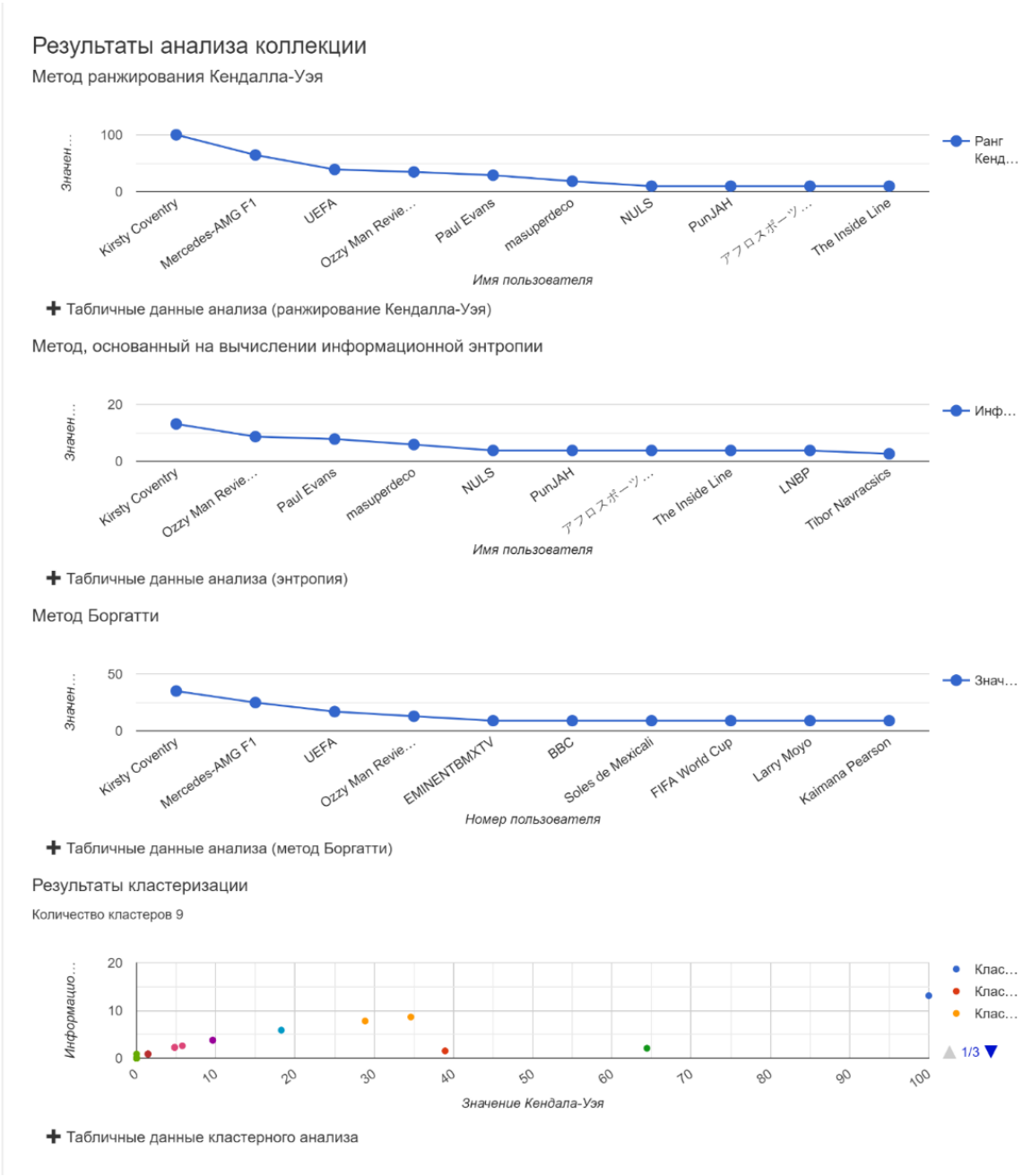

Рисунок Б.3 – Визуализация результатов идентификации пользователей до внесения изменений

Приложение В
(обязательное)

Свидетельство о государственной регистрации программы для ЭВМ


РОССИЙСКАЯ ФЕДЕРАЦИЯ

СВИДЕТЕЛЬСТВО

о государственной регистрации программы для ЭВМ

№ 2019610669

Идентификация пользователей – экспертов социальной сети Twitter в заданной предметной области

Правообладатель: *федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский Томский политехнический университет» (RU)*

Авторы: *Лунёва Елена Евгеньевна (RU), Банокин Павел Иванович (RU), Ефремов Александр Александрович (RU), Кондратьева Анна Александровна (RU)*

Заявка № 2018665342
Дата поступления 29 декабря 2018 г.
Дата государственной регистрации
в Реестре программ для ЭВМ *15 января 2019 г.*

Руководитель Федеральной службы по интеллектуальной собственности

Г.П. Ивлиев

Рисунок В.1 – Свидетельство о государственной регистрации программы для ЭВМ