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: Инженерная школа информационных технологий и робототехники
Направление подготовки: Информационные системы и технологии
Отделение школы (НОЦ): Отделение информационных технологий

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Кластеризация текстовых данных на основе методов машинного обучения

УДК 004.422.6:004.85

Студент

Группа	ФИО	Подпись	Дата
8И8А	Киселев Кирилл Игоревич		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	Иванова Юлия Александровна	к.т.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН	Рыжакина Татьяна Гавриловна	к.э.н		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель (ООД, ШБИП)	Мезенцева Ирина Леонидовна			

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	Цапко Ирина Валериевна	к.т.н.		

Томск – 2022 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код компетенции	Наименование компетенции
УК(У)-1	Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач
УК(У)-2	Способен определять круг задач в рамках поставленной цели и выбирать оптимальные способы их решения, исходя из действующих правовых норм, имеющихся ресурсов и ограничений
УК(У)-3	Способен осуществлять социальное взаимодействие и реализовывать свою роль в команде
УК(У)-4	Способен осуществлять деловую коммуникацию в устной и письменной формах на государственном языке Российской Федерации и иностранном(-ых) языке(-ах)
УК(У)-5	Способен воспринимать межкультурное разнообразие общества в социально-историческом, этическом и философском контекстах
УК(У)-6	Способен управлять своим временем, выстраивать и реализовывать траекторию саморазвития на основе принципов образования в течение всей жизни
УК(У)-7	Способен поддерживать должный уровень физической подготовленности для обеспечения полноценной социальной и профессиональной деятельности
УК(У)-8	Способен создавать и поддерживать в повседневной жизни и в профессиональной деятельности безопасные условия жизнедеятельности для сохранения природной среды, обеспечения устойчивого развития общества, в том числе при угрозе и возникновении чрезвычайных ситуаций и военных конфликтов
УК(У)-9	Способен проявлять предприимчивость в практической деятельности, в т.ч. в рамках разработки коммерчески перспективного продукта на основе научно-технической идеи
УК(У)-10	Способен принимать обоснованные экономические решения в различных областях жизнедеятельности
УК(У)-11	Способен формировать нетерпимое отношение к коррупционному поведению
ОПК(У)-1	Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности
ОПК(У)-2	Способен понимать принципы работы современных информационных технологий и программных средств, в том числе отечественного производства, и использовать их при решении задач профессиональной деятельности
ОПК(У)-3	Способен решать стандартные задачи профессиональной деятельности на основе информационной и библиографической культуры с применением информационно-коммуникационных технологий и с учетом основных требований информационной безопасности
ОПК(У)-4	Способен участвовать в разработке технической документации, связанной с профессиональной деятельностью с использованием стандартов, норм и правил
ОПК(У)-5	Способен инсталлировать программное и аппаратное обеспечение для информационных и автоматизированных систем
ОПК(У)-6	Способен разрабатывать алгоритмы и программы, пригодные для практического применения в области информационных систем и технологий

Код компетенции	Наименование компетенции
ОПК(У)-7	Способен осуществлять выбор платформ и инструментальных программно-аппаратных средств для реализации информационных систем
ОПК(У)-8	Способен применять математические модели, методы и средства проектирования информационных и автоматизированных систем
ПК(У)-1	Способен выполнять интеграцию программных модулей и компонент
ПК(У)-2	Способен выполнять работы и управлять работами по созданию (модификации) и сопровождению информационных систем
ПК(У)-3	Способен создавать техническую документацию на продукцию в сфере информационных технологий, управлять технической информацией
ПК(У)-4	Способен выполнять работы по обеспечению функционирования баз данных и обеспечению их информационной безопасности
ПК(У)-5	Способен проводить, оценивать и следить за выполнением концептуального, функционального и логического проектирования систем малого и среднего масштаба и сложности

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: Инженерная школа информационных технологий и робототехники
Направление подготовки: 09.03.02 Информационные системы и технологии
Отделение школы (НОЦ): Отделение информационных технологий

УТВЕРЖДАЮ
Руководитель ООП
_____ _____ Цапко И.В.
(Подпись) (Дата) (ФИО)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы

Студенту:

Группа	ФИО
8И8А	Киселев Кирилл Игоревич

Тема работы:

Кластеризация текстовых данных на основе методов машинного обучения	
Утверждена приказом директора (дата, номер)	03.02.2022 №34-1/с

Срок сдачи студентом выполненной работы:	06.06.2022

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.).	Разработка системы кластеризации поисковых запросов семантического ядра, позволяющей сократить время анализа поисковых запросов для поисковой оптимизации сайта.
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	1. Определение требований к системе; 2. Разработка нейросетевой части ИС; 3. Кластеризация данных и оценка получаемого результата; 4. Разработка веб-приложения ИС; 5. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение;

дополнительных разделов, подлежащих разработке; заключение по работе).	6. Социальная ответственность; 7. Заключение по работе.
Перечень графического материала (с точным указанием обязательных чертежей)	Презентация в формате *.pptx
Консультанты по разделам выпускной квалификационной работы (с указанием разделов)	
Раздел	**Консультант**
Финансовый менеджмент	Рыжакина Татьяна Гавриловна
Социальная ответственность	Мезенцева Ирина Леонидовна
Названия разделов, которые должны быть написаны на русском и иностранном языках:	
Все разделы должны быть написаны на русском языке.	

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	24.01.2022

Задание выдал руководитель:

Должность	ФИО	Учёная степень, звание	Подпись	Дата
Доцент ОИТ	Иванова Ю.А.	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И8А	Киселев Кирилл Игоревич		

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 – Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.03.02 Информационные системы и технологии
Уровень образования – Бакалавриат
Отделение школы (НОЦ) – Отделение информационных технологий
Период выполнения – весенний семестр 2021/2022 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:	06.06.2022

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
03.05.2022	Основная часть	75
19.05.2022	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	15
31.05.2022	Социальная ответственность	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	Иванова Юлия Александровна	к.т.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	Цапко Ирина Валериевна	к.т.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8И8А	Киселев Кирилл Игоревич

Школа	ИШИТР	**Отделение школы (НОЦ)**	Отделение информационных технологий
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*	*Работа с информацией, представленной в российских и иностранных научных публикациях, аналитических материалах, статических бюллетенях и изданиях, нормативно-правовых документах; анкетирование; опрос.*
2. *Нормы и нормативы расходования ресурсов*	
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения*	*Проведение предпроектного анализа. Определение целевого рынка и проведение его сегментирования. Выполнение SWOT-анализа проекта*
2. *Планирование и формирование бюджета научных исследований*	*Определение структуры работы. Расчет трудоемкости выполнения работ. Подсчет бюджета исследования*
3. *Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования*	*Рассчитать показатели финансовой эффективности, ресурсоэффективности и эффективности исполнения*
Перечень графического материала *(с точным указанием обязательных чертежей)***:**	
1. *Оценка конкурентоспособности технических решений* 2. *Матрица SWOT* 3. *Альтернативы проведения НИ* 4. *График проведения и бюджет НИ* 5. *Оценка ресурсной, финансовой и экономической эффективности НИ*	

Дата выдачи задания для раздела по линейному графику	03.02.2022

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН	Рыжакина Татьяна Гавриловна	к.э.н.		03.02.2022

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И8А	Киселев Кирилл Игоревич		03.02.2022

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО		
8И8А	Киселев Кирилл Игоревич		
Школа	**Инженерная школа информационных технологий и робототехники**	**Отделение (НОЦ)**	**Отделение информационных технологий**
Уровень образования	Бакалавриат	**Направление/ специальность**	09.03.02 Информационные системы и технологии

Тема ВКР:

Кластеризация текстовых данных с помощью методов машинного обучения	
Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места) при разработке проектного решения/при эксплуатации	*Объект исследования:* система кластеризации поисковых запросов семантического ядра *Область применения:* поисковая оптимизация *Рабочая зона:* офис *Размеры помещения:* 4*5 м. *Количество и наименование оборудования рабочей зоны:* персональный компьютер, рабочий стол, компьютерное кресло. *Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне:* изучение области поисковой оптимизации; определение набора данных для проведения работы; выбор методов обработки входных данных; определение методов кластеризации данных; разработка программного решения; проведение оценки полученного результата.
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности при разработке проектного решения: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	− «Трудовой кодекс Российской Федерации» от 30 декабря 2001 г. № 197-ФЗ (редакция, действующая с 1 марта 2022 года) − ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования
2. Производственная безопасность при разработке проектного решения:	**Опасные факторы:** 1. Производственные факторы, связанные с электрическим током, вызываемым разницей

– Анализ выявленных вредных и опасных производственных факторов	электрических потенциалов, под действие которого попадает работающий; **Вредные факторы:** 1. Производственные факторы, связанные с аномальными микроклиматическими параметрами на местонахождении работающего; 2. Отсутствие или недостаток необходимого освещения; 3. Нервно-психические перегрузки (умственное перенапряжение, перенапряжение анализаторов). **Требуемые средства коллективной и индивидуальной защиты от выявленных факторов:** обеспечение надлежащего проветривания помещения; обеспечение требуемого отопления в холодное время года; проведение регулярной влажной уборки помещения; использование увлажнителей воздуха в помещении; настройка яркости и контрастности монитора компьютера близким к уровню естественного освещения; установка дополнительных средств освещения при недостаточной освещенности рабочего места; 40-часовая рабочая неделя; наличие перерыва в течение рабочего дня; предоставление выходного дня каждую неделю; проводящие части, находящиеся под напряжением, не доступны; защитные меры предосторожности от прикосновения и повреждения проводящих частей.
3. Экологическая безопасность при разработке проектного решения	**Воздействие на литосферу:** утилизация отходов электрооборудования
4. Безопасность в чрезвычайных ситуациях при разработке проектного решения	**Возможные ЧС:** обрушение здания, аварии на коммунальных системах жизнеобеспечения населения, пожар **Наиболее типичная ЧС:** пожар
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Старший преподаватель (ООД, ШБИП)	Мезенцева Ирина Леонидовна			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8И8А	Киселев Кирилл Игоревич		

РЕФЕРАТ

В Выпускная квалификационная работа содержит: 79 страниц, 7 рисунков и 26 таблиц.

Ключевые слова: кластеризация, обработка естественного языка, поисковая оптимизация, семантическое ядро, преобразователи предложений.

Объектом исследования являются методы кластеризации текстовой информации с использованием нейросетей.

Цель данной работы является разработка системы кластеризации поисковых запросов семантического ядра.

В процессе исследования был проведен анализ методов обработки текстовых данных, определение алгоритмов кластеризации данных, проведена оценка результатов кластеризации, разработана клиентская и серверная часть веб-приложения.

Разрабатываемая система позволяет сократить время анализа поисковых запросов для семантического ядра сайта, при помощи объединения близких по смыслу поисковые запросы в удобные для работы группы.

Область применения: интернет-маркетинг.

Для развития проекта планируется дополнить систему методами определения коммерциализации поисковых запросов, собрать и обработать данные пользователей для улучшения текущих методов семантической кластеризации.

ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

ИС – информационная система.

НС – нейронная сеть.

SBERT – Sentense-BERT.

NLP – Natural Language Processing.

DBSCAN – density-based spatial clustering of applications with noise.

ВИ – вариант использования.

SEO – Search Engine Optimization.

HTTP – HyperText Transfer Protocol.

CSS – Cascading Style Sheets.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

С повсеместным проникновением интернета в повседневную жизнь и с постоянной конкуренцией за привлечение внимания клиентов важное значение приобретает функция продвижения сайтов.

Согласно исследованиям, было выявлено, что наиболее популярными способами продвижения бизнеса в интернете выступают поисковая оптимизация, маркетинг в социальных сетях и реклама, в которой рекламодатель платит за переход пользователя на сайт.

Результаты исследований показывают, что в долгосрочной перспективе поисковая оптимизация выступает самым эффективным методом. Согласно исследованиям, поисковая выдача определяет около 50% посещений интернет-ресурса, маркетинг в социальных сетях - около 5%, реклама - около 10% [10].

При этом важность функции продвижения затрагивает как коммерческие, так и информационные сайты. Для коммерческих ресурсов важен большой трафик клиентов, позволяющий обеспечить высокий уровень продаж. Информационные ресурсы не продают товары, но большое количество пользователей позволяет увеличить доход от размещаемой рекламы.

Таким образом большинство сайтов настраиваются с целью привлечения и удержания аудитории. Структура определяется согласно пользовательскому трафику, чтобы посетитель сайта переходил на интересующую его страницу и получал необходимую для него информацию.

Согласно исследованиям, 50% пользователей получают интересующую их информацию на первой странице поисковой системы, 35% пользователей рассматривают поисковые запросы на 2 и 3 страницах и только 15% доходят до четвертой. На последующих страницах интерес

пользователей резко падает и процент посещения значительно снижается [10].

Для продвижения позиции сайта в поисковых системах и улучшения его с позиции пользователей требуется поисковая оптимизация. Она позволяет поисковым системам корректно определять контент ресурса и верно представлять его для пользователей.

Важным, но трудоемким этапом поисковой оптимизации выступает разбиение семантического ядра на определенные группы, которые следует продвигать на одной странице. Для сокращения усилий, прилагающихся на определение данных групп, требуется автоматизация данного процесса.

Таким образом, целью данной работы является разработка системы кластеризации поисковых запросов семантического ядра.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Постановка требований к разрабатываемой системе;

2. Определение методов машинного обучения для обработки поисковых запросов;

3. Настройка нейросети для задачи поисковой оптимизации;

4. Определение алгоритма кластеризации данных;

5. Оценка результата кластеризации;

6. Разработка сервиса кластеризации поисковых запросов.

1 АНАЛИЗ ПРЕДМЕТНОЙ ОБЛАСТИ

Согласно комплексу стандартов на автоматизированные системы был проведен анализ разрабатываемой информационной системы по системы кластеризации поисковых запросов семантического ядра [8].

1.1 Назначение и цели создания системы

Кластеризация поисковых позволяет сформировать мнение о предполагаемом контенте разделов сайта, наиболее выгодных словах и фразах для продвижения в поисковых системах и оптимизации различных частей сайта.

Разрабатываемая система направлена на сокращение времени и вероятности ошибок при анализе важных для сайта поисковых запросов. Для этого используется процесс кластеризации, позволяющий автоматизировать процесс поисковой оптимизации, объединяющий близкие по смыслу поисковые запросы в удобные для работы группы.

В качестве потенциальных потребителей разрабатываемой системы выступают интернет-маркетологи.

Благодаря кластеризации поисковых запросов появляются возможности:

- Более точное понимание потребностей пользователей. Исследование тематической релевантности делает поисковую оптимизацию лучше для посетителей сайта: целью выступает определение потребности пользователя, а не выделение отдельных наиболее подходящих поисковых запросов.
- Получение наибольшего количества ключевых слов и фраз для продвижения. Становится проще ориентироваться в большом списке поисковых запросов, предварительно разделенном на смысловые группы.

- Отсеять неподходящие ключевые слова и фразы для продвижения. Можно рассмотреть вопрос исключения для формирования семантического ядра поисковых запросов, не объединённых в большие и точные кластеры, или проанализировать их дополнительно вручную.
- Определить как связаны между собой различные части сайта. Возможно, требуется объединить различный контент в один общий раздел сайта или необходимо разделить одну категорию на две меньшие.
- Увеличить видимость сайта в поисковых системах. Наиболее подходящие результаты поискового запроса позволят сделать сайт наиболее привлекательным для посетителей, что послужит в дальнейшем продвижением его позиций в результатах поисковой выдачи [10].

1.2 Требования к системе

Идентификация требований

В качестве подхода к кодификации требований используется вариант мнемонических идентификаторов. Формат идентификатора имеет вид

[А][ББ].[ВВ], где:

А — префикс

ББ, ВВ — двузначное число от 00 до 99

ББ — код первого уровня

ВВ — код второго уровня

Тип требования и его префикс:

Р — Показатели назначения

М — Требования к математическим моделям и численным методам

R — Требования к форматам файлов

S — Требование к информационной безопасности

F — Требование к функциям, выполняемым системой

A — Требования к архитектуре системы

L — Требования к лингвистическому обеспечению

T — Требования к аппаратному обеспечению

G — Требования к программному обеспечению

Показатели назначения

P01 Для кластеризации списка тысячи поисковых запросов время обработки не должно превышать 5 с.

P02 Система должна справляться с одновременной обработкой запросов ста пользователей.

Требования к математическим моделям и численным методам

M01 Система должна кластеризовывать список поисковых запросов на группы по следующему правилу: для одного из поисковых запросов одной группы количество совпадений первых десяти гиперссылок, выдаваемых в поисковой системе Google, должно быть больше, чем между поисковыми запросами разных групп.

M02 Система должна использовать нейросетевые подходы для кластеризации.

Требования к форматам файлов

R01 Загружаемый и выгружаемый списки поисковых запросов должны быть файлами в формате CSV.

R02 Загружаемый список поисковых запросов должен состоять из одного столбца, в котором записаны поисковые запросы.

R03 Выгружаемый список поисковых запросов должен состоять из столбца, в котором записаны поисковые запросы, и столбца, в котором указан номер кластера поискового запроса.

Требования к информационной безопасности

S01 Система не должна хранить загружаемый и выгружаемый списки поисковых запросов на постоянных устройствах хранения.

Требования к функциям, выполняемым системой

F01 Система должна позволять загрузить список поисковых запросов.

F02 Система должна кластеризовывать загружаемый список поисковых запросов.

F03 Система должна позволять выгрузить кластеризованный список поисковых запросов.

Требования архитектуре системы

A01 Система должна представлять собой веб-приложение.

A02 Кластеризация производится на стороне сервера.

Требования к лингвистическому обеспечению

L01 Система должна обеспечивать англоязычный интерфейса пользователя.

Требования к аппаратному обеспечению клиентской части

Требуется персональный компьютер/ноутбук со следующими характеристиками

T01 тактовая частота процессора: не менее 2 Ghz;

T02 оперативная память (ОЗУ): не менее 2 Gb;

T03 экран: не менее 14”, разрешение 1024×768.

Требования к аппаратному обеспечению серверной части

Требуется сервер со следующими характеристиками

T04 тактовая частота процессора: не менее 1 Ghz;

T05 оперативная память (ОЗУ): не менее 2 Gb;.

T06 жёсткий диск: не менее 24 Gb дискового пространства.

Требования к программному обеспечению клиентской части

G01 операционная система: Windows 7/8/10.

Требования к программному обеспечению серверной части

Не предъявляются.

2 РАЗРАБОТКА НЕЙРОСЕТЕВОГО КОМПОНЕНТА ИС

2.1 Обработка поисковых запросов

Выделенными популярными решениями решения задачи семантической кластеризации выступают нейронные сети GPT-n и Sentence-BERT (SBERT).

GPT-n является генеративной сетью-трансформером, которая позволяет предсказывать появление последующего слова в предложении. Серия GPT-n показывает очень многообещающие результаты для задач классификации, однако эти модели требуют огромных вычислительных ресурсов и сильно чувствительны к их настройке.

Для решения задачи семантической кластеризации была выбрана модификация предварительно обученной сети BERT — Sentence-BERT (SBERT). Использование данной архитектуры является наиболее подходящим решением, так как требуется меньшее количество вычислений, и сохраняется современный уровень точности для определения текстового сходства, заданный BERT [11].

SBERT — это сиамский бикодировщик, использующий объединение средних значений для кодирования и косинусное сходство или расстояние Манхэттена/Евклида для нахождения семантически сходных предложений. Это является очень эффективным, не требующим больших вычислительных ресурсов методом, который позволяет решить задачу поиска наиболее похожей пары в наборе из 10 000 предложений менее чем за 5 секунд [2].

SBERT можно легко адаптировать под конкретную задачу, что является положительным фактором для использования в данной работе.

Для настройки требуется сообщить сети, какие пары предложений похожи и должны быть близки в векторном пространстве, а какие пары отличаются и должны быть далеко в векторном пространстве. Для этого

требуется аннотировать пары предложений оценкой, указывающей на их сходство, по шкале от 0 до 1. Затем необходимо обучить сеть с помощью сиамской сетевой архитектуры.

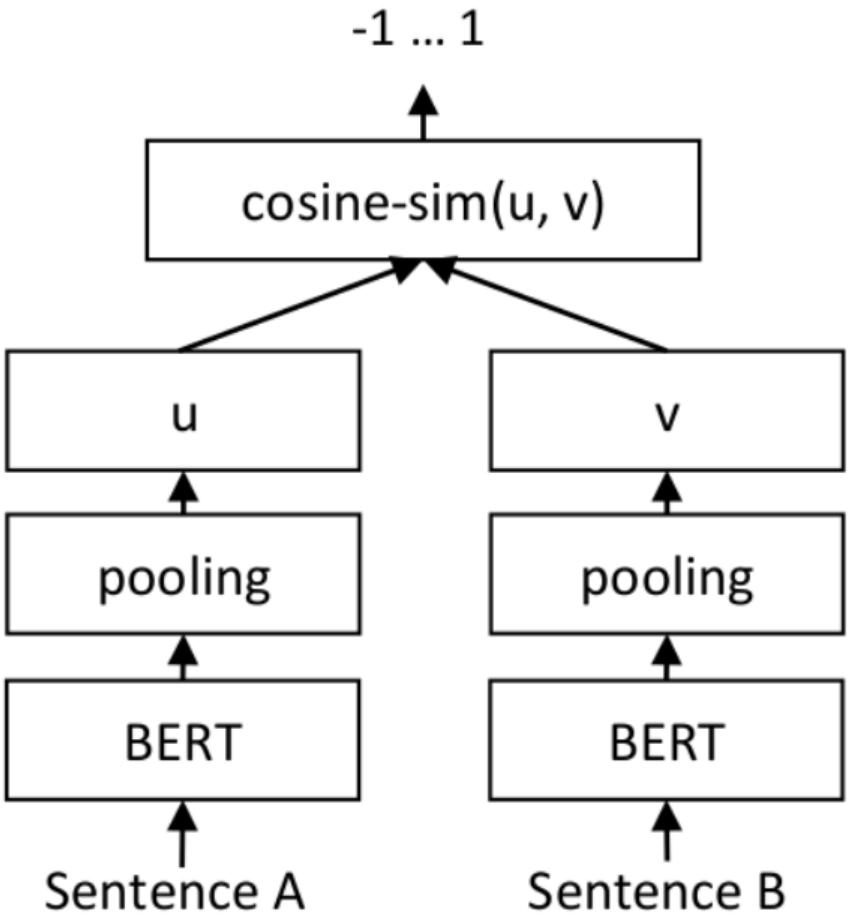

Рисунок 1 — Архитектура SBERT для определения показателей сходства предложений

Каждая пара предложений A и B пропускается через сеть, которая генерирует вложения u и v. Для данных вложений вычисляется косинусное сходство, и результат сравнивается с предварительно определенной оценкой. Это позволяет обеспечить простую и удобную настройку сети [3].

2.2 Предварительно обученные модели

Современным подходом решения задач по обработке естественного языка является использование предварительно обученных моделей. Это

определяется большими затратами времени и вычислительных ресурсов для обучения моделей с нуля.

В таблице 1 представлен обзор предварительно обученных моделей, предоставляемых библиотекой Sentence-Transformers. Они были тщательно оценены на предмет их качества для встроенных предложений (Performance Sentence Embeddings) и для встроенных поисковых запросов и абзацев (Performance Semantic Search) [1].

Таблица 1 — Список моделей Sentence-Transformers

Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
paraphrase-albert-small-v2	64.46	40.04	52.25	5000	43 MB
paraphrase-MiniLM-L3-v2	62.29	39.19	50.74	19000	61 MB
paraphrase-MiniLM-L6-v2	64.82	40.31	52.56	14200	80 MB
all-MiniLM-L6-v1	68.03	48.07	58.05	14200	80 MB
all-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB
multi-qa-MiniLM-L6-dot-v1	63.90	49.19	56.55	14200	80 MB
multi-qa-MiniLM-L6-cos-v1	64.33	51.83	58.08	14200	80 MB
paraphrase-MiniLM-L12-v2	66.01	43.01	54.51	7500	120 MB
all-MiniLM-L12-v1	68.83	50.78	59.80	7500	120 MB
all-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB
sentence-t5-base	67.84	44.63	56.23	2500	210 MB
gtr-t5-base	67.65	51.15	59.40	2500	210 MB
average_word_embeddings_komninos	51.13	21.64	36.39	22000	240 MB
paraphrase-TinyBERT-L6-v2	66.19	41.07	53.63	4500	240 MB
multi-qa-distilbert-dot-v1	66.67	52.51	59.59	4000	250 MB
multi-qa-distilbert-cos-v1	65.98	52.83	59.41	4000	250 MB

В качестве предварительно обученных моделей, настроенных для задачи семантической кластеризации, выступают all-, paraphrase- и sentence-модели.

Для рассмотрения были выбраны paraphrase-MiniLM-L3-v2, paraphrase-MiniLM-L6-v2, all-MiniLM-L6-v1 и all-MiniLM-L6-v2, имеющие требуемые значения показателя скорость.

Данные модели Sentence-transformers отображают предложения и абзацы в 384-мерном плотном векторном пространстве и могут

использоваться для таких задач, как кластеризация или семантический поиск. В таблице 2 представлена более подробная информация про данные модели. [1]

Таблица 2 — Сравнение предварительно обученных моделей

Название модели	Базовая модель	Максимальная длинна предложения	Размерность	Функции оценки	Размер, MB	Пулинг	Тренировочные данные
paraphrase-MiniLM-L3-v2	nreimers/MiniLM-L3-H384-uncased	128	384	косинусное сходство	61	Средний	~110 млн тренировочных пар
paraphrase-MiniLM-L6-v2		128			80		
all-MiniLM-L6-v1		128		косинусное сходство, скалярное произведение, евклидовое расстояние	80		~1,1 млрд тренировочных пар
all-MiniLM-L6-v2		256			80		

2.3 Данные для обучения моделей

Для обучения вышеперечисленных моделей использовалось объединения нескольких наборов данных. Наборы данных для обучения all-MiniLM-L6-v1 и all-MiniLM-L6-v2 представлены в таблице 3 [4].

Таблица 3 — Наборы данных для обучения all-MiniLM-L6-v1 и all-MiniLM-L6-v2

Наборы данных	**Количество обучающих кортежей**
Reddit comments (2015-2018)	726,484,430
S2ORC Citation pairs (Abstracts)	77,427,422
PAQ (Question, Answer) pairs	64,371,441
S2ORC Citation pairs (Titles)	52,603,982
S2ORC (Title, Abstract)	41,769,185
Stack Exchange (Title, Body) pairs	25,316,456
MS MARCO triplets	9,144,553
GOOAQ: Open Question Answering with Diverse Answer Types	3,012,496
Yahoo Answers (Title, Answer)	1,198,260
Code Search	1,151,414
COCO Image captions	828,395
SPECTER citation triplets	828,395
Yahoo Answers (Question, Answer)	681,164
Yahoo Answers (Title, Question)	659,896
SearchQA	582,261
Eli5	325,475
Flickr 30k	317,695
Stack Exchange Duplicate questions (titles)	304,525
AllNLI (SNLI and MultiNLI	277,230
Stack Exchange Duplicate questions (bodies)	250,519
Stack Exchange Duplicate questions (titles+bodies)	250,460
Sentence Compression	180,000
Wikihow	128,542
Altlex	112,696
Quora Question Triplets	103,663
Simple Wikipedia	102,225
Natural Questions (NQ)	100,231
SQuAD2.0	87,599
TriviaQA	73,346
Общее количество	**1,124,818,467**

Для адаптации модели под задачу поисковой оптимизации было произведено переобучение на составленном наборе данных, представляющем собой пары поисковых запросов и оценку от 0 до 1, определяющую их сходство.

Первым этапом определения оценки являлось определение доли совпадения первых десяти ссылок для пар поисковых запросов. В

дальнейшем рассматриваются только те пары, у которых было хоть одно совпадение.

	A	B	C
680	best ar 15 carry handle scope	best ar 15 carry handle scope mount	0.5
681	what is the best nigth vision rifle scope for the money	wich night vision scope is best	0.7
682	is vhf and cb the same	portable cb radio vs walkie talkie	0.1
683	can i bring cpap machine on a plane	can you take cpap machine on a plane	1
684	which e mountain bike to buy	wich electric maountain bike is best	0.6
685	who makes the best lock picks	wiorld's best lock picks	0.7
686	are fire sprinklers set off by smoke	are splinkers effective	0.2

Рисунок 2 — Определение схожести на основании поисковой выдачи

Следующим этапом выступала корректировка полученного значения. Так как значение доли совпадения не точно определяет оценку схожести между двумя предложениями, то данное значение было скорректировано на основании результатов, получаемых в предварительно обученной сети согласно формуле:

$$M_{\text{обуч.}} = 0{,}85 \cdot M_{\text{пр.сети}} + 0{,}15 \cdot M_{\text{поиск.}},$$

где $M_{\text{обуч.}}$ - метрика схожести, используемая в последующем обучении нейронной сети,

$M_{\text{пр.сети}}$ - метрика схожести, определяемая путем сравнения двух поисковых запросов в предварительно обученной нейронной сети,

$M_{\text{поиск.}}$ - метрика схожести, представляющая собой долю совпадения первых десяти ссылок для пары поисковых запросов.

Таким образом результирующий обучающий набор данных для задачи поисковой оптимизации примет вид:

	A	B	C
680	best ar 15 carry handle scope	best ar 15 carry handle scope mount	0.9556
681	what is the best nigth vision rifle scope for the money	wich night vision scope is best	0.9187
682	is vhf and cb the same	portable cb radio vs walkie talkie	0.1708
683	can i bring cpap machine on a plane	can you take cpap machine on a plane	0.9867
684	which e mountain bike to buy	wich electric maountain bike is best	0.826
685	who makes the best lock picks	wiorld's best lock picks	0.739
686	are fire sprinklers set off by smoke	are splinkers effective	0.0723

Рисунок 3 — Обучающий набор данных для задачи поисковой оптимизации

3 КЛАСТЕРИЗАЦИЯ ДАННЫХ И ОЦЕНКА ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

3.1 Кластеризация данных

В качестве предполагаемых методов были рассмотрены иерархическая кластеризация и DBSCAN. Определяющими факторами стали возможность применимости к нашим данным и настройка алгоритма без предварительного определения числа кластеров.

3.1.1 Иерархическая кластеризация

Иерархическая кластеризация — это группа алгоритмов кластеризации, позволяющих создать вложенные кластеры с помощью их последовательного слияния или разделения. Существует направления методов иерархической кластеризации:

- агломеративные методы, когда новые кластеры создаются путем объединения более мелких кластеров;
- дивизионные методы, когда новые кластеры создаются путем деления более крупных кластеров на более мелкие.

В данной работе использовался агломеративный подход. Критерии связывания (объединения более мелких кластеров) определяют метрику, используемую для стратегии слияния:

- Ward — минимизирует сумму квадратов разностей во всех кластерах;
- полное связывание сводит к минимуму максимальное расстояние между наблюдениями пар кластеров;
- среднее связывание минимизирует среднее расстояние между всеми наблюдениями пар кластеров;
- одиночная связь минимизирует расстояние между ближайшими наблюдениями пар кластеров.

Агломеративная кластеризация ведет к неравномерному распределению кластеров в связи с ее подходом, связанным с поглощением меньших групп большими кластерами. В данном случае одинарная связь — худшая стратегия, и Ward дает хорошие. Однако для неевклидовых показатели невозможно применение стратегии Ward, поэтому в качестве альтернативы было выбрано среднее связывание [6].

3.1.2 DBSCAN

Алгоритм DBSCAN рассматривает кластеры, как участки высокой плотности, разделенных районах с низкой плотностью. Из-за этого довольно общего представления кластеры, обнаруженные с помощью DBSCAN, могут иметь любую форму, в отличие от k-средних, которое предполагает, что кластеры имеют выпуклую форму. Данная особенность являлась важным фактором выбора данного алгоритма для дальнейшего рассмотрения.

Алгоритм DBSCAN является детерминированным, всегда генерируя одни и те же кластеры, когда им предоставляются одни и те же данные в одном порядке. Однако результаты могут отличаться, если данные предоставляются в другом порядке. Во-первых, даже если основные образцы всегда будут назначаться одним и тем же кластерам, метки этих кластеров будут зависеть от порядка, в котором эти образцы встречаются в данных. Во-вторых, что более важно, кластеры, которым назначены неосновные выборки, могут различаться в зависимости от порядка данных [6].

3.2 Меры оценки качества кластеризации

Оценка кластеризации является не тривиальной задачей, так как метод должен учитывать не абсолютные значения меток кластера, а определяет разделения данных согласно схожести членов одного кластера по сравнению элементами, не входящими в данный кластер.

3.2.1 Внешние меры оценки качества

Учитывая знания о правильном разделение на кластеры и назначениях кластеров алгоритмом кластеризации, используемыми в данной работе метриками, которые определяют сходство двух назначений, игнорируя перестановки, выступают:

1. Индекс Рэнда — метрика, оценивающая насколько много из тех пар элементов, которые находились в одном классе, и тех пар элементов, которые находились в разных классах, сохранили это состояние после кластеризации алгоритмом.

2. Взаимная информация — метрика двух случайных величин, описывающая количество информации, содержащееся в одной случайной величине относительно другой.

3. Метрики условного энтропийного анализа:

- однородность — каждый кластер содержит только членов одного класса;
- полнота — все члены данного класса относятся к одному кластеру;
- V-мера — гармоническое среднее значение однородности и полноты.

4. Индекс Фаулкса-Мэллоуса — это среднее геометрическое значение отзыва и точности, полученное между результатом алгоритма кластеризации и предварительно определенного набора проверки [7].

3.2.2 Внутренние меры оценки качества

Оценку структуры кластеров, опираясь лишь на нее, не используя внешней информации, осуществляют:

1. Коэффициент силуэта — метрика, определяющая насколько объект похож на свой кластер по сравнению с другими кластерами.

- Оценка ограничена от -1 за неправильную кластеризацию до +1 за высокоплотную кластеризацию. Баллы около нуля указывают на перекрывающиеся кластеры.
- Оценка выше, когда кластеры плотные и хорошо разделенные, что относится к стандартной концепции кластера.

2. Индекс Калински-Харабаса. Компактность основана на расстоянии от точек кластера до их центроидов, а разделимость - на расстоянии от центроид кластеров до глобального центроида.

- Оценка выше, когда кластеры плотные и хорошо разделенные, что относится к стандартной концепции кластера.

3. Индекс Дэвиса-Болдина — метрика, определяющая компактность как расстояние от объектов кластера до их центроидов, а отделимость - как расстояние между центроидами.

- В индексе вычисляются только количества и характеристики, присущие набору данных [7].

3.3 Оценка качества моделей

Проведена внешняя оценка качества иерархической кластеризации и DBSCAN для предварительно обученных моделей paraphrase-MiniLM-L3-v2, paraphrase-MiniLM-L6-v2, all-MiniLM-L6-v1 и all-MiniLM-L6-v2. Результаты данной оценки представлены в таблицах 4 и 5.

Таблица 4 — Внешняя оценка качества иерархической кластеризации

Название модели	Индекс Рэнда	Взаимная информация	Однородность	Полнота	V-мера	Индекс Фаулкса-Мэллоуса
paraphrase-MiniLM-L3-v2	0.865	0.469	0.651	0.472	0.547	0.265
paraphrase-MiniLM-L6-v2	0.871	0.509	0.689	0.501	0.580	0.299
all-MiniLM-L6-v1	0.874	0.516	0.694	0.508	0.587	0.335
all-MiniLM-L6-v2	0.875	0.516	0.702	0.505	0.587	0.329

Таблица 5 — Внешняя оценка качества DBSCAN

Название модели	Индекс Рэнда	Взаимная информация	Однородность	Полнота	V-мера	Индекс Фаулкса-Мэллоуса
paraphrase-MiniLM-L3-v2	0.778	0.367	0.527	0.447	0.484	0.226
paraphrase-MiniLM-L6-v2	0.784	0.414	0.533	0.488	0.509	0.276
all-MiniLM-L6-v1	0.774	0.383	0.518	0.461	0.489	0.240
all-MiniLM-L6-v2	0.770	0.403	0.513	0.489	0.501	0.280

Иерархический алгоритм кластеризации показал наилучшие результаты и был выбран для решения задачи поисковой оптимизации. Схожие результаты показали модели all-MiniLM-L6-v1 и all-MiniLM-L6-v2. В качестве используемой модели была выбрана all-MiniLM-L6-v2 из-за возможности обрабатывать входные данные большей длинны.

Выбранная модель была переобучена на описанном ранее наборе данных пар поисковых запросов. Проведена внутренняя оценка качества первоначальной и переобученной модели, представленная в таблице 6.

Таблица 6 — Внутренняя оценка качества переобученной модели

Название модели	Коэффициент силуэта	Индекс Калински-Харабаса	Индекс Дэвиса-Болдина
all-MiniLM-L6-v2	0.163	19.98	2.175
all-MiniLM-L6-v2 (переоб.)	0.170	22.89	1.966

Результатом переобучения стало формирование более плотных кластеров, сформированных при кластеризации поисковых запросов. Для системы была использована модель all-MiniLM-L6-v2 (переоб.).

4 РАЗРАБОТКА ВЕБ-ПРИЛОЖЕНИЯ ИС

Были описаны основные технические решения, сформированные в результате разработки данной системы.

4.1 Сценарий использования

Система имеет единственный сценарий использования, который подробно описан в варианте использования, представленным ниже.

ВИ «Кластеризация списка поисковых запросов»

Цель

Получить кластеризованный список поисковых запросов, в котором похожие по смыслу запросы объединены в одну группу

Акторы

Интернет-маркетолог

Стейкхолдеры

—

Предварительные условия / Начальное состояние

Система запущена и готова к работе

Активаторы

—

Основной сценарий

1. Пользователь загружает список поисковых запросов и инициирует кластеризацию

2. Система кластеризует загружаемый список поисковых запросов

3. Пользователь выгружает кластеризованный список поисковых запросов. Конец

Альтернативный сценарий

Предусловие: на шаге 1-3 основного сценария произошел сбой

Система сообщает о том, что произошел сбой и просит повторить шаги, начиная с первого. Конец.

4.2 Структура системы

Система использует двухзвенную архитектуру клиент-серверного взаимодействия с программной частью, отвечающую за функционирование пользовательского интерфейса, и серверной частью, отвечающей за кластеризацию загружаемых пользователем данных.

4.3 Задача «Кластеризация списка поисковых запросов»

Пользователь загружает список поисковых запросов. В браузере проверяется соответствие файла требуемому формату. Если проверка пройдена, пользователь инициирует кластеризацию. Файл отправляется на сервер. Список поисковых запросов из файла проходит нормализацию, затем он кластеризуется и отправляется в виде фала пользователю (рисунок 4)

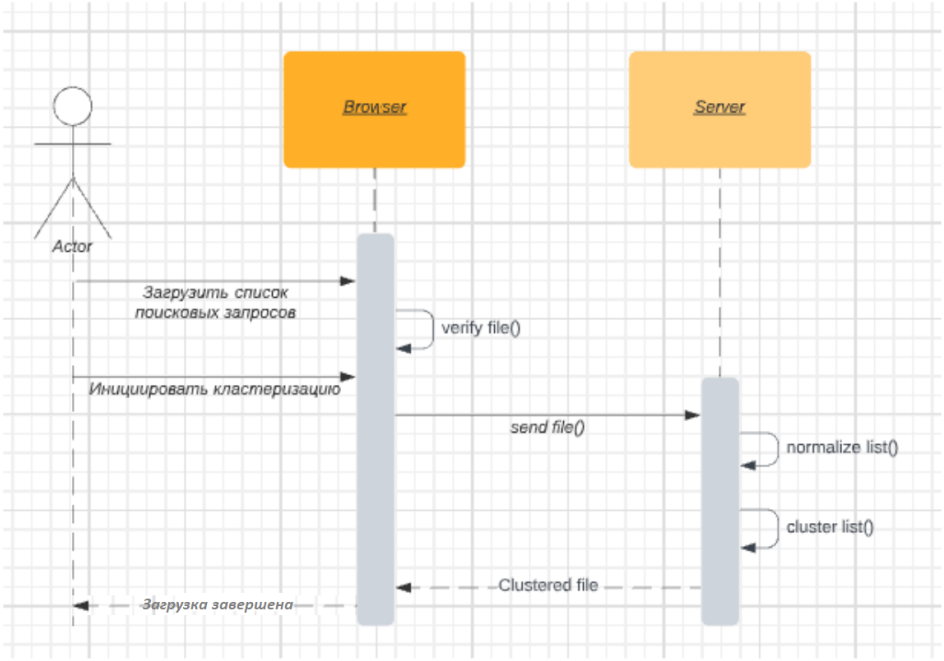

Рисунок 4 — Диаграмма последовательности для задачи «Кластеризация списка поисковых запросов»

4.4 Интерфейс пользователя

На рисунке 5 изображено главное окно приложения. Пользователю предлагается выбрать файл с перечнем поисковых запросов. После загрузки файла на сервер происходит обработка данных. Затем пользователю отправляется файл с поисковыми запросами, разбитыми на кластеры.

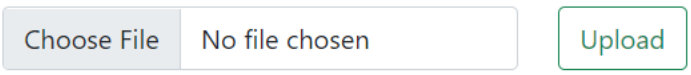

Рисунок 5 — Главное окно приложения

4.5 Состав задач и функций, реализуемых системой

Таблица 7 — Задачи и функции, реализуемые системой

Задачи, реализуемые системой	Функции, реализуемые системой
Загрузка списка поисковых запросов	Открытие стандартного окна для выбора файла
	Добавление списка поисковых запросов
	Верификация добавленного файла
Кластеризация загружаемого списка поисковых запросов	Отправление на сервер списка поисковых запросов
	Чтение отправленного списка
	Нормализация прочитанного списка
	Кластеризация нормализованного списка
Выгрузка кластеризованного списка поисковых запросов	Запись кластеризованного списка в файл
	Сохранение файла на сервере
	Загрузка файла на компьютер пользователя

4.6 Решения по составу программных средств

Система реализована в виде веб-приложения и предназначается для эксплуатации на компьютерах и ноутбуках.

Используемые технологий на стороне сервера:

- Язык программирования Python 3.9;
- Фреймворк Django 4.

Используемые технологий на стороне клиента:

- Языки HTML 5, CSS 3;
- Фреймворк Bootstrap 5.

5 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

В выпускной квалификационной работе рассматривается процесс кластеризации поисковых запросов с использованием методов машинного обучения.

Целью данного раздела является анализ перспектив проведения научного исследования.

Для достижения поставленной цели будут выполнены следующие задачи:

- оценка коммерческого потенциала и перспективности проведения научных исследований;
- определение возможных альтернатив проведения научных исследований, отвечающих современным требованиям в области ресурсоэффективности и ресурсосбережения;
- планирование научно-исследовательских работ;
- определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования.

5.1 Оценка коммерческого потенциала и перспективности проведения научных исследований

5.1.1 Потенциальные потребители результатов исследования

В качестве потенциальных потребителей выступают интернет-маркетологи.

Важным этапом их работы по продвижению сайта является поисковая оптимизация. Она используется для поднятия позиций сайта в результатах выдачи поисковых систем для увеличения его посещаемости и привлечения новых клиентов.

После сбора важных для сайта поисковых запросов, следует этап работы по их анализу. Для сокращения времени используется процесс кластеризации, позволяющий объединить близкие по смыслу поисковые запросы в удобные для работы группы.

5.1.2 Анализ конкурентных технических решений

В качестве конкурентных решений выделим следующие:

- Ручной выбор, когда процесс кластеризации выполняется вручную с помощью инструментов морфологического анализа;
- Кластеризация на основе результатов поисковой выдачи, когда поисковые запросы объединяются в группы согласно сайтам, находящимся в топе их поисковой выдачи.
- Кластеризация по семантической схожести, когда результатом является объединение в кластеры семантически близких поисковых запросов.

Анализ конкурентных технических решений рассчитывается по следующей формуле:

$$К = \sum Вi \cdot Бi, \quad (5.1)$$

где К – конкурентоспособность научной разработки,

Вi – вес показателя (в долях единицы),

Бi – балл -го показателя.

Анализ проведен с помощью оценочной карты. Оценочная карта для сравнения конкурентных технических решений представлена в таблице 5.1, где индексом «р» обозначен ручной выбор, индексом «п» – кластеризация на основе результатов поисковой выдачи, а индексом «с» – кластеризация по семантической схожести.

Таблица 5.1 – Оценочная карта для сравнения конкурентных технических решений

Критерии оценки	Вес критерия	Баллы			Конкурентоспособность		
		$Б_р$	$Б_п$	$Б_с$	$К_р$	$К_п$	$К_с$
Технические критерии оценки ресурсоэффективности							
1. Производительность труда	0,15	3	5	5	0,45	0,75	0,75
2. Удобство в эксплуатации	0,05	4	4	3	0,2	0,2	0,15
3. Точность измерения	0,2	3	4	5	0,6	0,8	1,0
4. Универсальность метода	0,15	5	3	4	0,75	0,45	0,6
5. Простота эксплуатации	0,1	3	4	3	0,3	0,4	0,3
Экономические критерии оценки эффективности							
1. Цена	0,1	5	4	4	0,5	0,4	0,4
2. Конкурентоспособность	0,15	4	3	5	0,6	0,45	0,75
3. Предполагаемый срок эксплуатации	0,05	4	5	3	0,2	0,25	0,15
4. Финансирование научной разработки	0,05	4	3	5	0,2	0,15	0,25
Итого	**1**	**35**	**35**	**37**	**3,8**	**3,85**	**4,35**

По результатам оценки конкурентных решений можно сказать, что наиболее эффективно применение кластеризации по семантической схожести. Данный метод имеет ряд технических преимуществ, которые ведут к увеличению производительности труда.

5.1.3 SWOT-анализ

SWOT-анализ применяют для оценки факторов, влияющих на продвижение проекта на рынок. На этапе SWOT-анализа необходимо определить сильные и слабые стороны проекта, а также рассмотреть возможности и угрозы для его реализации. SWOT-анализ представлен в таблице 5.2.

Таблица 5.2 – Матрица SWOT анализа

	Сильные стороны научно-исследовательского проекта: С1. Повышение производительности труда специалистов. С2. Актуальность и высокая технологичность проекта. С3. Гибкость и точность используемых алгоритмов. С4. Простой и быстрый запуск продукта	**Слабые стороны научно-исследовательского проекта:** Сл1. Малое количество потенциальных потребителей научной разработки Сл2. Сложности внедрения и поддержки продукта Сл3. Отсутствие полноценного финансирования разработки Сл4. Сложность оптимизации используемых алгоритмов
Возможности: В1. Автоматизация процессов поисковой оптимизации В2. Интеграция с другими инструментами В3. Снижение рутинной работы сотрудников В4. Расширение функционала продукта		
Угрозы: У1. Отсутствие спроса на программный продукт У2. Большая конкуренция на рынке У3. Санкции и ограничения в области информационных технологий У4. Вероятность появления более универсальных алгоритмов		

Следующий шаг состоит в выявлении соответствия сильных и слабых сторон научно-исследовательского проекта внешним условиям окружающей среды. Это соответствие или несоответствие должны помочь выявить степень необходимости проведения стратегических изменений.

Каждый фактор помечается либо знаком «+» (означает сильное соответствие сильных сторон возможностям), либо знаком «-» (что означает слабое соответствие); «0» – если есть сомнения в том, что поставить «+» или «-». Интерактивная матрица проекта представлена в табл. 5.3.

Таблица 5.3 - Интерактивная матрица сильных и слабых сторон и возможностей

		Сильные стороны				Слабые стороны			
Возможности проекта		С1	С2	С3	С4	Сл1	Сл2	Сл3	Сл4
	В1	+	0	+	-	+	-	0	0
	В2	+	+	-	+	0	+	+	0
	В3	+	-	+	-	+	0	-	+
	В4	-	+	0	+	-	+	+	+

Таблица 5.4 - Интерактивная матрица сильных сторон и слабых сторон и угроз

		Сильные стороны				Слабые стороны			
Угрозы проекта		С1	С2	С3	С4	Сл1	Сл2	Сл3	Сл4
	У1	+	+	-	-	+	0	-	0
	У2	+	+	+	-	+	+	+	-
	У3	+	0	-	+	+	+	0	-
	У4	0	+	+	-	0	-	+	+

Результатом анализа итеративных матриц проекта составляется итоговая матрица SWOT-анализа, представленная в таблице 5.5.

Таблица 5.5 - Итоговая матрица SWOT-анализа

	Сильные стороны научно-исследовательского проекта: С1. Повышение производительности труда специалистов. С2. Актуальность и высокая технологичность проекта. С3. Гибкость и точность используемых алгоритмов. С4. Простой и быстрый запуск продукта	**Слабые стороны научно-исследовательского проекта:** Сл1. Малое количество потенциальных потребителей научной разработки Сл2. Сложности внедрения и поддержки продукта Сл3. Отсутствие полноценного финансирования разработки Сл4. Сложность оптимизации используемых алгоритмов
Возможности: В1. Автоматизация процессов поисковой оптимизации В2. Интеграция с другими инструментами В3. Снижение рутинной работы сотрудников В4. Расширение функционала продукта	Интеграция с другими инструментами позволит лучше автоматизировать процесс поисковой автоматизации.	Расширение функционала негативно повлияет на конечный продукт. Требуется высокая квалификация для разработки требуемых алгоритмов.
Угрозы: У1. Отсутствие спроса на программный продукт У2. Большая конкуренция на рынке У3. Санкции и ограничения в области информационных технологий У4. Вероятность появления более универсальных алгоритмов	Точность алгоритмов, отвечающих за процесс кластеризации, должна выступать ключевым конкурентным фактором.	Самой большой угрозой проекта выступает большая конкуренция на рынке, которая может сделать решение нерентабельным.

5.2 Определение возможных альтернатив проведения научных исследований

Для оценки теоретически возможных вариантов, вытекающих из закономерностей строения объекта исследования, был применен морфологический подход. Выделены важные морфологические характеристики объекта исследования и раскрыты возможные варианты по каждой характеристике. Данные объединены в морфологическую матрицу, представленную в таблице 5.6.

Таблица 5.6 - Морфологическая матрица для кластеризации поисковых запросов

	1	**2**	**3**
А. Подход к решению задачи	Вручную с применением инструментов морфологического анализа	По совпадению результатов поисковой выдачи	Определение семантически близких поисковых запросов
Б. Представление данных	Таблица, сформированная по ключевым словам	Граф с уплотняющейся к центру структурой	Плотное векторное пространство
В. Технология представления данных	Синтаксический анализатор	Движок поисковых систем	Нейронные сети
Г. Алгоритм кластеризации	Логическая кластеризация	Лувенский метод	Иерархическая кластеризация
Д. Получаемый результат	Смысловые группы	Пронумерованные группы	Вложенные друг в друга подмножества

Наиболее подходящими решениями обладает кластеризация по семантической схожести. Привлекательными решениями других вариантов выступают представление данных в виде графа и наличие смысловых групп в получаемом результате.

5.3 Планирование работ по научно-техническому исследованию

5.3.1 Структура работ в рамках научного исследования

Разработка производится группой работников, состоящей из двух человек – научного руководителя и инженера (студента), которые выделены в качестве исполнителей.

Перечень этапов и работ, распределение исполнителей по данным видам работ приведен в таблице 5.7.

Таблица 5.7 – Перечень этапов, работ и распределение исполнителей

Основные этапы	№ раб	Содержание работ	Должность исполнителя
Выбор направления исследования	1	Обсуждение предметной области исследования	Научный руководитель, Инженер
	2	Анализ предметной области	Инженер
	3	Утверждение темы выпускной квалификационной работы	Научный руководитель, Инженер
	4	Изучение материалов по теме исследования	Инженер
Практическая часть исследования	5	Определение набора данных для проведения работы	Инженер
	6	Определение методов обработки входных данных	Научный руководитель, Инженер
	7	Определение методов кластеризации данных	Научный руководитель, Инженер
	8	Разработка программного решения	Инженер
Оценка результатов	9	Проведение оценки полученного результата	Инженер
Оформление работы	10	Составление раздела «Социальная ответственность»	Инженер
	11	Составление раздела «Финансовый менеджмент»	Научный руководитель
	12	Составление отчета по выпускной квалификационной работе	Инженер

5.3.2 Определение трудоемкости выполнения работ

Трудовые затраты в большинстве случаях образуют основную часть стоимости разработки, поэтому важным моментом является определение трудоемкости работ каждого из участников научного исследования.

Трудоемкость выполнения научного исследования оценивается экспертным путем в человеко-днях и носит вероятностный характер, который зависит от множества трудно учитываемых факторов. Для определения ожидаемого (среднего) значения трудоемкости $t_{ожi}$ используется следующая формула:

$$t_{ожi} = \frac{3t_{\min i} + 2t_{\max i}}{5}, \quad (5.2)$$

где $t_{ожi}$ – ожидаемая трудоемкость выполнения i-ой работы чел.-дн.;

$t_{\min i}$ – минимально возможная трудоемкость выполнения заданной i-ой работы, чел.-дн.;

$t_{\max i}$ – максимально возможная трудоемкость выполнения заданной i-ой работы, чел.-дн.;

Исходя из ожидаемой трудоемкости работ, определяется продолжительность каждой работы в рабочих днях T_p , учитывающая параллельность выполнения работ по нескольким исполнителями.

$$T_{pi} = \frac{t_{ожi}}{Ч_i}, \quad (5.3)$$

где T_{pi} – продолжительность одной работы, раб.дн.;

$t_{ожi}$ – ожидаемая трудоемкость выполнения одной работы, чел.-дн.;

$Ч_i$ – численность исполнителей, выполняющих одновременно одну и ту же работу на данном этапе, чел.

5.3.3 Разработка графика проведения научного исследования

Наиболее удобным и наглядным представлением проведения научных работ является построение ленточного графика в форме диаграммы Ганта.

Диаграмма Ганта – горизонтальный ленточный график, на котором работы по теме представляются протяженными во времени отрезками, характеризующимися датами начала и окончания выполнения данных работ.

Для удобства построение графика, длительность каждого из этапов работ из рабочих дней следует перевести в календарные дни. Для этого необходимо воспользоваться следующей формулой:

$$T_{ki} = T_{pi} \cdot k_{кал}, \quad (5.4)$$

где T_{ki} – продолжительность выполнения i-й работы в календарных днях;

T_{pi} – продолжительность выполнения i-й работы в рабочих днях;

$k_{кал}$ – коэффициент календарности.

Коэффициент календарности определяется по следующей формуле:

$$k_{кал} = \frac{T_{кал}}{T_{кал} - (T_{вых} + T_{пр})}, \quad (5.5)$$

где $T_{кал}$ – количество календарных дней в году;

$T_{вых}$ – количество выходных дней в году;

$T_{пр}$ – количество праздничных дней в году.

Расчет коэффициента календарности:

$$k_{кал} = \frac{365}{365 - 118} = 1{,}48$$

Таблица 5.8 – Временные показатели проведения научного исследования

Название работы	Трудоёмкость работ									Исполнители	Длительность работ в рабочих днях T_{pi}			Длительность работ в календарных днях T_{ki}		
	t_{min}, чел–дни			t_{max}, чел–дни			$t_{ожi}$, чел– дни									
	Варианты исполнения										Варианты исполнения					
	1	2	3	1	2	3	1	2	3		1	2	3	1	2	3
Обсуждение предметной области исследования	1	1	1	2	2	2	1,4	1,4	1,4	Научный руководитель, Инженер	0,7	0,7	0,7	1	1	1
Анализ предметной области	1	1	2	2	3	3	1,4	1,8	2,4	Инженер	1,4	1,8	2,4	2,1	2,7	3,6
Утверждение темы выпускной квалификационной работы	1	1	1	1	1	1	1	1	1	Научный руководитель, Инженер	0,5	0,5	0,5	0,7	0,7	0,7
Изучение материалов по теме исследования	3	2	3	4	3	5	3,4	2,4	3,8	Инженер	3,4	2,4	3,8	5	3,6	5,6
Определение набора данных для проведения работы	1	1	1	3	3	3	1,8	1,8	1,8	Инженер	1,8	1,8	1,8	2,7	2,7	2,7
Определение методов обработки входных данных	6	8	9	8	10	12	6,8	8,8	10,2	Научный руководитель, Инженер	3,4	4,4	5,1	5	6,5	7,5
Определение методов кластеризации	5	4	6	7	7	9	5,8	5,2	7,2	Научный руководитель, Инженер	2,9	2,6	3,6	4,3	3,8	5,3

данных																
Разработка программного решения	6	9	7	8	12	9	6,8	10,2	7,8	Инженер	6,8	10,2	7,8	10,1	15,1	11,5
Проведение оценки полученного результата	2	2	3	3	2	3	2,4	2	3	Инженер	2,4	2	3	3,6	3	4,4
Составление раздела «Социальная ответственность»	2	2	2	3	3	3	2,4	2,4	2,4	Инженер	2,4	2,4	2,4	3,6	3,6	3,6
Составление раздела «Финансовый менеджмент»	3	3	3	4	4	4	3,4	3,4	3,4	Инженер	3,4	3,4	3,4	5	5	5
Составление отчета по выпускной квалификационной работе	2	2	3	3	3	4	2,4	2,4	3,4	Инженер	2,4	2,4	3,4	3,6	3,6	5

На основе таблицы 5.8 был разработан календарный план выполнения работ для научного исследования по кластеризации данных на основе семантической схожести (вариант исполнения 3). Календарный план представлен в виде диаграммы Ганта на рисунке 5.1. Занятость исполнителей представлена на рисунке 5.2.

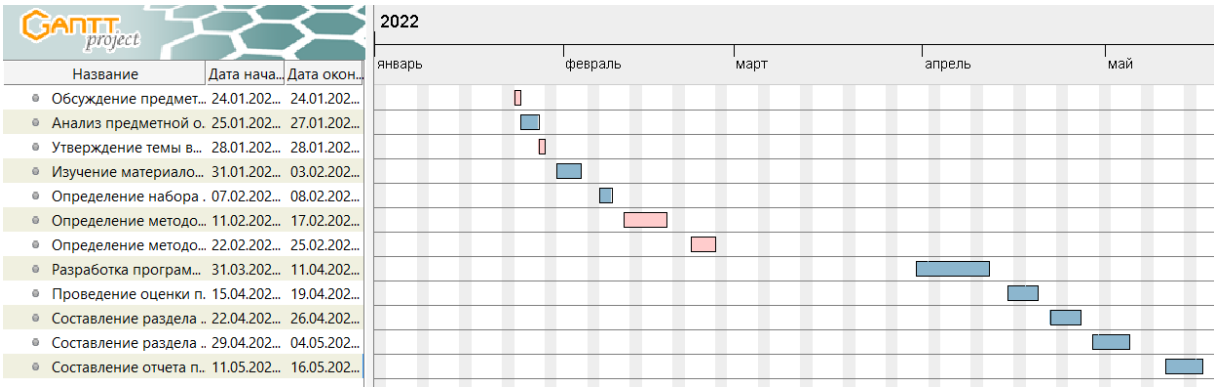

Рисунок 5.1 – Календарный план проведения работ

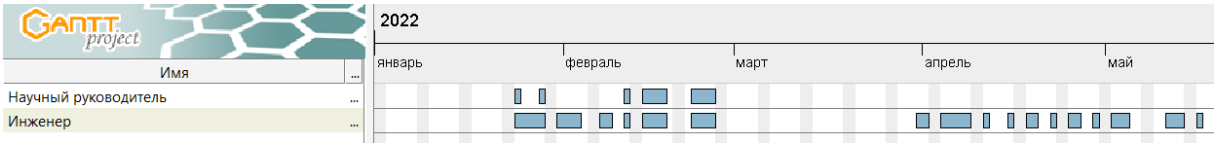

Рисунок 5.2 – Занятость исполнителей для работ

5.4 Бюджет научно-технического исследования (НТИ)

При планировании бюджета НТИ должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением. В процессе формирования бюджета НТИ используется следующая группировка затрат по статьям:

- материальные затраты НТИ;
- затраты на специальное оборудование для научных (экспериментальных) работ;
- основная заработная плата исполнителей темы;
- дополнительная заработная плата исполнителей темы;
- отчисления во внебюджетные фонды (страховые отчисления);
- затраты научные и производственные командировки;

- контрагентные расходы;
- накладные расходы.

5.4.1 Расчет материальных затрат НТИ

При планировании бюджета научно-техническое исследование должно быть обеспечено полное и достоверное отражение всех видов расходов, связанных с его выполнением.

Расчет материальных затрат осуществляется по формуле:

$$З_М = (1 + k_T) \cdot \sum_{i=1}^{m} Ц_i \cdot N_{расхi}\ , (5.6)$$

где m – количество видов материальных ресурсов, потребляемых при выполнении научного исследования;

$N_{расхi}$ – количество материальных ресурсов i-го вида, планируемых к использованию при выполнении научного исследования (шт., кг, м, м2 и т.д.);

$Ц_i$ – цена приобретения единицы i-го вида потребляемых материальных ресурсов (руб./шт., руб./кг, руб./м, руб./м2 и т.д.);

k_T – коэффициент, учитывающий транспортно-заготовительные расходы.

Научно-техническое исследование выполняется на персональных компьютерах и не требует дополнительных физических материалов и их доставки. Для расчета материальных затрат будет использоваться электроэнергия и интернет-соединение.

В среднем за один час энергопотребление компьютера составляет 0,1 кВт/час. За рабочий день в 8 часов с учетом освещения потребление электроэнергии составляет примерно 1кВт. Стоимость 1 кВт электроэнергии

в Томской области составляет 3,85 руб. Стоимость интернет-соединения за сутки составляет 14,5 руб.

Таблица 5.9– Материальные затраты

Наименование	Единица измерения	Количество			Цена за ед., руб.	Затраты на материалы, ($З_М$), руб.		
		Исполнение				Исполнение		
		1	2	3		1	2	3
Интернет-соединение	День	32	35	38	14,5	464	508	551
Электроэнергия	кВт	32	35	38	3,85	123	135	146
Итого, руб.						587	643	697

5.4.2 Расчет затрат на специальное оборудование для научных работ

В данную статью включают все затраты, связанные с приобретением специального оборудования (приборов, контрольно-измерительной аппаратуры, стендов, устройств и механизмов), необходимого для проведения работ по конкретной теме. Определение стоимости спецоборудования производится по действующим прейскурантам, а в ряде случаев по договорной цене. При приобретении спецоборудования необходимо учесть затраты по его доставке и монтажу в размере 15% от его цены. Расчет затрат по данной статье представлен в таблице 5.10.

Таблица 5.10 – Расчет бюджета затрат на приобретение спецоборудования для научных работ

Наименование	Единица измерения	Количество			Цена за ед., тыс. руб.	Затраты на оборудование, (Зм), тыс. руб.		
		Исполнение				Исполнение		
		1	2	3		1	2	3
Персональный компьютер	Шт.	1	1	1	50	50	50	50
IDE для разработки PyCharm	Шт.	1	1	1	3	3	3	3
Итого:						53	53	53

5.4.3 Основная заработная плата исполнителя темы

Статья включает основную заработную плату работников, непосредственно занятых выполнением проекта, (включая премии, доплаты) и дополнительную заработную плату и рассчитывается по формуле:

$$З_{зп} = З_{осн} + З_{доп}, \quad (5.7)$$

где $З_{осн}$ – основная заработная плата;

$З_{доп}$ – дополнительная заработная плата (12–20 % от $З_{осн}$).

Основная заработная плата руководителя рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot Т_{р}, \quad (5.8)$$

где $З_{осн}$– основная заработная плата одного работника;

$Т_{р}$ – продолжительность работ, выполняемых научно-техническим работником, раб. дн.;

$З_{дн}$ – среднедневная заработная плата работника, руб.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_м \cdot М}{F_д}, \quad (5.9)$$

где $З_м$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года:

при отпуске в 24 раб. дня М =11,2 месяца, 5 – дневная неделя;

при отпуске в 48 раб. дней М=10,4 месяца, 6 – дневная неделя;

$F_д$ – действительный годовой фонд рабочего времени научно–технического персонала, раб. Дн.

Таблица 5.11 – Баланс рабочего времени

Показатели рабочего времени	Научный руководитель	Инженер
Календарное число дней	365	365
Количество нерабочих дней - выходные дни - праздничные дни	118	118
Потери рабочего времени - отпуск - невыходы по болезни	 48 0	 72 0
Действительный годовой фонд рабочего времени	199	175

Месячный должностной оклад работника (руководителя):

$$З_м = З_{тс} \cdot (1 + k_{пр} + k_д) \cdot k_р, \quad (5.10)$$

где $З_{тс}$ – заработная плата по тарифной ставке, руб.;

$k_{пр}$ – премиальный коэффициент, равный 0,3 (т.е. 30 процентов от $З_{тс}$);

$k_д$ – коэффициент доплат и надбавок составляет примерно 0,2 – 0,5;

$k_р$ – районный коэффициент, равный 1,3 (для Томска).

На основании открытых вакансий Томского политехнического университета должностной оклад младшего научного сотрудника (инженера) составляет 11900 рублей в месяц. Должностной оклад научного руководителя – 30300 рублей в месяц. Расчет основной заработной платы для варианта исполнения 3 представлен в таблице 5.12.

Таблица 5.12 – Расчет основной заработной платы

Исполнители	$З_{тс}$, руб.	$k_{пр}$	$k_д$	$k_р$	$З_м$, руб.	$З_{дн}$, руб.	$Т_р$, раб. дн.			$З_{осн}$, руб.		
							Варианты исполнения					
							1	2	3	1	2	3
Научный руководитель	30300	0,3	0,4	1,3	66963	3769	8	9	12	30152	33921	45228
Инженер	11900	0,3	0,2	1,3	26299	1803	32	35	38	57696	63105	68514
										87848	97026	113742

5.4.4 Расчет дополнительной заработной платы

Дополнительная заработная плата учитывает величину предусмотренных Трудовым кодексом РФ доплат за отклонение от нормальных условий труда, а также выплат, связанных с обеспечением гарантий и компенсаций (при исполнении государственных и общественных обязанностей, при совмещении работы с обучением, при предоставлении ежегодного оплачиваемого отпуска и т.д.).

Расчет дополнительной заработной платы рассчитывается по формуле:

$$З_{доп} = k_{доп} \cdot З_{осн}\text{ , (5.11)}$$

где $k_{доп}$ – коэффициент дополнительной заработной платы, принятый на стадии проектирования за 0,15.

5.4.5 Отчисления во внебюджетные фонды

В данной статье расходов отражаются обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников.

Величина отчислений во внебюджетные фонды определяется исходя из формулы:

$$З_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}), \quad (5.12)$$

где $k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды (пенсионный фонд, фонд обязательного медицинского страхования и пр.).

Размер страховых взносов установлен равным 30%.

Отчисления во внебюджетные фонды представлены в таблице 5.13.

Таблица 5.13 – Отчисления во внебюджетные фонды

Исполнитель	Основная заработная плата, руб.			Дополнительная заработная плата, руб.		
	Варианты исполнения					
	1	2	3	1	2	3
Руководитель проекта	30152	33921	45228	4523	5088	6784
Студент	57696	63105	68514	8654	9466	10277
Коэффициент отчислений во внебюджетные фонды	0,3					
Итого						
Исполнение 1	30308					
Исполнение 2	33474					
Исполнение 3	39241					

5.4.6 Накладные расходы

Накладные расходы учитывают прочие затраты организации, не попавшие в предыдущие статьи расходов. Их величина определяется по формуле:

$$З_{накл} = (\sum статей) \cdot k_{нр}, \quad (5.13)$$

где $k_{нр}$ – коэффициент, учитывающий накладные расходы.

Величина коэффициента накладных расходов можно равна 16%. Значения накладных расходов приведены в таблице 5.14.

5.4.7 Формирование бюджета затрат научно-исследовательского проекта

Рассчитанная величина затрат научно–исследовательской работы является основой для формирования бюджета затрат проекта. Определение бюджета затрат на научно–исследовательский проект приведено в таблице 5.14.

Таблица 5.14 –Расчет бюджета затрат НТИ

Наименование статьи	Сумма, руб.			Примечание
	Исполнение			
	1	2	3	
1. Материальные затраты НТИ	587	643	697	Пункт 4.4.1
2. Затраты на специальное оборудование для научных (экспериментальных) работ	53000	53000	53000	Пункт 4.4.2
3. Затраты по основной заработной плате исполнителей темы	87848	97026	113742	Пункт 4.4.3
4. Затраты по дополнительной заработной плате исполнителей темы	13177	14554	17061	Пункт 4.4.4
5. Отчисления во внебюджетные фонды	30308	33474	39241	Пункт 4.4.5
6. Затраты на научные и производственные командировки	-	-	-	Отсутствуют
7. Контрагентские расходы	-	-	-	Отсутствуют
8. Накладные расходы	29587	31792	35799	Пункт 4.4.6
9. Бюджет затрат НТИ	214507	230479	259540	

5.5 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин: финансовой эффективности и ресурсоэффективности.

Интегральный показатель финансовой эффективности научного исследования определяется как:

$$I_{\text{фин.р}}^{\text{исп.}i} = \frac{\Phi_{pi}}{\Phi_{max}}, \quad (5.14)$$

$I_{\text{фин.р}}^{\text{исп.}i}$ – интегральный финансовый показатель разработки;

Φ_{pi} – стоимость i-го варианта исполнения;

Φ_{max} – максимальная стоимость исполнения научно-исследовательского проекта.

Значения интегральных показателей финансовой эффективности приведены в таблице 4.15.

Интегральный показатель ресурсоэффективности вариантов исполнения объекта исследования можно определить следующим образом:

$$I_{pi} = \sum_{i=1}^{n} a_i \times b_i, \quad (5.15)$$

где I_{pi} – интегральный показатель ресурсоэффективности для i-го варианта исполнения разработки;

a_i – весовой коэффициент i-го варианта исполнения разработки;

b_i^a, b_i^p – бальная оценка i-го варианта исполнения разработки, устанавливается экспертным путем по выбранной шкале оценивания;

n – число параметров сравнения.

Таблица 5.15 – Сравнительная оценка характеристик вариантов исполнения проекта

Объект исследования / Критерии	Весовой коэффициент параметра	Исп.1	Исп.2	Исп.3
1. Рост производительности труда пользователя	0,3	2	4	5
2. Удобство в эксплуатации	0,1	5	4	4
3. Точность результата кластеризации	0,25	4	3	5
4. Надежность программного решения	0,1	3	4	3
5. Использование сторонних программных разработок	0,25	3	2	4
Итого	1	3,15	3,25	4,45

Интегральный показатель эффективности вариантов исполнения разработки ($I_{испi}$) определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя.

Значения интегральных показателей эфективности вариантов исполнения разработки приведены в таблице 4.15.

Сравнение интегрального показателя эффективности вариантов исполнения разработки позволит определить сравнительную эффективность проекта и выбрать наиболее целесообразный вариант из предложенных.

Сравнительная эффективность проекта ($Э_{ср}$):

$$Э_{ср} = \frac{I_{исп2}}{I_{исп1}}, \quad (5.16)$$

Оценки сравнительной эффективности проектов приведены в таблице 5.16.

Таблица 5.16 – Сравнительная эффективность разработки

№	Показатели	Исп.1	Исп.2	Исп.3
1	Интегральный финансовый показатель разработки	0,83	0,89	1
2	Интегральный показатель ресурсоэффективности разработки	3,15	3,25	4,45
3	Интегральный показатель эффективности	3,8	3,65	4,45
4	Сравнительная эффективность вариантов исполнения	0,85	0,82	1

Исходя из сравнительной эффективности вариантов исполнения, можно сделать вывод, что реализация технологии в третьем исполнении является более эффективным вариантом решения задачи, поставленной в данной работе с позиции финансовой и ресурсной эффективности.

Вывод по разделу

В ходе написания раздела были определены потенциальные потребители решения, проведен анализ конкурентных технических решений с помощью оценочной карты. Составлен SWOT-анализ с выявлением сильных и слабых сторон проекта, а также возможностей и угроз.

Были определены возможные альтернативы проведения научных исследований, для которых была определена трудоемкость, исходя из составленного перечня этапов и работ. Выполнено планирование работ с помощью диаграммы Ганта.

Был определен бюджет проекта, а именно были рассчитаны материальные затраты, затраты на специальное оборудование, затраты на основную и дополнительную заработные платы исполнителей, затраты на отчисления во внебюджетные фонды и накладные расходы.

Были определены показатели эффективности исследования, основываясь на которых установлена наилучшая эффективность для проведения научно-технического исследования в рамках выпускной квалификационной работы.

6 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

Введение

Целью выпускной квалификационной работы является разработка системы кластеризации поисковых запросов семантического ядра.

Областью применения разрабатываемого решения является поисковая оптимизация. Она используется для поднятия позиций сайта в результатах выдачи поисковых систем для увеличения его посещаемости и привлечения новых клиентов.

Система направлена на сокращение времени анализа важных для сайта поисковых запросов. Для этого используется процесс кластеризации, позволяющий объединить близкие по смыслу поисковые запросы в удобные для работы группы.

В качестве потенциальных пользователей выступают интернет-маркетологи.

Рабочими процессами, связанными с объектом исследования, осуществляющимися на рабочем месте, выступают:

- изучение области поисковой оптимизации;
- определение набора данных для проведения работы;
- выбор методов обработки входных данных;
- определение методов кластеризации данных;
- разработка программного решения;
- проведение оценки полученного результата.

Разработка системы выполнялась на одном рабочем месте – в офисном помещении площадью 20 м² (4*5 м) с использованием следующего оборудования рабочей зоны: персональный компьютер, рабочий стол, компьютерное кресло.

6.1 Правовые и организационные вопросы обеспечения безопасности

6.1.1 Правовые нормы трудового законодательства

Трудовые отношения между работодателем и работником регулируются с помощью «Трудового кодекса Российской Федерации» от 30 декабря 2001 г. № 197-ФЗ (редакция, действующая с 1 марта 2022 года) [12].

Основные положения, применимые для разработки текущего проектного решения:

1. Продолжительность рабочего времени (ТК РФ Статья 91)

Нормальная продолжительность рабочего времени не может превышать 40 часов в неделю.

2. Перерывы для отдыха и питания (ТК РФ Статья 108)

В течение рабочего дня (смены) работнику должен быть предоставлен перерыв для отдыха и питания продолжительностью не более двух часов и не менее 30 минут, который в рабочее время не включается. Правилами внутреннего трудового распорядка или трудовым договором может быть предусмотрено, что указанный перерыв может не предоставляться работнику, если установленная для него продолжительность ежедневной работы (смены) не превышает четырех часов.

3. Выходные дни (ТК РФ Статья 111)

Всем работникам предоставляются выходные дни (еженедельный непрерывный отдых). При пятидневной рабочей неделе работникам предоставляются два выходных дня в неделю, при шестидневной рабочей неделе - один выходной день.

4. Обработка персональных данных работника (ТК РФ Статья 86)

Обработка персональных данных работника может осуществляться исключительно в целях обеспечения соблюдения законов и иных нормативных правовых актов, получении образования, контроля количества и качества выполняемой работы.

5. Заработная плата (ТК РФ Статья 135)

Заработная плата работнику устанавливается трудовым договором в соответствии с действующими у данного работодателя системами оплаты труда.

6.1.2 Основные эргономические требования к правильному расположению и компоновке рабочей зоны

Исходя из того, что разработка системы выполнялась на одном рабочем месте, и в течение длительного времени человек проводил время за персональным компьютером, должны проводиться соответствующие мероприятия по организации рабочего места в соответствии с требованиями ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования [13].

Конструкцией производственного оборудования и рабочего места должно быть обеспечено оптимальное положение работающего, которое достигается регулированием высоты рабочей поверхности, сиденья и пространства для ног.

Допускается проектировать и изготовлять оборудование с нерегулируемыми параметрами рабочего места. Высоту рабочей поверхности и сидения определяют согласно таблицы 6.1.

Таблица 6.1 – Числовые значения параметров высоты рабочей поверхности и сидения при организации рабочего места за ПЭВМ

Пол работающего	Высота рабочей поверхности, мм	Высота сиденья, мм
Женщины	630	400
Мужчины и женщины	655	420
Мужчины	680	430

Очень часто используемые средства отображения информации, требующие точного и быстрого считывания показаний, следует располагать в вертикальной плоскости под углом ±15° от нормальной линии взгляда и в горизонтальной плоскости под углом ±15° от сагиттальной плоскости.

6.2 Производственная безопасность при разработке проектного решения

Согласно ГОСТ 12.0.003-2015 Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация [14], неблагоприятные производственные факторы по результирующему воздействию на организм работающего человека подразделяют:

- на вредные производственные факторы, то есть факторы, приводящие к заболеванию, а также усугубляющие уже имеющиеся заболевания;
- опасные производственные факторы, то есть факторы, приводящие к травме, в том числе смертельной.

Производственные факторы, возникающие при разработке проектного решения, представлены в таблице 6.2.

Таблица 6.2 – Возможные опасные и вредные производственные факторы на рабочем месте за ПЭВМ

Факторы (ГОСТ 12.0.003- 2015)	Нормативные документы
Производственные факторы, связанные с аномальными микроклиматическими параметрами на местонахождении работающего	СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания [15]
Отсутствие или недостаток необходимого освещения	СП 52.13330.2016 Естественное и искусственное освещение [16]
Нервно-психические перегрузки (умственное перенапряжение, перенапряжение анализаторов)	Трудовой кодекс Российской Федерации от 30 декабря 2001 г. № 197-ФЗ (редакция, действующая с 1 марта 2022 года) [12]
Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которых попадает работающий	ГОСТ 12.1.019-2017 Система стандартов безопасности труда. Электробезопасность. Общие требования и номенклатура видов защиты [17]

6.2.1 Производственные факторы, связанные с аномальными микроклиматическими параметрами на местонахождении работающего

Использование персональных компьютеров может привести к нарушению параметров микроклимата: повышению температуры и снижению относительной влажности в рабочем помещении. Это может привести к снижению концентрации и работоспособности, быстрой утомляемости, а также к развитию заболеваний органов дыхания и сердечно-сосудистой системы. Для предотвращения негативных последствий следует соблюдение оптимальных показателей микроклимата в помещении.

Нормативные показатели микроклимата регламентируются СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и/или безвредности для человека факторов среды обитания [15].

Разработка проектного решения за компьютером относится к категории Iа по уровню энерготрат организма. Допустимые величины

параметров микроклимата на рабочих местах для данной категории работ представлены в таблице 6.3.

Таблица 6.3 – Допустимые величины параметров микроклимата на рабочих местах для категории работ Iа

Период года	Температура воздуха, °С		Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с	
	Диапазоны				Для диапазонов температур	
	Ниже оптимального значения	Выше оптимального значения			Ниже оптимального значения	Выше оптимального значения
Холодный	20,0-21,9	24,1-25,0	19,0-26,0	15-75	0,1	0,1
Теплый	21,0-22,9	25,1-28,0	20,0-29,0	15-75	0,1	0,2

Основными мерами поддержания подходящего микроклимата рабочего места выступают:

- обеспечение надлежащего проветривания помещения;
- обеспечение требуемого отопления в холодное время года;
- проведение регулярной влажной уборки помещения;
- использование увлажнителей воздуха в помещении.

6.2.2 Отсутствие или недостаток необходимого освещения

Отсутствие или недостаток необходимого освещения является вредным производственным фактором, который может вызвать утомление и снижение работоспособности, а также вызывать усталость глаз, которая может привести к ухудшению зрения.

Причиной недостаточной освещенности рабочего места могут выступать малая мощность осветительных приборов или неправильная организация рабочего места.

Требования к освещению рабочей зоны при разработке проектного решения, характеризующиеся СП 52.13330.2016 Естественное и искусственное освещение [16], представлены ниже:

Искусственное освещение: освещенность 400 лк при системе комбинированного освещения и 200 лк при системе общего освещения; объединенный показатель дискомфорта UGR не более 25; коэффициент пульсации $K_п$ не более 15 %;

Совместное освещение: КЕО $e_н$ при верхнем освещении или комбинированном освещении 3,0 %, при боковом освещении – 1,2 %.

Мероприятиями по обеспечению необходимого уровня освещенности выступают:

- настройка яркости и контрастности монитора компьютера близким к уровню естественного освещения;
- установка дополнительных средств освещения при недостаточной освещенности рабочего места

6.2.3 Нервно-психические перегрузки (умственное перенапряжение, перенапряжение анализаторов)

В следствие интенсивной умственной работы за компьютером возможно возникновение ряда нервно-психических перегрузок, таких как умственное перенапряжение и перенапряжение анализаторов. В связи с этим возможны значительные изменения кровяного давления и пульса, ведущие к сердечно-сосудистым заболеваниям.

Для снижения вероятности возникновения нервно-психических перегрузок требуется соблюдение режима рабочего времени согласно Трудовому кодексу Российской Федерации от 30 декабря 2001 г. № 197-ФЗ (редакция, действующая с 1 марта 2022 года) [12]:

1. Нормальная продолжительность рабочего времени не может превышать 40 часов в неделю.

2. В течение рабочего дня (смены) работнику должен быть предоставлен перерыв для отдыха и питания продолжительностью не более двух часов и не менее 30 минут, который в рабочее время не включается.

3. Всем работникам предоставляются выходные дни (еженедельный непрерывный отдых). При пятидневной рабочей неделе работникам предоставляются два выходных дня в неделю, при шестидневной рабочей неделе - один выходной день.

6.2.4 Производственные факторы, связанные с электрическим током, вызываемым разницей электрических потенциалов, под действие которых попадает работающий

В процессе эксплуатации персонального компьютера существует опасность поражения электрическим током, так как работающий может коснуться комплектующих компьютера, находящихся под напряжением. Поражение электрическим током при работе с компьютером может привести к появлению ожогов, механическим повреждениям тканей и сосудов.

Нормы защиты от поражения электрическим током регламентируются ГОСТ 12.1.019-2017 Система стандартов безопасности труда. Электробезопасность. Общие требования и номенклатура видов защиты [17].

Согласно стандарту, проводящие части, находящиеся под опасным рабочим, наведенным, остаточным напряжением, не должны быть доступными, а доступные проводящие части не должны находиться под опасным напряжением при нормальных условиях (при отсутствии повреждения), а также в случае единичного повреждения.

Защиту при нормальных условиях (защиту от прямого прикосновения) обеспечивают посредством основной защиты, а защиту при условиях единичного повреждения (защиту при косвенном прикосновении) обеспечивают посредством защиты при повреждении.

Усиленные защитные меры предосторожности обеспечивают защиту от прямого прикосновения и защиту при повреждении.

6.3 Экологическая безопасность при разработке проектного решения

Среда разработки проектного решения относится к пятому классу промышленных объектов согласно СанПиН 2.2.1./2.1.1.1200-03 Санитарно-защитные зоны и санитарная классификация предприятий, сооружений и иных объектов [18]. Для работ требуется компьютер, который имеет воздействие на литосферу при его утилизации.

Мероприятием по снижению воздействия разработки проектного решения на окружающую среду выступает утилизация вышедшего из употребления технологического оборудования с помощью специализированных на этом компаний. Перед утилизацией выполняется разбор оборудования, отделение деталей, подходящих для переработки и вторичного использования.

6.4 Безопасность в чрезвычайных ситуациях при разработке проектного решения

Чрезвычайная ситуация – это состояние, при котором в результате возникновения источника ЧС на объекте, определенной территории или акватории нарушаются нормальные условия жизни и деятельности людей, возникает угроза их жизни и здоровью, наносится ущерб имуществу населения, народному хозяйству и природной среде.

При разработке проектного решения возможно возникновение ряда чрезвычайных ситуаций техногенного характера: обрушение здания, аварии на коммунальных системах жизнеобеспечения населения, пожар.

Наиболее типичной чрезвычайной ситуацией при работе за компьютером выступает пожар, возникающий в случае замыкания

электропроводки оборудования, возникновения неисправности электросетей, не соблюдения мер пожаробезопасности. Был определен следующий класс возможного пожара при работе за ПЭВМ: пожары горючих веществ и материалов электроустановок, находящихся под напряжением (Е), согласно Федеральному закону от 22.07.2008 N 123-ФЗ (ред. от 30.04.2021) "Технический регламент о требованиях пожарной безопасности" [19].

Согласно ГОСТ 12.1.004-91 Система стандартов безопасности труда (ССБТ). Пожарная безопасность. Общие требования [20], противопожарная защита должна достигаться:

- наличием доступа к средствам пожаротушения;
- применением установок пожарной сигнализации и пожаротушения;
- устройствами, ограничивающими распространения пожара;
- организацией с помощью технических средств, включая автоматические, своевременного оповещения и эвакуации людей;
- применением средств коллективной и индивидуальной защиты людей от опасных факторов пожара;
- применением средств противодымной защиты.

Мероприятия для обеспечения пожарной безопасности:

- организация обучения работающих правилам пожарной безопасности;
- реализация правил пожарной безопасности;
- разработка мероприятий по действиям рабочих на случай возникновения пожара и организацию эвакуации людей.

При возникновении пожара или признаков горения необходимо:

1. Прекратить работу и обратиться в противопожарную службу по номеру 01, назвав адрес происшествия, место пожара и свою фамилию.

2. По возможности принять меры по эвакуации людей в соответствии с планом эвакуации здания.

3. К тушению пожара следует приступать только в случае, если нет угрозы для жизни и здоровья и существует возможность. В противном случае необходимо покинуть опасную зону, плотно прикрыв за собой двери.

Вывод по разделу

В ходе написания раздела были рассмотрены правовые и организационные вопросы обеспечения безопасности при разработке проектного решения, а именно изучены правовые нормы трудового законодательства и определены организационные мероприятия при компоновке рабочей зоны.

Был рассмотрен вопрос производственной безопасности при разработке проектного решения: выявлены вредные и опасные факторы, определены источники их возникновения, негативные последствия, нормы и мероприятия направленные на снижения их влияния. На рабочем месте соблюдались все установленные нормы.

При разработке проектного решения отсутствовали нарушения экологической безопасности.

Были определены возможные чрезвычайные ситуации при разработке проектного решения и описаны средства и мероприятия для обеспечения пожаробезопасности и действия в случае его возникновения, как наиболее типичного варианта чрезвычайной ситуации.

В отношении опасности поражения людей электрическим током помещение, в котором проводились работы, относится к первой категории – помещения без повышенной опасности, в которых отсутствуют условия, создающие повышенную или особую опасность.

Исполнитель работ относится к группе I персонала по электробезопасности.

Работы относятся к категории Iа по уровню энерготрат организма.

Рабочее помещение по взрывопожарной и пожарной опасности относится к категории В2 (пожароопасность).

Рабочее место относится к объектам IV категории, оказывающим минимальное негативное воздействие на окружающую среду.

ЗАКЛЮЧЕНИЕ

В результате выполнения выпускной квалификационной работы была разработана информационная система кластеризации поисковых запросов семантического ядра.

Для обработки поисковых запросов была выбрана предварительно обученная модель all-MiniLM-L6-v2 библиотеки SBERT. Составлены данные и проведено дополнительное обучение для решения задачи кластеризации семантического ядра.

Выбран иерархический алгоритм кластеризации. Проведена оценка результатов кластеризации, на основе которых были утверждены наиболее подходящая модель и алгоритм кластеризации.

Согласно поставленным техническим требованиям разработана и реализована клиентская и серверная часть информационной системы.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. SentenceTransformers Documentation [Электронный ресурс]. – Режим доступа: https://www.sbert.net/. – Дата доступа: 20.05.2022.

2. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks [Электронный ресурс]. – Режим доступа: https://arxiv.org/pdf/1908.10084.pdf. – Дата доступа: 20.05.2022.

3. Hugging Face. Sentence Transformers. [Электронный ресурс]. – Режим доступа: https://huggingface.co/sentence-transformers. – Дата доступа: 20.05.2022.

4. Sentence-transformers. Datasets. [Электронный ресурс]. – Режим доступа: https://public.ukp.informatik.tu-darmstadt.de/reimers/sentence-transformers/datasets/. – Дата доступа: 20.05.2022.

5. Semantic Search with S-BERT is all you need. [Электронный ресурс]. – Режим доступа: https://medium.com/mlearning-ai/semantic-search-with-s-bert-is-all-you-need-951bc710e160. – Дата доступа: 20.05.2022.

6. Clustering — scikit-learn 1.1.1 documentation. [Электронный ресурс]. – Режим доступа: https://scikit-learn.org/stable/modules/clustering.html. – Дата доступа: 20.05.2022.

7. Университет ИТМО. Оценка качества в задаче кластеризации. [Электронный ресурс]. – Режим доступа: https://neerc.ifmo.ru/wiki/index.php?title=Оценка_качества_в_задаче кластеризации. – Дата доступа: 20.05.2022.

8. Комплекс стандартов на автоматизированные системы [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/1200006924. – Дата доступа: 20.05.2022.

9. Django documentation. [Электронный ресурс]. – Режим доступа: https://docs.djangoproject.com/en/4.0/. – Дата доступа: 20.05.2022.

10. Organic Channel Share Expands to 53.3% of Traffic [Электронный ресурс]. – Режим доступа: https://www.brightedge.com/resources/research-reports/content-optimization. – Дата доступа: 20.05.2022.

11. Sentence Transformer Fine-Tuning (SetFit): Outperforming GPT-3 on few-shot Text-Classification while being 1600 times smaller [Электронный ресурс]. – Режим доступа: https://towardsdatascience.com/sentence-transformer-fine-tuning-setfit-outperforms-gpt-3-on-few-shot-text-classification-while-d9a3788f0b4e. – Дата доступа: 20.05.2022.

12. Трудовой кодекс Российской Федерации» от 30 декабря 2001 г. № 197-ФЗ (редакция, действующая с 1 марта 2022 года) [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/901807664. – Дата доступа: 20.05.2022.

13. ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/1200003913. – Дата доступа: 20.05.2022.

14. ГОСТ 12.0.003-2015 Система стандартов безопасности труда (ССБТ). Опасные и вредные производственные факторы. Классификация [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/1200136071. – Дата доступа: 20.05.2022.

15. СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/573500115. – Дата доступа: 20.05.2022.

16. СП 52.13330.2016 Естественное и искусственное освещение [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/456054197. – Дата доступа: 20.05.2022.

17. ГОСТ 12.1.019-2017 Система стандартов безопасности труда. Электробезопасность. Общие требования и номенклатура видов защиты [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/1200161238. – Дата доступа: 20.05.2022.

18. СанПиН 2.2.1./2.1.1.1200-03 Санитарно-защитные зоны и санитарная классификация предприятий, сооружений и иных объектов [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/902065388. – Дата доступа: 20.05.2022.

19. Федеральный закон от 22.07.2008 N 123-ФЗ (ред. от 30.04.2021) Технический регламент о требованиях пожарной безопасности [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/902111644. – Дата доступа: 20.05.2022.

20. Согласно ГОСТ 12.1.004-91 Система стандартов безопасности труда (ССБТ). Пожарная безопасность. Общие требования [Электронный ресурс]. – Режим доступа: https://docs.cntd.ru/document/9051953. – Дата доступа: 20.05.2022.