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: 09.03.04 «Программная инженерия»
Отделение школы (НОЦ): Отделение информационных технологий

БАКАЛАВРСКАЯ РАБОТА

Тема работы
Алгоритмическое и программное обеспечение кластеризации научных текстов по тематикам

УДК: 004.421:004.415.2:001.891.3

Студент

Группа	ФИО	Подпись	Дата
8К71	Коцин Денис Олегович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Савельев Алексей Олегович	к.т.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Маланина Вероника Анатольевна	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Черемискина Мария Сергеевна	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Чердынцев Евгений Сергеевич	к.т.н.		

Томск – 2021 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код результата	**Результат обучения (выпускник должен быть готов)**
Р1	Применять базовые и специальные естественнонаучные и математические знания в области информатики и вычислительной техники, достаточные для комплексной инженерной деятельности.
Р2	Применять базовые и специальные знания в области современных информационных технологий для решения инженерных задач.
Р3	Ставить и решать задачи комплексного анализа, связанные с созданием аппаратно-программных средств информационных и автоматизированных систем, с использованием базовых и специальных знаний, современных аналитических методов и моделей.
Р4	Разрабатывать программные и аппаратные средства (системы, устройства, блоки, программы, базы данных и т. п.) в соответствии с техническим заданием и с использованием средств автоматизации проектирования.
Р5	Проводить теоретические и экспериментальные исследования, включающие поиск и изучение необходимой научно-технической информации, математическое моделирование, проведение эксперимента, анализ и интерпретация полученных данных, в области создания аппаратных и программных средств информационных и автоматизированных систем.
Р6	Внедрять, эксплуатировать и обслуживать современные программно-аппаратные комплексы, обеспечивать их высокую эффективность, соблюдать правила охраны здоровья, безопасность труда, выполнять требования по защите окружающей среды.
Р7	Использовать базовые и специальные знания в области проектного менеджмента для ведения комплексной инженерной деятельности.
Р8	Владеть иностранным языком на уровне, позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты комплексной инженерной деятельности.
Р9	Эффективно работать индивидуально и в качестве члена группы, состоящей из специалистов различных направлений и квалификаций, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре организации.
Р10	Демонстрировать знания правовых, социальных, экономических и культурных аспектов комплексной инженерной деятельности.
Р11	Демонстрировать способность к самостоятельному обучению в течение всей жизни и непрерывному самосовершенствованию в инженерной профессии.

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 (специальность): 09.03.04 «Программная инженерия»
Отделение школы (НОЦ): Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Чердынцев Е.С.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Бакалаврской работы

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

Студенту:

Группа	ФИО
8К61	Коцину Денису Олеговичу

Тема работы:

Алгоритмическое и программное обеспечение кластеризации научных текстов по тематикам	
Утверждена приказом директора (дата, номер)	№32-2/с от 01.02.2021 г.

Срок сдачи студентом выполненной работы:	16.06.2021 г.

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.).*	Работа направлена на реализацию методов обработки естественного языка с целью кластеризации научных текстов по тематикам. Результатов является программное обеспечение оценки сходства корпусов научных текстов для целей информационной поддержки процессов организации научных исследований.
Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Обзор предметной области 2. Проектирование 3. Разработка системы поиска данных о публикациях и кластеризации 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение 5. Социальная ответственность

Перечень графического материала *(с точным указанием обязательных чертежей)*	1. Диаграмма в нотации IDEF0 2. Диаграмма в нотации EPC 3. Диаграмма в нотации DFD 4. Диаграмма логической схемы базы данных 5. Рисунки, демонстрирующие результаты

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Маланина В.А.
Социальная ответственность	Черемискина М.С.

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	01.03.2021 г.

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Савельев Алексей Олегович	к.т.н		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8К71	Коцин Денис Олегович		

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 (специальность): 09.03.04 «Программная инженерия»
Уровень образования: Бакалавр
Отделение школы (НОЦ): Отделение информационных технологий
Период выполнения: осенний / весенний семестр 2020 / 2021 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:	12.06.2020 г.

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
6.05.2021	Предметная область и бизнес-требования	15
13.05.2021	Проектирование системы	25
28.05.2021	Реализация системы	25
1.06.2021	Результат практического исследования	15
3.05.2021	Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	10
5.05.2021	Социальная ответственность	10

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Савельев Алексей Олегович	к.т.н.		

СОГЛАСОВАНО:
Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Чердынцев Евгений Сергеевич	к.т.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8К71	Коцину Денису Олеговичу

Школа	Инженерная школа информационных технологий и робототехники	Отделение школы (НОЦ)	ОИТ
Уровень образования	Бакалавриат	**Направление/специальность**	09.03.04 Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*Амортизационные затраты на оборудование – 9 212 руб.* *Затраты на основную заработную плату – 110451 руб.* *Затраты на дополнительную заработную плату – 16567 руб.* *Затраты на отчисления во внебюджетные фонды – 38 360 руб.* *Накладные расходы – 28319 руб.* *Бюджет затрат НИ – 202909 руб.*
2. *Нормы и нормативы расходования ресурсов*	*Бюджет проекта не более 250000 руб.,* *в том числе затраты на оплату труда не более 150000 руб.* *Значение показателя интегральной ресурсоэффективности - не менее 3 баллов из 5.*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Районный коэффициент – 1,3* *Коэффициент дополнительной заработной платы – 0,15* *Коэффициент отчислений во внебюджетные фонды – 0,302* *Коэффициент накладных расходов – 0,16*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	

1. *Оценка коммерческого потенциала, перспективности и альтернатив проведения НИ с позиции ресурсоэффективности и ресурсосбережения*	*1.Описание потенциальных потребителей продукта* *2. QuaD-анализ* *3.SWOT-анализ*
2. *Планирование и формирование бюджета научных исследований*	*1.Описание структуры работ в рамках научного исследования.* *2. Определение трудоемкости выполнения работ и разработка графика проведения научного исследования.* *3. Подсчет бюджета проекта*
3. *Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальной и экономической эффективности исследования*	*Оценка интегрального показателя эффективности разработки*
Перечень графического материала *(с точным указанием обязательных чертежей)***:**	
1. *Оценка конкурентоспособности технических решений* 2. *Матрица SWOT* 3. *График проведения и бюджет НИ* 4. *Оценка ресурсной, финансовой и экономической эффективности НИ*	

Дата выдачи задания для раздела по линейному графику	01.03.2021

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН ШБИП ТПУ	Маланина Вероника Анатольевна	к.э.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8К71	Коцин Денис Олегович		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА

«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8К71	Коцину Денису Олеговичу

Школа	Инженерная школа информационных технологий и робототехники	Отделение (НОЦ)	ОИТ
Уровень образования	Бакалавриат	Направление/специальность	09.03.04 Программная инженерия

Тема ВКР:

Алгоритмическое и программное обеспечение кластеризации научных текстов по тематикам	
Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	− Объект исследования – система кластеризации и визуализации кластеров научных статей; − Область применения – анализ научных текстов; − Рабочее место – рабочий стол с персональным компьютером в аудитории Кибернетического центра ТПУ;
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности: – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	− Рабочее место при выполнении работ сидя регулируется ГОСТом 12.2.032-78; − СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания; − Трудовой кодекс РФ;
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов	− Отклонение показателей микроклимата;

2.2. Обоснование мероприятий по снижению воздействия	− Недостаточная освещенность рабочей зоны; − Отсутствие или недостаток естественного света; − Повышенный уровень шума на рабочем месте; − Повышенный уровень электромагнитных излучений;
3. Экологическая безопасность:	Анализ воздействия на литосферу: − Утилизация компьютеров, смартфонов, оргтехники и бумаги; Анализ воздействия на гидросферу: − Утилизация компьютеров, смартфонов, оргтехники и бумаги; − Перерасход воды Анализ воздействия на атмосферу: − Перерасход электричества, производимого угольными электростанциями
4. Безопасность в чрезвычайных ситуациях:	Возможные чрезвычайные ситуации: − Пожар.

Дата выдачи задания для раздела по линейному графику	01.03.2021

Задание выдал консультант:

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Ассистент	Черемискина Мария Сергеевна	-		

Задание принял к исполнению студент:

Группа	**ФИО**	**Подпись**	**Дата**
8К71	Коцин Денис Олегович		

Реферат

Выпускная квалификационная работа содержит 67 страниц, 11 рисунков, 21 таблицу, 3 приложения и 9 литературных источников.

Ключевые слова: кластеризация, кластерный анализ, алгоритмы кластеризации, парсинг, веб-приложение.

Целью работы является разработка веб-приложения для сбора информации и её последующего хранения, анализа и кластеризации на основе названия и аннотации. Объектом исследования является веб-приложение для кластеризации данных, полученных с помощью широко используемой поисковой системы Google Scholar.

В первом разделе представлен обзор предметной области и основных алгоритмов, используемых в приложении.

Во втором разделе изложено описание процесса проектирования веб-приложения для кластеризации научных публикаций.

Третий раздел содержит информацию о непосредственной реализации разрабатываемого программного обеспечения, а также отображены результаты разработки.

Четвертый раздел представляет собой выполненное задание по разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение». Определены основные расходы на процесс разработки, а также оценен потенциал проделанной работы.

В пятом разделе содержится выполненное задание по разделу «Социальная ответственность». В данном разделе были определены требования к процессу разработки согласно стандартам ГОСТ и СанПиН, кроме того, были определены факторы, вредные для окружающей среды.

Определения, обозначения, сокращения, нормативные ссылки

1. Веб-приложение – это любая компьютерная программа, которая выполняет определенную функцию, используя в качестве клиента веб-браузер.
2. БД – это хранилище для большого количества систематизированных данных, с которыми можно производить определённые действия.
3. Методология IDEF0 – метод описания и формализации бизнес-процессов.
4. Событийная цепочка процессов (EPC) – тип блок-схемы, используемой для бизнес-моделирования.
5. Среда выполнения – вычислительное окружение, необходимое для выполнения компьютерной программы и доступное во время выполнения компьютерной программы.
6. Фреймворк – программное обеспечение, облегчающее разработку и объединение разных модулей программного проекта.
7. Python – высокоуровневый язык программирования общего назначения, ориентированный на повышение производительности разработчика и читаемости кода
8. Django — это фреймворк для создания веб-приложений с помощью языка программирования Python.
9. SQL – (Structured Query Language) – язык, используемый для работы с базами данных.
10. СУБД – специализированная программа, предназначенная для организации и ведения базы данных.

Оглавление

Введение

Любое научное исследование начинается с разработки методологии и поиска актуальной информации в интернете. Следовательно, для исследователя важно в короткое время найти связанные с задачей публикации, при этом исключив лишнюю информацию. С связи с развитием глобальной сети Интернет, количество публикуемой научной информации растёт каждый день в огромных масштабах. Только по запросу “Tomsk Polytechnic University” можно найти практически 40000 статей, из которых 4920 были написаны за 2020 год, в то время как в области биомедицины ежедневно публикуется около 1800 новых статей. Всего же такой сервис, как Google Scholar, покрывает порядка 389 миллионов статей [1].

Без предварительной классификации статей поиск документов человеком может опираться только на проверку вхождения ключевых слов из запроса в текст статьи, что сильно затрудняет процесс нахождения актуальной для конечного пользователя информации. Ручной процесс классификации также неэффективен с точки зрения времени, затрачиваемого на обработку. В решении данной задачи может помочь такое направление искусственного интеллекта, как обработка естественного языка.

Одним из вариантов решения данной проблемы информационной перегрузки является автоматическая кластеризация документов. Благодаря ней пользователи могут разбивать данные на тематические и подтематические категории, с легкостью отбрасывая малорелевантные группы. Таким образом, пользователь сможет сузить границы поиска информации и собрать необходимые данные в наиболее короткий промежуток времени.

Глава 1. Обзор предметной области:

1.1 Постановка задачи:

С помощью алгоритмического и программного обеспечения кластеризации научных текстов по тематикам разбить массив научных публикаций на несколько кластеров, содержимое которых схоже по тематике. Благодаря этому можно сократить зону поиска информации и ускорить процесс сбора информации при организации научных исследований.

На диаграмме IDEF0 изображена модель, изображающая структуру системы (Рисунок 1):

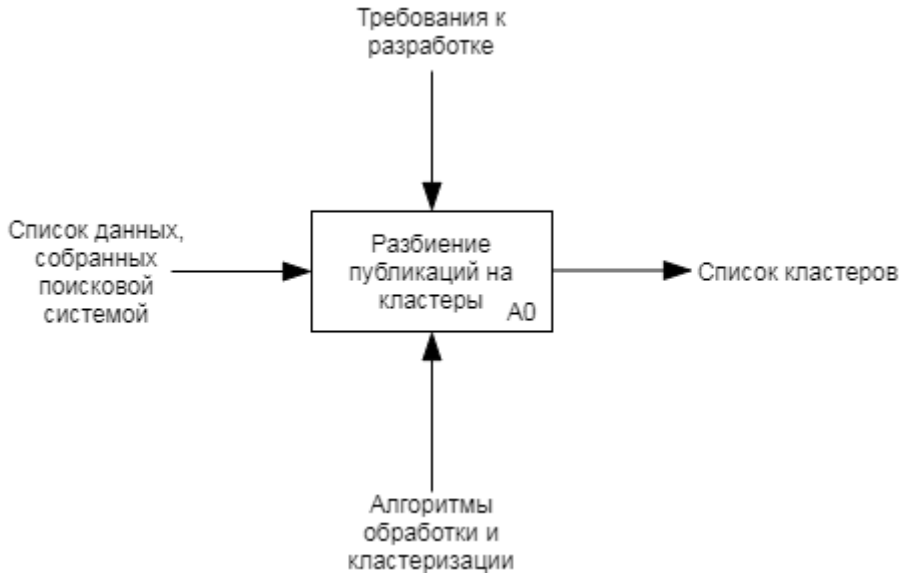

Рисунок 1 – Модель работы программы

1.2 Описание предметной области исследования:

Объектом исследования является веб-сайт для кластеризации данных, полученных с помощью широко используемой поисковой системы Google Scholar.

Данная система позволяет проводить поиск научных публикаций из разных сфер науки по полным текстам, а также проводить индексацию по различным показателям. На данный момент она является самым популярным сервисом для поиска публикаций. В список публикаций, которые содержит этот сервис, входят статьи, диссертации,

книги, рефераты, а также отчёты, опубликованные в научных изданиях. Впрочем, полные тексты публикаций представляются далеко не всегда, так как доступ к ним является платным. Также Google Scholar предоставляет информацию о цитируемости статей и авторов, что является важнейшим показателем представленности автора в информационной среде мирового научного сообщества.

Несмотря на то, что использование сервиса абсолютно бесплатно, Google Scholar не предоставляет доступа к API, при этом замедляя процесс автоматизированного поиска, либо вызывает CAPTCHA.

Данная платформа предоставляет множество полей, содержащих информацию о публикации, такие как:

- Заголовок
- Аннотация
- Количество цитирований
- Год выпуска
- Список авторов
- Издание
- Ссылка на статью
- Ссылки на статьи, цитирующие данную публикацию

Кроме того, с помощью данного сервиса можно получить подробную информацию об авторах, такую как:

- Имя автора
- Домен электронной почты
- Интересы автора
- Принадлежность автора к университету, либо другому научному сообществу

Из них наиболее полезными для решения задачи кластеризации являются заголовок и аннотация, так как именно они применяются для дальнейшей обработки и анализа.

Для дальнейшего анализа такой неструктурированной информации, как текст, необходимо предварительно его обработать, для чего была разработана следующая последовательность.

1. Токенизация
2. Удаление пунктуации и стоп-слов
3. Удаление статей на языках, отличных от английского
4. Лемматизация
5. TF-IDF-векторизация

В первую очередь, проводится токенизация, то есть массивы текстов разбиваются на отдельные слова и знаки пунктуации. Затем удаляются стоп-слова и пунктуация, так как данные элементы не несут полезной информации для последующей кластеризации. Так как в научных текстах используется множество разных форм одного слова, также целесообразно привести их все к единой форме с помощью лемматизации.

Последним пунктом является TF-IDF-векторизация [3]. TF-IDF-векторизация является метрикой, которая способна определить важность слов в тексте с помощью следующей формулы:

$$TFIDF(w) = c(w) * \log \frac{D}{d(w)}$$

Где:

c(w) – количество вхождений слова в документ относительно длины документа

D – общее число документов

d(w) – количество документов, в которые входит слово

В результате работы данного алгоритма получается набор векторов, каждый из которых представляет собой заголовок и аннотацию некоторой статьи. Полученную матрицу можно использовать для дальнейшего анализа посредством кластеризации.

Кластеризация – это процесс разбиения некоторого множества объектов на группы, схожие по некоторому признаку и называемые кластерами. Главное отличие кластерного анализа от обыкновенной классификации заключается в том, что перечень групп задан не чётко и определяется во время работы алгоритма. Также системы кластеризации не используют тезаурусы и онтологии, вместо этого применяется обучение без учителя [2].

Кластеризация научных публикаций является важной проблемой в области библиометрии. Методы кластеризации регулярно применяются в библиометрической литературе для определения областей исследований или научных областей. Эти методы, например, используются для группировки публикаций в кластеры на основе списка наиболее значимых слов или же их отношений в сети цитирования. Кластеризация данных представляет собой ценный инструмент анализа данных в современных приложениях машинного обучения и интеллектуального анализа данных. Во многих случаях кластеризация используется для получения первых сведений о данных в процессе анализа и для решения ряда реальных проблем, например, как моделирование тем в интеллектуальном анализе текста

В качестве алгоритма кластеризации в ходе работы применяется метод К средних. Данный метод представляет возможность проводить кластеризацию с высокой скоростью даже на больших объемах данных, что подходит для решения поставленной задачи.

Как правило алгоритм К средних состоит из следующих этапов

− выбираются начальные центры кластеров (по существу, это набор наблюдений, которые находятся далеко друг от друга

- каждый объект формирует кластер из одного, а его центр - значение переменных для этого объекта;

− для каждого объекта устанавливается его ближайший кластер, определенный в терминах расстояния до центроида;

– находятся центроиды кластеров, которые были сформированы;

– пересчитывается расстояние от каждого объекта до каждого центроида и обращаем внимание на объекты, которые не находятся в том кластере, к которому они ближе всего;

– процесс продолжается до тех пор, пока центроиды не станут относительно стабильными [4]

Пример графического представления данного алгоритма изображен на рисунке 2 [5]:

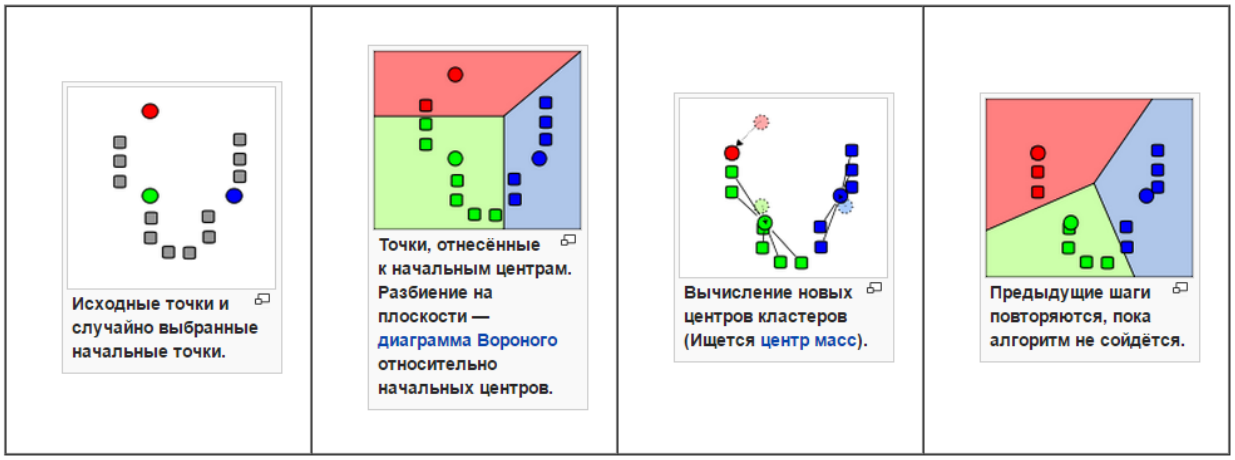

Рисунок 2 – графическое представление алгоритма К средних

Алгоритм k-means описывается с помощью следующего выражения:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

Где k-число кластеров, Si – полученные кластеры, i =1,2,…,k и μi – центры масс (центроиды) векторов xj ϵ Si .

Получив на вход некоторую матрицу, в первую очередь необходимо определить оптимальное количество кластеров, на которые будет по тематикам разбит текст. С этой целью в кластер-анализе вводится понятие критерия качества разбиения, определённого на множестве всех разбиений. Критерий зависит от объёмов кластеров и расстояний

между объектами, вошедшими в отдельные кластеры. Наилучшим разбиением считается то разбиение S* из всех S, на котором достигается экстремум (минимум или максимум) выбранного критерия качества.

Для этого необходимо несколько раз провести кластеризацию на разное число кластеров и определить для каждого сумму квадратов оценки ошибок. Место, где кривая графика заметно сгладится, называется локтем и соответствует оптимальному числу кластеров [6].

Глава 2. Проектирование

После анализа предметной области следующим важным этапом является проектирование информационной системы. Под проектированием подразумевается формирование архитектуры проекта, разработка моделей процессов, а также баз данных.

2.1 Выбор средств разработки

В начале разработки было необходимо определиться с технологиями, которые будут применяться в процессе. В первую очередь требовалось выбрать язык программирования в связке с фреймворком для бэкэнда. Для решения данной задачи была составлена матрица морфологического анализа, представленная в таблице 1.

Таблица 1 – Определение набора технологий для реализации приложения

Критерий	Весовой коэффицент критерия	Вариант		
		Python + Django	PHP+Laravel	NodeJS
Опыт работы	0.4	5	2	1
Качество документации	0.2	4	5	5
Наличие библиотек для работы с данными	0.3	5	2	3
Удобство для веб-разработки	0.1	4	3	4
Итого	1	4.5	3	3.25

В результате анализа было принято решение применять язык программирования Python и фреймворк Django. Ключевыми причинами для принятия данного решения является большое разнообразие средств обработки и анализа естественного языка, а также наличие опыта разработки с помощью данного фреймворка.

В ходе разработки использовалось несколько библиотек, предлагающих готовые решения для ключевых алгоритмов в данном программном обеспечении. Библиотека NLTK применяется для реализации обработки естественного языка, а именно для поиска стоп-слов, токенизации и последующей лемматизации. Библиотека scikit-learn предоставляет средства векторизации, а также множество уже реализованных алгоритмов машинного обучения, классификации и кластеризации. Наиболее важными элементами данной библиотеки были модули TF-IDF-векторизации, а также непосредственно алгоритм кластеризации методом K средних.

Для удобства работы с сервисом Google Scholar была использована библиотека Scholarly. Данная библиотека позволяет взаимодействовать с сервисом без использования API. Кроме того, с библиотекой предоставляются средства для использования VPN, адреса для которого в свою очередь получаются с помощью библиотеки Beautiful Soup.

В качестве базы данных используется реляционная СУБД MySQL, так как данная база данных позволяет считывать информацию быстрее, чем аналоги [7].

2.2 Требования к информационной системе по сбору и кластеризации информации

На основе обзора предметной области и анализа средств разработки были составлены требования для последующего проектирования и разработки программного обеспечения.

Функциональные требования:

1. Пользовательский интерфейс должен отображать список кластеров, созданных пользователем.
2. Для каждого кластера должен быть указан список ключевых слов.
3. При выборе кластера должен отобразиться список статей, который принадлежит к данному кластеру.

4. Для каждой из публикаций должны быть указаны заголовок, аннотация, список авторов, а также ссылка на публикацию.
5. Необработанные данные должны быть токенизированы.
6. Из данных должны быть удалены стоп-слова и пунктуация.
7. Статьи на языках, отличных от английского, должны быть удалены.
8. Слова в данных должны быть приведены в нормальную форму.
9. Система должна автоматически определить оптимальное количество кластеров.
10. Программа должна автоматически подключать VPN в случае блокировки сервисом.
11. Программа должна автоматически собирать данные из сервиса по запросу пользователя.
12. Пользователь должен иметь возможность зарегистрироваться на сайте.
13. Пользователь должен иметь возможность войти в свой профиль.
14. Пользователь должен иметь возможность удалить кластеры, созданные по своему запросу.
15. Должен отображаться список завершенных и незавершенных задач по добавлению статей в базу данных.

Нефункциональные требования:

1. Данные должны быть получены с помощью сервиса Google Scholar
2. В качестве данных для кластеризации должны использоваться заголовки и аннотации публикаций.
3. Программа должна работать без сбоев и аварийных завершений работы.
4. Программа должна быть написана на языке программирования Python.
5. Сбор данных должен производиться асинхронно.
6. Должна использоваться база данных MySQL.

2.3 Проектирование бизнес-процесса

На основе требований была разработана EPC-диаграмма, демонстрирующая весь бизнес-процесс поиска и кластеризации информации о научных публикациях. В данной диаграмме изображены все этапы поиска, обработки и представления данных пользователю.

Основное ветвление происходит на этапе поиска информации в системе Google Scholar, так как Google периодически блокирует автоматизированные запросы, следовательно, необходимо использование VPN или динамического IP. Если система обнаруживает, что доступа к Google Scholar нет, то производится поиск бесплатных VPN и попытка присоединения к одной из них. Попытки происходят до того момента, пока список сетей не закончится, либо подключение не будет успешным. В случае, если список сетей закончился, задача поиска откладывается на некоторый промежуток времени, после которого процесс начинается вновь. Необходимости заводить таблицу в базе данных для списка VPN нет, так как их список постоянно меняется.

После сбора информации о научных публикациях она проверяется на предмет дубликатов, после чего вносится в базу данных.

Далее пользователь создает запрос на разбиение полученных в базе данных текстов на тематики. Информация, соответствующая запросу пользователя, достается из базы данных, после чего проходит предобработку. Под предобработкой подразумевается удаление статей на иностранных языках, лемматизация и токенизация.

Затем обработанные данные поступают на вход алгоритма кластеризации, в результате работы которого получается набор кластеров, сгенерированных по тематикам.

В результате работы программного обеспечения пользователь имеет возможность просмотреть кластеры, созданные по его запросу, а также их содержание. Содержанием кластеров являются статьи, собранные на первом этапе данного алгоритма, и вся собранная о них информация. На рисунке 3 изображена EPC-

диаграмма, демонстрирующая бизнес-процесс “Алгоритмическое и программное обеспечение кластеризации текстов по тематикам”.

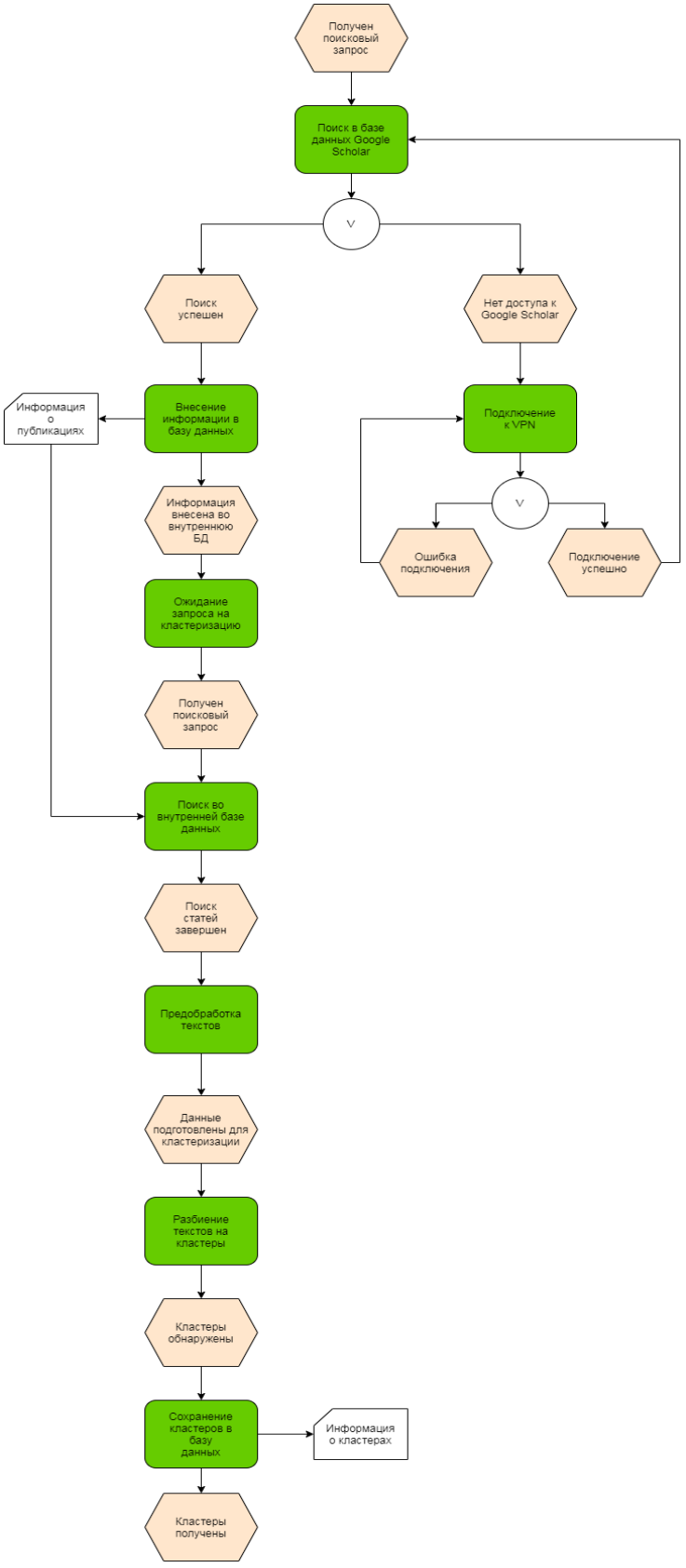

Рисунок 3 – Диаграмма EPC

2.4 Проектирование потоков данных

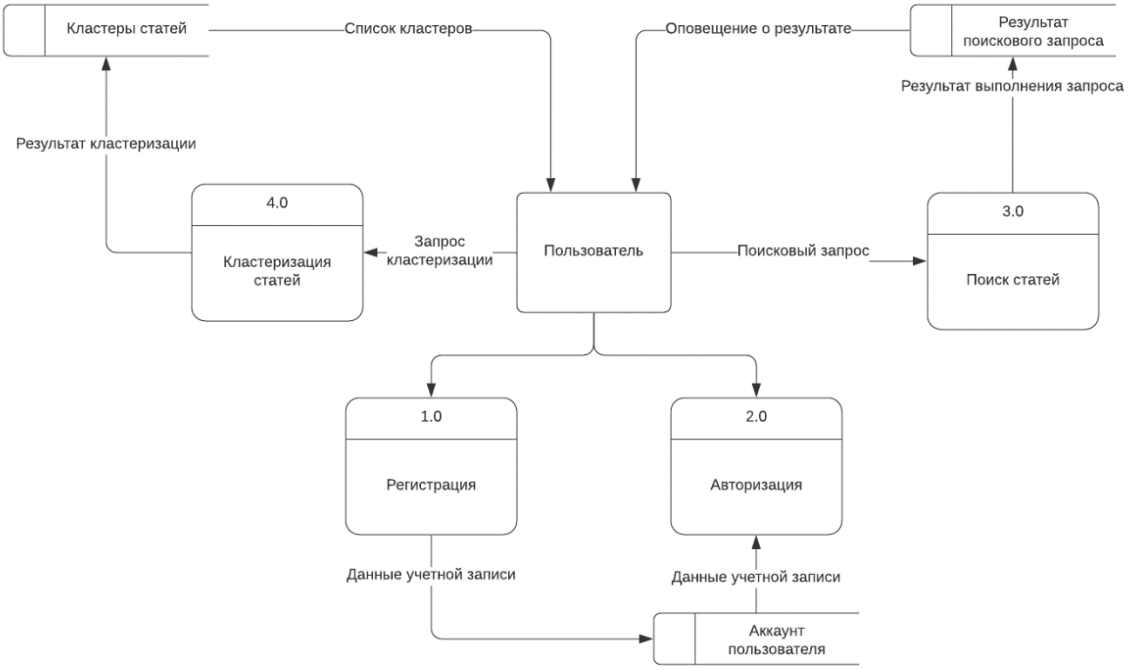

Рисунок 4 – Диаграмма потоков данных

Диаграмма потоков, изображенная на рисунке 4, наглядно отображает течение информации в пределах процесса или системы. Для изображения входных и выходных данных, точек хранения информации и путей ее передвижения между источниками и пунктами доставки в таких диаграммах применяются стандартные фигуры, такие как прямоугольники и круги, а также стрелки и краткие текстовые метки [8].

Данная диаграмма разработана в нотации Гейна-Сарсона и описывает потоки данных между пользователем и различными модулями системы, такими как модуль поиска статей, кластеризации, а также регистрации и авторизации.

2.5 Проектирование базы данных

Также была спроектирована логическая модель базы данных для хранения необходимых элементов. Для выполнения данной работы необходимо 4 таблицы.

Основная информация, которую необходимо хранить – это информация о пользователе, публикациях, а также о кластерах.

Также с учётом того, что задача поиска статей может занимать большое количество времени и попыток, а также выполняться несколькими пользователями одновременно, необходимо реализовать таблицу, в которой будет храниться информация о выполняемых на данных момент задачах. Логическая модель базы данных изображена на рисунке 5. Более подробное описание сущностей можно увидеть в приложении А.

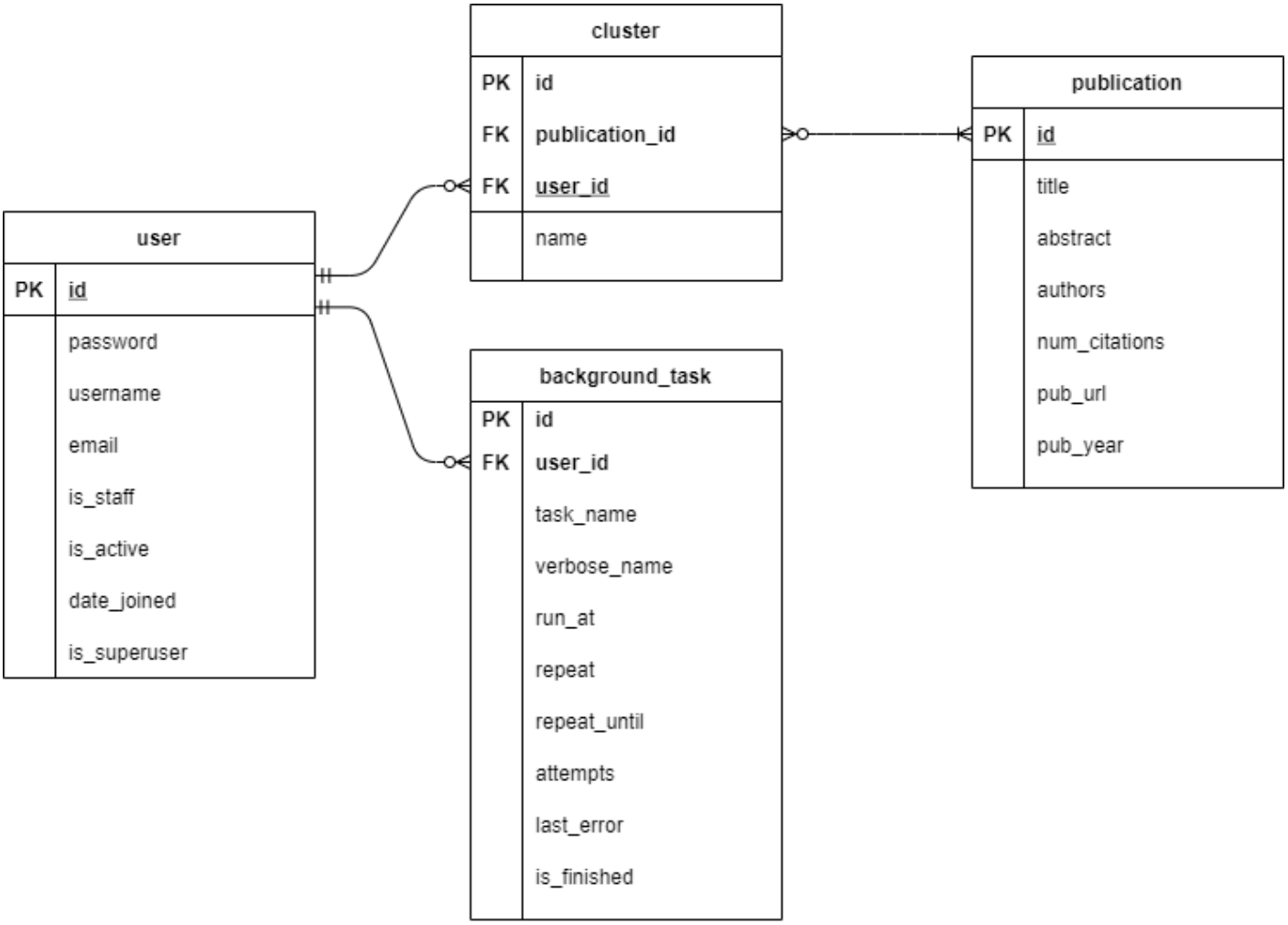

Рисунок 5 – Логическая модель данных предметной области

Глава 3. Разработка системы поиска данных о публикациях и кластеризации

На основе результатов проектирования было разработано 3 модуля программного обеспечения.

3.1 Формы регистрации и логины

Для получения доступа к системе и хранения информации о кластерах, созданных пользователем, в первую очередь необходимо пройти процесс регистрации и входа в профиль. Формы регистрации и авторизации изображены на рисунках 6 и 7.

Рисунок 6 – страница регистрации

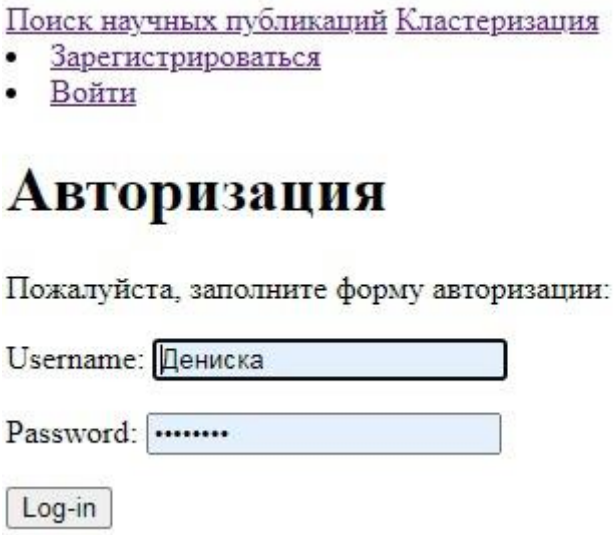

Рисунок 7 – страница авторизации

3.2 Поиск публикаций в интернете

На рисунке 8 изображена форма поиска информации в сети интернет с помощью Google Scholar.

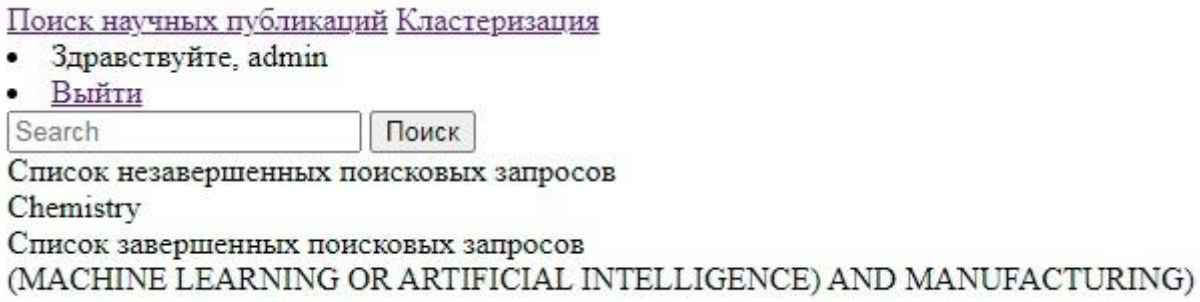

Рисунок 8 – страница поиска информации о публикациях в интернете

Так как поиск может заниматься продолжительное время, в результате отправления пользователем поискового запроса создается задача, которая в дальнейшем поступает в очередь для дальнейшей обработки. Необработанная на данный момент задача находится в списке незавершенных задач и переносится в список завершенных задач после окончания исполнения. Данный процесс был реализован посредством библиотеки Django background tasks. При создании новой задачи она помещается в таблицу базы данных background_tasks, в которой хранится вся информация о текущей задаче. В

результате выполнения задачи она помечается как успешно выполненная и отображается у пользователя в соответствующем разделе.

В ходе работы данного модуля существует вероятность появления ошибок, которые невозможно предупредить на этапе разработки. К ним относятся перебои соединения с сетью, отсутствие доступных VPN, проблемы со стороны сервиса Google Scholar. В случаях, когда данные ошибки происходят, выполнение задачи откладывается на некоторый срок. По истечение указанного срока система вновь попытается выполнить задачу. Листинги с кодом, относящимся к данному модулю, находятся в приложении Б.

3.3 Модуль кластеризации

Для запуска процесса разбиения на тематики существует отдельная форма. В результате запроса из базы данных получаются публикации, удовлетворяющие поисковому запросу, после чего из них извлекается информация о заголовке и аннотации. Данная информация предварительно обрабатывается, затем запускается процесс кластеризации.

Первым делом проводится удаление стоп-слов, токенизация и лемматизация заголовков и аннотаций статей. Соответствующий фрагмент кода указан (Листинг 1).

```
lemmatizer = WordNetLemmatizer()
stop_words = set(stopwords.words('english'))
stop_words.add("˜")
stop_words.add("˝")
words = word_tokenize(text_file)
filtered = [re.sub('[0-9]+', '', w) for w in words if w not in stop_words]
lemmas = [lemmatizer.lemmatize(t) for t in filtered]
```

Листинг 1 – реализация алгоритма предобработки

Далее необходимо провести TF-IDF-векторизацию. В ходе данного процесса строки преобразовываются в векторы, которые в свою очередь получит на вход алгоритм кластеризации (Листинг 2).

```
vectorizer = TfidfVectorizer(max_df=0.5, max_features=10000,
                             min_df=2, stop_words='english',
```

```
                                    use_idf=True, tokenizer=tokenize_and_stem)
X = vectorizer.fit_transform(data['text'])
```

Листинг 2 – алгоритм векторизации

Система автоматически определяет оптимальное количество кластеров, после чего проводит анализ методом К средних (Листинг 3).

```
sse = []
for k in range(1, 20):
    km = KMeans(n_clusters=k, init='k-means++', max_iter=100, n_init=1,
                verbose=False)
    km.fit(X)
    sse.append(km.inertia_)
kl = KneeLocator(range(1, 20), sse, curve="convex", direction="decreasing")
true_k = kl.elbow
km = KMeans(n_clusters=true_k, init='k-means++', max_iter=100, n_init=1,
                verbose=False)
```

Листинг 3 – реализация алгоритма предобработки

В результате анализа получается некоторое число кластеров, которые добавляются в базу данных и привязываются к данному пользователю. Таким образом, только у пользователя есть доступ к результатам кластерного анализа. Результат запроса изображен на рисунке 9.

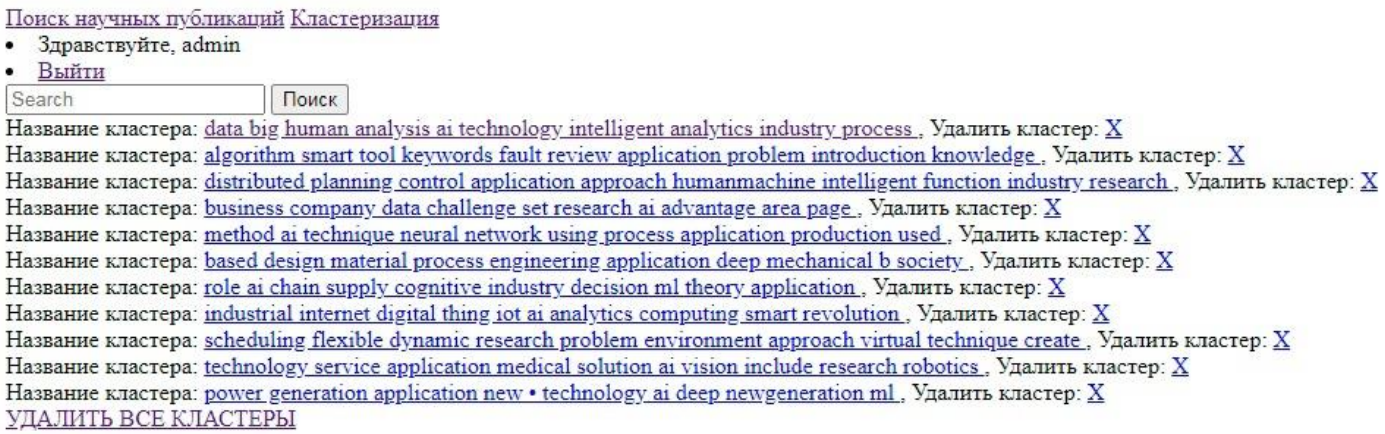

Рисунок 9 – страница кластеризации

Название каждого кластера состоит из ключевых слов, которые понятны пользователю. Так как кластеры принадлежат пользователю, пользователь может удалить информацию о любом кластере, который ему не нужен, либо же удалить информацию о всех кластерах.

При нажатии на название кластера пользователь переходит на страницу, включающую в себя список статей, принадлежащих данному кластеру (Рисунок 10).

Рисунок 10 – страница с содержимым кластера

На странице кластера указано имя кластера, а также список статей. При нажатии на название статьи, являющееся ссылкой, происходит переход на страницу с информацией о статье (Рисунок 11).

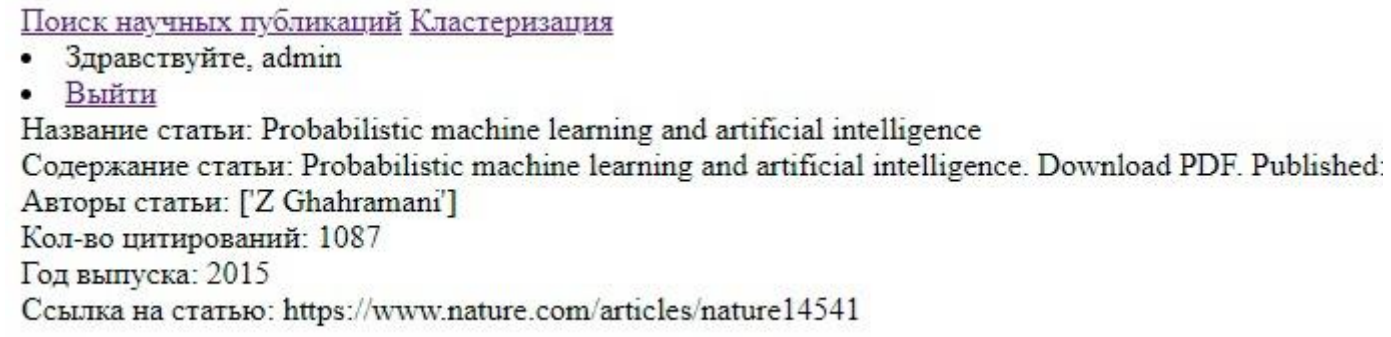

Рисунок 11 – страница с информацией о публикации

Глава 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

4.1 потенциальные потребители результатов исследования

Любое научное исследование начинается с анализа уже существующих научных публикаций и достижений, доступных в уже существующих библиографических базах данных. Для решения данной задачи существует множество сервисов, такие как Google Scholar, SCOPUS или Microsoft Academic Search. Однако, процесс поиска затрудняется тем, что количество статей, собранных сервисами, составляет сотни миллионов и с каждым днём это число увеличивается. Таким образом, разработка программного обеспечения, которое смогло бы дополнительно разбивать результаты поисковых запросов на тематики, помогло бы исследователям снизить нагрузку на начальном этапе деятельности и упростить процесс написания научных трудов.

Таким образом, потенциальными потребителями данного продукта являются такие виды исследователей, как студенты, преподаватели высших учебных заведений, а также члены различных научных обществ. В первую очередь данная система разрабатывается для того, чтобы упростить процесс написания научных работ студентами и работниками Томского Политехнического Университета.

Одной из особенностей данного проекта является то, что несмотря на обилие научных работ, связанных с кластеризацией тестов, включая научных публикации, на данный момент крайне проблематично найти сервис, который смог бы собирать информацию о научных публикациях по пользовательскому запросу, разбивая результаты по тематикам.

4.2 SWOT-анализ

SWOT-анализ (показан в таблице 1) – один из инструментов стратегического планирования, является простым и качественным инструментом оценивания конкурентоспособности. Данная методика применяется для анализа внутренней и

внешней среды проекта. Результаты первого этапа SWOT-анализа представлены в таблице 2.

Таблица 2 – Матрица SWOT

	Сильные стороны: С1. Проект не использует платные сервисы при поиске публикаций. С2. Отсутствие известных аналогов проекта. С3. Простота использования. С4. Высокая скорость процесса кластеризации. С5. Возможность интегрировать поиск по альтернативным библиографическим БД.	**Слабые стороны:** Сл1. Сложность разработки Сл2. Отсутствие изначального объема публикаций в базе данных Сл3. Длительное ожидание сбора информации о статьях с помощью сервиса Google Scholar Сл4. Отсутствие рекламной кампании Сл5. Необходимость периодического изменения IP-адреса из-за ограничения на кол-во запросов к Google Scholar.
Возможности: В1. Популяризация сервиса среди студентов разных вузов. В2. Увеличение популярности научных исследований В3. Попадание продукта на первые строчки выдачи поисковых систем	Благодаря высокой скорости процесса разбиения на тематики и низкой стоимости данный сервис может стать популярным среди студентов разных вузов. Благодаря отсутствию популярных аналогов проекта заинтересованные люди в первую очередь могут заинтересоваться данным сервисом.	Увеличение популярности сервиса и научных исследований позволит привлечь больший капитал и перейти на более стабильную БД.
Угрозы: У1. Пониженная заинтересованность в альтернативных средствах поиска У2. Появление более совершенного продукта У3. Запрет на использование данных сервисом Google Scholar У4. Запрет на использование VPN сервисом Google Scholar	С5 позволяет решить У3 и У4, но данное решение увеличит стоимость проекта, поэтому переход на альтернативные БД рекомендуется провести после получения прибыли. На У1 со стороны разработчика повлиять невозможно.	Основной угрозой являются проблемы со сторонними сервисами Google Scholar и бесплатными сервисами, предоставляющими VPN или динамический диапазон IP-адресов. Для уменьшения рисков имеет смысл заранее заполнить базу данных некоторым объемом статей.

По результатам SWOT-анализа были выявлены сильные и слабые стороны научной разработки, а также ее угрозы и возможности. Некоторые слабые стороны можно компенсировать возможностями, а угрозы – сильными сторонами.

4.3 Технология QuaD

Технология QuaD позволяет оценить качество и перспективность разработки программного продукта для того, чтобы определить целесообразность вложения денежных средств. Результаты оценки представлены в таблице 3.

Таблица 3 – Оценочная карта для сравнения конкурентных технических решений

Критерий оценки	Вес критерия	Баллы	Макс. Балл	Относительное значение (3/4)	Средневзвешенное значение (5/2)
1	2	3	4	5	6
Показатели оценки качества работы					
Интуитивно-понятный интерфейс	0,1	80	100	0,8	0,08
Кроссплатформенность	0,05	50	100	0,5	0,025
Простота ввода в эксплуатацию	0,1	60	100	0,6	0,06
Рациональные методы обработки данных	0,1	85	100	0,85	0,085
Корректная визуализация результатов	0,1	90	100	0,9	0,09
Язык написания программы	0,1	70	100	0,7	0,07
Показатели оценки коммерческого потенциала разработки					
Цена	0,2	70	100	0,7	0,14
Предполагаемый срок эксплуатации	0,1	80	100	0,8	0,08
Отсутствие аналогов	0,05	80	100	0,8	0,04
Конкурентоспособность продукта	0,1	90	100	0,9	0,09
Итого	1			7,55	0,76

Анализ, проведённый с помощью технологии QuaD, позволяет нам сделать вывод о том, что перспективность разработки программного продукта выше среднего, так как средневзвешенное значение показателя качества и перспективности научной разработки составляет 76.

4.4 Планирование разработки

4.4.1 Структура работ в рамках разработки

Распределение работ по исполнителям представлено в таблице 4.

Таблица 4 – Перечень этапов, работ и распределение исполнителей

Основные этапы	№ раб	Содержание работы	Должность исполнителя
Разработка технического задания	1	Составление и утверждение технического задания	Руководитель
Проектирование разработки	2	Определение целей исследования	Руководитель
	3	Составление календарного плана	Разработчик
	4	Составление диаграмм	Разработчик
Разработка приложения	5	Определение средств разработки	Разработчик
	6	Анализ подходящих алгоритмов для обработки результатов	Разработчик
	7	Разработка макетов приложения	Разработчик
	8	Программирование, отладка приложения	Разработчик
	9	Тестирование	Разработчик
Внедрение приложения	10	Развертывание ИС в сети	Разработчик
Оформление отчета по ВКР	11	Составление пояснительной записки	Разработчик

4.4.2 Определение трудоемкости выполнения работ

Важным моментом при разработке является определение трудоёмкости работ каждого из участников научного исследования.

Для определения ожидаемого значения трудоемкости используется следующая формула:

$$T_{ож} = \frac{3 * \mathrm{T}_{min} + 2 * T_{max}}{5}$$

Где

tож i - – ожидаемая трудоемкость выполнения i-ой работы человеко-дней

$t$min i – минимально возможная трудоемкость выполнения заданной i-ой работы, человеко-дней.;

$t$max i– максимально возможная трудоемкость выполнения заданной i-ой работы, человеко-дней.

Так как параллельные процессы во время разработки отсутствовали Tpi = tож i;

Округление длительности работ в рабочих и календарных днях производится по математическим правилам.

Коэффициент календарности, используемый при расчете длительности работ в календарных днях, на 2021 год равен 1,48.

Согласно плану, длительность работ составит 60 календарных дней.

Результаты расчетов трудоемкости работ представлены в таблице 5.

Таблица 5 – Временные показатели проведения научного исследования

Название работы	Трудоемкость			Длительность работ в рабочих днях Tpi	Длительность работ в календарных днях Tki
	tmin	tmax	tожид		
1. Составление и утверждение технического задания (Руководитель)	2	4	2,8	3	4
2. Определение целей исследования (Руководитель)	2	4	2,8	3	4
3.Составление календарного плана (Разработчик)	2	4	2,8	3	4
4.Составление диаграмм (Разработчик)	1	3	1,8	2	3
5.Определение средств разработки (Разработчик)	1	3	1,8	2	3
6.Анализ подходящих алгоритмов для обработки результатов (Разработчик)	2	4	2,8	3	4
7.Разработка макетов приложения (Разработчик)			,2	3	4
8.Программирование, отладка приложения (Разработчик)	5	9	6,6	7	10
9.Тестирование (Разработчик)	4	7	5,2	5	7
10.Развертывание ИС в сети (Разработчик)	2	4	2,8	3	4
11. Составление пояснительной записки (Разработчик)	7	12	9	9	13

4.5 Бюджет научно-технического исследования (НТИ)

4.5.1 Расчет амортизационных затрат

В процессе разработки программного продукта использовался персональный компьютер, а также ноутбук, купленные заранее. Затраты на специальное оборудование приведены в таблице 6.

Таблица 6 – Затраты на специальное оборудование

Наименование оборудования	Кол-во единиц	Цена за 1 ед. оборудования, руб.	Общая стоимость, руб.	Амортизационные отчисления
Ноутбук	1	20 000	20 000	1 100
Персональный компьютер	1	100 000	100 000	5 500
Монитор	2	13 000	26 000	1 430
Windows 10 Pro	1	21 500	21 500	1182
Итого:			168 700	9 212

Амортизационные отчисления для рассматриваемого проекта включают в себя амортизацию используемого оборудования за время выполнения работы. Срок полезного использования офисных машин составляет 3 года, когда как время написания ВКР – 2 месяца. Таким образом, норму амортизации можно рассчитать по следующей формуле:

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

Таким образом, годовые амортизационные отчисления для ноутбука:

$$A_г = 20000 * 0{,}33 = 6600 \text{ рублей}$$

Ежемесячные амортизационные отчисления для ноутбука:

$$A_г = 6600/12 = 550 \text{ рублей}$$

Аналогичным образом рассчитаны амортизационные отчисления для всего оборудования.

4.5.2 Основная заработная плата исполнителей темы

Таблица 7 – Баланс рабочего времени

Показатели рабочего времени	Студент	Руководитель
Календарные дни	365	
Нерабочие дни (праздники/выходные)	74	
Потери рабочего времени -отпуск -невыходы по болезни	56	56
Действительный годовой фонд рабочего времени	235	235

Затраты на заработную плату рассчитываются по следующей формуле:

$$Зп = Зосн + Здоп, \text{ где}$$

Здоп – дополнительная заработная плата, руб;

Зосн – основная заработная плата, руб.

$$Зосн = Здн * Тр * (1 + Кпр + Кд) * Кр, \text{ где}$$

Здн – среднедневная заработная плата, руб.;

Кпр – премиальный коэффициент;

Кд – коэффициент доплат и надбавок;

Кр – районный коэффициент (для Томска 1,3)

Тр – продолжительность работ, выполняемых работником, раб. дни Среднедневная заработная плата:

$$Здн = \frac{Зм * М}{Fд}$$

Зм – месячный оклад работника, руб;

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

Fд – действительный годовой фонд рабочего времени персонала, раб. дн.

На основе найденных показателей можно произвести расчёт заработной платы, при этом месячный оклад студента равняется 30000, а руководителя – 45000 рублей.

Планирование основной заработной платы приведено в приложении В.

Таблица 8 – Расчет основной заработной платы

Исполнители	Здн, руб	Кпр	Кд	Кр	Тр	Зосн
Разработчик	1327,6	-	-	1,3	52	89 745
Руководитель	1991	-	-	1,3	8	20 706
Итого:						110 451

Дополнительная заработная плата исполнителей темы

Таблица 9 – Расчет дополнительной заработной платы

Исполнитель	Основная заработная плата, руб.	Коэффициент дополнительной заработной платы	Дополнительная заработная плата, руб
Разработчик	89 745	0,15	13 461
Руководитель	20 706		3 106
Итого:			16 567

Отчисления во внебюджетные фонды (страховые отчисления)

Таблица 10 – Расчет отчислений во внебюджетные фонды

Исполнитель	Основная заработная плата + дополнительная, руб.	Коэффициент отчислений во внебюджетные фонды	Сумма отчислений
Разработчик	103 206	0,302	31 168
Руководитель	23 812		7 191
Итого:			38 360

Накладные расходы

Накладные расходы учитывают прочие затраты организации, не попавшие в предыдущие статьи расходов: печать и ксерокопия материалов, оплата услуг связи, электроэнергии и т.д.

Знакл = (сумма статей 1-4) * Кнр, где

Кнр – коэффициент, учитывающий накладные расходы (16%)

Таблица 11 – Бюджет затрат НТИ

Наименование статьи	Сумма, руб	Примечание
Амортизационные затраты на оборудование	9 212	Таблица 6
Затраты на основную заработную плату	110 451	Таблица 7
Затраты на дополнительную заработную плату	16 567	Таблица 8
Затраты на отчисление во внебюджетные фонды	38 360	Таблица 9
Накладные расходы	28 319	16% от суммы статей 1-4
Бюджет затрат НТИ	202 909	Сумма статей 1-5

4.6 Определение ресурсной (ресурсосберегающей), финансовой, бюджетной, социальное 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 финансовый показатель разработки определяется как:

$$I_{\text{ф}}^{p} = \frac{\Phi_p}{\Phi_{max}},$$

где

$\Phi_{\text{р}}$ – стоимость исполнения, руб;

Φ_{max} – максимальная стоимость исполнения научно-исследовательского проекта (в т.ч. аналоги).

Полученная величина интегрального финансового показателя разработки отражает соответствующее численное увеличение бюджета затрат разработки в разах.

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

$$\text{I} = \sum_n a \times b \quad (6)$$

где

a – весовой коэффициент параметра;

b – бальная оценка параметра для аналога и разработки, устанавливается экспертным путем по выбранной шкале оценивания;

n – число параметров сравнения.

Результаты расчета интегрального показателя ресурсоэффективности приведены в таблице 12.

Таблица 12 – Расчет интегрального показателя ресурсоэффективности

Объект исследования/Критерии	Весовой коэффициент параметра	Исп.1
Способствует росту производительности труда пользователя	0,25	5
Функциональность	0,1	3
Простота эксплуатации	0,1	4
Экономия времени	0,20	5
Надежность	0,15	3
Точность	0,2	4
Итого	1	

$$I_{исп} = 0{,}25*5+0{,}1*3+0{,}1*4+0{,}2*5+0{,}15*3+0{,}2*4=4.2$$

Таким образом можно сделать вывод, что данная разработка является привлекательной для инвесторов, так как разрабатываемое программное обеспечение не имеет популярных аналогов, при этом помогая исследователям сэкономить время и силы на начальном этапе деятельности. Кроме того, интегральный показатель ресурсоэффективности выше 4, следовательно результат работы можно считать положительным.

Вывод по разделу

В разделе исследовательской работы «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» был проведен анализ целесообразности разработки проекта, а также планирование предстоящих работ. Для этого был

проведён SWOT-анализ, благодаря чему были определены сильные и слабые стороны проекта, а также учтены риски. Общая длительность работы над проектом ориентировочно составляет 60 календарных дней, а потенциальная стоимость разработки данного программного обеспечения равна 202 909 рублей.

Глава 5. Социальная ответственность

Введение

Данная научно-исследовательская работа представляет собой алгоритмическое и программное обеспечение кластеризации научных текстов по тематикам. Результатом работы является программный продукт, который способен обеспечить оценку сходства корпусов научных текстов посредством обработки естественного языка. В процессе работы будут спроектирована и разработана совокупность алгоритмов и программных модулей для автоматизированного поиска научных публикаций в системе Google Scholar, извлечения и хранения информации из данного сервиса, а также последующего обнаружения кластеров схожих по смыслу статей. Практическое назначение разрабатываемого продукта - для информационная поддержка процессов организации научных исследований. Впрочем, алгоритм кластеризации является универсальным для любых англоязычных статей, поэтому при подключении дополнительных модулей сбора информации в базу данных функционал данной системы можно с легкостью расширять.

Данное программное обеспечение разрабатывалось на персональном компьютере, оборудованном клавиатурой, мышью, а также двумя мониторами в домашних условиях, так как в условиях коронавируса рекомендуется оставаться дома. Продолжительная работа за компьютером также может нести вред для здоровья. Опасность для зрения несёт использование экрана и работа с крупным текстом, значительно утомляя глаза. Кроме того, неправильная цветопередача может вызывать некоторую дезориентацию, ухудшение зрения, а также головные боли. Также работа выполняется в сидячем положении, что способствует ухудшению осанки, появлению болевых ощущений в области позвоночника, а также нарушениям кровообращения.

5.1 Правовые и организационные вопросы обеспечения безопасности

Разработка данного приложения проводилась вне помещений Томского Политехнического Университета, так как в настоящее время рекомендуется вести работу удалённо в целях предотвращения распространения коронавируса. В качестве искусственного источника света использовались 4 лампы накаливания, обеспечивающие достаточно яркое освещение для комфортной работы за компьютером. Также помещение оборудовано компьютерным столом, которое является рабочим местом разработчика. В качестве операторского кресла использовалось классическое компьютерное кресло с регулирующейся высотой, уровнем наклона спинки и высотой подлокотников.

Далее приведены некоторые наиболее важные пункты ГОСТ 12.2.032-78 «Система стандартов безопасности труда (ССБТ) [9]. Рабочее место при выполнении работ сидя. Общие эргономические требования», а также СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и безвредности для человека факторов среды обитания» [10]:

- Подвижность кресла относительно пола или другой поверхности, на которой оно установлено, может не ограничиваться. В случае необходимости обеспечения строго определенного положения человека - оператора по отношению к средствам отображения информации и органам управления, а также в случае, если трудовая деятельность человека - оператора сопряжена с силовыми и резкими движениями, кресло должно быть фиксировано. При этом, в зависимости от характера трудовой деятельности оператора, должна быть обеспечена возможность изменения положения кресла или сиденья в горизонтальной плоскости с фиксацией его в нужном положении. При необходимости подвижность кресла должна задаваться также вращением кресла на 180 - 360° вокруг вертикальной оси опорной конструкции кресла с фиксацией в нужном положении.

- Пользователь должен иметь возможность наклонить или повернуть видеодисплей таким образом, чтобы сохранить ненапряженную рабочую позу независимо от высоты уровня глаз с минимальными прилагаемыми усилиями, и при этом на экране не должно возникать раздражающих отражений и бликов.
- Угол обзора (оптимальный угол 0°) не должен превышать 40° по всей активной площади экрана. Специфические ограничения на расстояние до экрана и углы зрения и обзора должны рассматриваться с учетом применяемого пользователем метода коррекции зрения и его возраста.
- Оптимальное расстояние наблюдения для офисной работы в положении сидя составляет 600 мм

Таким образом, рабочее место, оборудованное дома, исполняет большую часть требований, указанных в вышеупомянутых стандартах.

Также были исследованы условия использования Google на предмет законности использования данных, предоставляемых Google Scholar. Несмотря на то, что корпорация Google не предоставляет официального доступа к API интерфейсу, единственный запрет, который может относиться к теме выполняемой работе заключается в запрете на копирование, изменение, распространение, продажу и сдачу в аренду элементов сервисов и программ, принадлежащих Google. Однако разрабатываемое программное обеспечение собирает исключительно информацию, лежащую в свободном доступе и не принадлежащую Google.

В результате рассмотрения данного соглашения был сделан вывод о том, что дальнейшая разработка программы не нарушает условий пользования сервисами Google.

5.2 Производственная безопасность

В данном пункте производится анализ вредных и опасных факторов, которые могут возникнуть на этапах выполнения данной работы.

Все выявленные вредные и опасные факторы для каждого из этапов, которым подвергается разработчик, представлены в Таблице 13. Также представлен список нормативных документов, на основе которых регулируются допустимые показатели.

Таблица 13 – Возможные опасные и вредные факторы

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Разработка	Внедрение	Эксплуатация	
Отклонение показателей микроклимата	+	+	+	СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания
Недостаточная освещенность рабочей зоны	+	+	+	СП 52.13330.2016 «Естественное и искусственное освещение»
Повышенный уровень шума на рабочем месте	+	+	+	СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания
Отсутствие или недостаток естественного света	+	+	+	СНиП 23-05-95* Естественное и искусственное освещение

Продолжение

Повышенный уровень электромагнитных излучений	+	+	+	СанПиН 1.2.3685-21 Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания

Исходя из данной таблицы можно сделать вывод, что на разработчика программного обеспечения на всем протяжении работ воздействовали исключительно физические и психологические факторы, когда как химические исключаются. Причиной длительности воздействия каждого из факторов является то, что весь процесс деятельности разработчика проводится в одном помещении.

5.2.1 Анализ опасных и вредных производственных факторов

В данном пункте в соответствии с порядком в Таблице 13 подробнее рассматриваются источники опасных и вредных факторов, их влияние на организм человека, а также допустимые нормы в соответствии со стандартами СанПиН, СНиП и СП. Данная информация продемонстрирована в Таблице 14.

Таблица 14 – Допустимые величины показателей микроклимата на рабочих местах

Фактор	**Источник**	**Воздействие**	**Допустимые нормы**
Отклонение показателей микроклимата	Отсутствие кондиционеров и увлажнителей воздуха	Вялость, усталость, снижаются концентрация и внимание	Таблица 15

Продолжение

Недостаточная освещенность рабочей зоны	Отсутствие или недостаточная мощность осветительных приборов	Усталость глаз, ухудшение зрения	Освещенность на рабочей поверхности от системы общего искусственного освещения 200-300 лк.
Повышенный уровень шума на рабочем месте	Использование персональных компьютеров, наличие центральной вентиляции	Снижение работоспособности, повышение утомляемости	Предельно допустимый уровень звука 55 дБА. Уровни звукового давления для источников постоянного шума указаны в Таблице 4
Отсутствие или недостаток естественного света	Отсутствие окон	Усталость глаз, ухудшение зрения	КЕО не ниже 1.2%
Повышенный уровень электромагнитных излучений	Компоненты персональных компьютеров и ноутбуков	Возможно возникновение рака	Напряженность электростатического поля не более 20 кВ/м

Нормы микроклимата включат в себя множество параметров, такие как температура воздуха в помещении, поверхностей, относительная влажность и скорость

движения воздуха. Допустимые величины показателей микроклимата продемонстрированы в Таблице 15.

Таблица 15 – Допустимые величины показателей микроклимата на рабочих местах

Период года	**Температура воздуха, °С**	**Температура поверхностей, °С**	**Относительная влажность воздуха, %**	**Скорость движения воздуха, м/с**
Холодный	22-24	21-25	15-75	0,1
Теплый	23-25	22-26	15-75	0,1

Нормы звукового давления также отличаются на разных октавных полосах, предельно допустимые уровни звукового давления указаны в Таблице 16 для дома, так как разработка велась удалённо.

Таблица 16 – Предельно допустимые уровни звукового давления, уровни звука и эквивалентные уровни звука для инженера-программиста

Вид трудовой деятельности, рабочее место	**Уровни звукового давления, дБ в октавных полосах со среднегеометрическими частотами, Гц**						
	31,5	63	125	250	500	1000	2000
Жилые комнаты квартир, домов стационарных организаций Социального обслуживания, организации для детей-сирот и детей, оставшихся без попечения родителей, спальные помещения в школах-интернатах, дошкольных	79	63	52	45	39	35	32

образовательных организациях, домов отдыха, пансионатов							

5.2.2 Обоснование мероприятий по снижению уровней воздействия опасных и вредных факторов

Мероприятия по снижению воздействия показателей микроклимата вне допустимых значений

Для восстановления и поддержания показателей микроклимата в пределах допустимых значений необходимо провести ряд следующих мероприятий:

- Оборудовать помещение кондиционером, системами обогрева, увлажнения воздуха, а также вентиляции.
- Защитить фасад здания от солнца с помощью штор, навесов, жалюзи и т.д.
- Своевременная влажная уборка помещения.
- Размещение рабочего места исследователя должно быть рациональным.

Мероприятия по снижению воздействия отсутствия или недостатка естественного света, а также плохой освещенности рабочего места:

- Сокращение рабочего времени в соответствии с длиной светового дня.
- Ремонт помещения светлых тонах
- Установка более мощного освещения в правильном положении и необходимом количестве.
- Своевременная чистка оконных стёкол.

Мероприятия по снижению воздействия повышенного шума в помещении:

Для уменьшения воздействия шума рекомендуется использовать следующие методы:

- Экранирование рабочих мест посредством установки перегородок между рабочими местами
- Установка оборудования, производящего минимальный шум
- Регулярное техническое обслуживание оборудования с целью снижения влияния загрязнения на воспроизводимый шум.
- Оборудование помещения звукоизолирующими пластиковыми окнами.

Повышенный уровень электромагнитных излучений сокращается при выполнении следующих пунктов:

- Сокращение времени, проводимого за компьютером
- Выключение компьютера вне времени его использования
- Прекращение использования мониторов с электронно-лучевой трубкой
- Расположение монитора в углу помещения так, чтобы стены поглощали излучение.

Также для того, чтобы избежать поражения электрическим током, следует соблюдать следующие рекомендации:

- При обнаружении неисправности в работе электроприбора немедленно прекратить его эксплуатацию
- Не заниматься самостоятельной починкой электроприборов при отсутствии соответствующих навыков
- Не пользоваться электроприборами при отсутствии необходимого защитного заземления

5.3 Экологическая безопасность

В данном подразделе рассматривается характер воздействия процесса разработки и использования разрабатываемого приложения на окружающую среду.

Единственным косвенным источником загрязнения атмосферы, который был обнаружен в ходе анализа является потребление электроэнергии. Так как потребляемая

энергия получается в следствии работы угольных электростанций, выбросы полученного в процессе горения углекислого газа наносят прямой вред окружающей среде. Тем не менее, возможности сокращения влияния разработчика на процесс загрязнения атмосферы обнаружено не было, так как даже сокращение потребления электроэнергии не уменьшит количества вырабатываемого электростанциями электричества. Тем не менее, следует сократить частоту использования электроприборов без необходимости, а также применять энергосберегающие лампы.

К источникам негативного влияния на гидросферу относятся перерасход воды, используемой разработчиком, а также тепловое и химическое загрязнение воды в следствии работы предприятий и гидроэлектростанций. Рекомендуется также уменьшить потребление воды и электричества без необходимости, например не оставлять воду включенной при чистке зубов или бритье, а также следить за состоянием сантехники.

Основным источником загрязнения гидросферы и литосферы является неправильная утилизация отходов продуктов, использующихся в процессе разработки, а именно канцелярии и макулатуры, деталей персонального компьютера, а также организационной техники.

Такие отходы, как пластик, алюминий и бумагу следует сортировать и утилизировать соответствующим образом. Вторично переработанные, данные материалы можно использовать вновь, тем самым сокращая негативное влияние на окружающую нас среду.

В производстве компьютерной и организационной техники используется большое количество разнообразных материалов, часть которых может оказать крайне негативное влияние на окружающую среду при неправильной переработке. По этой причине переработку данных приборов следует возложить на предприятия, обладающие соответствующими техническими условиями, персоналом, а также полигонами, при

помощи которых все компоненты сложных электроприборов будут извлечены и переработаны надлежащим образом.

Батарейки и лампы также следует сдавать специализирующимся на переработке данных отходов компаниям, так как неправильная переработка данных элементов наносит большой вред природе.

5.4 Безопасность в чрезвычайных ситуациях

Процесс разработки алгоритмического и программного обеспечения для кластеризации научных текстов по тематикам подразумевает собой постоянное использование электроприборов. Несоблюдение правил электробезопасности может повлечь за собой возникновение пожара.

Во избежание таким ситуаций необходимо следовать следующим профилактическим мероприятиям по предупреждению пожара:

- Не храните в доме бензин, керосин, легковоспламеняющиеся жидкости.
- Приобретите хотя бы один огнетушитель.
- Не оставляйте без присмотра включенные электрические и газовые плиты, чайники, утюги, приёмники, телевизоры, обогреватели.
- Следите за исправностью электропроводки, розеток.
- Не включайте в одну розетку несколько бытовых электрических приборов (особенно большой мощности).
- Не разогревайте на открытом огне краски, лаки и т. п.

Действия при пожаре в квартире:

1. Сообщите о пожаре в пожарную охрану по телефонам «112», «01» (с сотового тел. 01*, 112).

2. Если нет опасности поражения электротоком, приступайте к тушению пожара водой, или используйте плотную (мокрую ткань).

3. При опасности поражения электротоком отключите электроэнергию.

4. Горючие жидкости тушить водой нельзя (тушите песком, землёй, огнетушителем, если их нет, накройте плотной смоченной в воде тканью).

5. При пожаре ни в коем случае не открывайте форточки и окна.

6. Если вам не удаётся своими силами ликвидировать пожар, выйдите из квартиры, закрыв за собой дверь, и немедленно сообщите о пожаре соседям и жильцам выше-ниже находящихся квартир.

7. Встретьте пожарных и проведите их к месту пожара.

8. При высокой температуре, сильной задымлённости необходимо передвигаться ползком, так как температура у пола значительно ниже и больше кислорода.

9. При невозможности эвакуироваться из квартиры через лестничную площадку, когда пути эвакуация отрезаны, необходимо выйти на балкон, закрыв за собою дверь, и звать на помощь прохожих.

Вывод по разделу

В заключении данного раздела можно сделать вывод, что помещение, в котором производится разработка алгоритмического и программного обеспечения кластеризации научных текстов по тематикам удовлетворяет требованиям и нормам, представленным в соответствующих документах, что способствует комфортному и безопасному выполнению работы. Также было описано влияние процесса разработки на окружающую среду и меры, необходимые для уменьшения влияния негативных факторов на организм человека и окружающую среду.

Заключение

В результате выполнения выпускной квалификационной работы было разработано алгоритмическое и программное обеспечение кластеризации научных текстов по тематикам.

Были разработаны функции извлечения информации о научных публикациях с помощью сервиса Google Scholar, подготовки полученной информации к кластеризации, также реализована кластеризация публикаций на основе заголовков и аннотаций методом К средних.

В ходе разработки был проведён обзор выбранного алгоритма кластеризации, включая необходимые средства предобработки текста, а также анализ существующих средств разработки. Далее был спроектирован бэкенд приложения для кластеризации научных публикаций по тематикам, также была разработана система хранения публикаций и информации о кластерах в базе данных. Была выполнена программная реализация веб-приложения, после чего были описаны результаты работы.

В дальнейшем, использование разработанного веб-приложения позволит упростить и ускорить поиск информации на первичном этапе анализа предметной области, тем самым позволяя более эффективно использовать время и ресурсы пользователя.

На этапе финансового анализа были выявлены конкурентные черты разработки собственного решения, бюджет и сроки реализации.

На этапе анализа данных социальной ответственности было отмечено отсутствие нарушений при выполнении выпускной квалификационной работы по различным аспектам в области безопасности.

Список использованных источников

1. Google Scholar | Wikipedia [Электронный ресурс]. 2021// URL: https://en.wikipedia.org/wiki/Google_Scholar (дата обращения: 06.06.2021)
2. A clustering approach for topic filtering within systematic literature reviews [Электронный ресурс]. 2021 // URL: https://www.sciencedirect.com/science/article/pii/S2215016120300510 (дата обращения: 06.06.2021)
3. TF-IDF с примерами кода: просто и понятно [Электронный ресурс]. 2021 // URL: http://nlpx.net/archives/57 (дата обращения: 06.06.2021)
4. Statistics: Cluster Analysis [Электронный ресурс]. 2007 // URL: http://www.statstutor.ac.uk/resources/uploaded/clusteranalysis.pdf (дата обращения: 06.06.2021)
5. Метод k-средних | Wikipedia [Электронный ресурс]. 2021// URL: https://ru.wikipedia.org/wiki/Метод_k-средних (дата обращения: 06.06.2021)
6. Tutorial: How to determine the optimal number of clusters for k-means clustering [Электронный ресурс]. 2021 // URL: https://blog.cambridgespark.com/how-to-determine-the-optimal-number-of-clusters-for-k-means-clustering-14f27070048f (дата обращения: 06.06.2021)
7. PostgreSQL или MySQL: какая из этих реляционных СУБД лучше впишется в ваш проект [Электронный ресурс]. 2021 // URL: https://www.lucidchart.com/pages/ru/диаграмма-dfd (дата обращения: 06.06.2021)
8. Урок по диаграммам DFD | Lucidchart [Электронный ресурс]. 2021// URL: https://en.wikipedia.org/wiki/Google_Scholar (дата обращения: 06.06.2021)
9. ГОСТ 12.2.032-78 Система стандартов безопасности труда (ССБТ). Рабочее место при выполнении работ сидя. Общие эргономические требования // Электронный фонд правовой и нормативно-технической документации [Электронный ресурс]. 2021. URL: http://docs.cntd.ru/document/1200003913 (дата обращения: 06.06.2021)

10. СанПиН 1.2.3685-21 "Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания" // Электронный фонд правовой и нормативно-технической документации [Электронный ресурс]. 2020. URL: https://docs.cntd.ru/document/573500115 (дата обращения: 06.06.2021)

Приложение А

Таблица А.1 – user

id	Идентификатор пользователя	integer
password	Пароль пользователя	varchar(128)
username	Имя пользователя	varchar(150)
Email	Почтовый адрес пользователя	varchar(254)
is_staff	Является ли пользователь модератором	bool
is_active	Активность аккаунта	bool
date_joined	Дата создания аккаунта	datetime
is_superuser	Является ли пользователь администратором	bool

Таблица А.2 – publication

id	Идентификатор публикации	integer
title	Заголовок публикации	varchar(200)
abstract	Аннотация публикации	text
authors	Список авторов публикации	text
num_citations	Количество цитирований	integer
pub_url	Ссылка на публикацию	text
pub_year	Дата публикации	integer

Таблица А.3 – cluster

id	Идентификатор кластера	integer
user_id	Идентификатор пользователя	integer
name	Название кластера	varchar(200)

Таблица А.4 – background_task

id	Идентификатор задачи	integer
user_id	Идентификатор пользователя	integer
task_name	Название задачи	varchar(190)
verbose_name	Читаемое имя задачи	varchar(255)
run_at	Время запуска задачи	datetime
repeat	Количество повторений	integer
repeat_until	Дата окончания повторения	datetime
attempts	Текущая попытка	integer
last_error	Последняя ошибка выполнения	text
is_finished	Закончено ли выполнение	bool

Приложение Б

Scholar_parse.py

```
def get_publications(input_line, logger, mode="None", single_counter=0):
    old_pubs = Publication.objects.values('title', 'pub_url')
    pg = ProxyGenerator()
    free_failure_counter = 0
    if mode == "Free":
        try:
            pg.FreeProxies()
            logger.debug("Используем FreeProxies")
            scholarly.use_proxy(pg)
            logger.debug("Первичная настройка прокси завершена.")
        except Exception as e:
            logger.debug(e)
    elif mode == "Single":
        proxies = get_proxies()
        pg.SingleProxy(http=proxies[single_counter],
https=proxies[single_counter])
        scholarly.use_proxy(pg)
    elif mode == "None":
        pass

    output = []
    while True:
        try:
            search_results = scholarly.search_pubs(input_line)
            break
        except Exception as e:
            logger.debug(e)
            if mode == "Free" and free_failure_counter <= 5:
                free_failure_counter += 1
                try:
                    pg.FreeProxies()
                    scholarly.use_proxy(pg)
                except Exception as e:
                    logger.debug(e)
            elif mode == "Single":
                single_counter += 1
                mode = "Single"
                pg.SingleProxy(http=proxies[single_counter],
https=proxies[single_counter])
                scholarly.use_proxy(pg)
            else:
                return False
    i = 0
    while i < 400:
        try:
            duplicate = False
            item = next(search_results)
            title = item.get('bib').get('title')
            pub_irl = item.get('pub_url')
            for item in old_pubs:
```

```
                if pub_irl == item.get('pub_url'):
                    duplicate = True
                    logger.info('Duplicate')
                    break
            if duplicate == False:
                output.append(item)
            logger.debug(i)
            i += 1
        except Exception as e:
            logger.debug(e)
            if mode == "Free" and free_failure_counter <= 5:
                free_failure_counter += 1
                try:
                    pg.FreeProxies()
                    scholarly.use_proxy(pg)
                except Exception as e:
                    logger.debug(e)
            else:
                if single_counter < len(proxies):
                    single_counter += 1
                    mode = "Single"
                    pg.SingleProxy(http=proxies[single_counter],
https=proxies[single_counter])
                    scholarly.use_proxy(pg)
                else:
                    return output
    return output


def get_proxies():  # получением списка прокси
    url = 'https://free-proxy-list.net/'
    headers = {"Accept-Language": "en-US, en;q=0.5"}
    response = get(url)
    html_soup = BeautifulSoup(response.text, 'html.parser')
    body = html_soup.find("table").find("tbody").find_all("tr")
    proxies = []
    for i in range(0, len(body)):
        c = body[i].findAll('td')[0].text + ":" + body[i].findAll('td')[1].text
        proxies.append(c)
    return proxies
```

Tasks.py

```
def add_to_db(input_line):
    logger.debug("Стартуем")
    context = get_publications(input_line, logger)
    logger.debug("получили публики")
    #progress_recorder = ProgressRecorder(self)
    for item in context:
        try:
            author = item.get('bib').get('author')
        except:
            author = 'Empty'
        title = item.get('bib').get('title')
        try:
```

```
            pub_year = item.get('bib').get('pub_year')
        except:
            pub_year = 'Empty'
        try:
            abstract = item.get('bib').get('abstract')
        except:
            abstract = 'Empty'
        try:
            pub_url = item.get('pub_url')
        except:
            pub_url = 'Empty'
        try:
            num_citations = item.get('num_citations')
        except:
            num_citations = 'Empty'
        Publication.objects.create(title=title, authors=author,
pub_year=pub_year, abstract=abstract,
                                   pub_url=pub_url,
num_citations=num_citations)
```

Приложение В

Таблица В.1 – Планирование основной заработной платы

Название работы	Длительность работ в календарных днях		Заработная плата, приходящаяся на один чел.- день, руб.		Всего заработная плата по тарифу (окладам)	
	Разработчик	Руководитель	Разработчик	Руководитель	Разработчик	Руководитель
Составление и утверждение технического задания	0	4	1327,6	1991	0	7964
Определение целей исследования	0	4	1327,6	1991	0	7964
Составление календарного плана	4	0	1327,6	1991	5308	0
Составление диаграмм	3	0	1327,6	1991	3982,8	0
Определение средств разработки	3	0	1327,6	1991	3982,8	0
Анализ подходящих алгоритмов для обработки результатов	4	0	1327,6	1991	5308	0
Разработка макетов приложения	4	0	1327,6	1991	5308	0

Продолжение

Программирование, отладка приложения	10	0	1327,6	1991	13276	0
Тестирование	7	0	1327,6	1991	9293,2	0
Развертывание ИС в сети	4	0	1327,6	1991	5308	0
Составление пояснительной записки	13	0	1327,6	1991	17258	0
Итого:					69035	15928