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 _09.04.02 Информационные системы и технологии_
Отделение школы (НОЦ) _Отделение информационных технологий_

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Алгоритмическое и программное обеспечение анализа схожести наборов медицинских данных по их описанию и содержимому

УДК 004.62:004.732:61

Студент

Группа	**ФИО**	**Подпись**	**Дата**
8ИМ9М	Соколовский Дмитрий Евгеньевич		

Руководитель ВКР

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент ОИТ	Аксенов С.В.	к.т.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент ОСГН	Гончарова Н.А.	к.э.н.		

По разделу «Социальная ответственность»

Должность	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент ООД ШБИП	Сечин А.А.	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	**ФИО**	**Ученая степень, звание**	**Подпись**	**Дата**
Доцент ОИТ	Савельев А.О.	к.т.н.		

Томск – 2021 г.

ЗАПЛАНИРОВАННЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ

Код результата	Результат обучения
Общие по направлению подготовки (специальности)	
Р1	Применять глубокие математические и профессиональные знания основ построения информационных технологий и систем, достаточные для решения научных и профессиональных задач производства. Знать современные проблемы и методы прикладной информатики и научно-технического развития информационно-коммуникационных технологий.
Р2	Ставить и решать инновационные задачи анализа с использованием глубоких фундаментальных и специальных знаний, аналитических методов и сложных моделей в условиях неопределенности и определять методы и средства их эффективного решения, нормализовывать задачи прикладной области. Применять полученные знания для решения нечетко определенных профессиональных задач, стоящих в области внедрения новейших технологий в сфере прикладной информатики.
Р3	Выполнять инновационные проекты с применением глубоких и принципиальных знаний, оригинальных методов проектирования для достижения новых результатов, обеспечивающих конкурентные преимущества в условиях жестких экономических, экологических, социальных и других ограничений. Применять современные методы и инструментальные средства прикладной информатики для автоматизации и информатизации решения прикладных задач различных классов и создания ИС.
Р4	Проводить инновационные профессиональные исследования, включая критический анализ данных из мировых информационных ресурсов, сложный эксперимент, формулировку выводов в условиях неоднозначности с применением глубоких и принципиальных знаний и оригинальных методов для достижения требуемых результатов. Способен проводить маркетинговый анализ ИКТ и вычислительного оборудования для рационального выбора инструментария автоматизации и информатизации прикладных задач.
Р5	Способен организовывать работы по моделированию прикладных ИС и реинжинирингу прикладных и информационных процессов предприятия и организации. Способен управлять проектами по информатизации прикладных задач и созданию ИС предприятий и организаций.
Р6	Способен использовать передовые методы оценки качества, надежности и информационной безопасности ИС в процессе эксплуатации прикладных ИС; использовать международные информационные ресурсы и стандарты в информатизации предприятий и организации; использовать информационные сервисы для автоматизации прикладных и информационных процессов; интегрировать компоненты и сервисы информационных систем.

Код результата	Результат обучения
Универсальные компетенции	
Р8	Использовать глубокие знания по проектному менеджменту для ведения инновационной инженерной деятельности с учетом юридических аспектов защиты интеллектуальной собственности. Способен использовать углубленные знания правовых и этических норм при оценке последствий своей профессиональной деятельности, при разработке и осуществлении социально значимых проектов.
Р9	Активно владеть иностранным языком на уровне, позволяющем работать в иноязычной среде, разрабатывать документацию, презентовать и защищать результаты инновационной инженерной деятельности. Демонстрировать глубокие знания социальных, этических и культурных аспектов инновационной инженерной деятельности, компетентность в вопросах устойчивого развития.
Р10	Эффективно работать индивидуально, в качестве члена и руководителя группы, состоящей из специалистов различных направлений и квалификаций, демонстрировать ответственность за результаты работы и готовность следовать корпоративной культуре организации.
Р11	Самостоятельно учиться и непрерывно повышать квалификацию в течение всего периода профессиональной деятельности.

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

Школа _Инженерная школа информационных технологий и робототехники_
Направление подготовки _09.04.02 Информационные системы и технологии_
Отделение школы (НОЦ) _Отделение информационных технологий_

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Савельев А.О.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

Студенту:

Группа	ФИО
8ИМ9М	Соколовскому Дмитрию Евгеньевичу

Тема работы:

Алгоритмическое и программное обеспечение анализа схожести наборов медицинских данных по их описанию и содержимому	
Утверждена приказом директора (дата, номер)	От 29.04.2021 г. № 119-32/с

Срок сдачи студентом выполненной работы:	15.06.2021 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.).*	Объектом исследования является анализ схожести наборов медицинских данных. Предметом исследования являются алгоритмы исследования схожести наборов медицинских данных, с помощью инструментов машинного обучения. Область применения: возможность применения в различных областях, в частности, в области исследования медицинских наборов данных.

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 литературы по теме исследования 2. Алгоритмическое обеспечение поиска схожести наборов медицинских данных с помощью методов машинного обучения 3. Программное обеспечение поиска схожести наборов медицинских данных с помощью языка программирования python 4. Тестирование разработанного решения
Перечень графического материала	Презентация Microsoft Office PowerPoint

Консультанты по разделам выпускной квалификационной работы

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Гончарова Н.А.
Социальная ответственность	Сечин А.А.

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

Обзор литературы по теме исследования

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	Аксенов С.В.	к.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ9М	Соколовский Дмитрий Евгеньевич		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ИМ9М	Соколовский Дмитрий Евгеньевич

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	Магистратура	Направление/специальность	Информационные системы и технологии

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»: работа с информацией, представленной в российских и иностранных научных публикациях, аналитических материалах, статистических бюллетенях и изданиях, нормативно-правовых документах.	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*Использовать действующие ценники и договорные цены на потребленные материальные и информационные ресурсы, а также указанную в МУ величину тарифа на эл. энергию.*
2. *Нормы и нормативы расходования ресурсов*	—
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Действующие ставки единого социального налога и НДС (см. МУ).*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*Проведение предпроектного анализа*
2. *Разработка устава научно-технического проекта*	—
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*Построение плана-графика выполнения НИ, составление соответствующей сметы затрат НИ, оценка рисков.*
4. *Определение ресурсной, финансовой, экономической эффективности*	*Определение экономической эффективности НИ.*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. График проведения и бюджет НИ *2. Перечень работ и продолжительность их выполнения* *3. Оценка экономической эффективности НИ*	

Дата выдачи задания для раздела по линейному графику	22.02.2021

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Гончарова Наталья Александровна	К.э.н		22.02.2021

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ9М	Соколовский Дмитрий Евгеньевич		22.02.2021

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ИМ9М	Соколовский Дмитрий Евгеньевич

Школа	Инженерная школа информационных технологий и робототехники	**Отделение (НОЦ)**	Информационных технологий
Уровень образования	Магистратура	**Направление/специальность**	09.03.02 Информационные системы и технологии

Тема ВКР:

Алгоритмическое и программное обеспечение анализа схожести наборов медицинских данных по их описанию и содержимому	
Исходные данные к разделу «Социальная ответственность»:	
1. Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика, рабочая зона) и области его применения	Целью выпускной квалификационной работы является алгоритмическое и программное обеспечение анализа схожести наборов медицинских данных. Актуальность работы заключается в том, что созданное алгоритмическое и программное обеспечение, могут использоваться не только исследователями в области медицинских или иных данных, но в перспективе и медицинскими учреждениями для устранения разнообразности наборов медицинских данных, путем анализа их схожести, тем самым упрощая анализ и восприятие различной медицинской информации в целом.
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.	− Специальные правовые нормы трудового законодательства при работе с компьютером и орг. техникой (Трудовой кодекс РФ, СанПиН 2.2.2/2.4.1340-03 Гигиенические требования к персональным электронно-вычислительным машинам и организации работы); – СанПиН 2.2.4.548–96. Гигиенические требования к микроклимату производственных помещений; − требования к организации рабочих мест пользователей (ГОСТ 12.2.032-78 «ССБТ.

	Рабочее место при выполнении работ сидя. Общие эргономические требования», – ГОСТ 12.1.003–83 ССБТ. Шум – Влияние реализации проекта на организацию рабочего места медицинского сотрудника, как пользователя ПК.
2. Производственная безопасность: 2.1. Анализ выявленных вредных и опасных факторов 2.2. Обоснование мероприятий по снижению воздействия	Вредные производственные факторы, создаваемые объектом исследования: - Электромагнитные излучения. Опасные производственные факторы, создаваемые объектом исследования: - Поражение электрическим током. Вредные производственные факторы, возникающие на рабочем месте: - Микроклимат; - Освещенность; - Монотонность работы. Опасные производственные факторы, возникающие на рабочем месте: - Возникновение пожара.
3. Экологическая безопасность:	– При разработке опасность представляет утилизация компонентов ПЭВМ
4. Безопасность в чрезвычайных ситуациях:	– Перечень возможных ЧС: пожары и взрывы; обрушение зданий; ураганы, ливни, заморозки; наводнения, паводки; эпидемии Наиболее типичная ЧС: пожары – Мероприятия профилактике и недопущению пожаров согласно нормативным документам: НПБ 105-03; ППБ 01–03.

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Сечин Андрей Александрович	К.т.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ИМ9М	Соколовский Дмитрий Евгеньевич		

РЕФЕРАТ

Выпускная квалификационная работа содержит 109 страниц, 22 рисунка, 23 таблицы, 30 источников, 2 приложения.

Ключевые слова: машинное обучение, наборы медицинских данных, телемедицина, электронные медицинские записи, дерево решений, описания, python, medical data, machine learning, telemedicine, LDA, decisionTreeClassifier, kaggle, description, electronic medical records, datasets.

Объектом исследования является анализ схожести наборов медицинских данных.

Предметом исследования являются алгоритмы исследования схожести наборов медицинских данных, с помощью инструментов машинного обучения.

Цель работы – повышение скорости поиска наборов данных по медицинской тематике на медицинских сервисах.

В первой главе исследования приведен обзор литературы по тематике исследования, выбраны способы и методы, с помощью которых будет проводиться исследование и разработка. Во второй главе приведено алгоритмическое обеспечение поиска схожести наборов медицинских данных с помощью методов машинного обучения. Предложена последовательность действие для создания программного обеспечения. Проведена оценка эффективности работы алгоритма. В третьей главе проведен процесс программного обеспечения поиска схожести наборов медицинских данных с помощью языка программирования python. В четвертой главе проведен процесс успешного тестирования программного кода. В результате исследования были созданы алгоритмы анализа схожести наборов медицинских данных по их описанию и содержимому, разработан программный код по данным алгоритмам. Осуществлена визуализация программного кода в программную оболочку.

Область применения: возможность применения в различных областях, в частности, в области исследования медицинских наборов данных.

Экономическая эффективность работы заключается в малом количестве конкурентов, в данном направлении исследований.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ, НОРМАТИВНЫЕ ССЫЛКИ

В данной работе применены следующие термины с соответствующими определениями:

Kaggle: платформа машинного обучения, управляемое сообществом. Он содержит много учебников, которые охватывают сотни реальных проблем ML. Конечно, качество данных может варьироваться, но все они полностью свободны. Также можно загрузить собственную базу данных в библиотеку.

LDA: метод моделирования темы был предложен Дэвидом Блей, Эндрю Ыном и Майклом Джорданом в 2003 году. LDA относится к семейству генеративных вероятностных моделей, в которых темы представлены вероятностями возникновения каждого слова в каждом наборе.

В работе использованы следующие обозначения и сокращения:

LDA – Latent Dirichlet Allocation;

AWS – Amazon Web Services;

NLTK – Natural Language Toolkit;

ЭМР – электронные медицинские записи;

МИС – медицинская информационная система.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Информатизация здравоохранения привела к созданию большого количества медицинских информационных систем (МИС) в рамках нацпроекта «Здоровье» и реализации региональных пилотных проектов. Ежегодно подводятся итоги конкурса «Лучшая медицинская информационная система».

Данные системы собирают и хранят разнообразную медицинскую информацию о пациентах и процессе их лечения. После анализа нескольких систем оказалось, что нет готовых решений для работы с нестандартизированными медицинскими наборами данных или наборами данных с отсутствующими элементами, которые зачастую используются при обмене данными между этими МИС. Данная проблема в настоящее время является актуальной.

Актуальность работы заключается в том, что созданный алгоритм и программное обеспечение, которое будет улучшаться и дополняться, может использоваться не только исследователями в области медицинских или иных данных, но в перспективе и в медицинских учреждениях для устранения разнообразности наборов медицинских данных, путем анализа их схожести, тем самым упрощая анализ и восприятие различной медицинской информации в целом.

Основной целью данной работы является повышение скорости поиска наборов данных по медицинской тематике на медицинских сервисах.

В ходе работы был разработан алгоритм и программное обеспечение для облегчения анализа схожести наборов медицинских данных, так же был визуализирован модуль сравнения на схожесть, автоматизирован способ сбора данных со специализированных сайтов, который в дальнейшем будет расширяться и интегрироваться с различными базами данными медицинских веществ. Приведен план по дальнейшей разработке и улучшению функционала.

1 Обзор литературы по теме исследования

Многие медицинские учреждения и организации не используют концептуальный подход к организации и управлению качеством данных, особенно в долгосрочной перспективе. Значение медицинских записей и основанных на них данных растет с течением времени. Даже внедрение электронных медицинских записей (ЭМР) не упростило обработку данных в режиме реального времени в надлежащей степени, поскольку функциональность используемого программного обеспечения весьма ограничена.

Вот основные проблемы с обработкой медицинских данных:

- различные уровни качества электронных медицинских записей;
- отсутствие совместимости, а также сложность клинических систем;
- сложность процесса сбора, поиска и анализа данных;
- необходимость обработки неполных или отсутствующих данных;
- охват и выборка данных;
- нормативные требования и бюрократические процессы.

Тему нашего исследования можно косвенно отнести к решению многих проблем, названных выше, в частности к необходимости обработки неполных или отсутствующих данных, а также сложности процесса сбора, поиска и анализа данных.

1.1 Специализированные сайты с наборами данных

На международном уровне наиболее популярными считаются следующие сайты:

- Kaggle;

Kaggle ежедневно обновляется энтузиастами и содержит одну из крупнейших библиотек баз данных в интернете [1].

Kaggle — это платформа машинного обучения, управляемое сообществом. Он содержит много учебников, которые охватывают сотни

реальных проблем ML. Конечно, качество данных может варьироваться, но все они полностью свободны. Также можно загрузить собственную базу данных в библиотеку.

Имеется много ресурсов для обучения информатике, от Datacamp до Udacity, все для изучения науки о данных. Но если вы тот человек, который любит учиться, делая, то Kaggle, возможно, лучшая платформа для улучшения ваших навыков через практические исследовательские проекты.

Kaggle, который позиционирует себя как "ваш дом для науки о данных", первоначально был сайт конкурса машинного обучения, но ресурсы науки о данных теперь можно найти там же. Стоит отметить несколько основных особенностей Kaggle:

Наборы данных: Многие наборы данных различных типов и размеров, которые можно скачать бесплатно. Здесь вы можете найти интересные данные, чтобы узнать или проверить свои навыки моделирования.

Kaggle (The Kaggle Team 2018) является платформой для прогностический моделирования и аналитики соревнований, где участники соревнуются, чтобы произвести лучшую прогностический модель для данного набора данных. Он хорошо известен своими конкурсами (например, Родос 2011), некоторые из которых приходят с богатыми денежными призами (например, Говард 2013). Есть также учебные соревнования (Agarwal 2018), призванные помочь новичкам отточить свои навыки сбора данных. Победители, как правило, должны делиться своим кодом, а иногда и вновь возникающие алгоритмы внедряются в широкое сообщество, например, глубокие нейронные сети (Hinton и Dahl 2012) и XGBoost (Chen and Guestrin 2016).

В 2015 году Kaggle InClass был представлен в качестве платформы самообслуживания для проведения соревнований. Эти соревнования могут быть частными, ограничены членами университетского курса, и легко настроить. Это возможность для преподавателей предоставить возможность для студентов объективно проверить свои изучения прогностический моделирования. Как соревнование, с независимой четкой метрикой производительности, наряду с

динамической доской лидеров, студенты могут видеть, как их прогнозы моделей соотносятся с моделями, производимыми другими студентами. Возможность сделать несколько представлений в течение нескольких недель позволяет им опробовать подходы к улучшению своих моделей. В этой статье рассматриваются образовательные преимущества проведения конкурсов прогностический моделирования в классе по производительности, вовлеченности и интересам [2].

- Поиск наборов данных от Google;

Dataset Search является надежным источником информации для исследований [3]. В нем все наборы данных отсортированы по:

- актуальность;
- формат файла;
- тип лицензии;
- тема;
- последнее обновление.

Базы данных загружаются здесь различными международными организациями, такими как Всемирная организация здравоохранения, Statista и Гарвард.

- Реестр открытых данных на AWS;

В реестре открытых данных на AWS любой желающий может поделиться пакетом данных или найти тот, который им нужен [4]. С помощью инструментов Amazon Data Analytics вы можете проводить исследования на основе данных, которые вы найдете. Создатели этих баз данных включают Facebook данные для добра, НАСА космического закона соглашения, и Космический телескоп Институт космических исследований.

- Открытые наборы данных Microsoft Azure;

Открытые наборы данных Azure регулярно обновляются и доступны разработчикам приложений и исследователям [5]. Они содержат данные правительства США, другие статистические и научные данные, а также информацию из онлайн-сервисов, которые корпорация Майкрософт собирает о своих пользователях.

Кроме того, Azure предлагает пользователям набор инструментов, которые помогут им создавать свои собственные облачные базы данных, перенести рабочие нагрузки на Azure при сохранении полной совместимости серверов S'L и создавать мобильные и веб-приложения, управляемые данными.

- R / наборы данных;

В SubReddit DataSet любой желающий может опубликовать базы данных с открытым исходным кодом [6]. Посмотрите туда, чтобы найти прохладный набор данных и сделать некоторые интересные исследования с ним.

- Репозиторий машинного обучения UCI;

UCI предлагает более 500 различных наборов данных, которые охватывают такие темы, как банковский маркетинг, оценка автомобиля, диагностика рака легких, и многое другое [7]. Вы можете сортировать пакеты данных по:

- стандартные задачи;
- типы данных;
- повторное использование;
- предмет;
- Библиотеки CMU.

Университет Карнеги-Меллона имеет свою собственную коллекцию общедоступных наборов данных, которые можно использовать для исследований. Там вы найдете подробные базы данных по американской

культуре, музыке и истории, которые ни один другой агрегатор не предоставляет [8].

- Открытые базы данных на Github.

Это большой набор наборов данных с открытым исходным кодом, разделенных на отрасли [9].

1.2 Методы тематического моделирования

Моделирование тем является неконтролируемым методом машинного обучения для определения тем в сборнике документов. Чем отличие от обычной кластеризации? Цель кластеризации состоит в том, чтобы разделить тело документов на группы, тогда цель тематического моделирования состоит в том, чтобы выделить основные темы из набора заявлений. Самое главное, кластеризация является дедуктивным, и тема моделирования является индуктивным.

Существуют различные методы моделирования темы [10]:

1. Латентное Размещение Дирихле (LDA) [11].

2. Латентный семантический анализ (LSA) [12].

3. Неотрицательно матричное разложение (NNMF) [13].

Для наших исследований метод скрытого развертывания Dirichlet является отличным решением.

Этот метод моделирования темы был предложен Дэвидом Блей, Эндрю Нг и Майклом Джорданом в 2003 году. LDA относится к семейству генеративных вероятностных моделей, в них представлены различные темы, которые обуславливаются вероятностями возникновения слов в каждом наборе данных, в том числе медицинских. Также наборы могут содержать комбинации тем. Уникальной особенностью моделей LDA является то, что темы не должны быть разными, а слова могут отображаться в нескольких темах; это придает

определенную двусмысленность обсуждаемым темам, что может быть полезно для того, чтобы справиться с гибкостью языка. Рисунок 1 иллюстрирует этот метод.

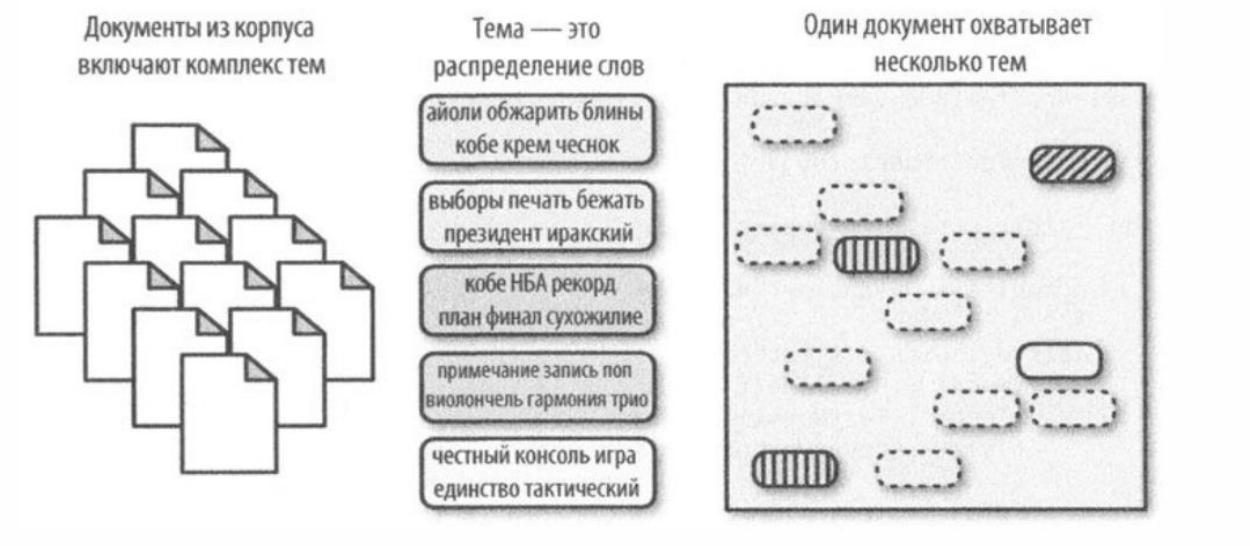

Рисунок 1 - Метод распределения Latent Dirichlet [14]

Блей (Blei) и его коллеги обнаружили, что Latent Dirichlet, семейство непрерывных дистрибутивов (способ измерения группировки по распределениям), является удобным способом для выявления тем, которые появляются в корпусе, а также появляются в различных комбинациях в каждом документе в корпусе. Латентное размещение дирихле (LDA) дает нам определенное слово, с его помощью можно попытаться определить наиболее вероятную тему, распределение слов в каждой теме, различные сочетания их в наборах данных. Чтобы использовать методы моделирования тем в приложении, необходимо создать пользовательский конвейер, который экстраполирует темы из неструктурированных текстовых данных и способ сохранить лучшую модель.

Конвейер моделирования темы будет выглядеть так:

1. Загрузка корпуса

2. Предварительная переработка текста

2.1 Удаление стоп-слов

2.2 Удалить знаки препинания

2.3 Лемматизация слов

3. Создание словаря

4. Выбор оптимального количества тем

С началом и неконтролируемым ростом цифровых документов, Автоматическая тема извлечения считается активной темой исследования. Для обработки документов в литературе были представлены различные алгоритмы для извлечения темы с использованием методов моделирования распределения и классификации. Среди различных методов моделирования извлечения темы Latent Dirichlet Распределение (LDA) является одним из важных алгоритмов для идентификации темы. Несмотря на то, что LDA является популярным методом для идентификации темы, он превращается в трудности в определении параметров модели и, страдает с поиском степени сходства и семантической обработки. Для решения этих задач предлагается новый метод автоматического извлечения темы. Соответственно, этот метод называется, семантический скрытый dirichlet распределение (SLDA) предлагается путем расширения LDA в семантической образом. Включены новые математические вычисления, в которых параметры модели оцениваются с использованием новой степени членства в процессе Semantic Latent Dirichlet вместе с семантическим показателем сходства. Эксперименты проводятся с двумя различными базами данных, и он отметил, что SLDA превзошел, показывая лучше по сравнению с существующими LDA в коэффициенте Jaccord Coefficient [15].

Реализация с gensim [16].

Модели Gensim имеют более настраиваемые параметры, чем scikit-learn. Gensim был первоначально разработан как библиотека моделирования темы. Библиотеки, используемые в этой реализации.

- gensim - содержит все алгоритмы моделированиятем;
- pandas - необходимые для работы с языковымкорпусом;
- nltk - содержит алгоритмы лемматизации и словарь стоп-слов;
- pyLDAvis - плагин для визуализации моделиLDA;
- matplotlib - библиотека визуализации.

1.3 Машинное обучение и анализ данных

Тренировать нейронную сеть можно по-разному: с учителем, без учителя, с подкреплением. Но как выбрать оптимальный алгоритм и чем они отличаются? Существует несколько способов сбора мебели ИКЕА. Каждый из них ведет к собранному дивану или стулу. Но в зависимости от мебели и ее компонентов, один метод будет более разумным, чем другие.

У вас есть инструкция и все части, которые вам нужны? Просто следуй инструкциям. Ну, как там? Вы можете выбросить руководство и работать по своему усмотрению. Но если вы путаете процедуру, это до вас, чтобы решить, что делать с этой кучей деревянных болтов и досок.

То же самое с глубоким обучением. Разработчик предпочтет алгоритм с определенным методом обучения, учитывая тип данных и поставленную задачу. На рисунке 2 показан результат тренировки нейронной сети - кластеризация изображений.

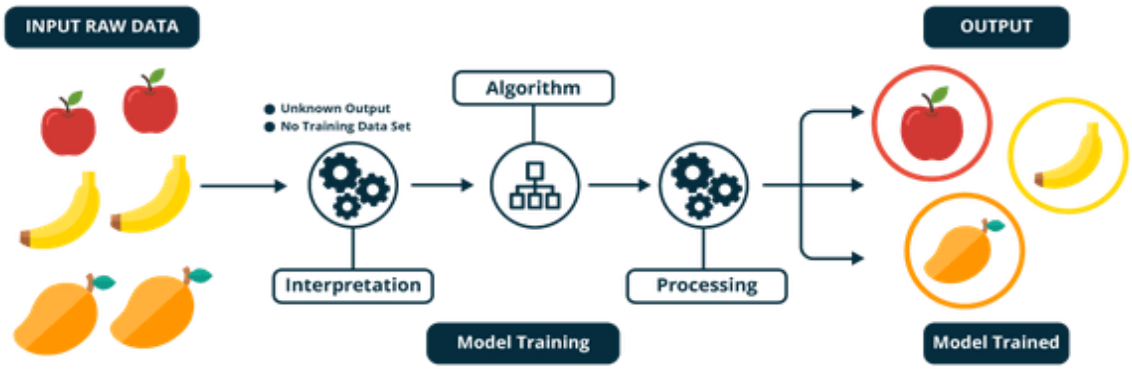

Рисунок 2 - Результат обучения нейронной сети [17]

Результатом обучения нейронной сети является кластеризация изображений.

При контролируемом обучении нейронная сеть тренируется на помеченном наборе данных и предсказывает ответы, которые используются для оценки точности алгоритма на учебных данных. При неконтролируемом обучении модель использует неразмеченные данные, из которых алгоритм самостоятельно пытается извлечь функции и зависимости.

Частичное обучение учителей – это что-то между ними. Он использует небольшое количество помеченных данных и большой набор неразмеченных данных. Укрепление обучения обучает алгоритм с помощью системы вознаграждения. Агент получает обратную связь в виде вознаграждений за то, что поступает правильно. Аналогичным образом выгоняют животных.

Для каждого метода обучения рассмотрим примеры данных и подходящих для него задач.

1.3.1 Обучение с учителем

Контролируемое обучение требует полного набора помеченных данных для обучения модели на всех этапах ее разработки.

Наличие полностью помеченного набора данных означает, что каждый пример в учебном наборе соответствует ответу, который должен получить алгоритм. Таким образом, помеченный набор данных цветочных фотографий будет обучать нейронной сети, где изображены розы, ромашки или нарциссы. Когда сеть получает новую фотографию, она сравнивает ее с примерами из набора учебных данных для прогнозирования ответа.

В основном контролируемое обучение используется для решения двух типов задач: классификации и регрессии.

При проблемах классификации алгоритм предсказывает дискретные значения, соответствующие числам классов, к которым принадлежат объекты. В наборе учебных данных с фотографиями животных каждое изображение будет иметь соответствующую метку - "кошка", "коала" или "черепаха". Качество

алгоритма оценивается по тому, насколько точно он может правильно классифицировать новые фотографии с коалами и черепахами.

С другой стороны, задачи регрессии связаны с непрерывными данными. Один из примеров, линейная регрессия, вычисляет ожидаемое значение переменной y с учетом конкретных значений x.

Более утилитарные задачи машинного обучения включают в себя множество переменных. Например, нейронная сеть, которая предсказывает цену квартиры в Сан-Франциско в зависимости от ее района, местоположения и доступности общественного транспорта. Алгоритм выполняет работу эксперта, который рассчитывает цену квартиры на основе тех же данных.

Таким образом, контролируемое обучение наиболее подходит для задач, когда есть впечатляющий набор надежных данных для обучения алгоритма. Но это не всегда так. Недостаток данных является наиболее распространенной проблемой в машинном обучении.

1.3.2 Обучение без учителя

Идеально помечены и чистые данные не легко найти. Поэтому иногда перед алгоритмом стоит задача найти ранее неизвестные ответы. Здесь происходит обучение без учителя.

В неконтролируемом обучении модель имеет набор данных, и нет четкого указания на то, что с ней делать. Нейронная сеть пытается самостоятельно найти корреляции в данных, извлекая полезные функции и анализируя их.

Кластеризация. Даже без особых знаний получив фотографии животных, мы сможем их разделить на группы похожих животных. Именно в этом примере и происходит кластеризация - наиболее распространенная задача для неконтролируемого обучения. Алгоритм собирает похожие данные, находит общие черты и сгруппировывает их вместе.

Обнаружение аномалий. Банковские учреждения могут отслеживать подозрительные операции со счетами клиентов. Например, будет

подозрительным, если банковская карта, будет использоваться в разных странах, в одно и тоже время. Аналогичным образом, неконтролируемое обучение используется для поиска выбросов в данных.

Ассоциации. Выберите телефон, защитное стекло, наушники из интернет-магазина, и сайт будет рекомендовать вам добавить чехол или аксессуары для данной модели. Это и есть пример ассоциаций.

Автоинкодеры. Автоинкодеры берут входные данные, кодируют их, а затем пытаются воссоздать исходные данные из полученного кода. Есть не так много реальных ситуаций, когда используется простой автоинкодер. Но добавить слои и возможности расширяются: с помощью шумных и оригинальных версий изображений для обучения, автоинкодеры могут удалить шум из видео данных, изображений или медицинских сканирований для улучшения качества данных.

При неконтролируемом обучении трудно рассчитать точность алгоритма, поскольку в данных отсутствуют правильные ответы" или метки. Но помеченные данные часто ненадежны или слишком дороги для получения. В таких случаях предоставление модели свободы поиска зависимостей может привести к хорошим результатам.

1.3.3 Обучение с частичным привлечением учителя

Полу контролируемое обучение характеризуется его названием: набор учебных данных содержит как помеченные, так и нет данные. Этот метод особенно полезен, когда трудно извлечь важные функции из данных, или когда это утомительная задача, чтобы разметить все объекты.

Частичное обучение учителей часто используется для решения медицинских проблем, когда небольшое количество помеченных данных может привести к значительному повышению точности.

Этот метод машинного обучения является общим для анализа медицинских изображений, таких как КТ или МРТ. Опытный радиолог может

разметить небольшое подмножество сканирований, которые показывают опухоли и заболевания. Но ручная разметка всех сканирований является слишком трудоемкой и дорогостоящей задачей. Тем не менее, нейронная сеть может извлекать информацию из небольшой доли помеченных данных и повысить точность прогнозирования по сравнению с моделью, которая тренируется исключительно на неразмеченных данных.

Популярный метод обучения, который требует небольшой набор помеченных данных, заключается в использовании генеративной состязательной сети, или GAN [18] [19]. Две нейронные сети состязаются друг с другом в игре (в виде игры с нулевой суммой, где выигрыш одного агента – это потеря другого агента).

Наиболее популярными являются контролируемые методы обучения, когда модель (классификатор машины) построена из корпуса помеченных данных (учебный образец), который затем применяется к новым, неразмеченным текстам.

Традиционные методы машинного обучения такого типа, такие как наивный классификатор Байеса, деревья принятия решений, машины поддержки векторов, логистическая регрессия и т. д.

1.3.4 Выводы по разделу

Проанализировав литературу, по тематике данной работы, были выделены некоторые пункты, по каждому подпункту выбраны методы, платформы и решения, который мы посчитали более эффективными и которые будут использоваться в процессе исследования. В частности, сбор наборов медицинских данных будет производиться из платформы Kaggle. Тематическое моделирование будет использовать метод латентного размещения дирихле, реализованное с помощью gensim, а анализ схожести этих данных будет осуществляться с помощью машинного обучения с учителем, в совокупности с алгоритмическим и программным обеспечением на языке python [20].

2 Алгоритмическое обеспечение поиска схожести наборов медицинских данных с помощью методов машинного обучения

Для начала мы анализируем интервалы каждого элемента, получая значение по пересечению и объединению.

Для этого мы собираем несколько медицинских выборок с различными веществами и вручную добавляем в google colab для дальнейшей работы с ними. Далее мы отсекаем некорректные для обучения столбцы (значения столбцов должны содержать как минимум больше 10 уникальных значений и быть вещественными или целочисленными).

Для обучения модели нужно найти или сгенерировать как можно больше как элементов одинаковых, так и разных.

После получения сгенерированных выборок создается алгоритм функции для сравнения и вывода значений по пересечению и объединению интервалов, алгоритм функции представлен на рисунке 3, обозначения переменных представлено в таблице 1.

Таблица 1 Переменные алгоритма сравнения

Переменная в алгоритме	Обозначение
MinA	Начало первого интервала
MaxA	Конец первого интервала
MinB	Начало второго интервала
MaxB	Конец второго интервала

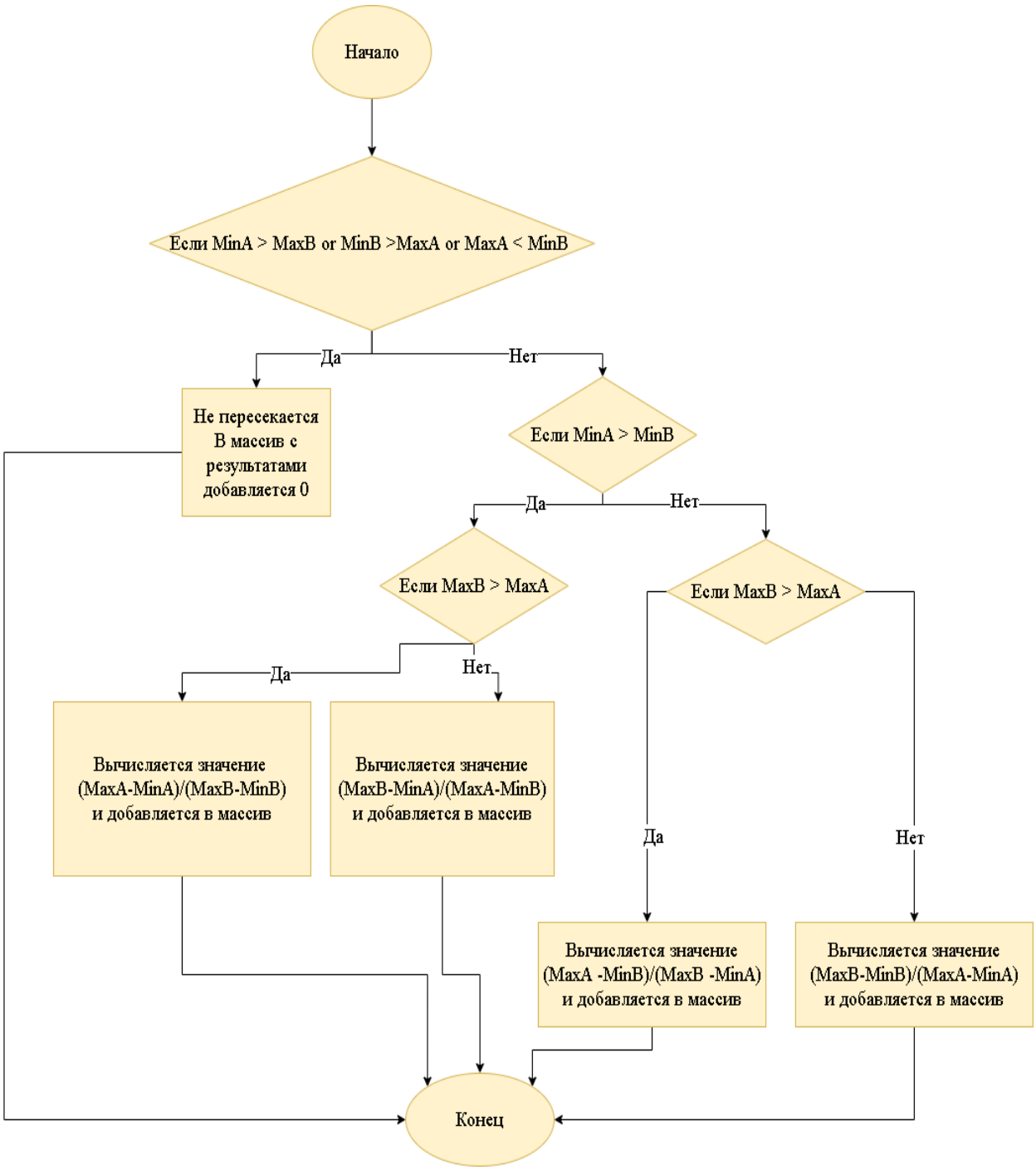

Рисунок 3 – Алгоритм функции сравнения интервалов

Чтобы данный алгоритм заработал и начал выводить нужный нам результат, мы должны создать еще один алгоритм. Он будет получать минимальные и максимальные значения интервалов, которые получены из входных наборов медицинских данных и если они пересекаются, то запускать алгоритм функции. В итоге, с помощью данного алгоритма, представленного на

рисунке 4, мы получим необходимые массивы одинаковых и разных элементов для обучения модели.

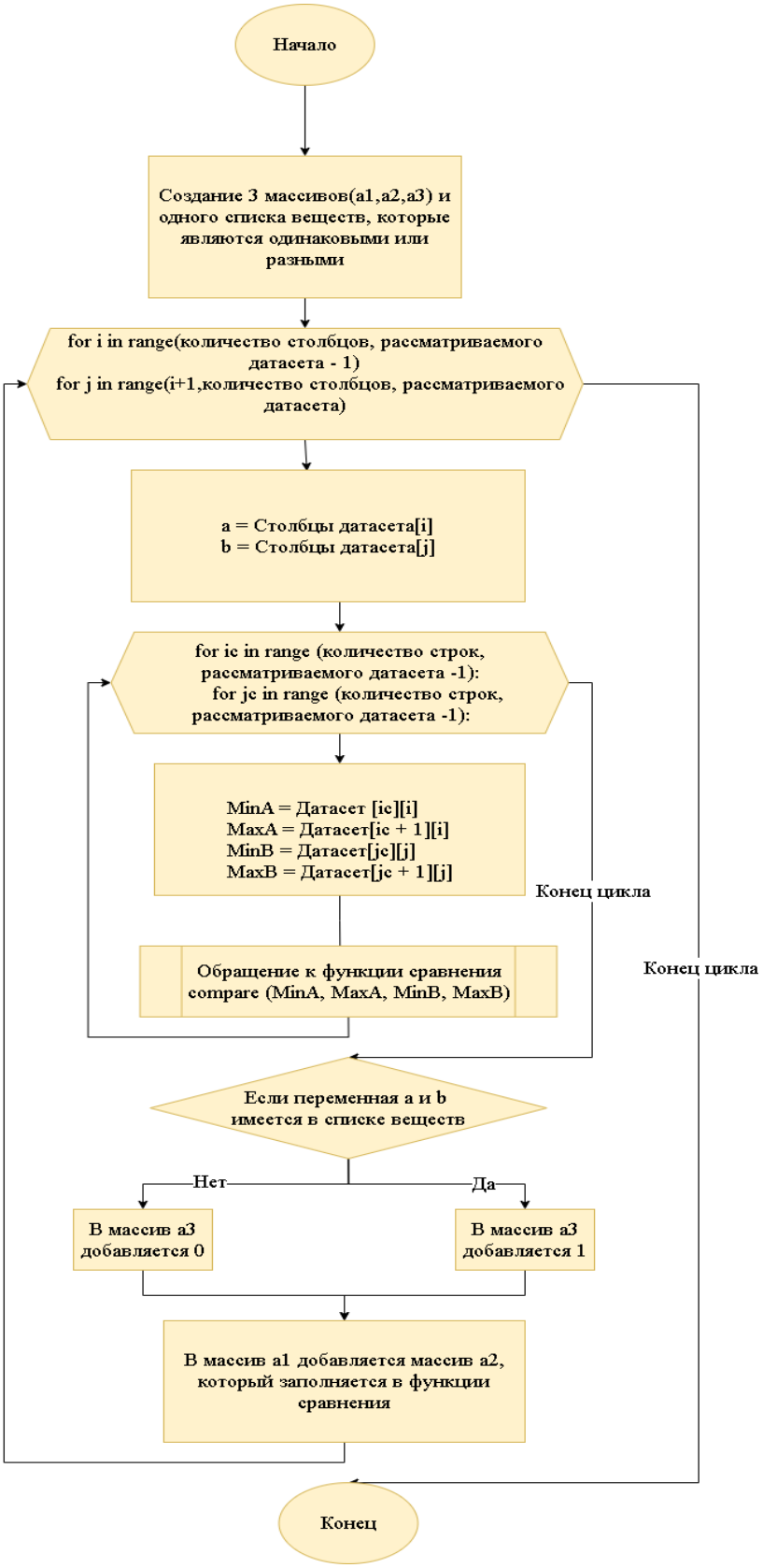

Рисунок 4 – Алгоритм получения массивов значений одинаковых и разных элементов

Для дальнейших исследований мы можем автоматизировать сбор данных с сайта Kaggle, алгоритм сбора представлен на рисунке 5.

Рисунок 5– Алгоритм сбора метаданных и описаний

Получив описания, наборов данных, мы должны их обработать перед обучением lda модели и обучить модель, которая выведет нам тематику по документам. После этого определить доминирующую тему и отобрать

медицинские наборы данных. На рисунке 6 представлен алгоритм, который будет производить данные действия.

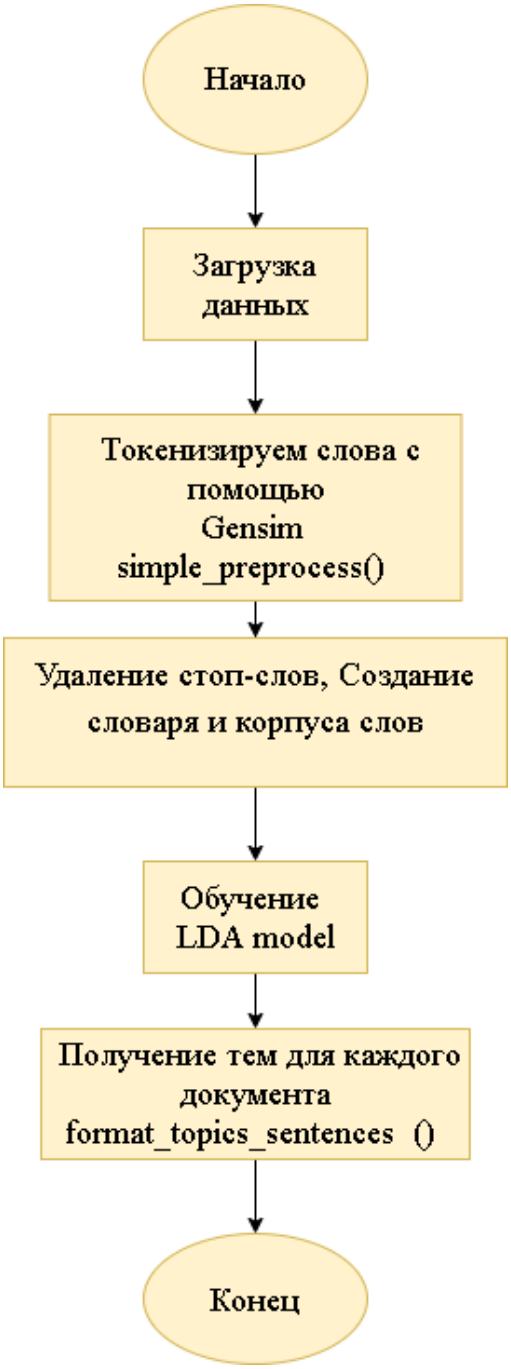

Рисунок 6– Алгоритм предобработки текста для обучения lda модели и получение доминирующей темы

По итогам данного раздела была произведена разработка алгоритмического обеспечения для анализа схожести наборов медицинских данных по их описанию и содержимому, для дальнейшего создания программного обеспечения и получения наглядного результата.

3 Программное обеспечение поиска схожести наборов медицинских данных с помощью языка программирования python

По уже имеющимся алгоритмам, пишем программный код, который будет реализован на языке python. Для начала подключаем нужные для работы библиотеки и модули, основные из которых:

– gensim для тематического моделирование;

– pandas для работы с корпусом наборов данных;

– nltk для лемматизации и формирования стоп-слов;

– pyLDAvis для визуализации модели LDA;

– matplotlib для визуализации;

– seaborn для построения графиков;

– numpy для построения массивов;

– math модуль для работы с числами;

– pickle для сохранения и загрузки сложных объектов;

– numpy для работы с массивами;

– docx для работы с файлами формата docx;

– json для работы с файлами формата json;

– модули для подключения google drive;

– Sklearn для обучения модели.

Для генерации выборок разделяем одну выборку, на несколько частей, сначала одинаковых веществ, а позже и разных.

Программный код функции сравнения представлен в таблице 2.

Таблица 2 Функция сравнения интервалов

```
def compare (MinA, MaxA, MinB, MaxB):

 if (MinA > MaxB) or (MinB > MaxA) or (MaxA < MinB):
  print ('Не пересекается', 0, '|', a, '-', b)
  a2.append(0)
 else:
   if (MinA > MinB):
      if (MaxB > MaxA):
       p = (MaxA-MinA) / (MaxB-MinB)
       print (round(p, 2), '|', a, '-', b)
       a2.append(round(p, 2))
      else:
       p = (MaxB-MinA)/ (MaxA - MinB)
       print (round(p, 2), '|', a, '-', b)
       a2.append(round(p, 2))
   else:
       if (MaxB > MaxA):
        p = (MaxA-MinB) / (MaxB - MinA)
        print (round(p, 2), '|', a, '-', b)
        a2.append(round(p, 2))
       else:
        p =(MaxB-MinB) / (MaxA - MinA)
        print (round(p, 2), '|', a, '-', b)
        a2.append(round(p, 2))
```

Обучение модели происходит с помощью DecisionTreeClassifier(), после ранее проведенного исследования, мы выяснили, что обучение выполняется на 2 интервалах, а не на 4.

Имея хорошо обученную модель, мы уже можем улучшать код, который выполнялся для обучения, для его более быстрой работы, а также для уменьшения количества строк и более автоматизированной работы в целом.

Для начала мы создаем папку на google диске, и добавляем туда другие медицинские выборки, не учувствовавшие в обучении. Это делается для того, чтобы не добавлять файлы вручную постоянно, а в дальнейшем даже иметь возможность делать их парсинг с определенного сайта и добавлять сразу на

google диск. Далее обновляем код, для создания таблицы квантилей по каждому файлу с выборкой перед дальнейшим слиянием, а не после него. Это делается для предотвращения потенциальных ошибок

После этого получаем все отфильтрованные столбцы по квантилям в одном датафрейме, с которыми можно работать.

Запускаем функцию compare, которая уже автоматизирована и сокращает данные до сотых, для наглядного представления.

Также был улучшен основной код, который собирает данные для сравнения, для того чтобы не отрабатывали, полностью не совпадающие элементы и не записывались в массив a1, что уменьшает время работы. А также добавлена запись сравниваемых элементов в списки для удобного доступа к ним в дальнейшем. Данный программный код представлен в таблице 3.

Таблица 3 Программный код отбора данных для функции сравнения

```
a1 = []
for i in range(dfboth.shape[1] -1):
  for j in range(i+1, dfboth.shape[1]):
    a2 = []
    a = df2.columns[i]
    b = df2.columns[j]
    mina = df2[a].min()
    maxa = df2[a].max()
    minb = df2[b].min()
    maxb = df2[b].max()
    if (mina > maxb) or (minb > maxa):
      print ('не совпадает полностью')
    else:
      for ic in range (dfboth.shape[0]-1):
       for jc in range (dfboth.shape[0]-1):
        MinA = dfboth[ic][i]
        MaxA = dfboth[ic + 1][i]
        MinB = dfboth[jc][j]
        MaxB = dfboth[jc + 1][j]
        compare (MinA, MaxA, MinB, MaxB)
      a1.append(a2)
```

Для парсинга датасетов с сайта Kaggle.com, скачиваем все нужные нам библиотеки и подключаемся к google drive. На google drive создаем несколько папок для работы (Kaggle, metadata, description, datasets). В папку Kaggle будет помещен файл Kaggle.json, скачиваемый из личного кабинета для дальнейшей работы с сайтом. В папку metadata будут помещены все метаданные скачанных датасетов, в папку description будут сохраняться описания данных датасетов. В папку datasets соответственно датасеты.

Далее мы делим текст на слова, создав функцию (sent_to_words), и записываем все полученные слова в переменную (data_words), параллельно удаляя знаки препинания (deacc).

Представляем датасет в виде одного списка токенов (которые уже обработаны стеммером), записывая их в переменную (total_tokens) и находим частоту вхождения каждого токена (переменная fdist). Далее выделяем нечасто встречаемые слова, если слово встречается меньше пяти раз, записываем его в переменную (infreqWords).

Тоже самое проделываем и со словами, встречающимися очень часто (переменная topWords).

Так же можем вручную добавить слова, которые совсем не нужны нам в анализе в переменную (stops). Далее соединяем все этим переменные в общую переменную стоп слов(stop_words). Строим модели биграмм и триграмм (bigram_mod и trigram_mod). Создаем функции для стоп-слов(remove_stopwords), биграмм(make_bigrams), триграмм(make_trigrams) и лемматизации(lemmatization).

Далее поочередно запускаем функции, удаляем стоп-слова, сформировываем биграммы, инициализируем модель 'en', оставив только компонент теггера (для эффективности), делаем лемматизацию, получая лематизированный список данных.

Далее создаем lda_model из предполагаемых двух тем, настройка модели представлена в таблице 4.

Таблица4 Параметры lda модели

Переменная	Значение	Выполняемая функция
corpus	corpus	корпус данных
id2word	id2word	словарь
num_topics	2	число тем
random_state	100	управляет перемешиванием данных
update_every	1	определяет, как часто параметры модели должны обновляться
chunksize	100	количество документов, которые будут использоваться в каждом обучающем чанке
passes	10	общее количество проходов обучения
alpha	auto	гиперпараметр, которые влияет на разреженность тем
per_word_topics	True	установка этого значения в True позволяет извлекать наиболее вероятные темы для данного слова.

Обучение модели происходит пошагово.

На первом шаге для каждого документа d выбирается случайный вектор распределения тем θd из распределения Дирихле с параметром α.

На втором шаге выбирается тема tdi (в классической модели LDA количество тем фиксировано изначально) из мультиномиального распределения с параметром θd. Наконец согласно выбранной теме tdi выбирается слово wdi из распределения φt, которое является распределением Дирихле с параметром β.

Таким образом, порождающая модель слова w из документа d представляется в виде:

$$p\ (w|d, \theta, \varphi) = \sum_t p(w|t, \varphi_t)\ p(t|d,\theta_d), \quad (1)$$

где $\theta \sim Dir(\alpha)$;

$\varphi \sim Dir(\beta)$;

α и β — задаваемые так называемые гиперпараметры распределения Дирихле.

В моделях LDA каждый документ состоит из нескольких тем. Но, как правило, только одна из тем является доминирующей. Функция format_topics_sentences, извлекает эту доминирующую тему для каждого предложения и показывает вес темы и ключевых слов в хорошо отформатированном выводе.

Визуальная оболочка программного кода разработки, а также руководство для пользователя, написанная с помощью средств библиотеки tkinter, представлена в приложении Б.

Основные переменные программного кода представлены в таблице 5.

Таблица 5 Основные переменные программного кода

Названия переменных	Описание переменной
ds_name	Имена датасетов при парсинге
fileName	Имена файлов с описаниями
fullText	Полный текст одного описания
dfcorpus	Тексты всех описаний
sent_to_words	Функция - Деление текста на слова
data_words	Данные слов из текстов
total_tokens	Данные в виде одного списка токенов
fdist	Частота вхождения каждого токена
infreqWords	Нечасто встречаемые слова
topWords	Часто встречаемые слова
stops	Стоп слова, введенную вручную
stop_words	Итоговые стоп слова
bigram_mod, trigram_mod	Модели биграмм и триграмм
remove_stopwords	Функция – удаление стопслов
make_bigrams, make_trigrams	Функция – Создание биграмм и триграмм
lemmatization	Функция – лемматизация
data_lemmatized	Лематизированный список данных
corpus	Список частоты слов в тексте
lda_model	Модель LDA
cloud	Wordcloud
format_topics_sentences	Функция – создания датафрейма и определение тематики для каждого текста
medicalTopic	Имена датасетов, относящиеся к медицинской теме
aI	Массив итоговых значений одинаковых элементов
bI	Массив итоговых значений разных элементов
compare	Функция – Сравнение элементов
clf	Обученная модель
X	Массив значений
y	Массив нулей и единиц
y1	Результат работы модели

4 Тестирование разработанного решения

Для того чтобы проверить написанный программный код, протестируем его и посмотрим на результаты.

Для тестирования были взяты данные с сайта Kaggle, из которых были сгенерированы выборки для обучения модели, которая будет сравнивать атрибуты наборов медицинских данных по их содержимому.

Пример полученных сгенерированных выборок для обучения модели представлен в таблице 6.

Таблица 6 Генерация выборок и их квантили

	glucose1_1	**glucose1_2**	**glucose1_3**	**glucose1_4**
0.01	56.1096	56.2993	56.8864	56.3745
0.25	77.0400	77.8275	77.7000	76.4325
0.50	92.8150	93.3200	92.5900	90.2000
0.75	116.7075	118.2850	113.9600	111.8050
0.99	243.5870	243.5305	239.2934	238.1368

Для наглядности выводится диаграмма boxplot для сгенерированных выборок (glucose1_1, glucose1_2, glucose 1_3, glucose 1_4), как на рисунке 7.

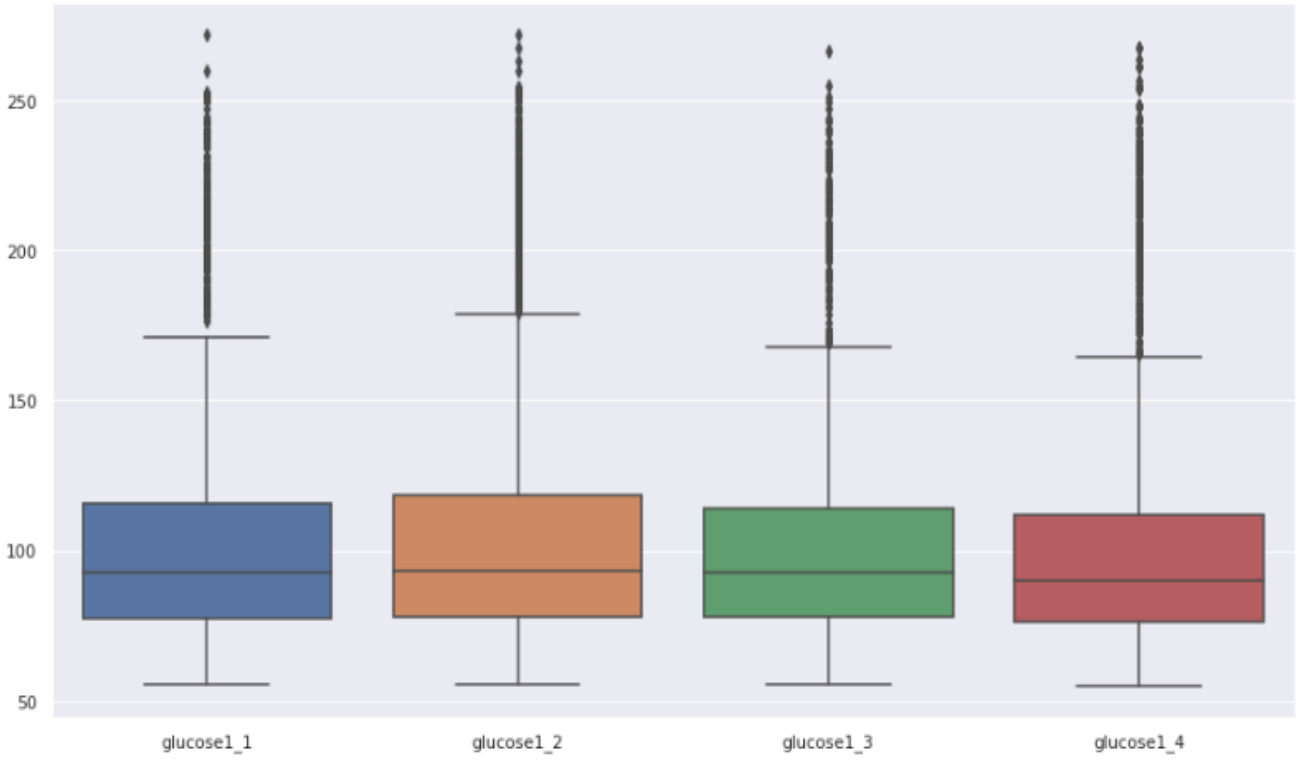

Рисунок 7 – Boxplot для сгенерированных выборок (glucose1_1, glucose1_2, glucose 1_3, glucose 1_4)

Далее запускается такой код, в котором сравниваются предварительно известные одинаковые элементы и в массив a3 добавляются единицы, означающие одинаковость веществ, это нужно для дальнейшего обучения модели. А все реальные значения добавляются в массив a1.

После этого повторяем, для достаточного количества одинаковых элементов, соединяем все столбцы и получаем итоговый массив aI, со всеми рассчитанными значениями. Длина массива aI равна 204, часть этого массива представлена на рисунке 8.

```
aI
```

```
[[0.87, 0, 0, 0, 0.09, 0.75, 0.0, 0, 0, 0.0, 0.88, 0.04, 0, 0, 0, 0.78],
 [0.6, 0.07, 0, 0, 0, 0.27, 0.62, 0.03, 0, 0, 0, 0.8, 0, 0, 0, 0.04],
 [0.94, 0, 0, 0, 0.04, 0.83, 0, 0, 0, 0.03, 0.93, 0, 0, 0, 0.0, 0.98],
 [0.9, 0.03, 0, 0, 0, 0.92, 0.0, 0, 0, 0.0, 0.84, 0.04, 0, 0, 0, 0.79],
 [0.62, 0.31, 0.01, 0, 0, 0, 0.45, 0.13, 0, 0, 0, 0.44, 0, 0, 0, 0.16],
 [0.95, 0.02, 0, 0, 0, 0.91, 0.02, 0, 0, 0, 0.89, 0, 0, 0, 0.02, 0.91],
 [0.96, 0.02, 0, 0, 0, 0.91, 0.02, 0, 0, 0, 0.9, 0, 0, 0, 0.02, 0.85],
 [0.95, 0, 0, 0, 0.02, 0.95, 0.0, 0, 0, 0.0, 1.0, 0.0, 0, 0, 0.0, 0.95],
 [0.97, 0, 0, 0, 0.02, 0.95, 0.0, 0, 0, 0.0, 0.97, 0.01, 0, 0, 0, 0.89],
 [0.72, 0, 0, 0, 0.02, 0.91, 0, 0, 0, 0.02, 0.82, 0, 0, 0, 0.04, 0.8],
 [0.67, 0.1, 0, 0, 0, 0.67, 0.05, 0, 0, 0, 0.83, 0, 0, 0, 0.03, 0.85],
 [0.84, 0.07, 0, 0, 0, 0.8, 0.02, 0, 0, 0, 0.87, 0.02, 0, 0, 0, 0.96],
 [0.65, 0.33, 0, 0, 0, 0.01, 0.58, 0.06, 0, 0, 0, 0.26, 0, 0, 0, 0.44],
 [0.86, 0, 0, 0, 0.07, 0.72, 0, 0, 0, 0.05, 0.88, 0.01, 0, 0, 0, 0.73],
 [0.7, 0.03, 0, 0, 0, 0.89, 0, 0, 0, 0.01, 0.84, 0, 0, 0, 0.04, 0.75],
 [0.9, 0.04, 0, 0, 0, 0.84, 0.03, 0, 0, 0, 0.87, 0.02, 0, 0, 0, 0.89],
 [0.8, 0, 0, 0, 0.02, 0.83, 0, 0, 0, 0.06, 0.67, 0, 0, 0, 0.07, 0.86],
 [0.97, 0, 0, 0, 0.02, 0.95, 0.0, 0, 0, 0.0, 0.99, 0, 0, 0, 0.0, 0.89],
 [0.84, 0, 0, 0, 0.09, 0.75, 0.0, 0, 0, 0.0, 1.0, 0.0, 0, 0, 0.0, 0.99],
 [0.84, 0, 0, 0, 0.09, 0.75, 0.0, 0, 0, 0.0, 0.93, 0.02, 0, 0, 0, 0.69],
 [0.64, 0.06, 0, 0, 0, 0.81, 0, 0, 0, 0.03, 0.94, 0, 0, 0, 0.0, 0.99],
 [0.73, 0.02, 0, 0, 0, 0.91, 0.02, 0, 0, 0, 0.85, 0.03, 0, 0, 0, 0.79],
 [0.86, 0.03, 0, 0, 0, 0.82, 0.07, 0, 0, 0, 0.68, 0.06, 0, 0, 0, 0.78],
 [0.87, 0.02, 0, 0, 0, 0.91, 0.02, 0, 0, 0, 0.83, 0.03, 0, 0, 0, 0.83],
 [0.63, 0.04, 0, 0, 0, 0.35, 0.61, 0, 0, 0, 0.01, 0.71, 0, 0, 0, 0.08],
 [0.6, 0.04, 0, 0, 0, 0.35, 0.56, 0.02, 0, 0, 0, 0.53, 0, 0, 0, 0.18],
 [0.99, 0.0, 0, 0, 0.0, 1.0, 0.0, 0, 0, 0.0, 0.96, 0, 0, 0, 0.01, 0.97],
 [0.88, 0.0, 0, 0, 0.0, 0.93, 0.04, 0, 0, 0, 0.93, 0.0, 0, 0, 0.0, 0.8],
 [0.81, 0, 0, 0, 0.03, 0.87, 0, 0, 0, 0.03, 0.82, 0, 0, 0, 0.04, 0.85],
 [0.6, 0.33, 0.01, 0, 0, 0, 0.48, 0.12, 0, 0, 0, 0.37, 0, 0, 0, 0.19],
 [0.57, 0.07, 0, 0, 0, 0.31, 0.61, 0.01, 0, 0, 0, 0.55, 0, 0, 0, 0.17],
```

Рисунок 8 – Массив значений aI

Так же соединяем все массивы с единицами и получаем массив ai со всеми единицами.

Далее повторяем все тоже самое, но для разных элементов, с небольшим изменением кода и получаем массив bI, часть представлена на рисунке 9.

```
bI
```

```
[[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.63, 0.14, 0, 0],
 [0, 0, 0, 0, 0, 0, 0, 0, 0.03, 0, 0, 0, 0.02, 0.01, 0.01, 0.01],
 [0, 0, 0, 0.14, 0, 0, 0, 0.05, 0, 0, 0, 0.07, 0, 0, 0, 0.28],
 [0.0, 0.0, 0.0, 0.08, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0.06, 0.06, 0.11, 0.6, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0.19, 0.22, 0.33, 0, 0, 0, 0.09, 0.03, 0, 0, 0, 0.05, 0, 0, 0, 0.14],
 [0, 0, 0, 0, 0.09, 0.08, 0.09, 0.65, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0.33, 0.23, 0, 0, 0, 0.05, 0.27, 0.31, 0, 0, 0, 0.23],
 [0.24, 0.52, 0.11, 0, 0, 0, 0.58, 0.06, 0, 0, 0, 0.46, 0, 0, 0, 0.25],
 [0, 0, 0, 0.45, 0, 0, 0, 0.43, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0.43, 0.03, 0, 0, 0, 0.04, 0, 0, 0, 0.05, 0, 0, 0, 0.35],
 [0, 0.14, 0, 0, 0, 0.05, 0, 0, 0, 0.07, 0, 0, 0, 0.23, 0.02, 0],
 [0, 0.2, 0.15, 0, 0, 0, 0.11, 0, 0, 0, 0.16, 0, 0, 0, 0.53, 0.02],
 [0, 0, 0, 0, 0.0, 0.0, 0.01, 0.02, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0, 0, 0.1, 0, 0, 0, 0.79, 0, 0, 0],
 [0, 0, 0, 0.09, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0.82, 0, 0, 0, 0.09, 0.56, 0, 0, 0, 0.17, 0.5, 0, 0, 0, 0.09, 0.4],
 [0.05, 0.05, 0.06, 0.38, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0.24, 0.11, 0.08, 0.14, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0.3, 0, 0, 0, 0.22, 0.33, 0, 0, 0, 0.34, 0.29, 0],
 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.09, 0, 0, 0],
 [0.16, 0.14, 0.14, 0.44, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.1, 0, 0, 0],
 [0.73, 0, 0, 0, 0.03, 0.38, 0, 0, 0, 0.58, 0.01, 0, 0, 0, 0.31, 0.18],
 [0, 0, 0, 0, 0.01, 0.01, 0.04, 0.46, 0, 0, 0, 0, 0, 0, 0, 0],
 [0.43, 0, 0, 0, 0.16, 0, 0, 0, 0.14, 0.05, 0, 0, 0, 0.75, 0, 0],
 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.11, 0, 0, 0],
 [0, 0, 0, 0, 0.01, 0.01, 0.01, 0.13, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0.36, 0, 0, 0, 0.3, 0.5, 0, 0, 0, 0.06, 0.48, 0.12],
 [0, 0, 0, 0.37, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0.35, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.37, 0, 0, 0],
 [0, 0, 0, 0.63, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0.38, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0, 0.35, 0, 0, 0, 0.18, 0, 0, 0, 0.22, 0, 0, 0, 0.03],
 [0, 0, 0, 0.12, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0, 0, 0.24, 0.27, 0, 0, 0, 0.64, 0, 0, 0, 0, 0, 0, 0, 0],
 [0.02, 0.01, 0.01, 0.03, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
 [0.22, 0, 0, 0, 0.66, 0, 0, 0, 0.04, 0.61, 0.25, 0, 0, 0, 0.02, 0.22],
```

Рисунок 9 – Массив значений bI

После этого мы получаем все массивы с нулями, соединяем их и получаем массив bi со всеми нулями.

Теперь у нас есть все значения для обучения модели, объединяем массивы aI, bI и отдельно ai и bi.

Массивы aI и bI состоят из элементов, полученных с помощью сравнения атрибутов наборов данных и представляют собой коэффициент схожести, каждого интервала, а массивы ai и bi, состоят из нулей и единиц соответственно.

Для обучения делим выборки на тестовые и тренировочные и обучаем.

На рисунке 10 представлено дерево алгоритма обучения модели с четырьмя интервалами, обозначения переменных представлено в таблице 7.

Таблица 7 Переменные дерева алгоритма обучения модели с четырьмя интервалами

Переменная	Обозначение	Значение
Gini	Коэффициент неоднородности	Описывает насколько классы перепутаны.
Samples	Количество образцов	Показывает, количество образцов, использующихся в обучении.
Value	Вероятности классов	Показывает вероятность классов [Класс А, Класс Б]
Class	Класс	Победивший класс (1 – одинаковые или 0 – разные элементы)
Пример: 0 <= 0.515 означает, что узел разделен таким образом, что все выборки, где характеристика 0 ниже 0,515, идут к левому дочернему элементу, а образцы, где характеристика выше 0,15 к правому дочернему элементу.		

Индекс Gini вычисляется по формуле [21]:

$$\mathrm{Gini}(Q) = 1 - \sum_{i=1}^{n} p_j^2, \qquad (2)$$

где Q — результирующее множество;

n — число классов в нём;

p_j— вероятность i-го класса (выраженная как относительная частота примеров соответствующего класса).

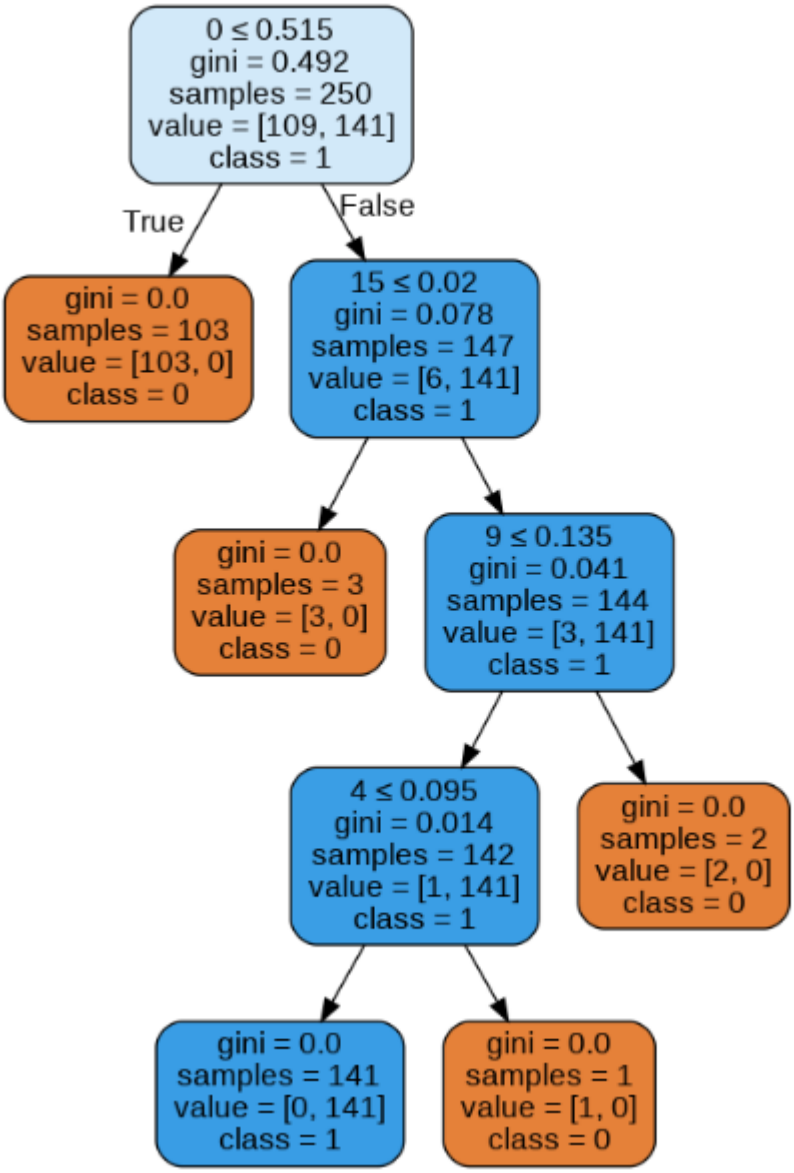

Рисунок 10 – Дерево алгоритма обучения модели с четырьмя интервалами

Матрица неточности и отчет по классификации модели с четырьмя интервалами представлены на рисунке 11.

```
Confusion Matrix:
[[45  0]
 [ 0 63]]
Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        45
         1.0       1.00      1.00      1.00        63

    accuracy                           1.00       108
   macro avg       1.00      1.00      1.00       108
weighted avg       1.00      1.00      1.00       108

Accuracy: 1.0
```

Рисунок 11– Матрица неточности и отчет по классификации модели с четырьмя интервалами

Проведя исследование сравнивая обученные модели с 2, 4 и 5 интервалами, выяснилось, что модель с двумя интервалами, работает настолько же эффективно как и модель с четырьмя, это мы видим сравнивая рисунок 11 и 12 (матрица неточности и отчет по классификации модели с двумя интервалам).

Модель с двумя интервалами, нашла схожие элементы в таком же количестве. Так же и время работы программы намного меньше у модели с двумя интервалами (4 секунды против 14 секунд у модели с четырьмя интервалами).

```
Confusion Matrix:
[[47  0]
 [ 0 55]]
Classification Report:
              precision    recall  f1-score   support

         0.0       1.00      1.00      1.00        47
         1.0       1.00      1.00      1.00        55

    accuracy                           1.00       102
   macro avg       1.00      1.00      1.00       102
weighted avg       1.00      1.00      1.00       102

Accuracy: 1.0
```

Рисунок 12 – Матрица неточности и отчет по классификации модели с двумя интервалами

Сам алгоритм работы модели с двумя интервалами представлен на рисунке 13.

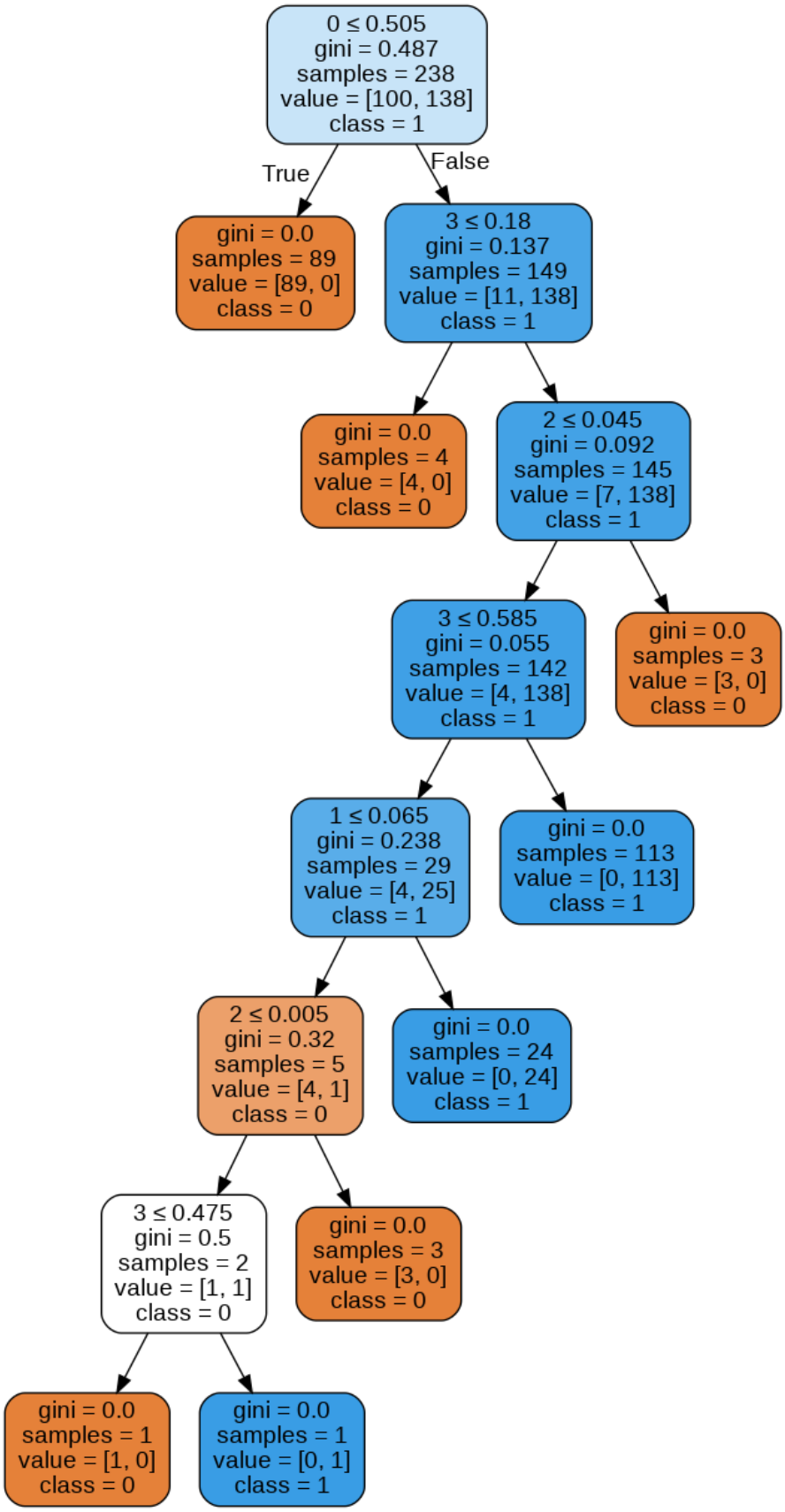

Рисунок 13 – Дерево алгоритма работы модели с двумя интервалами

Делаем вывод, что для данной программы сравнение по двум интервалам будет также эффективно, как и по четырем, но время работы существенно

снижается. Поэтому в дальнейшем будет использоваться алгоритм модели с двумя интервалами с именем (clf.pkl).

Выбрав данные, которые не использовались в обучении модели, но относящиеся к этой же тематике, и запустив уже имеющиеся алгоритмы (без условия и добавления массивов нулей и единиц), а также запустив обученную модель получаем результат, который был получен из модели с двумя интервалами, представленный в таблице 8.

Таблица 8 Результат работы алгоритма модели с двумя интервалами

Схожие элементы в наборах данных
Glucose - time
Glucose - avg_glucose_level
bgr - Blood Glucose Random (mgs/dL)
bu - Blood Urea (mgs/dL)
sc - homa
sc - Serum Creatinine (mgs/dL)
sod - Sodium (mEq/L)
pot - Potassium (mEq/L)
hemo - Hemoglobin (gms)
wc - White Blood Cells (cells/cmm)
rc - Potassium (mEq/L)
rc - Red Blood Cells (millions/cmm)
homa - Serum Creatinine (mgs/dL)

После получения успешной модели запустим программный код для сбора данных. Вывод поискового запроса представлен на рисунке 14. Далее получаем метаданные и описания из них. Пример метаданных датасета представлен в таблице 9.

ref	title	size	lastUpdated	downloadCount
uciml/pima-indians-diabetes-database	Pima Indians Diabetes Database	9KB	2016-10-06 18:31:56	188341
fedesoriano/stroke-prediction-dataset	Stroke Prediction Dataset	67KB	2021-01-26 19:29:28	31032
openfoodfacts/world-food-facts	Open Food Facts	109MB	2017-09-18 12:27:58	44977
sulianova/cardiovascular-disease-dataset	Cardiovascular Disease dataset	742KB	2019-01-20 01:28:23	27796
cdc/national-health-and-nutrition-examination-survey	National Health and Nutrition Examination Survey	7MB	2017-01-26 20:11:45	12248
dileep070/heart-disease-prediction-using-logistic-regression	Logistic regression To predict heart disease	58KB	2019-06-07 06:12:56	11912
kandij/diabetes-dataset	Diabetics prediction using logistic regression	9KB	2019-05-07 05:18:45	6771
saurabh00007/diabetescsv	diabetes.csv	9KB	2017-11-13 11:42:48	14320
johndasilva/diabetes	diabetes	12KB	2018-04-25 19:17:46	6364
mathchi/diabetes-data-set	Diabetes Data Set	9KB	2020-08-05 21:27:01	3315
vikasukani/diabetes-data-set	Diabetes Data Set	12KB	2020-08-08 11:23:25	1399
colearninglounge/chronic-kidney-disease	Chronic Kidney Disease	10KB	2020-08-26 03:33:35	922
christofel04/cardiovascular-study-dataset-predict-heart-disea	Cardiovascular Study Dataset	75KB	2020-09-22 12:24:18	1021
blackbee2016/6-months-daily-diabetes-history	6 Months Daily Diabetes Measures	11KB	2018-12-29 23:03:51	597
salihacur/diabetes	diabetes	9KB	2020-03-21 13:50:58	815
rahulsah06/machine-learning-for-diabetes-with-python	Machine Learning for Diabetes with Python	9KB	2019-09-20 08:04:12	747
edubrq/diabetes	Diabetes	9KB	2018-06-28 19:04:27	964
saurabhtayal/diabetic-patients-readmission-prediction	Diabetic Patients' Re-admission Prediction	3MB	2020-08-27 15:05:37	564
peterboos/determine-insuline-intake	Determine Insulin intake for a diabetic	20KB	2019-02-26 21:43:36	388
bhadaneeraj/cardio-vascular-disease-detection	Cardio Vascular Disease Detection	750KB	2020-06-01 13:12:51	936

Рисунок 14 – Вывод поискового запроса «glucose»

Таблица 9 Метаданные набора данных

Переменная (Обозначение)	Значение переменной
Subtitle (подзаголовок)	"To classify the patients to be healthy or suffering from cardiovascular disease"
Description (описание)	"# **Cardiovascular Disease Detection**\n\n## **Citation:**\n1.\tLicense:\tUnknown\n2. \tDomain: Public\n3. \tDataset owner: Svetlana Ulianova\n4. \tDate created: 2019-01-20 \n\n\n\n**## Features:**\n1.\tAge | Objective Feature | age | int (days)\n2. \tHeight | Objective Feature | height | int (cm) |\n3.\tWeight | Objective Feature | weight | float (kg) |\n4. \tGender | Objective Feature | gender | categorical code |\n5. \tSystolic blood pressure | Examination Feature | ap_hi | int |\n6. \tDiastolic blood pressure | Examination Feature | ap_lo | int |\n7. \tCholesterol | Examination Feature | cholesterol | 1: normal, 2: above normal, 3: well above normal |\n8. \tGlucose | Examination Feature | gluc | 1: normal, 2: above normal, 3: well above normal |\n9. \tSmoking | Subjective Feature | smoke | binary |\n10. \tAlcohol intake | Subjective Feature | alco | binary |\n11. \tPhysical activity | Subjective Feature | active | binary |\n12. \tPresence or absence of cardiovascular disease | Target Variable | cardio | binary |\n\ nAll of the dataset values were collected at the moment of medical examination.\n\n**
Title (Заголовок)	"Cardio Vascular Disease Detection"
id_no (Номер набора данных)	687373
Keywords (Ключевые слова)	["health conditions", «classification", beginner", "heart conditions","exploratory data analysis", "binary classification"]
Id набора данных	"bhadaneeraj/cardio-vascular-disease-detection"

После сбора описаний данных, добавляем в корпус данных переменная (dfcorpus), как на рисунке 15.

```
'## Context\n\nThis dataset is originally from the National Institute of D
'### Context\n\nAccording to the World Health Organization (WHO) stroke is
'- Want to play with food ? We now have a list of AI tasks that have a rea
'#### Data description\n\nThere are 3 types of input features:\n\n - *Obje
'# Context \n\nThe [National Health and Nutrition Examination Survey (NHAN
"**LOGISTIC REGRESSION - HEART DISEASE PREDICTION**\n\n**Introduction**\nW
'The data was collected and made available by "National Institute of Diabe
'',
'Dataset of diabetes, taken from the hospital Frankfurt, Germany\n\ndiabet
'### Context\n\nThis dataset is originally from the National Institute of
'',
"## Description\nThe data was taken over a 2-month period in India with 25
'## This Dataset is for HME Workshop in Oct 3, 2020\n\n### Introduction\nW
'### Context\n\nMy grandma has been suffering from diabetes for over 15 ye
'- Number of Instances: 768\n- Number of Attributes: 8 plus class \n \nFor
'About one in seven U.S. adults has diabetes now, according to the Centers
'',
'**Dataset name**: Diabetes 130-US hospitals for years 1999-2008 Data Set\
'### Context\nA friend of mine has a very unstable form of diabetics, not
'# **Cardiovascular Disease Detection**\n\n## **Citation:**\n1.\tLicense:\
```

Рисунок 15 – Корпус описаний из метаданных

Далее если описание меньше 3 слов, алгоритм добавляет ключевые слова в пустые места в dfcorpus, извлекая их из метаданных.

В итоге мы получаем полный корпус для комфортной и более точной работы с ним, на рисунке 16.

```
['## Context\n\nThis dataset is originally from the National Institute of Diabet
 '### Context\n\nAccording to the World Health Organization (WHO) stroke is the
 '- Want to play with food ? We now have a list of AI tasks that have a real-wor
 '#### Data description\n\nThere are 3 types of input features:\n\n - *Objective
 '# Context \n\nThe [National Health and Nutrition Examination Survey (NHANES)](
 "**LOGISTIC REGRESSION - HEART DISEASE PREDICTION**\n\n**Introduction**\nWorld
 'The data was collected and made available by "National Institute of Diabetes a
 ['health', 'heart conditions', 'healthcare'],
 'Dataset of diabetes, taken from the hospital Frankfurt, Germany\n\ndiabetes',
 '### Context\n\nThis dataset is originally from the National Institute of Diabe
 ['diabetes'],
 "## Description\nThe data was taken over a 2-month period in India with 25 feat
 '## This Dataset is for HME Workshop in Oct 3, 2020\n\n### Introduction\nWorld
 '### Context\n\nMy grandma has been suffering from diabetes for over 15 years.
 '- Number of Instances: 768\n- Number of Attributes: 8 plus class \n \nFor Each
 'About one in seven U.S. adults has diabetes now, according to the Centers for
 ['health conditions',
  'classification',
  'beginner',
  'heart conditions',
  'exploratory data analysis',
  'binary classification'],
 '**Dataset name**: Diabetes 130-US hospitals for years 1999-2008 Data Set\n\n**
 '### Context\nA friend of mine has a very unstable form of diabetics, not sure
 '# **Cardiovascular Disease Detection**\n\n## **Citation:**\n1.\tLicense:\tUnkn
```

Рисунок 16 – Полный корпус данных

Далее программа очищает текст, токенизирует его, удаляет стоп слова, создаем биграммы и триграммы, а также производит лемматизацию. Создается словарь id2word и получается частота слов в текстах и записывается в корпус (corpus).

Читаемый вид корпуса первого текста представлен в таблице 10.

Таблица 10 Corpus для первого текста

Слова	Частота слов в тексте
'also'	1
'classification'	2
'diabete'	1
'disease'	1
'female'	1
'instance'	1
'kidney'	1
'national'	1
'patient'	1
'pima'	1
'selection'	1
'several'	1
'use'	1

После работы lda модели мы получаем темы, входящие в них термины и их вес в теме. Результат представлен в таблице 11.

Таблица 11 Термины и их вес

Основные слова в теме	Вес слов в теме
Тема: Продукты и еда	
Food	0.013
Vitamin	0.010
Ingredient	0.008
Insulin	0.008
Dose	0.007
Product	0.006
Label	0.005
Learn	0.005
Use	0.005
World	0.005
Тема: Медицина	
Diabetes	0.011
Medical	0.010
Patient	0.010
Health	0.008
Diabete	0.008
Smoke	0.007
Readmission	0.007
Pressure	0.006
Disease	0.006
weight	0.006

Видим, что получились две темы, первая тема связана с продуктами питания, а вторая тема является медицинской.

Наглядно это представляется в виде облака слов на рисунке 17.

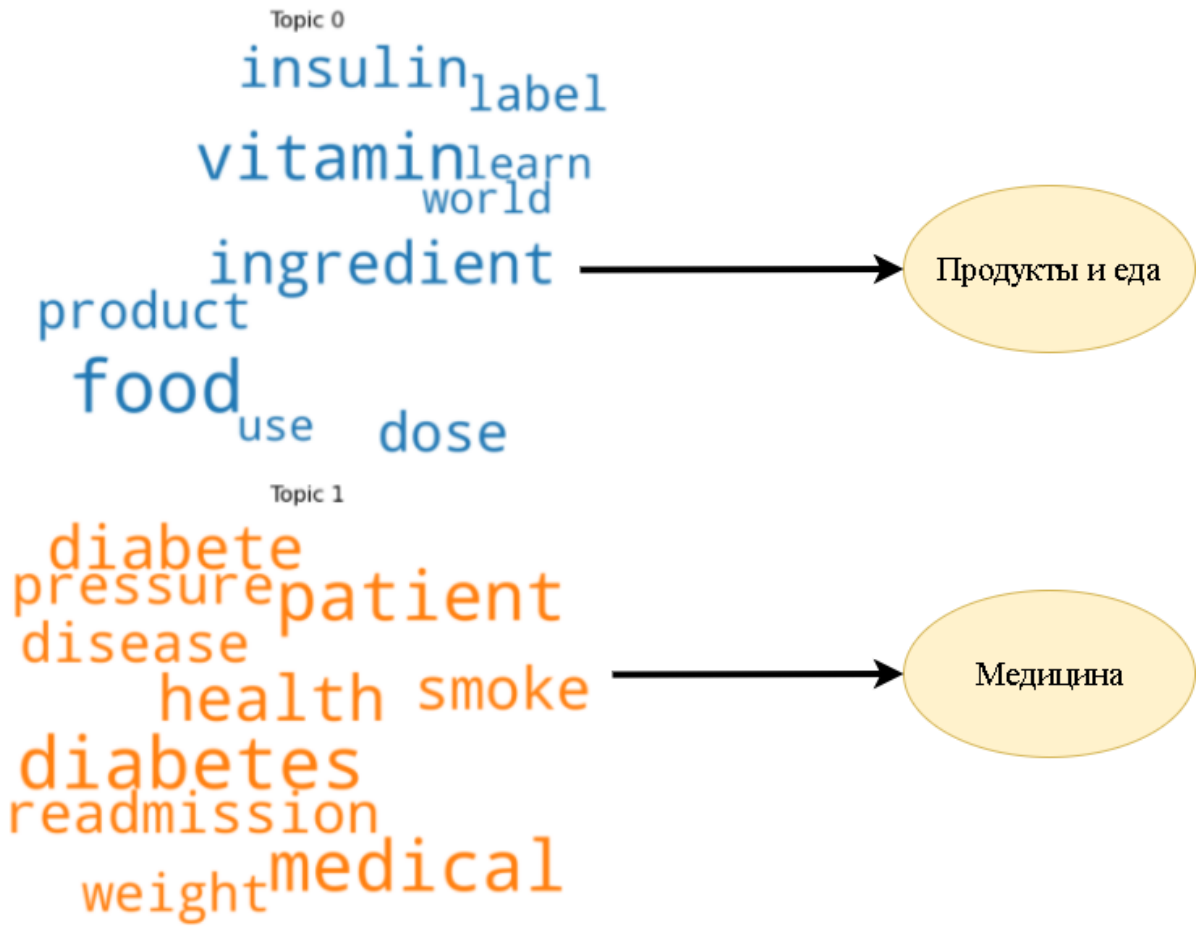

Рисунок 17 – Облако слов

В моделях LDA каждый документ состоит из нескольких тем. Но, как правило, только одна из тем является доминирующей. Приведенная функция format_topics_sentences, извлекает эту доминирующую тему для каждого предложения и показывает вес темы и ключевых слов в хорошо отформатированном выводе.

В результате мы получили распределение документов по темам, как на рисунке 18.

Document_No		fileName	\
0	0	descriptions-uciml-pima-indians-diabetes-datab...	
1	1	descriptions-fedesoriano-stroke-prediction-dat...	
2	2	descriptions-openfoodfacts-world-food-facts.docx	
3	3	descriptions-sulianova-cardiovascular-disease-...	
4	4	descriptions-cdc-national-health-and-nutrition...	
5	5	descriptions-dileep070-heart-disease-predictio...	
6	6	descriptions-kandij-diabetes-dataset.docx	
7	7	descriptions-saurabh00007-diabetescsv.docx	
8	8	descriptions-johndasilva-diabetes.docx	
9	9	descriptions-mathchi-diabetes-data-set.docx	
10	10	descriptions-vikasukani-diabetes-data-set.docx	
11	11	descriptions-colearninglounge-chronic-kidney-d...	
12	12	descriptions-christofel04-cardiovascular-study...	
13	13	descriptions-blackbee2016-6-months-daily-diabe...	
14	14	descriptions-salihacur-diabetes.docx	
15	15	descriptions-rahulsah06-machine-learning-for-d...	
16	16	descriptions-edubrq-diabetes.docx	
17	17	descriptions-saurabhtayal-diabetic-patients-re...	
18	18	descriptions-peterboos-determine-insuline-inta...	
19	19	descriptions-bhadaneeraj-cardio-vascular-disea...	

	Dominant_Topic	Topic_Perc_Contrib	\
0	1.0	1.00	
1	1.0	1.00	
2	0.0	1.00	
3	1.0	1.00	
4	1.0	1.00	
5	1.0	1.00	
6	1.0	1.00	
7	1.0	0.99	
8	0.0	0.79	
9	1.0	1.00	
10	1.0	1.00	
11	1.0	1.00	
12	1.0	1.00	
13	0.0	0.99	
14	1.0	1.00	
15	0.0	0.99	
16	1.0	0.99	
17	1.0	1.00	
18	0.0	1.00	
19	1.0	1.00	

Рисунок 18– Распределение тем

Мы получили переменную medicalTopic, в которой содержатся имена файлов с описанием, относящиеся к медицинской тематике.

Далее после редактирования данного списка, как на рисунке 19, мы получаем переменную medicalTopics, в которой содержится имена датасетов с медицинской тематикой.

```
['uciml/pima-indians-diabetes-database',
 'fedesoriano/stroke-prediction-dataset',
 'sulianova/cardiovascular-disease-dataset',
 'cdc/national-health-and-nutrition-examination-survey',
 'dileep070/heart-disease-prediction-using-logistic-regression',
 'kandij/diabetes-dataset',
 'saurabh00007/diabetescsv',
 'mathchi/diabetes-data-set',
 'vikasukani/diabetes-data-set',
 'colearninglounge/chronic-kidney-disease',
 'christofel04/cardiovascular-study-dataset-predict-heart-disea',
 'salihacur/diabetes',
 'edubrq/diabetes',
 'saurabhtayal/diabetic-patients-readmission-prediction',
 'bhadaneeraj/cardio-vascular-disease-detection']
```

Рисунок 19 – Имена медицинских датасетов

Результатом данного программного кода являются файлы csv готовые для использования в анализе на схожесть, на рисунке 20.

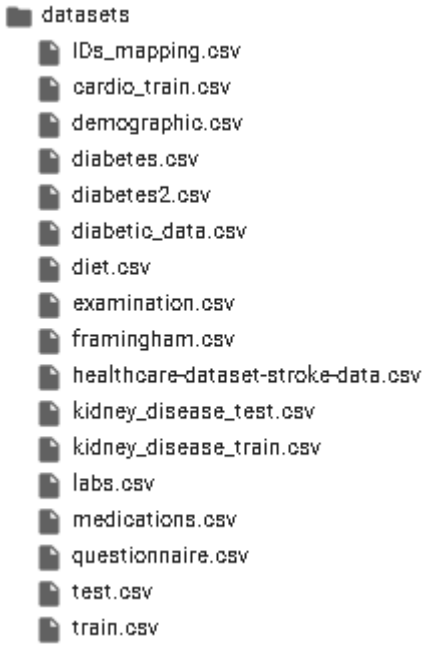

Рисунок 20 – Полученные наборы данных

Как мы видим алгоритм в большинстве случаев подбирает правильные темы. Но будет ли он определять так же точно, если вместо описания, анализировать более меньшие списки ключевые слова.

Для этого мы вытащим из метаданных и запишем в корпус данных, только ключевые слова, как на рисунке 21.

```
[['earth and nature', 'health', 'diabetes', 'healthcare', 'india'
 ['health',
  'health conditions',
  'public health',
  'healthcare',
  'binary classification'],
 ['earth and nature', 'alcohol', 'nutrition'],
 ['health', 'heart conditions', 'healthcare'],
 ['earth and nature',
  'health',
  'health conditions',
  'healthcare',
  'nutrition',
  'drugs and medications'],
 ['health',
  'health conditions',
  'heart conditions',
  'healthcare',
  'regression',
  'logistic regression'],
 ['computer science', 'biology', 'diabetes', 'logistic regression
 ['diabetes'],
 ['diabetes'],
 ['education', 'health', 'diabetes'],
 ['diabetes', 'regression', 'binary classification', 'logistic re
 ['health', 'health conditions', 'heart conditions'],
 ['earth and nature',
  'health',
  'health conditions',
  'heart conditions',
```

Рисунок 21 – Ключевые слова

И пройдя по такому же коду, как и ранее мы получили темы. По полученным темам на рисунке 22, видно, что они определяются не совсем корректно и одних ключевых слов недостаточно для корректного определения тем. Поэтому дальнейший запуска кода можно не производить.

Рисунок 22 – Облако слов по ключевым словам

5 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

Информатизация здравоохранения привела к созданию большого количества медицинских информационных систем (МИС), которые собирают и хранят медицинскую информацию о пациентах и процессе их лечения.

Целью работы является алгоритмическое и программное обеспечение анализа схожести наборов медицинских данных по их описанию и содержимому, а также атрибутам.

Актуальность работы заключается в том, что созданное алгоритмическое и программное обеспечение, могут использоваться не только исследователями в области медицинских или иных данных, но в перспективе и медицинскими учреждениями для устранения разнообразности наборов медицинских данных, путем анализа их схожести, тем самым упрощая анализ и восприятие различной медицинской информации в целом.

Но перед тем, как приступать к разработке мы должны провести финансовый менеджмент и провести несколько расчетов для выявления сметы затрат НИ, проведение оценки рисков, определение экономической эффективности НИ и т. д.

5.1 Проведение предпроектного анализа

5.1.1 Потенциальные потребители результатов исследования

Целевым рынком являются медицинские учреждения, работающие с информацией о пациентах, а также компании, занимающиеся схожими операциями. Целевой сегмент рынка — это непосредственно медицинские работники различных медицинских учреждений. Изучив рынок, мы пришли к выводу, что конкурентов в конкретной области в России в данный момент нет, так как в нашей стране только начинается освоение стандартов обмена передачи информации и вытекающие из нее подтемы.

5.1.2 SWOT-анализ

SWOT-анализ заключается в выявлении сильных и слабых сторон проекта, возможностей для дальнейшего развития и угроз существованию и развитию; направлен на исследование внутренней и внешней среды проекта.

Составим итоговую матрицу SWOT-анализа, представленную в таблице 12.

Таблица 12 Матрица SWOT-анализа

	Сильные стороны научно-исследовательского проекта: С1. Проект уникальный и конкуренты отсутствуют. С3. Перспективный с точки зрения инвестирования проект. С4. Унифицированность разработки.	**Слабые стороны научно-исследовательского проекта:** Сл3. Для использования продукта требуется навыки программирования для последующего обучения модели и навыки анализа данных.
Возможности: В1. Появление дополнительного спроса на данный продукт. В2. Широкий спектр применения данной разработки в клиниках. В3. Возможность использования проведения анализа данных .	Уникальность проекта и отсутствие конкурентов определяет перспективы проекта. Данные перспективы отражены в возможности внедрения данного модуля в программное обеспечение в другие клиники.	Снижение спроса на данную технологию могут помешать продвижению разработки и привлечению дополнительных средств с другихисточников для развития проекта.Подобное снижение может отодвинуть дату начала проекта или вовсе закрыть его.
Угрозы: У1. Снижение спроса в связи с появлением более простых в применении технологий.	В условиях нынешней конкуренции в разработке, возможны трудности с продвижением продукта.	При недостаточном финансировании разработки могут занять большее время, что приведёт к снижению спроса. Помимо этого, могут появиться разработки конкурентов, которые будут более привлекательными для потребителя, что приведёт к снижению спроса и потере доверия.

5.2 Организация и планирование работ

В данном пункте составляется полный перечень проводимых работ, определяются их исполнители и рациональная продолжительность. Наглядным результатом планирования работ является сетевой, либо линейный график реализации проекта. Так как число исполнителей редко превышает двух в большинстве случаев предпочтительным является линейный график. Для построения линейного графика хронологически упорядоченные вышеуказанные данные сведены в таблицу 13.

В соответствии с видами работ участниками планирования выбраны:

1) Научный руководитель (НР);

2) Исполнитель ВКР (И).

Таблица 13 – Перечень работ и продолжительность их выполнения

Этапы работы	**Исполнители**	**Загрузка исполнителей**
Постановка целей и задач	НР	НР – 100%
Разработка календарного плана	НР, И	НР – 100% И – 20%
Подбор и изучение материалов по тематике	НР, И	НР – 30% И – 100%
Обсуждение литературы	НР, И	НР – 30% И – 100%
Изучение материалов по данной тематике	И	И – 100%
Реализация алгоритма получения файлов для дальнейшего анализа	И	И – 100%
Реализация алгоритмов получения описаний датасетов.	И	И – 100%
Реализация алгоритмов оценки схожести медицинских данных.	И	И – 100%
Тестирование и отладка реализованного комплекса	НР, И	НР – 100% И – 100%
Описание мероприятий по социальной ответственности	И	И – 100%
Описание ресурсоэффективности и ресурсоснабжения исследования	И	И – 100%
Составление графической оболочки	НР, И	НР – 30% И – 100%
Составление отчета о проделанной работе	И	И – 100%
Оформление графического материала	И	И – 100%
Подведение итогов	И	И – 100%

5.2.1 Продолжительность этапов работ

Расчет продолжительности этапов работ может осуществляться двумя методами:

– технико-экономическим;

– опытно-статистическим.

Первый применяется в случаях наличия достаточно развитой нормативной базы трудоемкости планируемых процессов, что в свою очередь обусловлено их высокой повторяемостью в устойчивой обстановке. Так как исполнитель работы зачастую не располагает соответствующими нормативами, то используется опытно-статистический метод, который реализуется двумя способами:

– аналоговый;

– экспертный.

Аналоговый способ привлекает внешней простотой и около нулевыми затратами, но возможен только при наличии в поле зрения исполнителя НИР не устаревшего аналога, т. е. проекта в целом или хотя бы его фрагмента, который по всем значимым параметрам идентичен выполняемой НИР. В большинстве случаев он может применяться только локально – для отдельных элементов (этапов работы).

Экспертный способ используется при отсутствии вышеуказанных информационных ресурсов и предполагает генерацию необходимых количественных оценок специалистами конкретной предметной области, опирающимися на их профессиональный опыт и эрудицию. Для определения вероятных (ожидаемых) значений продолжительности работ $t_{ож}$ применяется следующая формула:

$$t_{ож} = \frac{3 \times t_{min} + 2 \times t_{max}}{5}, \quad (3)$$

где t_{min} – минимальная продолжительность работы, дн.;

t_{max} – максимальная продолжительность работы, дн.

5.2.2 Разработка графика проведения научного исследования

Расчет продолжительности выполнения каждого этапа в рабочих днях ($T_{РД}$) ведется по следующей формуле:

$$T_{РД} = \frac{t_{ож}}{K_{ВН}} \times K_{Д}, \quad (4)$$

где $t_{ож}$ – продолжительность работы, дн.;

$K_{ВН}$ – коэффициент выполнения работ, учитывающий влияние внешних факторов на соблюдение предварительно определенных длительностей, в частности, возможно $K_{ВН} = 1$;

$K_{Д}$ – коэффициент, учитывающий дополнительное время на компенсацию непредвиденных задержек и согласование работ ($K_{Д} = 1{,}2$).

Расчет продолжительности этапа в календарных днях ведется по формуле.

$$T_{КД} = T_{РД} \times T_{К}, \quad (5)$$

где $T_{КД}$ – продолжительность выполнения этапа в календарных днях;

$T_{К}$ – коэффициент календарности, позволяющий перейти от длительности работ в рабочих днях к их аналогам в календарных днях, и рассчитываемый по формуле:

$$T_{К} = \frac{T_{КАЛ}}{T_{КАЛ} - T_{ВД} - T_{ПД}}, \quad (6)$$

Для шестидневной рабочей недели $T_{К} = 1{,}22$ ($T_{КАЛ} = 366$, $T_{ВД} = 52$, $T_{ПД} = 14$).

Все рассчитанные значения представлены в таблице 14. Диаграмма Ганта представлена в таблице 15.

Таблица 14 – Трудозатраты на выполнение проекта

Этап	**Исполнители**	**Продолжительность работ, дни**			**Трудоемкость работ по исполнителям чел.-дни**			
					$T_{РД}$		$T_{КД}$	
		t_{min}	t_{max}	$t_{ож}$	НР	И	НР	И
1	2	3	4	5	6	7	8	9
Постановка целей и задач	НР	2	4	2,8	3,36	–	4,1	–
Разработка календарного плана	НР, И	2	4	2,8	3,36	0,672	4,1	0,1
Подбор и изучение материалов по тематике	НР, И	9	15	11,4	4,1	13,68	5	16,69
Обсуждение литературы	НР, И	3	6	4,2	1,51	5,04	1,84	6,15
Изучение материалов по данной тематике	И	4	6	4,8	–	5,76	–	7,03
Реализация алгоритма получения файлов для дальнейшего анализа	И	6	9	7,2	–	8,64	–	10,54
Реализация алгоритмов получения описаний датасетов.	И	3	6	4,2	–	5,04	–	6,15
Реализация алгоритмов оценки схожести медицинских данных.	И	4	6	4,8	–	5,76	–	7,03
Тестирование и отладка реализованного комплекса	НР, И	8	12	9,6	11,52	11,52	14,05	14,05
Описание мероприятий по социальной ответственности	И	6	12	8,4	–	10,8	–	12,3
Описание ресурсоэффективности и ресурсоснабжения исследования	И	5	8	6,2	–	7,44	–	9,08
Составление графической оболочки	НР, И	7	9	7,8	2,8	9,36	3,43	11,42
Составление отчета о проделанной работе	И	6	12	8,4	–	10,08	–	12,3
Оформление графического материала	И	4	6	4,8	–	5,76	–	7,03
Подведение итогов	НР, И	5	8	6,2	4,46	7,44	5,45	9,08
Итого:					**31,11**	**106,27**	**37,97**	**128,93**

Таблица 15 – Линейный график работ

Этап	НР	И	Янв		Фев			Мар			Апр			Май			Июн
			10	20	30	40	50	60	70	80	90	100	110	120	130	140
1	4,1	–														
2	4,1	0,08														
3	5	16,69														
4	1,84	6,15														
5	–	7,03														
6	–	10,54														
7	–	6,15														
8	–	7,03														
9	14,05	14,05														
10	–	12,3														
11	–	9,08														
12	3,43	11,42														
13	–	12,3														
14	–	7,03														
15	5,45	9,08														

– научный руководитель (НР);

– исполнитель ВКР (И).

5.3 Расчет сметы затрат на выполнение проекта

В состав затрат на создание проекта включается величина всех расходов, необходимых для реализации комплекса работ, составляющих содержание данной разработки. Расчет сметной стоимости ее выполнения производится по следующим статьям затрат:

– материалы и покупные изделия;
– заработная плата;
– социальный налог;
– расходы на электроэнергию (без освещения);
– амортизационные отчисления;
– командировочные расходы;
– оплата услуг связи;
– арендная плата за пользование имуществом; прочие (накладные расходы) расходы.

5.3.1 Расчет затрат на материалы

К данной статье расходов относится стоимость материалов, покупных изделий, полуфабрикатов и других материальных ценностей, расходуемых непосредственно в процессе выполнения работ над объектом проектирования. Сюда же относятся специально приобретенное оборудование, инструменты и прочие объекты, относимые к основным средствам, стоимостью до 40 000 руб. включительно. Цена материальных ресурсов определяется по соответствующим ценникам или договорам поставки. Кроме того, статья включает так называемые транспортно-заготовительные расходы, связанные с транспортировкой от поставщика к потребителю, хранением и прочими процессами, обеспечивающими движение (доставку) материальных ресурсов от поставщиков к потребителю. Сюда же включаются расходы на совершение сделки куплипродажи (т.н. транзакции). Приближенно они оцениваются в процентах к отпускной цене закупаемых материалов, как правило, это 5–20 %. Исполнитель работы самостоятельно выбирает их величину в указанных границах.

Результаты расчета материальных затрат представлены в таблице 16.

Таблица 16– Расчет затрат на материалы

Наименование материалов	**Цена за ед., руб.**	**Количество**	**Сумма, руб.**
Бумага для принтера формата А4	250	1 уп.	250
Картридж для принтера черный	1290	1 шт.	1290
Итого:			**1540**

Допустим, что ТЗР составляют 5 % от отпускной цены материалов, тогда расходы на материалы с учетом ТЗР равны $C_{мат}$ = 1540 × 1,05 = 1617 руб.

5.3.2 Расчет заработной платы

Определим заработную плату научного руководителя и исполнителя проекта на основе трудоемкости выполнения каждого этапа и величины месячного оклада.

Также необходимо определить среднедневную тарифную заработную плату по следующей формуле:

$$ЗП_{дн-т} = \frac{МО}{25}, \quad (7)$$

Где МО – месячный оклад. Расчет производится с учетом того, что в году 300 рабочих дней и, следовательно, в месяце в среднем 25 рабочих дней при шестидневной рабочей неделе.

Расчет затрат на заработную плату приведен в таблице 17. Затраты времени взяты из расчета затраченных рабочих дней с округлением до целого. Также необходимо учесть в составе заработной платы премии, дополнительные зарплаты и районной надбавки, для этого используются следующие коэффициенты соответственно: $К_{ПР}$ = 1,1; $Кд_{опЗП}$ = 1,188 (для шестидневной рабочей недели); $К_Р$ = 1,3. Таким образом, для перехода от тарифной заработной платы к полному

заработку, необходимо тарифную заработную плату умножить на интегральный коэффициент $К_И$ = 1,1 ·1,188 ·1,3 = 1,699.

Таблица 17 – Затраты на заработную плату

Исполнитель	**Оклад, руб./мес**	**Среднедневная ставка, руб./раб.день**	**Затраты времени, раб.дни**	**Коэффициент**	**Фонд з/платы, руб.**
НР	33664	1346,56	31	1,699	70922,0
И	9489	379,56	106	1,699	68356,5
Итого:					**139278,5**

5.3.3 Расчет затрат на социальный налог

Затраты на единый социальный налог (ЕСН), включающий в себя отчисления в пенсионный фонд, на социальное и медицинское страхование, составляют 30 % от полной заработной платы по проекту: $C_{соц} = C_{зп} \cdot 0{,}3$.

Итак, $C_{соц}$ = 139278,5·× 0,3 = 41783,55 руб.

5.3.4 Расчет затрат на электроэнергию

Затраты на электроэнергию, потраченную на работу используемого оборудования, рассчитываются по формуле:

$$Сэл.об = Роб \times tоб \times Цэ, \qquad (8)$$

где $Р_{об}$ – мощность, потребляемая оборудованием, кВт; $t_{об}$ – время работы оборудования, час; $Ц_э$ – тариф на 1 кВт·ч (для ТПУ – 6,59 руб./кВт/ч).

Время работы оборудования вычисляется на основе данных таблицы 2 ($Т_{РД}$ исполнителя) из расчета, что продолжительность рабочего дня равна 8 часов:

$$tоб = Трд \times Кt, \qquad (9)$$

где $К_t \leq 1$ – коэффициент использования оборудования по времени, равный отношению времени его работы в процессе выполнения проекта к $Т_{РД}$.

Мощность, потребляемая оборудованием, определяется по формуле:

$$Р_{об} = Р_{ном} \times К_с, \qquad (10)$$

где $Р_{ном}$ – номинальная мощность оборудования, кВт; $К_с \leq 1$ – коэффициент нагрузки, зависящий от средней степени использования номинальной мощности.

Расчет затрат на электроэнергию для технологических целей приведен в таблице 18.

Таблица 18 – Затраты на электроэнергию технологическую

Наименование оборудования	**Время работы оборудования $t_{об}$, час**	**Потребляемая мощность $Р_{об}$, кВт**	**Затраты $Э_{об}$, руб**
Персональный компьютер	106·8·0,9	0,3	1508,85
Струйный принтер	1	0,1	0,66
Итого:			**1509,51**

5.3.5 Расчет амортизационных расходов

Рассчитаем амортизацию используемого оборудования за время выполнения проекта по следующей формуле:

$$С_{АМ} = \frac{Н_А \times Ц_{ОБ} \times t_{рф} \times n}{F_Д}, \qquad (11)$$

где $Н_А$ – годовая норма амортизации единицы оборудования; $Ц_{ОБ}$ – балансовая стоимость единицы оборудования с учетом ТЗР; $F_Д$ – действительный годовой фонд времени работы соответствующего оборудования; $t_{рф}$ – фактическое время работы

оборудования в ходе выполнения проекта; n – число задействованных однотипных единиц оборудования.

Для ПК принимаем $F_Д = 300 \cdot 8 = 2400$ часов (для 300 рабочих дней в году), для принтера $F_Д = 300 \cdot 0{,}02 = 6$ часов. Как и в предыдущем пункте для ПК $t_{рф}$ принимается равным $95{,}4 \cdot 8 = 763{,}2$; для принтера – 1. Показатель $Н_А$ определяется на основе срока полезного использования оборудования, который берется из постановления правительства. Рассматриваемое оборудование: ПК и принтер, относится группе «Машины офисные прочие», срок полезного использования которого от 2 до 3 лет, в данном случае возьмем 3 года. $Н_А$ является обратной величиной и в данном случае $Н_А = 0{,}33$. Выполним расчеты амортизационных расходов для ПК и принтера соответственно.

$$С_{АМ\ ПК} = \frac{0{,}33 \cdot 50000 \cdot 763{,}2 \cdot 1}{2400} = 5247{,}04\ руб$$

$$С_{АМ\ принтер} = \frac{0{,}33 \cdot 5000 \cdot 1 \cdot 1}{6} = 275\ руб.$$

Итого начислено амортизации:

$$С_{АМ} = 5247{,}04 + 275 = 5522{,}04\ руб.$$

5.3.6 Расчет расходов, учитываемых непосредственно на основе платежных (расчетных) документов (кроме суточных)

Сюда относятся:

– командировочные расходы, в т.ч. расходы по оплате суточных, транспортные расходы, компенсация стоимости жилья;

– арендная плата за пользование имуществом;

– оплата услуг связи;

– услуги сторонних организаций.

Для выполнения работ командировочные расходы, арендная плата за пользование имуществом и оплата услуг сторонних организаций не предусмотрены.

5.3.7 Расчет прочих расходов

В статье «Прочие расходы» отражены расходы на выполнение проекта, которые не учтены в предыдущих статьях, их следует принять равными 10% от суммы всех предыдущих расходов, т.е.

$$С_{проч} = ((С_{мат} + С_{зп} + С_{соц} + С_{эл.об} + С_{ам}) \cdot 0{,}1), \qquad (12)$$

В данном случае:

$$С_{проч} = (1617 + 139278{,}5 + 41783{,}55 + 1509{,}51 + 5522{,}04) \cdot 0{,}1$$

$$= 18971{,}06 \text{ руб.}$$

5.3.8 Расчет общей себестоимости разработки

Проведя расчет по всем статьям сметы затрат на разработку, можно определить общую себестоимость проекта «Изучение методов сегментации анатомических структур сердца». Расчет общей себестоимости разработки приведен в таблице 19.

Таблица 19 – Смета затрат на разработку проекта

Статья затрат	**Условные обозначения**	**Сумма, руб.**
Материальные и покупные изделия	$С_{мат}$	1617
Основная заработная плата	$С_{зп}$	139278,5
Отчисления в социальные фонды	$С_{соц}$	41783,55
Расходы на электроэнергию	$С_{эл}$	1509,51
Амортизационные отчисления	$С_{ам}$	5522,04
Непосредственно учитываемые расходы	$С_{нр}$	0
Прочие расходы	$С_{проч}$	18971,06
Итого:		**208 681,7**

Таким образом, затраты на разработку составили *С* = **208 681,7** руб.

5.3.9 Расчет прибыли

Прибыль от реализации проекта в зависимости от конкретной ситуации (масштаб и характер получаемого результата, степень его определенности и коммерциализации, специфика целевого сегмента рынка и т.д.) может определяться различными способами. Если исполнитель работы не располагает данными для применения «сложных» методов, то прибыль следует принять в размере 5–20 % от полной себестоимости проекта. В данном случае прибыль составляет 20 868,17 руб. (10 %) от расходов на разработку проекта.

5.3.10 Расчет НДС

НДС составляет 20% от суммы затрат на разработку и прибыли. В данном случае это (208681,7+20868,17) · 0,2 = 45 909,97

5.3.11 Цена разработки ВКР

Цена разработки ВКР равна сумме полной себестоимости, прибыли и НДС, в данном случае:

$$Ц_{НИР} = 208\,681{,}7 + 20868{,}17 + 45\,909{,}97 = 275\,459{,}8 \text{ руб.}$$

5.4 Оценка экономической эффективности проекта

Результатом реализации рассматриваемого проекта станет приложение, позволяющее медицинскому персоналу быстрее обрабатывать полученную информацию, тем самым ускоряя процесс анализа полученных медицинских данных. Решение может быть использовано как бюджетными, так и частными медицинскими учреждениями. В зависимости от этого можно рассматривать два вида эффективности проекта: бюджетный и коммерческий соответственно.

В первом случае возможна экономия бюджетных средств различных уровней, благодаря дополнительному источнику дохода непосредственно медицинских учреждений, путем более эффективной работы медицинского персонала, не отвлекаясь на смежные операции и уменьшение рабочего времени. Во втором случае финансовый эффект возможен благодаря непосредственно

получению дополнительного дохода от увеличения предоставленных услуг, путем уменьшения времени на анализ медицинских данных.

Немаловажным показателем качества выполненной работы является показатель экономического эффекта. Экономическая эффективность обусловлена факторами экономического эффекта. В рассматриваемой НИ прямого экономического эффекта нет. Факторы его получения заключаются в следующем:

– сокращение времени обработки медицинских данных работником,

– сокращение врачебных ошибок,

– увеличение эффективности работы медицинского персонала, путем сокращения времени на смежную работу, и более быстрого анализа.

Для получения количественной оценки эффекта, а, следовательно, и эффективности результатов работы нужно провести специальное дополнительное исследование.

5.5 Вывод по разделу

В ходе разработки части дипломной работы, затрагивающей финансовую и ресурсную эффективность, была проведена оценка потребителей. Также был проведен SWOT-анализ, анализ конкурентных решений, что позволило выявить слабые и сильные стороны разрабатываемого проекта и найти пути улучшения конкурентоспособности продукта. Также были рассмотрены статьи затрат на реализацию проекта, проведена оценка экономической эффективности.

Количественная оценка экономического эффекта не может быть дана, так как нет статистических данных о точном количестве проанализированной информации на схожести. Исследование проводилось лишь на тестовых данных и ранее аналоги не использовалось в медицинских учреждениях.

6 Социальная ответственность

6.1 Введение

Целью работы является алгоритмическое и программное обеспечение анализа схожести наборов медицинских данных по их описанию и содержимому. Так как уровень развития здравоохранения и доступность медицинских услуг во многом определяет качество жизни населения, данная работа имеет высокую социальную значимость.

Актуальность работы заключается в том, что созданное алгоритмическое и программное обеспечение, могут использоваться не только исследователями в области медицинских или иных данных, но в перспективе и медицинскими учреждениями для устранения разнообразности наборов медицинских данных, путем анализа их схожести, тем самым упрощая анализ и восприятие различной медицинской информации в целом. Проектирование решения, его дальнейшие разработка и использование осуществляется в основном на автоматизированном рабочем месте.

В связи с этим нужно рассмотреть вопросы обеспечения производственной и экологической безопасности, а также безопасности в чрезвычайных ситуациях во время разработки и конечной эксплуатации приложения.

Целю написания данного раздела является принятие проектных решений, исключающих несчастные случаи в производстве, защиту здоровья и снижение вредных воздействий на окружающую среду. В ходе данного исследования необходимо изучить возможные вредные и опасные факторы, влияющие на исполнителей при разработке и эксплуатации программного продукта и разработать решения для минимизации их влияния.

6.2 Правовые и организационные вопросы обеспечениябезопасности

6.2.1 Правовые нормы трудового законодательства для рабочейзоны оператора ПЭВМ

Согласно трудовому кодексу, продолжительность рабочего дня не должна

превышать 24 часов в неделю для работников до 16 лет, 35 часов для работников в возрасте от 16 до 18 лет или являющихся инвалидами I или II групп. В остальных случаях рабочая неделя должна длиться не более 40часов.

Так как работа программиста относится к категории работ, требующей постоянного взаимодействия с ПЭВМ, то рекомендуется организовывать перерывы длительностью 10–15 минут через каждые 40–60 минут работы. Кроме того, продолжительность непрерывной работы с ЭВМ не должна превышать 60 минут. При работе в ночную смену, с 22:00 до 6:00 необходимо увеличивать длительность перерывов на 30%. Во время таких перерывов рекомендуется выполнять комплекс упражнений для снижения нервно-эмоционального напряжения, утомления зрительного анализатора, устранения влияния гиподинамии и гипокинезии, предотвращения развития позотонического утомления. Помимо этого, организацией должен быть предоставлен перерыв длиною не менее 30 минут для приема пищи.

Также организация обязана предоставлять ежегодный отпуск продолжительностью 28 календарных дней. Дополнительные отпуска предоставляются работникам, занятым на работах с вредными или опасными условиями труда, работникам имеющими особый характер работы,работникам с ненормированным рабочим днем и работающим в условиях Крайнего Севера и приравненных к нему местностях [22].

6.3 Производственная безопасность

6.3.1Анализ вредных и опасных факторов

Согласно ГОСТ [23] опасные и вредные факторы согласно природе воздействия выделяют в следующие группы:

физические; химические; биологические; психофизиологические.

При работе с персональными электронно-вычислительными машинами (ПЭВМ) возможны воздействия факторов каждой из этой группы. В таблице 20 представлены опасные и вредные факторы, которые могут возникнуть при работе программиста с ПЭВМ.

Таблица 20 Опасные и вредные факторы

Источник фактора, наименование видов работ	Факторы (по ГОСТ 12.0.003-74)		Нормативные документы
	Вредные	Опасные	
Работа за ПЭВМ	– Повышенный уровень шума на рабочем месте; – Повышенная или пониженная температура воздуха рабочей зоны; – Повышенная или пониженная влажность воздуха; – Повышенный уровень электромагнитных излучений; – Отсутствие или недостаток естественного света; – Недостаточная освещенность рабочей зоны; – Повышенная яркость света; – Пониженная контрастность;	– Повышенное значение напряжения;	– СанПиН 2.2.2/2.4.1340–03. Санитарно-эпидемиологические правила и нормативы «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы»; – СанПиН 2.2.4.548–96. Гигиенические требования к микроклимату
	– Прямая и отраженная блескость; – Повышенная пульсация светового потока; – Умственное перенапряжение; – Перенапряжение анализаторов; – Монотонность труда; – Эмоциональные перегрузки.		производственных помещений; – ГОСТ 12.1.003–83 ССБТ. Шум. Общие требования безопасности; – "Трудовой кодекс Российской Федерации" от 30.12.2001 N 197-ФЗ (ред. от 05.02.2018). ГОСТ 12.2.032-78 «ССБТ Рабочее место при выполнении работ сидя»

6.3.1.1 Физические перегрузки

Для деятельности разработчиков характерны физические перегрузки, связанные с рабочей позой, длительной концентрацией и умственном напряжении (вызванном в том числе информационной нагрузкой). В СанПиН 2.2.2/2.4.1340-03 [24] предусмотрены меры для минимизации воздействия данного вредного фактора.

В стандарте прописаны нормы организации рабочего процесса, связанного с использованием электронно-вычислительной техники, в частности – организации перерывов. В данном случае, труд проектировщиков и разработчиков системы можно отнести к группе I категории В, так как они работают с ПЭВМ более 4 часов за смену, не имея возможности сменить вид деятельности. Для них регламентировано суммарное время перерывов не менее 90 минут при восьмичасовой смене и 140 минут – при двенадцатичасовой. В случаях, когда характер работы требует постоянного взаимодействия с видео дисплейным терминалом (набор текстов или ввод данных и т.п.) с 79 напряжением внимания и сосредоточенности, при исключении возможности периодического переключения на другие виды трудовой деятельности, не связанные с ПЭВМ, рекомендуется организация перерывов на 10–15 мин через каждые 45–60 мин работы. При возникновении у работающих с ПЭВМ зрительного дискомфорта и других неблагоприятных субъективных ощущений, несмотря на соблюдение санитарно-гигиенических и эргономических требований, рекомендуется применять индивидуальный подход с ограничением времени работы с ПЭВМ.

6.3.1.2 Микроклимат

Для уменьшения негативного влияния микроклимата помещения рекомендуется придерживать следующих показателей температуры и влажности помещения, в зависимости от времени года. Ниже представлена таблица 21 оптимальных показателей микроклимата для категории работ 1а [25].

Таблица 21 Оптимальные величины показателей микроклимата на рабочих местах производственных помещений

Период года	Температура воздуха, °С	Температура поверхностей, °С	Относительная влажность воздуха, %	Скорость движения воздуха, м/с
Холодный	22–24	21–25	60–40	0,1
Теплый	23–25	22–26	60–40	0,1

В таблице 22 представлены допустимые величины показателей микроклимата на рабочих местах производственных помещений.

Таблица 22 Допустимые величины показателей микроклимата на рабочих местах производственных помещений

Период года	Температура воздуха, °С		Температура поверхностей, °С	Относительная влажность воздуха,	Скорость движения воздуха, м\с	
	Диапазон ниже оптимальных величин	Диапазон выше оптимальных величин			Ниже оптимальных величин, не более	Выше оптимальных величин, не более
Холодный	20,0–21,9	24,1–25,0	19–26	15–75	0,1	0,1
Теплый	21,0–22,9	25,1–28,0	20–29	15–75	0,1	0,2

В результате измерений была получена температура воздуха, равная 26 °С в теплое время года, что превышает оптимальное значение. Однако, согласно СанПиН 2.2.4.548–96, данный показатель находится в допустимом пределе 25,1–28,0 °С .

6.3.1.3 Освещение

Для снижения зрительной нагрузки при работе за компьютером важно подобрать комфортное освещение.

Источники освещения делятся на два основных типа: естественные и искусственные. Как правило, естественное освещение используется в дневное время суток, а искусственное преимущественно в темное время суток. Однако при недостаточном освещении оба этих типа могут комбинироваться.

Рабочее место в данной работе организовано таким образом, что тыльная сторона монитора обращена к окну. Размер рабочего места 1,0х0,7м., размер окна 2,5х1,7 м.

Индекс помещения рассчитывается по следующей формуле [26]:

$$И_{п} = \frac{ab}{(h_1 - h_2)(a + b)}, \quad (13)$$

Где $И_П$ – индекс помещения, h_1 – высота помещения, h_2 – высота рабочего стола, a – длина помещения, b – ширина помещения. Так длина помещения составила 5,12 м, ширина – 3,5 м, высота помещения – 2,5 м, а высота стола 0,75 м. Таким образом, индекс помещения равен:

$$И_{п} = \frac{3{,}5 * 5{,}12}{(2{,}5 - 0{,}75)(3{,}5 + 5{,}12)} = 1{,}188$$

Далее была рассчитана освещенность по формуле [27]:

$$E = \frac{К_{св} * К_{л} * СП_{л} * U}{S * k_з * 100}, \quad (14)$$

где $К_{СВ}$ – количество светильников, $К_Л$ – количество ламп всветильнике, $СП_Л$ – световой поток лампочки, U – коэффициент использования, S – площадь, $k_З$ – коэффициент запаса.

Поскольку потолок белого цвета, стены и пол – светлые, то коэффициент отражения потолка и стен равен 0,7, для пола – 0,3. Согласно таблице, коэффициент использования помещения U равен 55.

В качестве основных источников освещения используются 3 светильника, каждый из которых содержит по 2 светодиодные лампы.

Мощность лампы 13 W, цветовая температура 4500 K. Коэффициент запаса равен 1,2. Таким образом, освещенность равна:

$$E = \frac{3 * 2 * 1200 * 55}{17{,}92 * 1{,}2 * 100} = 184 \text{ лк}$$

Нормативные значения для операций с высоким уровнем зрительной (объекты от 0,3 до 0,5 мм) работы составляют от 200 до 400 лк [28]. Таким образом, для достаточного освещения рабочего места необходимо либо работать в светлое время суток, либо использовать настольную лампу. Впроцессе работы в качестве дополнительного точечного источника света выступала настольная лампа, оснащенная источником света с цветовой температурой 3000 K и мощностью 9 W, расположенная слева от монитора.

6.3.1.4 Шум

Важное значение имеет показатель уровня шума в помещении. Повышенный уровень шума негативно воздействует на нервную и слуховую системы человека, приводя к различным заболеваниям, а также снижая работоспособность. К основным источникам шума при работе с ПЭВМ можно отнести шум систем охлаждения ПЭВМ, шум работающего жесткого диска. К другим источникам шума относятся уличный шум, бытовой шум.

Согласно ГОСТ [29] уровень шума не должен превышать 50 дБ. Для снижения уровня шума рекомендуется проводить мероприятия по техническому обслуживанию ПЭВМ.

6.3.1.5 Электробезопасность

В связи с тем, что работа выполняется с помощью ПЭВМ, то соблюдение правил электробезопасности имеет непосредственное отношение к работе. К основным источникам электрического воздействия, находящимся непосредственно рядом с рабочим местом, относятся ПЭВМ, настольная лампа, электрические розетки.

На территории России электроприборы включены в сеть под напряжением 220 В, с частотой 50 Гц, что является опасным фактором воздействия на организм человека. Поражающими факторами электрического тока являются термическое, биологическое и электролитическое воздействия.

6.3.1.6 Электромагнитные излучения

Компьютер, как и многие электроприборы, является источником электромагнитного излучения. Воздействие электромагнитного излучения на организм определятся различными параметрами, такими как напряженность поля, поток энергии, частота колебаний. В таблице 23 приведены временныедопустимые уровни электромагнитного поля (ЭМП).

Таблица 23. Временные допустимые уровни ЭМП, создаваемых ПЭВМ на рабочих местах

Наименование параметров		Временные допустимые уровни электромагнитного поля
Напряженность электрического поля	5 Гц–2 кГЦ	25 В/м
	2 кГЦ–400 кГц	2,5 В/м
Плотность магнитного потока	5 Гц–2 кГЦ	250 нТл
	2 кГЦ–400 кГц	25 нТЛ
Напряженность электростатического тока		15 В/м

6.3.1.7 Пожарная безопасность

Пожары являются опасным фактором, который может привести к потере информации, хранящейся на ПЭВМ, а также, что немало важно, причинить вред здоровью человека. Поэтому меры противопожарной безопасности помогут избежать негативных последствий.

Главными вероятными источниками пожара могут стать неисправная электропроводка, поврежденные электроприборы и легковоспламеняющиеся вещества, например бумага.

Для предотвращения возможных пожаров используются исправные электроприборы, а также сетевые фильтры с плавким предохранителем. В качестве мер быстрого реагирования используются дымовые датчики, а также пожарная сигнализация.

Также с сотрудниками должен проводиться инструктаж по действиям при возникновении данной чрезвычайной ситуации. Во всех служебных помещениях должен присутствовать план эвакуации людей. После окончания работы все оборудование должно быть выключено, а сеть обесточена.

Для предотвращения пожара рабочее помещение должно быть оборудовано устройствами, предназначенными для локализации и ликвидации возгорания на начальной стадии – первичными средствами пожаротушения. К ним относятся огнетушители, вода, песок, пожарная сигнализация для извещения о наступлении пожара.

6.3.1.8 Опасность поражения электрическим током

Основные причины воздействия тока на человека: случайные проникновения или приближение на опасное расстояние к токоведущим частям, появление напряжения на металлических частях машин в результате повреждения изоляции.

Поражающее действие электрического тока зависит от значения и длительности протекания тока через тело человека, рода и частоты тока, индивидуальных свойств человека. Наиболее опасным для человека является ток с

частотой 20–100 Гц. Опасной величиной является ток, равный 0,001А, а смертельный 0,1А.

При поражении электрическим током могут возникать следующие виды воздействий: термическое (ожоги), электрическое, механическое и биологическое (паралич мышц).

Согласно ГОСТ Р 12.1.019–2009 для обеспечения защиты от поражения электрическим током при прикосновении к металлическим нетоковедущим частям, которые могут оказаться под напряжением в результате повреждения изоляции, применяют защитное заземление, систему защитных проводов, защитное отключение, электрическое разделение сети, контроль изоляции и пр.

Технические способы и средства применяют раздельно или в сочетании друг с другом так, чтобы обеспечивалась оптимальная защита при нормальном функционировании электроустановок и при возникновении аварийных ситуаций. Офисное помещение относится к категории помещений без повышенной опасности, однако необходимо соблюдать меры предосторожности при работе с компьютером. Так, не рекомендуются следующие действия:

1) закладывать провода и шнуры за газовые и водопроводные трубы, за батареи отопительной системы;

2) выдергивать штепсельную вилку из розетки за шнур, усилие должно быть приложено к корпусу вилки;

3) работать на средствах вычислительной техники и периферийном

оборудовании, имеющих нарушения целостности корпуса, нарушения изоляции проводов, неисправную индикацию включения питания, с признаками электрического напряжения на корпусе;

4) класть на средства вычислительной техники и периферийное оборудование посторонние предметы.

6.4 Экологическая безопасность

6.4.1 Анализ воздействия на окружающую среду

В данном подразделе рассматривается характер воздействия проектируемого решения на окружающую среду. Выявляются предполагаемые источники загрязнения окружающей среды, возникающие в результате разработки и реализации, предлагаемых в ВКР решений.

Создание и применение методов сегментации, а также работа за ПК не являются экологически опасными работами, потому объект, на котором производилось внедрение системы, а также объекты, на которых будет производиться ее использование пользователями ПК относятся к предприятиям пятого класса, размер селитебной зоны для которых равен 50 м.

Непосредственно методы, созданные и примененные в ходе выполнения выпускной квалификационной работы, не наносят вреда окружающей среде ни на стадиях разработки, ни на стадиях эксплуатации. Однако, средства, необходимые для разработки, внедрения и эксплуатации могут наносить вред окружающей среде.

В процессе выполнения работы возможны такие отходы, как бумага, неисправные детали ПЭВМ, неработающие электролампы.

В состав компонентов ПЭВМ входят такие загрязняющие вещества, как ртуть, входящая в состав жидкокристаллических экранов, мышьяк и бериллий, используемые при производстве плат, свинец, применяемый для пайки, поливинилхлорид, используемый для изготовления изоляции кабелей. Сейчас некоторые из данных веществ запрещены для использования, например свинец, но остальные используются до сих пор. Добыча данных материалов уже сама по себе наносит вред окружающей среде [30].

Поэтому ПЭВМ и ее компоненты по окончании срока службы необходимо утилизировать соответствующим образом, чтобы избежать дальнейшего негативного влияния на окружающую среду. Кроме того, необходимо сдавать макулатуру в специальные пункты приема.

В частности, современные ПК производят практически без использования вредных веществ, опасных для человека и окружающей среды. Исключением являются аккумуляторные батареи компьютеров и мобильных устройств. В аккумуляторах содержатся тяжелые металлы, кислоты и щелочи, которые могут наносить ущерб окружающей среде, попадая в гидросферу и литосферу, если они были неправильно утилизированы. Для утилизации аккумуляторов необходимо обращаться в специальные организации, специализировано занимающиеся приемом, утилизацией и переработкой аккумуляторных батарей.

Люминесцентные лампы, применяющиеся для искусственного освещения рабочих мест, также требуют особой утилизации, т. к. в них присутствует от 10 до 70 мг ртути, которая относится к чрезвычайно-опасным химическим веществам и может стать причиной отравления живых существ, а также загрязнения атмосферы, гидросферы и литосферы. Сроки службы таких ламп составляют около 5 лет, после чего их необходимо сдавать на переработку в специальных пунктах приема.

Юридические лица обязаны сдавать лампы на переработку и вести паспорт для данного вида отходов.

6.5 Безопасность в чрезвычайных ситуациях

6.5.1 Наиболее вероятная чрезвычайная ситуация

В ходе выполнения ВКР могут возникнуть чрезвычайные ситуации (ЧС) техногенного, экологического, стихийного и биолого-социального характера. Ниже представлены наиболее вероятные и опасные ЧС по каждой категории:

–техногенные: пожары;

– стихийные: ураганы;

– биолого-социальные: эпидемии.

Наиболее вероятной чрезвычайной ситуацией, которая может возникнуть в офисе во время разработки проекта, является пожар. Его могут вызвать следующие причины:

1) несоблюдение мер пожаробезопасности;

2) обрыв проводов;

3) замыкание электропроводки оборудования.

Существует комплекс мероприятий, позволяющих уменьшить вероятность возникновения пожара и более оперативно ликвидировать последствия:

1) регулярные проверки;

2) отключения оборудования при покидании рабочего места;

3) проведение инструктажа работников по действиям при пожаре;

4) проведение учебной тревоги два раза в год;

5) установка систем противопожарной сигнализации;

6) оборудование запасных выходов при пожаре;

7) создание плана эвакуации и размещение его экземпляров в доступных местах.

Для обеспечения пожарной безопасности необходимо выполнение комплекса организационных, режимных, технических и эксплуатационных мероприятий по предупреждению пожаров.

6.5.2 Меры по предупреждению чрезвычайной ситуации

В офисных помещениях, в котором происходила разработка и эксплуатация, присутствуют пыль, материалы и вещества, способные при взаимодействии с кислородом только гореть, поэтому данные помещения относятся к категории В.

К мерам, устраняющим возможные причины возникновения пожаров, относятся следующие мероприятия:

1) эксплуатационные – выбор и использование современных автоматических средств сигнализации, автоматических стационарных систем тушения пожаров, первичных средств пожаротушения, разработка методов и применение устройств ограничения распространения огня и т. п.

2) организационные – обучение сотрудников правилам пожарной

безопасности, разработка и реализация норм и правил пожарной безопасности, инструкций правильной эксплуатации рабочего оборудования, разработка планов эвакуации людей и т. д.

Пожар может нанести не только вред здоровью, но и материальный ущерб. Применимо к выполняемой работе в случае пожара могут быть уничтожены бумажные документы и\или электронные носители информации.

Для защиты информации рекомендуется использовать облачные хранилища данных для данных и документов. Для исходных кодов программ рекомендуется использовать системы контроля версий.

6.6 Выводы по разделу

В разделе «Социальная ответственность» рассматриваются вопросы соблюдения прав персонала на труд, выполнения требований к безопасности и гигиене труда, к промышленной безопасности, охране окружающей среды и ресурсосбережению. Целями данного раздела являются принятие проектных решений, исключающих несчастные случаи в производстве, и снижение вредных воздействий на окружающую среду.

В результате выполнения работ по данному разделу были проанализированы моральные, общественные и экологические возможные негативные последствия и ущерб здоровью человека в результате разработки, внедрения и использования рассматриваемого решения.

В рамках рассмотрения вопросов по правовой безопасности было отмечено, что для данной работы актуальны вопросы не только охраны жизни и здоровья работников проектируемой медицинской системы и её пользователей, но и вопросы защиты медицинских данных пользователей, которые регулируются законом отдельно.

Также были определены вредные и опасные факторы производства и эксплуатации предлагаемого решения. Все опасные факторы связаны с актуальной и для разработчиков, и для конечных пользователей спецификой работы, связанной с использованием ПЭВМ: с сосредоточенной работой с большим объемом информации и текста в статичной позе и с вредным излучением от непосредственно техники. Причиной возникновения опасных факторов так же в основном является использование компьютерной и офисной техники.

Для каждого фактора были рассмотрены причины его возникновения, влияние на человека и меры по предотвращению его возникновение или минимизации последствий воздействия. Был произведен расчет соответствия освещенности рабочего места и помещения установленным нормативам.

Минимальный уровень освещенности соблюдается, для повешения этого показателя до более оптимального используется дополнительное местное освещение.

В вопросах экологической защиты выявлен один основной вопрос по правильной утилизации отходов или их переработке.

Наиболее вероятной чрезвычайной ситуацией при разработке предлагаемого решения является пожар. Были рассмотрены основные меры по предотвращению возникновения ЧС, устранению и минимизации последствий.

Также отдельно стоит подчеркнуть то, что работа носит социальный характер, так как косвенно направлен на улучшение состояния здоровья граждан и, как следствие, качество их жизни.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы были выполнены все поставленные задачи, разработано решение (алгоритм и программное обеспечение) для анализа схожести наборов медицинских данных.

В процессе работы проводился обзор литературы по заданной тематике, поиск медицинских выборок для дальнейшей работы с ними. Далее создан алгоритм сравнения и написан программный код и подготовлены данные для обучения модели. В ходе обучения модели, были получены неплохие результаты. Проведено исследование, в котором выяснилось, что алгоритм сравнения с двумя интервалами работает эффективнее, чем с четырьмя интервалами. Далее, данная модель была проверена

Алгоритм был опробован на совсем разных выборках, не участвовавших в обучении. Перед тестированием программный код был автоматизирован и улучшен, для лучшего восприятия и уменьшения время работы.

Проверка модели оказалась успешной, хотя и не без ошибок. Поэтому было решено использовать дополнительные пункты анализа схожести, автоматизированный сбор данных с сайта Kaggle. Также было проведено исследование, в котором выяснилось, что для эффективной работы, одних ключевых слов недостаточно, поэтому они используются совместно с описанием, если вдруг оно отсутствует. Помимо того, что данный алгоритм используется в анализе схожести, параллельно он определяет тему документа, в частности относится ли он вообще к медицинской тематике и если относится, то к какому направлению ближе.

После добавления дополнительных алгоритмов результат вывода улучшился. Большинство одинаковых веществ определены верно, но есть и некоторые разные вещества, которые модель определила, как одинаковые. Это получилось, потому что некоторые вещества имеют похожие значения, зачастую даже слишком похожие, поэтому модель их определяет как одинаковые. Или,

наоборот, одинаковые элементы, определяются, как разные из-за того, что, например в разных выборках возраст людей может отличаться значительно.

В дальнейшем данный алгоритм будет улучшаться, будет вычисляться процент схожести, а также будет добавлен программный модуль, направленный на медицинских работников, который позволяет привести медицинскую терминологию к общему виду, позволяя при этом задавать каноничные именования, а также контролировать процесс замены терминов, стандартизировать медицинские термины, что сокращает вероятность появления недопустимых ошибок, а также структурирует медицинские данные для удобной работы с ними.

Модуль будет иметь универсальный характер, применение не ограничивается отраслью медицины и может использоваться везде, где требуется повышенная точность при работе с опечатками и сокращениями.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Kaggle [Электронный ресурс.] URL: https://www.kaggle.com/ (дата обращения 21.05.2021).
2. A Study on Student Performance, Engagement, and Experience with Kaggle InClass data Challenges [Электронный ресурс.] URL: https://www.tandfonline.com/doi/full/10.1080/10691898.2021.1892554 (дата обращения 21.05.2021).
3. Datasetsearch [Электронный ресурс.] URL: https://datasetsearch.research.google.com/ (дата обращения 21.05.2021).
4. Реестр открытых данных на AWS [Электронный ресурс.] URL: https://registry.opendata.aws/ (дата обращения 21.05.2021).
5. Открытые наборы данных Azure [Электронный ресурс.] URL: https://docs.microsoft.com/en-us/azure/azure-sql/public-data-sets (дата обращения 21.05.2021).
6. Reddit [Электронный ресурс.] URL: https://www.reddit.com/r/datasets/top/?sort=top&t=all (дата обращения 21.05.2021).
7. UCI [Электронный ресурс.] URL: https://archive.ics.uci.edu/ml/index.php (дата обращения 21.05.2021).
8. Библиотеки CMU [Электронный ресурс.] URL: https://guides.library.cmu.edu/az.php (дата обращения 21.05.2021).
9. Открытые базы данных на Github [Электронный ресурс.] URL: https://github.com/awesomedata/awesome-public datasets#machinelearning (дата обращения 21.05.2021).
10. Тематическое моделирование [Электронный ресурс.] URL: http://www.machinelearning.ru/wiki/index.php?title=Тематическое_моделирование (дата обращения 21.05.2021).
11. Латентное_размещение_Дирихле [Электронный ресурс.] URL: https://ru.wikipedia.org/wiki/Латентное_размещение_Дирихле (дата обращения 21.05.2021).

12. Латентно-семантический_анализ [Электронный ресурс.] URL: https://ru.wikipedia.org/wiki/Латентно-семантический_анализ (дата обращения 21.05.2021).

13. Неотрицательное_матричное_разложение [Электронный ресурс.] URL: https://ru.wikipedia.org/wiki/Неотрицательное_матричное_разложение (дата обращения 21.05.2021).

14. Тематическое моделирование в действии lda [Электронный ресурс.] URL: https://lambda-it.ru/post/tematicheskoe-modelirovanie-v-deistvii-lda (дата обращения 21.05.2021).

15. Semantic latent dirichlet allocation for automatic topic extraction [Электронный ресурс.] URL: https://www.tandfonline.com/doi/abs/10.1080/02522667.2016.1165000 (дата обращения 21.05.2021).

16. Genism [Электронный ресурс.] URL: https://radimrehurek.com/gensim/ (дата обращения 21.05.2021).

17. Обучение с учителем и без учителя [Электронный ресурс.] URL: https://neurohive.io/ru/osnovy-data-science/obuchenie-s-uchitelem-bez-uchitelja-s-podkrepleniem/ (дата обращения 21.05.2021).

18. GAN [Электронный ресурс.] URL: https://en.wikipedia.org/wiki/Generative_adversarial_network (дата обращения 21.05.2021).

19. Generative Adversarial Nets [Электронный ресурс.] URL: https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf (дата обращения 21.05.2021).

20. Python [Электронный ресурс.] URL: https://www.python.org (дата обращения 21.05.2021).

21. Gini [Электронный ресурс.] URL: https://wiki.loginom.ru/articles/gini-index.html (дата обращения 21.05.2021).

22. "Трудовой кодекс Российской Федерации" от 30.12.2001 N 197-ФЗ (ред. от 27.12.2018).

23. ГОСТ 12.0.003–74. ССБТ. Опасные и вредные производственные факторы. Классификация;

24. СанПиН 2.2.2/2.4.1340–03. Санитарно-эпидемиологическиеправила и нормативы «Гигиенические требования к персональным электронно-вычислительным машинам и организации работы»;

25. СанПиН 2.2.4.548–96. Гигиенические требования к микроклимату производственных помещений;

26. Таблица «Коэффициент использования светового потока» [Электронный ресурс] – URL: https://www.websor.ru/metodkoefi.html (дата обращения 08.05.2021);

27. Как самостоятельно выполнить расчет освещенности помещения [Электронный ресурс] / Электрика своими руками – URL: https://elektrika- svoimi-rykami.com/raschet-osveshheniya/raschet-osveshheniya (дата обращения 08.05.2021);

28. СНиП 23-05-95. Естественное и искусственное освещение;

29. ГОСТ 12.1.003–83 ССБТ. Шум. Общие требования безопасности;

30. Грязная и опасная сторона технологий [Электронный ресурс] /Мир ПК – URL: https://www.osp.ru/pcworld/2013/06/13035804 (дата обращения 08.05.2021)

Приложение А
(справочное)

Literature review on algorithmic and software for analyzing the similarity of medical datasets in terms of their description and content

Студент

Группа	ФИО	Подпись	Дата
8ИМ9М	Соколовский Дмитрий Евгеньевич		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ ИШИТР	Аксенов С.В.	к.т.н.		

Консультант-лингвист отделения иностранных языков ШБИП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИЯ ШБИП	Степура С.Н.	к.ф.н.		

1 Literature review on algorithmic and software for analyzing the similarity of medical datasets in terms of their description and content.

Many medical institutions and organizations do not use a conceptual approach to organize and manage data quality, especially in the long term. The value of medical records and the data based on them grows over time. Even the introduction of electronic medical records (EMR) did not simplify the processing of data in real time to the proper extent, because the functionality of the software used is very limited.

Here are the main problems with the processing of medical data:

- different levels of quality of electronic medical records;
- lack of compatibility, as well as the complexity of clinical systems;
- the complexity of the process of collecting, searching and analyzing data;
- the need to process incomplete or missing data;
- coverage and sampling of data;
- regulatory requirements and bureaucratic processes.

The topic of our research can be indirectly attributed to the solution of many of the problems mentioned above, in the case of the need for processing or missing data, as well as the complexity of the process of collecting, searching, and analyzing data.

1.1 Specialized sites with datasets

At the international level, the following sites are considered the most popular:

- Kaggle;

Kaggle is updated daily by enthusiasts and contains one of the largest database libraries on the Internet [1].

Kaggle is a community driven machine learning platform. It contains many tutorials that cover hundreds of real-life ML problems. Of course, the quality of the data may vary, but they are all completely free. It is also possible to load your own database into the library.

There are many data science training resources available, from Datacamp to Udacity, all for learning data science. But if you are the kind of person who loves to learn by doing, then Kaggle is arguably the best platform to improve your skills through hands-on research projects.

Kaggle, which bills itself as "your home for data science," was originally a machine learning competition site but data science resources can now be found there as well. Several main features of Kaggle are worth mentioning:

Datasets: Many datasets of various types and sizes that are free to download. Here you can find interesting data to learn or test your modeling skills.

Kaggle (The Kaggle Team 2018) is a platform for predictive modeling and analytics competitions where participants compete to produce the best predictive model for a given dataset. It is well known for its competitions (e.g., Rhodes 2011), some of which come with rich monetary prizes (e.g., Howard 2013). There are also learning competitions (Agarwal 2018), designed to help novices improve their data mining skills. Winners are typically expected to share their code, and occasionally newly emerged algorithms are introduced to the broad community, for example, deep neural networks (Hinton and Dahl 2012) and XGBoost (Chen and Guestrin 2016).

In 2015, Kaggle InClass was introduced, as a self-service platform to conduct competitions. These competitions can be private, limited to members of a university course, and are easy to setup. This is an opportunity for educators to provide a vehicle for students to objectively test their learning of predictive modeling. As a competition, with an independent clear performance metric, along with a dynamic leader board, students can see how their model predictions compare with the models produced by other students. Being able to make multiple submissions over a several week time frames enables them to try out approaches to improve their models. This article examines the educational benefits of conducting predictive modeling competitions in class on performance, engagement, and interest [2].

- Dataset Search from Google;

Dataset Search is a reliable source of information for research [3]. In it, all datasets are sorted by:

- relevance;
- file format;
- the type of license;
- topic;
- the latest update.

The databases are downloaded here by various international organizations such as the World Health Organization, Statista, and Harvard.

- Open Data Registry on AWS;

In the open data registry on AWS, anyone can share a data package or find the one they need [4]. And with Amazon Data Analytics tools, you can conduct research based on the data you find. The creators of these databases include Facebook's Data for Good, NASA's Space Act Agreement, and the Space Telescope Institute for Space Research.

- Open datasets Microsoft Azure;

Azure open datasets are regularly updated and made available to application developers and researchers [5]. They contain US government data, other statistical and scientific data, and information from online services that Microsoft collects about its users.

In addition, Azure offers users a set of tools to help them create their own cloud databases, migrate SQL workloads to Azure while maintaining full SQL Server compatibility, and build data-driven mobile and web applications.

- R / datasets;

In SubReddit DataSet anyone can publish open-source databases [6]. Look there to find a cool dataset and do some interesting research with it.

- UCI Machine Learning Repository;

The UCI offers over 500 different datasets that cover topics such as banking marketing, car valuation, lung cancer diagnostics, and more [7]. You can sort data packets by:

- standard tasks;
- data types;
- areas of use;
- subject;
- CMU libraries.

Carnegie Mellon University has its own collection of publicly available datasets that you can use for research. There you will find detailed databases on American culture, music, and history that no other aggregator provides [8].

- Open databases on Github.

This is a great collection of open-source datasets divided by industry [9].

1.2 Topic Modeling Techniques

Topic modeling is an unsupervised machine learning technique for defining topics in a collection of documents. What is the difference from conventional clustering? The purpose of clustering is to divide the body of documents into groups, and then the purpose of thematic modeling is to highlight the main topics from a set of statements. Most importantly, clustering is deductive and topic modeling is inductive.

There are various methods of topic modeling [10]:

1. Latent Dirichlet Allocation (LDA) [11].
2. Latent Sernantic Analysis (LSA) [12].
3. Non-Negative Matrix Factorization (NNMF) [13].

For our research, the Dirichlet Latent Deployment method is an excellent solution.

This topic modeling method was proposed by David Blei, Andrew Ng, and Michael Jordan in 2003. LDA belongs to a family of generative probabilistic models in which topics are represented by the probabilities of occurrence of each word in each set. Documents, in turn, can be presented as combinations of topics. A unique feature of LDA models is that topics do not have to be different, and words can appear in multiple topics; this imparts some ambiguity to the topics being defined, which can be useful for coping with the flexibility of the language. Figure 1 illustrates this method.

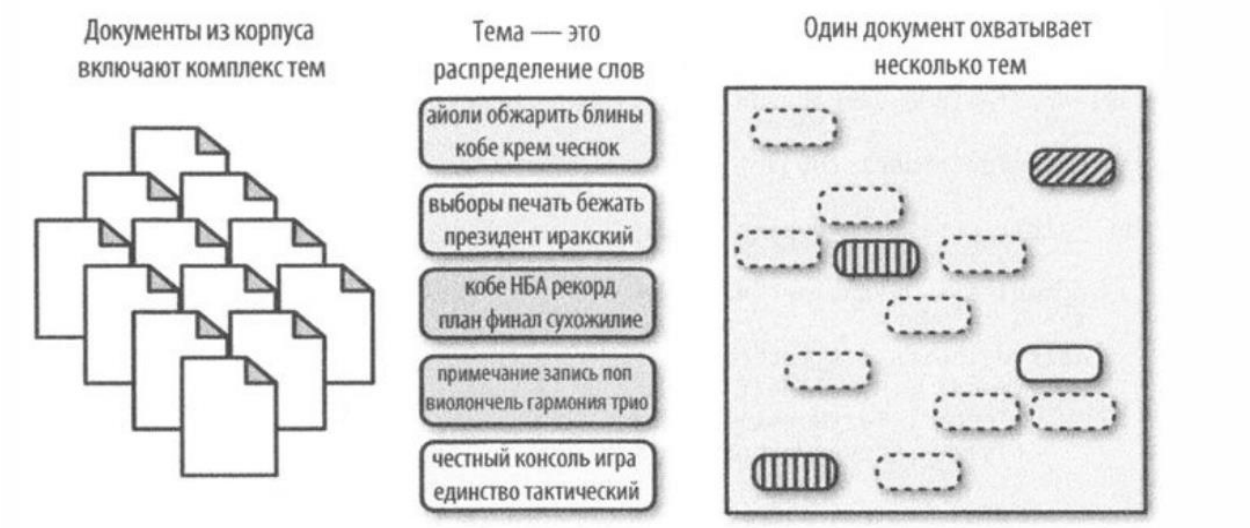

Figure 1 - Latent Dirichlet Allocation Method [14]

Blei and his colleagues found that the Dirichlet distribution, a family of continuous distributions (a way of measuring grouping by distributions), is a convenient way to identify topics that appear in a corpus, as well as appear in different combinations in each document in the corpus. In fact, Latent Dirichlet Allocation (LDA) gives us an observable word or lexeme by which we can try to determine the likely topic, the distribution of words in each topic, and the combination of topics in the document. To leverage topic modeling techniques in your application, you need to create a custom pipeline that extrapolates topics from unstructured text data and a way to save the best model.

The topic modeling pipeline would look like this:

1. Loading the case

2. Text preprocessing

2.1 Deleting stop words

2.2 Remove punctuation

2.3 Lemmatization of words

3. Creating a dictionary

4. Selection of the optimal number of topics

With the inception and uncontrollable growth of digital documents, Automatic Topic Extraction is found to be an active research topic. To handle the processing of documents, variety of algorithms was presented in literature for topic extraction using distribution Modeling and classification approaches. Among different modeling methods of topic extraction, Latent Dirichlet Allocation (LDA) is one of the important algorithms for Topic Identification. Even though LDA is popular technique for topic identification, it turns to difficulty in determining the model parameters and, suffers with finding the degree of similarity and semantic handling. To handle these challenges, a new method is proposed for automatic topic extraction. Accordingly, this method called, Semantic Latent Dirichlet Allocation (SLDA) is proposed by extending LDA in a semantic way. A new mathematical computation is included where the model parameters are estimated using new membership degree in Semantic Latent Dirichlet Process along with semantic similarity measure. The experimentation is carried out with two different databases and it observed that SLDA outperformed by showing better when compared with existing LDA in Jaccord Coefficient [15].

Implementation with gensim [16].

Gensim models have more configurable parameters than scikit-learn. Gensim was originally developed as a topic modeling library. Libraries used in this implementation.

- gensim - contains all topic modeling algorithms;

- pandas - required for working with the language corpus;

- nltk - contains lemmatization algorithms and a stop word dictionary;

- pyLDAvis - plugin for visualization of the LDA model;
- matplotlib - visualization library.

1.3 Machine learning and data analysis

You can train a neural network in different ways: with a teacher, without a teacher, with reinforcement. But how to choose the optimal algorithm and how do they differ? There are several ways to collect IKEA furniture. Each one leads to an assembled sofa or chair. But depending on the piece of furniture and its components, one method will be more sensible than others.

Do you have an instruction manual and all the parts you need? Just follow the instructions. Well, how is it? You can throw away the manual and work on your own. But if you confuse the procedure, it is up to you to decide what to do with this pile of wooden bolts and planks.

It is the same with deep learning. The developer will prefer an algorithm with a specific learning method, considering the type of data and the task at hand. Figure 2 shows the result of training a neural network - image clustering.

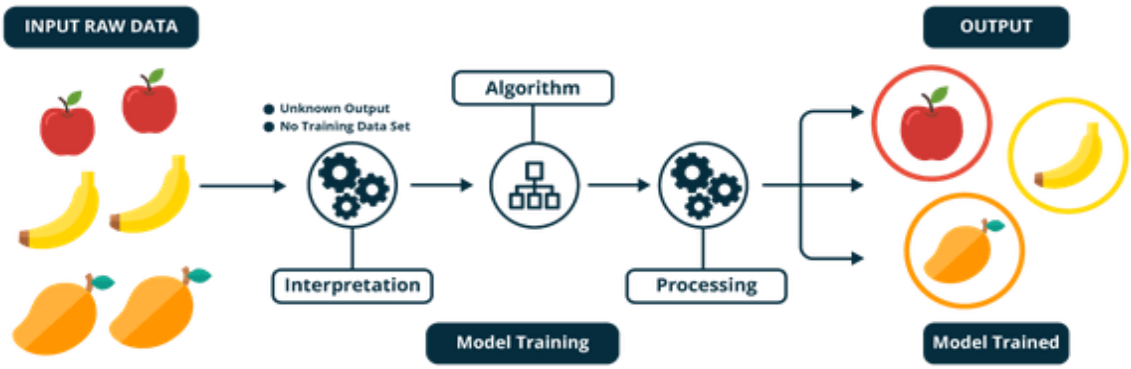

Figure 2 - The result of training a neural network [17]

The result of training a neural network is image clustering.

In supervised learning, the neural network trains on a labeled dataset and predicts responses, which are used to evaluate the accuracy of an algorithm on the training data. In unsupervised training, the model uses unlabeled data, from which the algorithm independently tries to extract features and dependencies.

Part-teacher learning is something in between. It uses a small amount of labeled data and a large set of unlabeled data. Reinforcement learning trains the algorithm using a reward system. The agent receives feedback in the form of rewards for doing the right thing. Animals are expelled in a similar way.

For each learning method, consider examples of data and tasks suitable for it.

Supervised Learning

Supervised learning requires a complete set of labeled data to train the model at all stages of its construction.

The presence of a fully labeled dataset means that each example in the training set corresponds to the answer that the algorithm should receive. Thus, a tagged dataset of flower photographs will train a neural network where roses, daisies or daffodils are depicted. When the network receives a new photo, it will compare it with examples from the training dataset to predict the answer.

Mainly supervised learning is used to solve two types of problems: classification and regression.

In classification problems, the algorithm predicts discrete values corresponding to the numbers of the classes to which the objects belong. In a training dataset with animal photos, each image will have a corresponding label - "cat", "koala" or "turtle". The quality of the algorithm is assessed by how accurately it can correctly classify new photos with koalas and turtles.

Regression tasks, on the other hand, are associated with continuous data. One example, linear regression, calculates the expected value of a variable y given specific x values.

More utilitarian machine learning tasks involve many variables. For example, a neural network that predicts the price of an apartment in San Francisco based on its

area, location, and public transport availability. The algorithm performs the work of an expert who calculates the price of an apartment based on the same data.

Thus, supervised learning is most suitable for tasks when there is an impressive set of reliable data for training the algorithm. But this is not always the case. Lack of data is the most common problem in machine learning.

1.3.2 Unsupervised Learning

Perfectly labeled and clean data is not easy to come by. Therefore, sometimes the algorithm is faced with the task of finding previously unknown answers. This is where learning without a teacher comes in.

In unsupervised learning, the model has a dataset, and there is no explicit indication of what to do with it. The neural network tries to independently find correlations in the data, extracting useful features and analyzing them.

Clustering. Even without the specialized knowledge of an expert ornithologist, you can look at a collection of photographs and divide them into groups by bird species, based on feather color, beak size or shape. This is where clustering lies - the most common task for unsupervised learning. The algorithm picks up similar data, finds common features, and groups them together.

Anomaly detection. Banks can detect fraudulent transactions by identifying unusual behavior in customer buying behavior. For example, it is suspicious if the same credit card is used in California and Denmark on the same day. Likewise, unsupervised learning is used to find outliers in data.

Associations. Choose diapers, applesauce, and a baby sippy cup from the online store, and the site will recommend that you add a bib and baby monitor to your order. This is an example of associations: some characteristics of an object are correlated with other characteristics. By considering a couple of the key attributes of an object, the model can predict others with which there is a connection.

Autoencoders. Autoencoders take input, encode it, and then try to recreate the initial data from the resulting code. There are not many real-life situations where a

simple autoencoder is used. But add layers and the possibilities expand: By using noisy and original versions of images for training, autoencoders can remove noise from video data, images, or medical scans to improve data quality.

In unsupervised learning, it is difficult to calculate the accuracy of an algorithm because the data lacks correct answers" or labels. But tagged data is often unreliable or too expensive to obtain. In such cases, giving the model the freedom to find dependencies can lead to good results.

1.3.3 Part-teacher learning

Semi-supervised learning is characterized by its name: the training dataset contains both tagged and unlabeled data. This method is especially useful when it is difficult to extract important features from the data, or when it is a tedious task to mark up all objects.

Part-teacher learning is often used to solve medical problems where a small amount of labeled data can lead to a significant increase in accuracy.

This machine learning method is common for the analysis of medical images such as CT scans or MRI scans. An experienced radiologist can mark up a small subset of scans that show tumors and diseases. But manually marking up all scans is too time-consuming and expensive task. However, a neural network can extract information from a small fraction of labeled data and improve prediction accuracy compared to a model that trains exclusively on untagged data.

A popular training method that requires a small set of labeled data is to use a generative adversarial network, or GAN [18] [19]. Is a class of machine learning frameworks designed by Ian Goodfellow and his colleagues in 2014. Two neural networks contest with each other in a game (in the form of a zero-sum game, where one agent's gain is another agent's loss).

The most popular are supervised learning methods, when a model (machine classifier) is built from a corpus of labeled data (training sample), which is then applied to new, unlabeled texts.

Traditional machine learning methods of this type such as naive Bayes classifier, decision trees, support vector machines, logistic regression, etc.

After analyzing the literature on the topic of this work, some points were highlighted, for each sub-point, methods, platforms, and solutions were selected, which we considered more effective, and which will be used in the research process.

In particular, the collection of medical data sets will be done from the Kaggle platform. Thematic modeling will use the latent dirichlet placement method implemented using gensim, and the analysis of the similarity of these data will be carried out using supervised machine learning, in conjunction with algorithmic and software in the python language [20].

Приложение Б

РУКОВОДСТВО ПОЛЬЗОВАТЕЛЯ

После алгоритмического и программного обеспечения анализа схожести наборов медицинских данных началась разработка визуальной части, добавив для начала алгоритм сравнения из имеющихся датасетов.

Главное окно программы представлено на рисунке 1.

Рисунок 1 – Главное окно программы

Нажав на кнопку файл, появится выпадающее меню с пунктом справка, в которой представлена последовательность для корректной работы программы, как на рисунке 2.

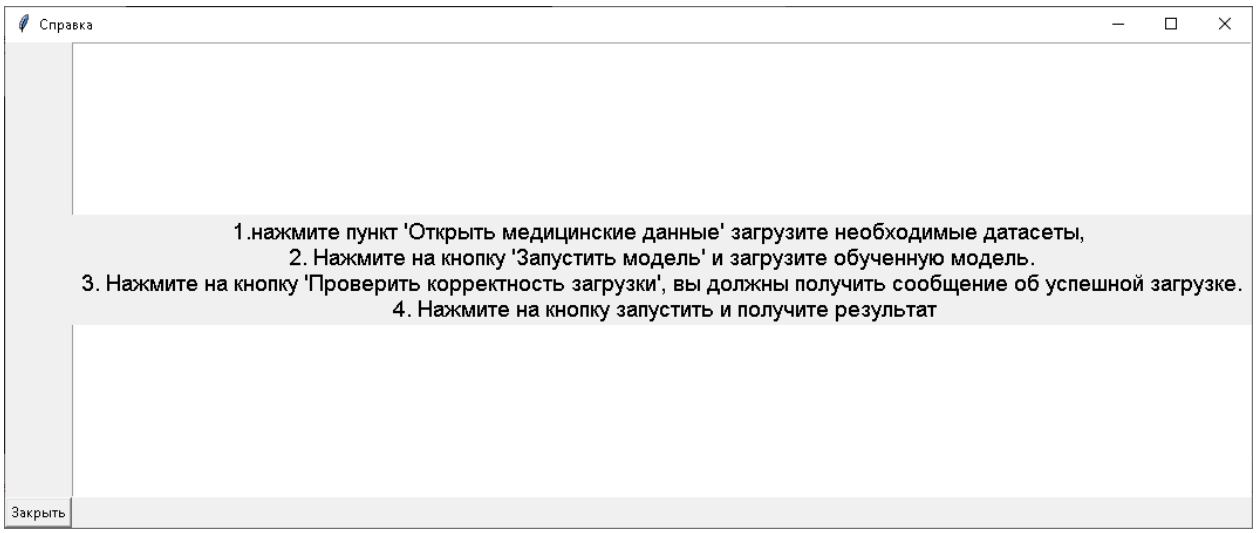

Рисунок 2 – Справка

Окно пункта «Открыть медицинские данные» представлено на рисунке 3.

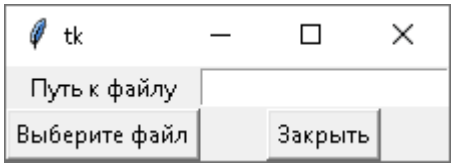

Рисунок 3 – Кнопка «Открыть медицинские данные»

Окно пункта «Запустить модель» представлено на рисунке 4.

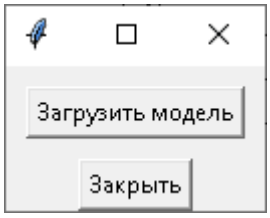

Рисунок 4 – Кнопка «Запустить модель»

После загрузки данных и модели нажимаем кнопку «Проверить корректность загрузки», если все загружено корректно должна появиться надпись, как на рисунке 5, если что-то не загрузилось, появится надпись как на рисунке 6.

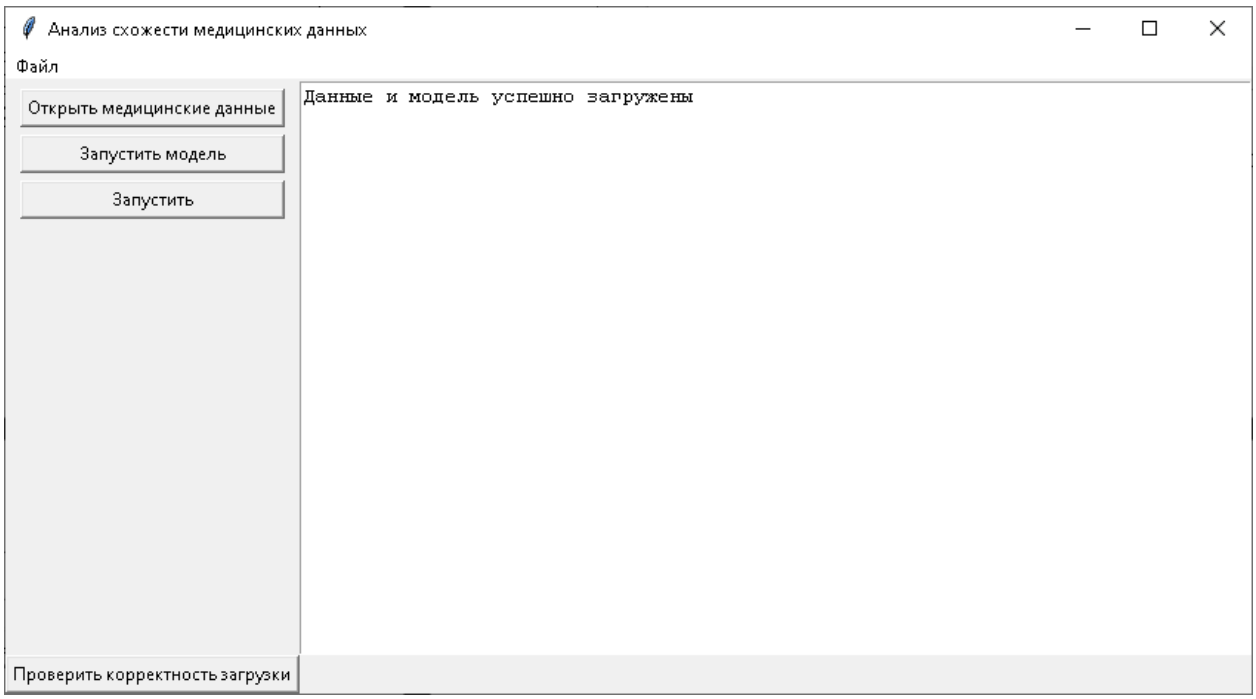

Рисунок 5 – Данные загружены

Рисунок 6 – Данные не загружены

Далее после всех манипуляций нажимаем на кнопку «Запустить», выполняется алгоритм и работа обученной модели. Мы получаем результат, который отображается в отдельном окне, как на рисунке 7

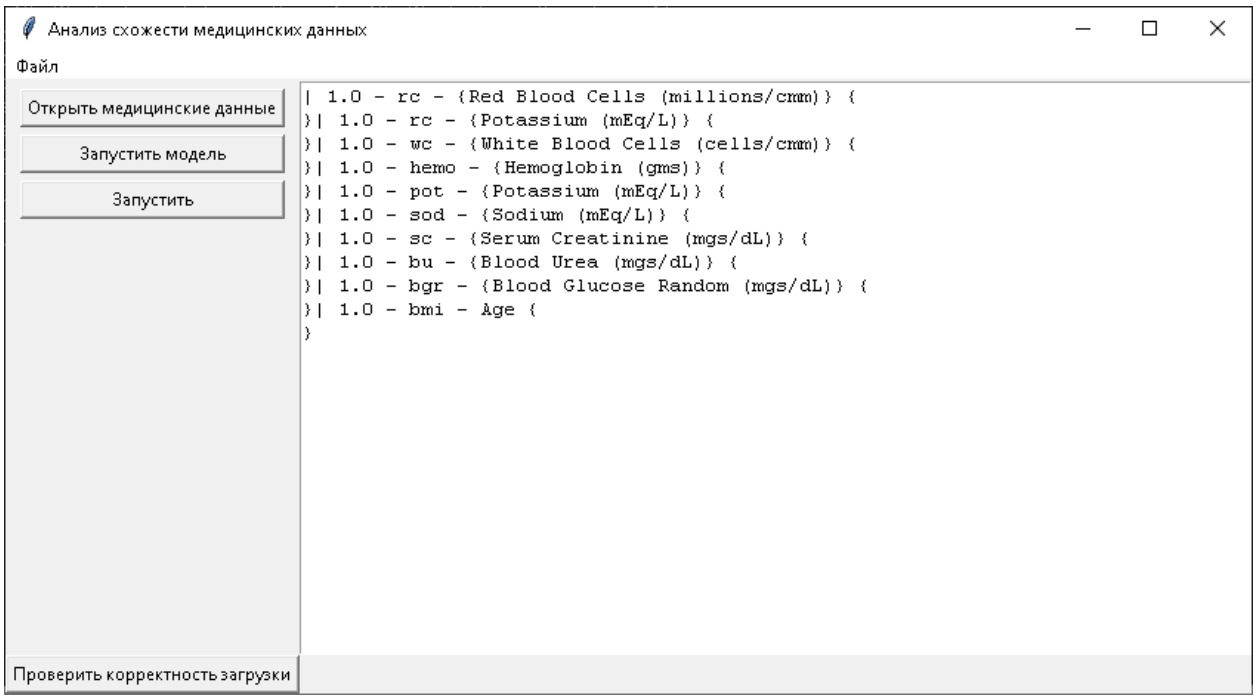

Рисунок 7– Результат работы программы