«Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)»

(СПбГЭТУ «ЛЭТИ»)

Направление	09.04.02 – Информационные системы и технологии
Профиль	Распределенные вычислительные комплексы систем реального времени
Факультет	ФКТИ
Кафедра	ИС

К защите допустить

Зав. кафедрой _____ Цехановский В.В.
подпись

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА

Тема: Методика анализа политик безопасности на основе онтологического представления предметной области

Студент		_____ *подпись*	Кузнецов М.Д.
Руководитель	к.т.н., доцент *(Уч. степень, уч. звание)*	_____ *подпись*	Новикова Е.С.
Консультант	к.э.н., доцент *(Уч. степень, уч. звание)*	_____ *подпись*	Жукова Т.Н.

Санкт-Петербург

2021

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Утверждаю

Зав. кафедрой ИС

_____ Цехановский В.В.

подпись

«____»_____ 2021 г.

Студент Кузнецов М.Д. Группа 5374

Тема работы: Методика анализа политик безопасности на основе онтологического представления предметной области.

Место выполнения ВКР: Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И.Ульянова (Ленина)

Исходные данные (технические требования): —

Содержание ВКР: В разделе «Анализ предметной области» произведен анализ литературы и работ в данной области, в разделе «Проектирование» проведено проектирование инструментария для сбора датасета, «Реализация» приведены некоторые аспекты реализации.

Перечень отчетных материалов: пояснительная записка, иллюстрационный материал.

Дополнительные разделы: «Составление бизнес-плана по коммерциализации результатов научно-исследовательской работы магистранта».

Дата выдачи задания «____»_____ 2021 г.

Дата представления ВКР к защите «____»_____ 2021 г.

Студент		_____ *подпись*	Кузнецов М.Д.
Руководитель	к.т.н., доцент *(Уч. степень, уч. звание)*	_____ *подпись*	Новикова Е.С.
Консультант	к.э.н., доцент *(Уч. степень, уч. звание)*	_____ *подпись*	Жукова Т.Н.

КАЛЕНДАРНЫЙ ПЛАН ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Утверждаю

Зав. кафедрой ИС

_____ Цехановский В.В.

подпись

«____»_____ 2021 г.

Студент Кузнецов М.Д. Группа 5374

Тема работы: Методика анализа политик безопасности на основе онтологического представления предметной области.

№ п\п	Наименование работ	Срок выполнения
1	Обзор литературы по теме работы	01.02 – 28.02
2	Анализ предметной области	01.03 – 31.03
3	Проектирование инструментария разметки	01.04 – 15.04
4	Реализация инструментария разметки	15.04 – 30.04
5	Оформление пояснительной записки	01.05 – 07.05
6	Оформление иллюстративного материала	07.05 – 15.05

Студент _____ Кузнецов М.Д.

подпись

Руководитель к.т.н., доцент _____ Новикова Е.С.

(Уч. степень, уч. звание) *подпись*

Консультант к.э.н., доцент _____ Жукова Т.Н.

(Уч. степень, уч. звание) *подпись*

РЕФЕРАТ

Поясн. зап. 98 стр., 28 рис., 14 табл., 33 ист., 1 прил.

АВТОМАТИЗИРОВАННАЯ ОБРАБОТКА ЕСТЕСТВЕННОГО ЯЗЫКА, ПОЛИТИКИ БЕЗОПАСНОСТИ, ПОЛЬЗОВАТЕЛЬСКИЕ СОГЛАШЕНИЯ

Объектом исследования являются способы эффективной автоматизированной формализации политик безопасности.

Цель работы – разработать эффективный план автоматизированных способов формализации политик безопасности на основе онтологического представления, разработать инструменты создания обучающей выборки для автоматизированной формализации политик безопасности.

Политики конфиденциальности предоставляют пользователям информацию о том, как их личные данные собираются, обрабатываются и передаются третьим лицам. Однако в большинстве случаев они написаны нечетко и непрозрачно. Поэтому важно сделать политику конфиденциальности ясной и прозрачной для конечного пользователя. В этой работе исследуется применение методов LSA, LDA, POS для обнаружения семантических тем, представленных в политике конфиденциальности. Также тестируется POS подход с пулами синонимов. Однако такие строгие способы обработки текста не очень точны. Использование методов глубокого обучения с онтологическим представлением предметной области делает возможной точную формализацию политики конфиденциальности. Для этого были созданы поисковый робот и инструмент аннотации. С помощью поисковый бота был получен набор данных из 592 политик конфиденциальности.

ABSTRACT

Privacy policies provide end users information about how they personal data collected, processed and shared with third parties. However, in major cases they are written in unclear and not transparent manner. So, it is important to make privacy policies clear and transparent to end user. In this work, application of the LSA, LDA, POS techniques to detect semantic topics presented in the privacy policy are investigated. Also POS with synonyms pools are tested. However, more strict ways of text processing are not very accurate. Using deep learning techniques with ontology representation of subject field making accurate privacy policy formalization possible. For that the crawler and annotation tool were created. Finally, the privacy policies dataset consisting of 592 was obtained with crawler.

ТЕРМИНЫ И ОПРЕДЕЛЕНИЯ

В настоящей пояснительной записке к выпускной квалификационной работе используются следующие термины с соответствующими определениями.

Датасет — набор данных для обучения моделей анализа естественного языка

Вэб-скрейпинг — это технология извлечения данных из вэб-страниц путем из скачивания и обработки

ПЕРЕЧЕНЬ СОКРАЩЕНИЙ И ОБОЗНАЧЕНИЙ

В настоящей пояснительной записке к выпускной квалификационной работе используются следующие сокращения и обозначения.

LSA — (от англ. Latent Semantic Search) латентно-семантический анализ

LDA — (от англ. Latent Dirichlet Allocation) латентное размещение Дирихле

POS — (от англ. Part Of Speech) разложение по частям речи

TF-IDF — (от англ. Term Frequency – Inverse Document Frequency) инверсная частотная характеристика документа

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В настоящее время персональные данные широко используются в предоставлении цифровых услуг, их персонализации и улучшении. Персональные данные – это любые данные, идентифицировать физическое лицо [1]. Таким образом, личные данные – это не только биометрическая информация, данные о состоянии здоровья человека, а также фото абонента услуги, местонахождение, информация о приложении и устройстве, которое можно использовать для отслеживания действий и информации о потребителе. Несколько массовых утечек персональных данных за последнее десятилетие привело к ужесточению законодательных требований во многих страны по всему миру. В настоящее время требуется, чтобы все личные данные обрабатывались надежно, а действия с ними были ясны и прозрачны для субъекта данных в соответствии с его или ее явно указанным согласием. Политики конфиденциальности поставщиков услуг, онлайн-согласие пользователей – единственные законные документы, сообщающие конечным пользователям, как собираются, обрабатываются их личные данные и передается третьим лицам. Однако в большинстве случаев эти документы написаны так, что их довольно сложно понять. И в настоящее время ситуация такова, что законодательные требования соблюдаются производителями продукции и поставщиков услуг, но конечные пользователи дают свое согласие без четкого понимания того, как обрабатываются их личные данные, потому что политика конфиденциальности и онлайн-согласие пользователя читаются редко из-за их сложности и низкой читабельности. Это ведет к ситуациям, когда конечные пользователи не знают о рисках для конфиденциальности связанных с использованием определенной услуги или устройства.

В настоящее время сфера информационных технологий является одной из самых быстрорастущих, в ней решается множество задач прикладного характера. Одной из прогрессивных технологий является технология глубокого

обучения. Полученные с помощью глубокого обучения модели способны решать широкий спектр прикладных задач. Однако, у данного подхода имеется существенный недостаток – необходимость датасета для обучения. Датасет играет критически важную роль в формировании результата в целом. Если качество датасета будет посредственным, либо он окажется недостаточно объемным, то поставленная задача не будет решена с адекватной точностью. В то же время сбор датасета – работа кропотливая и рутинная. Отличным решением является автоматизация данного процесса, возможно не полная, но частичная. Она абсолютно точно увеличит скорость сбора информации, что позволит за то же время собирать более объемные датасеты, и как следствие более точные модели будут получены после обучения.

На момент написания выпускной квалификационной работы актуальность данной работы является высокой, так как формализация политик безопасности открывает возможности для более простой и ясной формулировки политик безопасности, что уменьшит количество угроз персональным данным. Также становится возможной разрабатка методик расчета рисков потребления цифровых услуг и устройств.

Цель работы – разработать эффективный план автоматизированных способов формализации политик безопасности на основе онтологического представления, разработать инструменты создания обучающей выборки для автоматизированной формализации политик безопасности. В ходе выполнения предполагается реализация инструментов для сбора датасета, который будет применен для обучения классификатора. Классификатор позволит автоматизированно формализовать политики безопасности. По формализованному описанию политик станет возможной оценка рисков для персональных данных пользователей.

Для достижения данной цели необходимо:

– провести анализ предметной области;

– разработать методики сбора, очистки и разметки обучающей выбор-

ки;

– спроектировать инструментарий для построения обучающей выборки, обеспечивающей обучение классификатора с учетом онтологического представления предметной области;

– реализовать инструментарий для построения обучающей выборки, обеспечивающей обучение классификатора с учетом онтологического представления предметной области.

Выпускная квалификационная работа состоит из введения, четырех разделов и заключения. В первом разделе производится анализ предметной области. Во втором разделе описаны приемы и методики проектирования, аргументация их применения. В третьем разделе описан процесс разработки и полученные результаты. В четвертом разделе предложен план по коммерциализации научно-исследовательской работы магистранта.

1 Анализ предметной области

1.1 Вступление

Конфиденциальность в настоящее время вызывает растущую озабоченность. Это касается конфиденциальности личных данных. Персональные данные - это данные, идентифицирующие личность живого человека [1]. Широкое распространение оцифровки личных данных приводит к неожиданным и непредсказуемым угрозам конфиденциальности. Они могут быть как преднамеренными (т. Е. Организация намеренно собирает и использует личные данные для явно не заявленных целей), так и непреднамеренными (т. Е. Организация не обеспечивает требуемый уровень безопасности для личных данных) со стороны организации, которой пользователи предоставляют это.

Для защиты конфиденциальности пользователей было разработано несколько стандартов, в том числе Общий регламент ЕС по защите данных (GDPR) [1], COPPA [2], Правило конфиденциальности HIPAA [3] и другие. Эти стандарты обязывают организации четко указывать в политиках конфиденциальности, какие личные данные они используют, для каких целей, кем и как долго.

Существует множество проблем, связанных с политикой конфиденциальности: пользователи не всегда понимают прямые и побочные риски, связанные с обменом личными данными; правила обычно довольно длинные и непонятные, и пользователи не тратят времени на их чтение; не существует единого стандарта генерации политики конфиденциальности, несмотря на многочисленные попытки; информация, представленная в политике конфиденциальности, не соответствует реальной ситуации. Три первых проблемы взаимосвязаны. Хотя существуют исследования и подходы, посвященные определенным проблемам, не существует комплексного подхода, который позволял бы оценить риски политик конфиденциальности, представленных на естественном языке, и предоставить их пользователям в удобной фор-

ме.

В этой статье авторы рассматривают проблему расчета рисков конфиденциальности на основе анализа политики конфиденциальности, который позволит пользователям и организациям понять, какое влияние на конфиденциальность может оказать текст политики конфиденциальности. Предлагаемый подход основан на анализе текстов политики конфиденциальности, представленных сначала на естественном языке, создании и обработке онтологии политики конфиденциальности для каждой политики, указанной на естественном языке, и, наконец, вычислении рисков конфиденциальности с использованием сгенерированной онтологии. Предлагаемая онтология политики конфиденциальности, которая описывает различные сценарии использования данных, является ядром предлагаемого подхода.

Исследование было впервые представлено на 3-м Международном семинаре по атакам и защите для Интернета вещей (ADIoT 2020) [4]. В этой статье авторы расширили анализ соответствующего исследования и добавили сравнение существующих подходов к анализу политик конфиденциальности. Авторы предложили алгоритм расчета рисков конфиденциальности на основе политики конфиденциальности, тогда как в исследовании, представленном на ADIoT 2020, авторы рассматривали только частные случаи алгоритма. Кроме того, авторы добавили эксперимент, описывающий расчет рисков конфиденциальности для нескольких политик и их сравнение, тогда как в исследовании, представленном на ADIoT 2020, авторы продемонстрировали только один вариант использования.

Работа организована следующим образом. В разделе 2 анализируются родственные работы в этой области. В разделе 3 представлена разработанная методика оценки конфиденциальности. В разделе 4 описаны проведенные эксперименты. Статья завершается заключением и перспективами дальнейших исследований.

1.2 Обзор текущего состояния предметной области

Связанные работы можно разделить на три группы, учитывая проблемы политики конфиденциальности, изложенные во введении. Есть документы, связанные с анализом рисков конфиденциальности. Вторая группа работ связана с анализом политик, представленных на естественном языке, и их дальнейшим представлением в удобной форме. Для этого используются методы обработки естественного языка (NLP). И третья группа статей посвящена разработке единого стандарта политик конфиденциальности и их автоматизированной генерации. Для этого используются методы разработки формальных языков. Дальнейшие политики, определенные с использованием формальных языков, могут быть использованы для анализа или оценки частных рисков.

Как было сказано во Введении, эти три группы взаимосвязаны с точки зрения оценки рисков. Сначала тексты политики конфиденциальности, представленные на естественном языке, обрабатываются для формального определения политик конфиденциальности (с использованием некоторого формального языка), наконец, политики конфиденциальности, указанные на формальном языке, используются для расчета рисков конфиденциальности.

Анализ текстов политики конфиденциальности, представленных на естественном языке, рассматривается в статьях [5], [6], [9]. В статье [5] описан подход к автоматизированному извлечению и анализу политик конфиденциальности для приложений Android. Авторы используют подход TF-IDF (термин «частота и обратная частота документа») для построения вектора признаков из текста политик и классификатора машины опорных векторов (SVC) для обнаружения различных методов обработки данных, таких как контактный адрес электронной почты, Контактный номер телефона, местоположение GPS, Wi-Fi и т. Д. В правилах. Для обучения моделей авторы создали аннотированный корпус политик конфиденциальности APP-350 Corpus, до-

ступный по ссылке: https://www.usableprivacy.org.

В статье [6] описана семантическая структура PrivOnto для анализа политик конфиденциальности. PrivOnto использует в качестве входных данных набор аннотированных политик конфиденциальности и разработанную общую онтологию. Предлагаемая онтология представляет собой набор политик с определенными практиками в отношении данных с учетом конфиденциальности. Во-первых, эксперты проанализировали набор политик конфиденциальности и вручную аннотировали их, используя выделенные 11 категорий методов обработки данных (Собственный сбор / использование, сторонний обмен / сбор, выбор пользователя / контроль, доступ пользователя / редактирование / удаление, Хранение данных, Безопасность данных, Изменение политики, Не отслеживать, Международные и специальные аудитории, Другое). Эти категории служили основными концепциями для моделирования политик конфиденциальности. Исследователи аннотировали более 23000 практик обработки данных, извлеченных из 115 политик конфиденциальности, доступных по ссылке: https://www.usableprivacy.org. Затем аннотированный набор использовался для обучения фреймворка автоматизированному аннотированию. Авторы использовали краудсорсинг, машинное обучение и обработку естественного языка для автоматизированного аннотирования политик конфиденциальности для создания конкретных онтологий. Это исследование предлагает наиболее близкий к подходу подход, предложенный в данной статье, но авторы сосредотачиваются не только на выявлении практики обработки данных в тексте политик, но и на оценке рисков для персональных данных.

Онтологический подход к представлению политики конфиденциальности также предлагается в статьях [7], [8]. В [7] авторы разработали онтологию конфиденциальности PrOnto для проверки соответствия политики GDPR, но они генерируют онтологию вручную. В [8] предлагается подход к политике конфиденциальности, основанный на построении онтологии с использованием вопросов компетенции. Авторы данной статьи использовали этот подход

для разработки своей онтологии.

В документе [9] описывается подход машинного обучения к автоматическому обнаружению вариантов отказа от некоторых сборов и использования личных данных в текстах политики конфиденциальности и связанном расширении веб-браузера. Авторы [9] протестировали различные методы машинного обучения для анализа текста политики, такие как линейная регрессия и нейронные сети, и экспериментировали с различным набором функций. Ограничение подхода состоит в том, что для его применения требуется помеченный набор данных. Авторы реализовали разметку вручную. В статье [10] также рассматривается автоматическое обнаружение вариантов отказа в текстах политики конфиденциальности. Но авторы используют набор данных из статьи [6] для обучения своих моделей.

Разработка формальных языков для автоматизированной генерации и единой спецификации политик конфиденциальности рассматривается в статьях [11]–[15]. Формальный язык состоит из языкового алфавита и правил построения последовательностей с использованием символов алфавита, то есть языковой грамматики. Текст, указанный на таком языке, можно обработать математическими методами.

В документе [11] предлагается Платформа для корпоративных практик конфиденциальности (E-P3P), чтобы формализовать политику конфиденциальности на машиночитаемом языке. Этот язык может быть автоматически применен на предприятии с помощью механизма авторизации. Формализованная политика определяет, какие типы личной информации (PII), для каких целей и какими пользователями в организации могут быть использованы. Машиночитаемый язык включает терминологию и набор правил авторизации. Терминология включает категории данных, цели, пользователей данных, набор действий, набор обязательств и набор условий. Правила авторизации используются, чтобы разрешить или запретить действие. Аналогичный подход к управлению авторизацией и контролю доступа представлен в [12].

Предлагаемая модель состоит из пользователей / групп, используемых данных, целей доступа и режимов доступа. Он используется для обеспечения того, чтобы личная информация использовалась только для авторизации. Авторы [12] также предложили язык конфиденциальности, основанный на предложенной модели. Этот язык используется для формализации правил конфиденциальности и контроля доступа и автоматического применения этих правил с помощью системы контроля доступа. Предлагаемая модель ограничивается только контролем доступа с учетом аспектов конфиденциальности.

В статье [13] также используется подход, основанный на языке. Авторы [13] рассматривают принцип конфиденциальности, который гласит, что личные данные пользователя не могут использоваться для целей, отличных от той, для которой они были собраны, без согласия заинтересованного пользователя. Авторы [13] предполагают, что в большинстве случаев пользователи не имеют представления о том, как и для каких целей используется их личная информация. Чтобы решить эту проблему, авторы предлагают политику обработки данных (DHP), показывающую пользователям, кто и на каких условиях может обрабатывать их личные данные. Эта политика может быть разработана поставщиком услуг или пользователем с использованием разработанного языка DHP. Язык включает набор условий (а именно, получателей, действия, цели, PII, условия, положения и обязательства) и правил. Затем DHP применяется с использованием точек принятия решения по политике (принятие решения в отношении запроса доступа) и точек реализации политики (реализация решения) системы управления доступом. Минус в том, что такую политику нужно разрабатывать для каждого нового продукта.

В статье [14] предлагается язык под названием PILOT для спецификации политики конфиденциальности. Авторы также разработали инструмент, позволяющий оценивать риски, связанные с конфиденциальностью, если политика определяется с использованием предложенного языка. Преимущество подхода в том, что он позволяет оценить риски. Недостатком является то, что

такой подход не позволяет оценивать их автоматически, если политика не указана с использованием разработанного формального языка. Авторы предлагают пользователям сами определять политики конфиденциальности, а затем представляют риски разработанной политики. Из статьи также неясно, как определить все возможные риски, которые необходимы для оценки конкретного риска. В документе [15] предлагается многоуровневый язык конфиденциальности (LPL), который удовлетворяет следующим требованиям: различие между источником и получателем данных, создание политик конфиденциальности с учетом целей операций с данными, гарантия удобочитаемости на основе многоуровневых политик конфиденциальности. . К недостаткам этой работы можно отнести следующие: исследование не завершено, и предлагаемая формулировка сейчас не охватывает все аспекты конфиденциальности; компания должна определить свою политику конфиденциальности, используя LPL, прежде чем анализировать ее.

Как было упомянуто выше, оценка рисков конфиденциальности политик конфиденциальности, заданная с использованием формального языка PILOT, рассматривается в [14].

Отдельно следует отметить подходы, позволяющие рассчитывать риски конфиденциальности с учетом операций с персональными данными в анализируемой системе. Эти подходы связаны с четвертой проблемой конфиденциальности из Введения. Эти подходы не основаны непосредственно на политике конфиденциальности, но авторы проанализировали их как соответствующие исследования в области оценки рисков конфиденциальности.

NIST предложил методологию оценки рисков конфиденциальности (PRAM) [16], которая основана на ручной идентификации требований конфиденциальности к анализируемой системе и связанных с ними рисков конфиденциальности. Методология оценки включает оценку вероятности (по шкале от 0 до 10) и воздействия (с точки зрения различных затрат, которые следует суммировать) каждого риска, а затем расчет (как умножение воздействия и

вероятности) и определение приоритетности рисков.

В статье [17] предлагается подход к оценке рисков конфиденциальности. Он основан на деревьях вреда. Деревья построены на основе информации о системе, личных данных, соответствующих источниках риска, соответствующих событиях и их влиянии на конфиденциальность. Узлы дерева вреда представлены в виде троек, включающих персональные данные, компонент системы и источник риска. Корневой узел дерева повреждений соответствует нарушению конфиденциальности. Листовые узлы соответствуют использованию данных наиболее вероятным источником риска. Настройки конфиденциальности пользователей также учитываются при расчете вероятности нарушения конфиденциальности.

Результаты анализа соответствующих работ представлены в Таблице 1. Хотя существует множество исследований, посвященных анализу конфиденциальности и относящихся к трем упомянутым группам, нет комплексного исследования, охватывающего все три группы из анализа представленных политик конфиденциальности представлены с использованием естественного языка для автоматического расчета рисков конфиденциальности, кроме того, некоторые из них связаны с расчетами рисков, и не существует практических подходов, применяемых для расчета рисков конфиденциальности.

Таблица 1 – Сопутствующее сравнение работ

Описание аспектов конфиденциальности из политики конфиденциальности	**Формализация политики конфиденциальности**	**Оценка риска неприкосновенности частной жизни**	**Генерация онтологий**
- NLP: TF-IDF для построения вектора признаков; SVC для обнаружения практики конфиденциальности. - Аннотированный корпус политик конфиденциальности APP-350 Corpus. - Ограничено приложениями для Android.	–	–	–

Продолжение таблицы 1

Описание аспектов конфиденциальности из политики конфиденциальности	**Формализация политики конфиденциальности**	**Оценка риска неприкосновен-ности частной жизни**	**Генерация онтологий**
- Краудсорсинг, ML, НЛП. - Автоматическая аннотация политик конфиденциальности. - 115 аннотированных политик конфиденциальности.	Создайте онтологию для формального представления политик.	–	+
- Текст анализируется и онтология генерируется вручную. - Позволяет проверить соответствие политики GDPR.	Онтология.	–	Онтология PrOnto.
Построение онтологии политики конфиденциальности на основе ручной обработки текста.	Онтология.	–	- Ontology is generated man-ually. - Approach based on com-petence ques-tions.
- ML: линейная регрессия и нейронные сети. - Автоматическое определение вариантов отказа. - Требуется маркированный набор данных. Авторы разметили набор данных вручную.	–	–	–
- НЛП, модели включения фраз и модели машинного обучения (логистическая регрессия, линейная SVM, случайный лес, наивный байесовский алгоритм и ближайший сосед). - Автоматическое определение вариантов отказа. - Требуется маркированный набор данных. Авторы использовали набор данных из [7].	–	–	–

Продолжение таблицы 1

Описание аспектов конфиденциальности из политики конфиденциальности	**Формализация политики конфиденциальности**	**Оценка риска неприкосновенности частной жизни**	**Генерация онтологий**
–	- Машиночитаемый язык, включающий терминологию и набор правил авторизации (разрешить и запретить действия). - Позволяет формализовать политику, чтобы указать, какие типы PII, для каких целей и для каких пользователей могут использоваться.	–	–
–	- Язык конфиденциальности, основанный на модели, включающей пользователей / группы, данные, к которым осуществляется доступ, цели доступа и режимы доступа. - Позволяет формализовать правила контроля доступа и автоматизировать выполнение этих правил.	–	–
–	- Подход, основанный на языке DHP. Язык включает набор терминов и правил. - Позволяет показать пользователям, кто и на каких условиях может обрабатывать их личные данные, принимать и реализовывать решения относительно запроса доступа. - Политика должна разрабатываться для каждого нового продукта.	–	–
–	Подход на основе языка PILOT.	Позволяет оценить риски, связанные с конфиденциальностью, если политика указана с помощью PILOT.	–

Продолжение таблицы 1

Описание аспектов конфиденциальности из политики конфиденциальности	**Формализация политики конфиденциальности**	**Оценка риска неприкосновенности частной жизни**	**Генерация онтологий**
–	- Подход, основанный на LPL. - Позволяет различать источник и получатель данных. - Позволяет формировать политики конфиденциальности с учетом целей работы с данными. - Гарантировать удобочитаемость на основе многоуровневых политик конфиденциальности. - Предлагаемая формулировка не охватывает все аспекты конфиденциальности.	–	–
–	–	Качественная оценка на основе анкет. Непосредственно к политике конфиденциальности не применяется.	–
–	–	- Деревья вреда основаны. - Деревья вреда должны формироваться вручную.	–
Использование NLP для извлечения аспектов использования данных.	Онтология.	Автоматический расчет рисков конфиденциальности на основе онтологии.	Онтология P2Onto.

В данной статье авторы предлагают подход, который включает в себя сначала анализ текста политики конфиденциальности, представленной на естественном языке, генерацию и автоматическую обработку онтологии для каждой политики, указанной на естественном языке с использованием NLP, и окончательный расчет рисков конфиденциальности с использованием сгенерированных данных. онтология. На втором этапе авторы используют онтологию для формальной спецификации онтологии. Авторы вводят предло-

женную онтологию, которая является основой разработанного подхода. Основным вкладом этого исследования является новый подход к оценке рисков конфиденциальности, основанный на анализе политик конфиденциальности, определенных с использованием естественного языка и онтологии для политик конфиденциальности.

1.3 Расчет рисков на основе анализа политик безопасности

Входными данными для предлагаемой процедуры оценки рисков конфиденциальности является политика конфиденциальности, доступная конечному пользователю службы или устройства. Поскольку в большинстве случаев эти документы содержат информацию об использовании персональных данных в неструктурированной форме, необходимо создать формальное описание данных, представленных в их тексте, для применения любых дальнейших процедур оценки. Авторы предлагают использовать онтологию в качестве формального представления действий по обработке данных и их характеристик, необходимых для выполнения оценки риска. Выбор формализации на основе онтологий объясняется возможностью определения основных понятий, сущностей, их свойств и семантических отношений между ними как для человека, так и для машинного чтения и многократного использования. Таким образом, предлагаемый подход включает следующие шаги (рисунок 1):

1) Создание базовой многоразовой онтологии P2Onto, которая описывает основные аспекты сценариев использования персональных данных и служит основой для установления процедур расчета рисков.

2) Отображение текста политики конфиденциальности в базовую онтологию P2Onto.

3) Расчет оценки риска на основе сгенерированного онтологического представления и алгоритмов, указанных для онтологии P2Onto.

Таким образом, ключевым элементом предлагаемого подхода является онтология P2Onto, которая описывает различные аспекты обработки персо-

нальных данных, такие как сбор первой стороной, совместное использование третьей стороной и т. Д., И обеспечивает формальную основу для процедуры оценки риска, которая учитывает его концепции и категории при вычислении оценки риска. . Сопоставление людей с концепциями P2Onto является важной задачей, реализуемой с использованием методов естественного языка. Эта проблема активно исследуется [18], и в настоящее время она не входит в рамки данной статьи. Онтология P2Onto и алгоритмы оценки рисков подробно представлены в следующих подразделах.

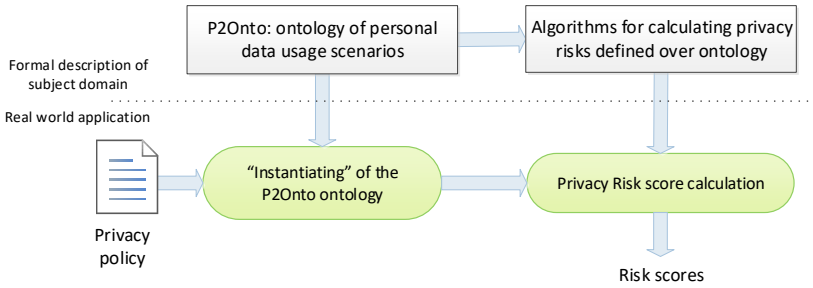

Рисунок 1 – Общая схема процедуры оценки риска неприкосновенности частной жизни

Онтология P2Onto призвана обеспечить формальную основу для процедуры оценки риска и может использоваться для проверки и объяснения полученных оценок риска. В нем описываются различные аспекты обработки персональных данных, участвующие в этом процессе субъекты и устанавливаются семантические отношения между ними. Согласно рабочему процессу проектирования онтологий на основе политик конфиденциальности, предложенному в [8], построение онтологии требует сначала идентификации основных сценариев использования персональных данных и установления их характеристик, соответствующих задаче анализа. Последнее можно сделать, сформулировав вопросы о компетентности для каждого аспекта конфиденциальности при обработке данных. На рисунке 2 показана схема потока проектирования онтологии P2Onto.

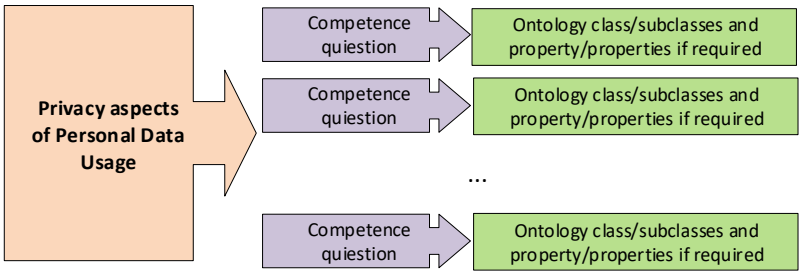

Рисунок 2 – P2Onto процесс проектирования онтологии

За основу были выбраны следующие сценарии и аспекты использования персональных данных:

Собственный сбор и использование данных. Этот аспект характеризует, какие личные данные собирает поставщик услуг, управляя устройством, вебсайтом или приложением, как они собираются, каковы правовые основания и цели сбора данных.

Сбор и обмен сторонними данными. Этот аспект характеризует все вопросы, касающиеся процедур обмена данными, включая форму обмена данными - агрегированные, анонимные или необработанные.

Безопасность данных. Этот аспект описывает механизмы безопасности, как технические, так и организационные, используемые для защиты данных.

Хранение данных. Этот аспект характеризует временные проблемы обработки и хранения персональных данных.

Агрегация данных. Этот аспект определяет, собирает ли поставщик услуг личные данные.

Настройки конфиденциальности. Эта практика определяет доступные инструменты и варианты для конечного пользователя, чтобы ограничить объем собираемых персональных данных (вопросы согласия / отказа при сборе персональных данных).

Контроль данных. Этот аспект относится к инструментам и механиз-

мам, предоставляемым пользователю для манипулирования личными данными - доступа, редактирования и удаления.

Уведомление о нарушении конфиденциальности. Этот аспект относится к инструментам и механизмам, которые поставщик услуг использует для информирования о нарушении конфиденциальности личных данных.

Изменение политики. Этот аспект относится к тому, какие инструменты и механизмы использует поставщик услуг для информирования конечного пользователя об изменениях в тексте конфиденциальности личных данных и возможных реакциях, доступных конечному пользователю.

Не отслеживать. Эта практика описывает, как обрабатываются сигналы отслеживания для онлайн-отслеживания и рекламы.

Международная и особая аудитория. В этом аспекте обсуждаются различные вопросы, связанные с обработкой персональных данных особой аудитории, такой как дети, и граждане определенных государств и регионов.

Они определены экспертами в предметной области, которые проанализировали существующие политики конфиденциальности и соответствующие правовые нормы и требования, такие как COPPA [2] и Правило конфиденциальности HIPAA [3], и широко используются в исследованиях [6], [18], [8].

Вопросы компетентности к соответствующим аспектам конфиденциальности сценариев использования персональных данных были сформулированы на основе руководящих принципов и анкет, предоставленных международными структурами оценки IoT безопасности, такими как IoTF, GSMA, в области оценки рисков конфиденциальности [19], [20]. Примеры некоторых сценариев использования персональных данных и их проблем с конфиденциальностью с соответствующими вопросами о компетентности и возможными ответами, предоставленными в форме классов онтологий, показаны в таблице 2. Мы ссылались на определения из GDPR, словарей конфиденциальности данных, разработанные W3C Data Privacy Vocabularies и Управляет группой сообщества [21] другими существующими онтологиями, такими как PROV-O

[22], [1] чтобы формализовать ответы на вопросы о компетенции и поддерживать единый подход.

Таблица 2 – Сценарии использования персональных данных и вопросы компетенции

Сценарий использования персональных данных	**Вопросы о компетенции**	**Примеры (возможные классы онтологии)**
First-party data collection and usage	Какие категории персональных данных собираются?	Данные учетной записи пользователя, данные устройства, данные приложений, данные о конкретных услугах, финансовая информация, данные отслеживания, конфиденциальные данные и т.д.
	Какой режим сбора данных?	Автоматически без согласия пользователя, автоматически, но с предоставленным согласием каждый раз, когда выполняется автоматический сбор, например когда пользователь явно разрешает приложению оценивать данные о местоположении или предоставляется пользователем напрямую, когда пользователь совершает платежи
	Какова цель сбора данных?	Предоставление услуг, включая дополнительные услуги, техническую поддержку и поддержку пользователей, аналитику и персонализацию, маркетинг и рекламу, персонализацию, безопасность, требования правовых норм
	Каково основание для сбора данных	Пользователь дал согласие, юридические требования, прочее
	Собираете ли вы данные от сторонних поставщиков услуг?	Нет, общедоступные источники, социальные сети, сторонние поставщики услуг и т.д.
	Кто является владельцем данных?	Пользователь, другие третьи стороны, например когда собираются данные о членах семьи, гостях
First-party data collection and usage	Какие категории личных данных передаются?	Данные учетной записи пользователя, данные устройства, данные приложений, данные о конкретных услугах, финансовая информация, данные отслеживания, конфиденциальные данные и т.д.
	В каком формате передаются данные?	Агрегированные, анонимные, оба, сырые или неопределенные
	Какова цель обмена?	Предоставление услуг, включая дополнительные услуги, техническую поддержку и поддержку пользователей, слияние и приобретение, маркетинг и рекламу, персонализацию, безопасность, требования правовых норм и т.д.
	Какова цель распространения?	Пользователь дал согласие, юридические требования, прочее

Продолжение таблицы 2

Сценарий использования персональных данных	**Вопросы о компетенции**	**Примеры (возможные классы онтологии)**
	Что такое третьи стороны?	Аффилированные лица и субподрядчики, юридические лица, другие третьи стороны
Data retention	Какие категории личных данных сохраняются?	Данные учетной записи пользователя, данные устройства, данные приложений, данные о конкретных услугах, финансовая информация, данные отслеживания, конфиденциальные данные и т.д.
	Какова цель удержания?	Предоставление услуг, включая дополнительные услуги, техническую поддержку и поддержку пользователей, аналитику и персонализацию, маркетинг и рекламу, персонализацию, безопасность, требования правовых норм
	Какова правовая основа удержания?	Пользователь дал согласие, юридические требования, прочее
	Какой срок хранения?	Неопределенный, указанный с явно указанным сроком хранения, неопределенный или другой
	Где они хранятся?	Устройство, облако

Тщательный анализ их сценариев использования персональных данных и их аспектов конфиденциальности позволил выделить четыре общих класса - Данные, Действия, Агент и Механизм - которые образуют основу для описания основных сценариев использования персональных данных, остальные классы используются для определения их свойств.

Данные – это суперкласс, который используется для определения категорий личных и неличных данных. Авторы следуют определению GDPR, чтобы указать типы персональных данных, которые описываются как «любая информация, относящаяся к идентифицированному или идентифицируемому физическому лицу («субъект данных»); идентифицируемое физическое лицо - это лицо, которое может быть идентифицировано прямо или косвенно, в частности, посредством ссылки на идентификатор, такой как имя, идентификационный номер, данные о местоположении, онлайн-идентификатор или один или несколько факторов, специфичных для физического, физиологиче-

ская, генетическая, ментальная, экономическая, культурная или социальная идентичность этого естественного человека»[**?**], [4]. Это позволило нам определить такие подклассы персональных данных, как User_Account_Data, которые включают информацию о входе в систему, аватар пользователя, электронную почту, физический адрес, User_Device_Info и User_App_Info, которые содержат данные о пользовательском устройстве и приложениях, такие как версия, модель, время обновления и т. Д., Соответственно. . Мы также обрисовали в общих чертах Tracking_Data, чтобы указать данные, которые могут использоваться для отслеживания пользователя, такие как IP-адрес, файлы cookie, отпечаток браузера, чтобы иметь возможность оценить сценарий «Не отслеживать», и представили новый подкласс Service_Data, который используется для указания конкретных данных. для обслуживания и работы устройства, например, блокировки и разблокировки, яркости лампы и т. д., которые могут использоваться для определения привычек и стиля жизни пользователя. Подробная иерархия классов данных, включая иерархию конфиденциальных данных, показана на рисунке 3.

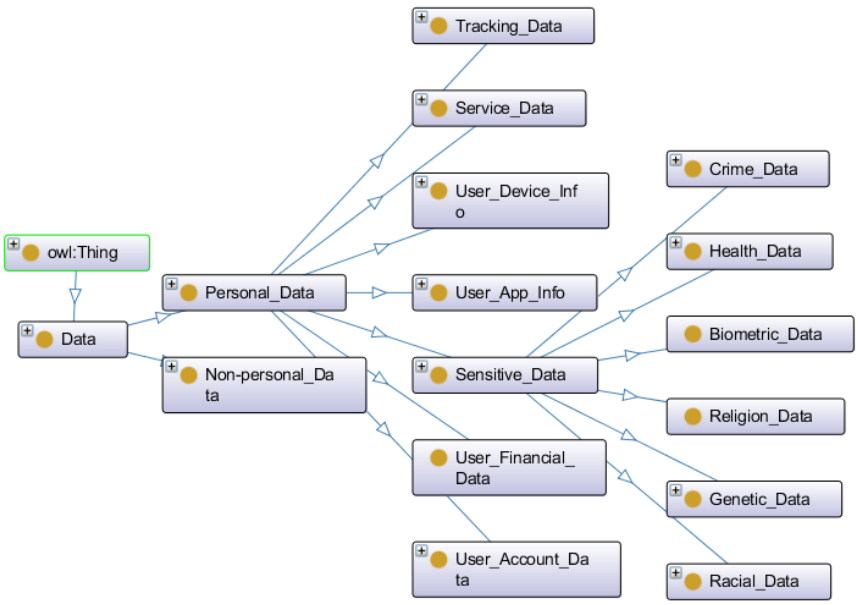

Рисунок 3 – Иерархия классов данных

Следует отметить, что класс Non_Personal_Data используется для опи-

сания неличных данных, возникающих при получении персональных данных посредством анонимизации или агрегации персональных данных. Знание того, сколько типов данных - идентифицируемых и нет - собираются о конкретном пользователе устройства, имеет важное значение в процедуре оценки рисков.

Как следует из списка аспектов конфиденциальности использования персональных данных, некоторые аспекты напрямую связаны с обработкой данных, например сбор, обработка, совместное использование, хранение или безопасность данных первой стороной, в то время как другие относятся к деятельности, которая косвенно связана с обработкой данных. такие как уведомления в случае изменения политики или нарушения данных, предоставление доступа, прав редактирования и стирания и т. д. По этой причине мы выделили два разных подкласса класса активности - Data_Activity и Control_Activity. На рисунке 4 показана иерархия подклассов Activity.

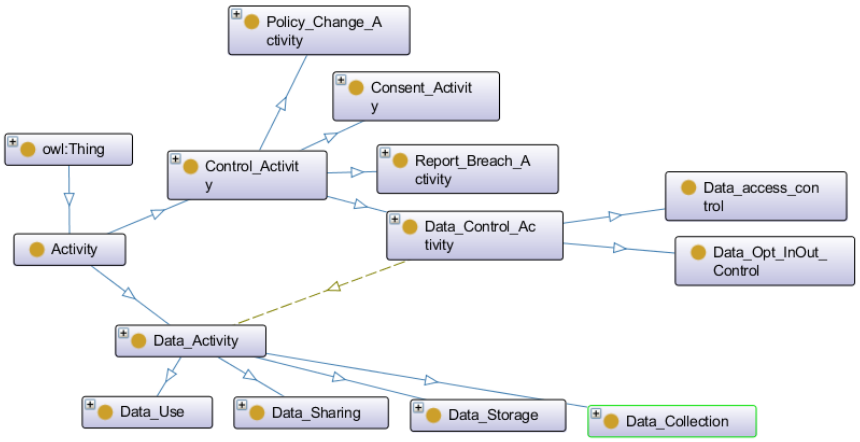

Рисунок 4 – Иерархия классов активности

Класс Data_Activity - это общий класс для определения различных типов операций по обработке данных. Несмотря на то, что эти действия имеют свои отличительные характеристики, можно выделить общие черты, такие как цель операций с данными, формат обрабатываемых данных - анонимные или необработанные, правовая основа для обработки данных и контролиру-

ющих лиц. На рисунке 5 показаны наиболее важные классы, относящиеся к деятельности по обработке данных. Цель обработки данных является важной концепцией при оценке рисков конфиденциальности, и мы выделяем следующие цели обработки данных: 1) Предоставление услуг, 2) Реклама и маркетинг 3) Аналитика и исследования, 4) Персонализация, 5) Безопасность, 6) Слияние и поглощение, 7) Соответствие законодательству 8) Другое, 9) Не определено, каждый из них представляет собой отдельный подкласс.

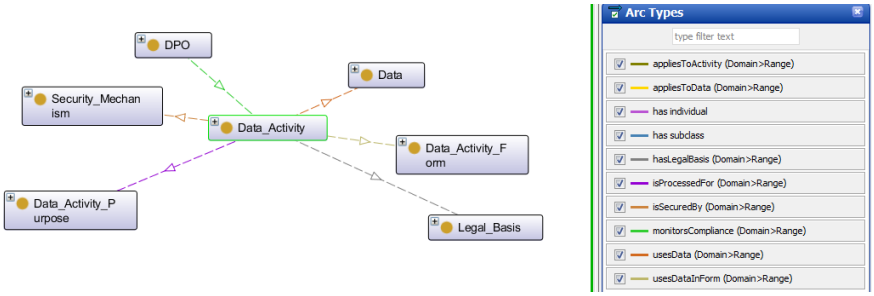

Рисунок 5 – Отношения между универсальным классом Data_Activity и другими классами, характеризующими различные аспекты конфиденциальности

Чтобы указать владельца данных, обработчика данных и ответственного за контроллер данных (DPO), а также других третьих сторон, участвующих в обработке данных, используется класс Agent (рисунок 6). Авторы предлагают повторно использовать эту концепцию из онтологии PROV-O, которая определяет концепт «Агент» как субъект, который «несет некоторую форму ответственности за происходящую деятельность, за существование сущности или за деятельность другого агента» [22]. Эта концепция позволяет указать случаи, когда данные собираются от третьих лиц, таких как социальные сети, общедоступные источники с открытым исходным кодом и другие третьи стороны. Он также используется для выявления случаев, когда данные собираются от посторонних лиц, то есть людей, которые не владеют устройством или услугой и с большой вероятностью не дают согласия на обработку данных.

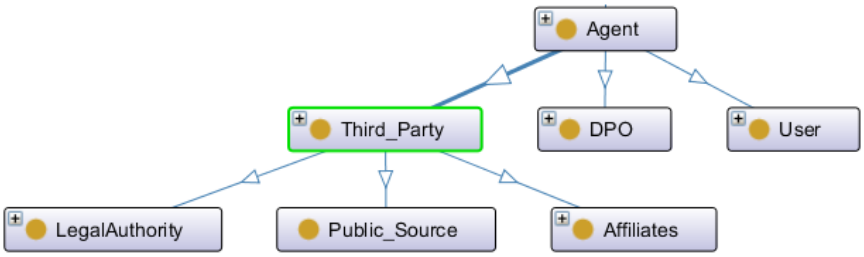

Рисунок 6 – Иерархия классов агентов

Класс Mechanism - это общий класс, который используется для определения различных инструментов, опций, механизмов и интерфейсов, поддерживающих реализацию различных действий - сбор данных, совместное использование, использование, уведомление в случае изменения политики или нарушения данных. Он используется для характеристики таких свойств, как режим обработки (автоматический или нет), детали реализации деятельности, такие как уведомление по электронной почте или на веб-сайте, доступ к данным через приложение или через конкретный запрос по почте и т. Д.

Все упомянутые выше классы связаны друг с другом с помощью свойств, которые определяют семантические отношения между ними. На рисунке 6 показаны основные концепции и свойства, относящиеся к сценарию использования сохранения данных. Сущности, отмеченные желтыми точками, являются классами, а объекты, отмеченные пурпурным ромбиком, - это отдельные лица, то есть отрывки текста, обнаруженные в политике конфиденциальности. Стрелки соответствуют свойствам, связывающим сущности, их цвет зависит от их типа. Конкретный сценарий, приведенный на Рисунке 7, соответствует сохранению данных для конкретной услуги (износ зубной щетки, оценка производительности и продолжительность чистки), которые хранятся в течение трех месяцев для управления зубной щеткой. Правовая основа для этой деятельности, а также формат хранения данных не были явно упомянуты в тексте, поэтому авторы использовали константу Not Defined для этих классов.

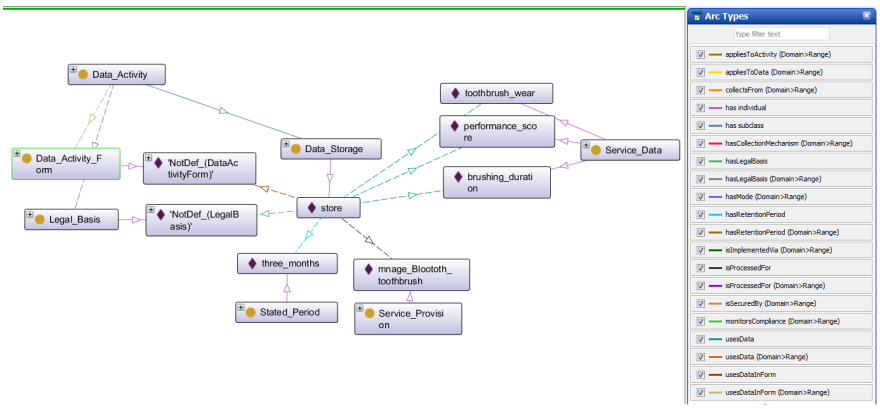

Рисунок 7 – Scenario representing the retention of the service specific data

Вопросы о компетенции из раздела 3.1 помогли нам определить основные концепции и свойства онтологии P2Onto.

Алгоритм оценки риска конфиденциальности принимает в качестве входных данных онтологию политики конфиденциальности, описывающую 11 сценариев использования (см. Раздел 3.1).

Основная идея алгоритма заключается в том, что типы персональных данных и их количество определяют основу оценки риска неприкосновенности частной жизни. Другие аспекты, указанные в политике конфиденциальности, такие как цель, правовая основа и варианты подписки / отказа, могут только увеличивать или уменьшать их. Формализующий алгоритм представлен в приложении А.1.

Расчет оценки риска конфиденциальности RiskScoreBase основан на критичности типов персональных данных и присвоенных им весах, представленных в Таблице 3. Идея алгоритма расчета риска конфиденциальности RiskScoreBase заключается в определении типа персональных данных с наивысшей критичностью. присутствовать в тексте политики конфиденциальности, а затем увеличивать его в зависимости от количества различных типов личных данных. Если в тексте политики присутствуют все типы персональных данных, то риски увеличиваются вдвое. Оценка риска увеличивается логарифмически, чтобы избежать быстрого роста оценки риска.

Таблица 3 – Категории персональных данных, их важность и вес

Классы и подклассы онтологии	**Категории (класс)**	**Критичность типа ПД**	**Влияние категории**
Data	NonPersonal Data	0	0,0
	Service Data	3	0,9
	Tracking Data	4	1,3
	User Device Data	2	0,4
	App Data	2	0,4
	Sensitive Data	5	2,2
	User Financial Info	4	1,7
	User Account Data	2	0,4
	Other	3	1,3
	Not defined	3	1,3

Общая оценка риска на основе анализа онтологии рассчитывается следующим образом:

$$PrivacyRiskScore = \sum_i w_i \cdot UsageScenarioRiskScore_i, \sum_i w_i = 1, \quad (1)$$

где wi - весовой коэффициент, определяющий влияние оценки риска i-го сценария использования данных. В текущей версии все wi равны друг другу, а общая оценка риска рассчитывается следующим образом:

$$PrivacyRiskScore = \frac{1}{n} \sum_{i=1}^{n} w_i \cdot UsageScenarioRiskScore_i, \quad (2)$$

где n – количество анализируемых сценариев использования данных.

Ниже приведен пример расчета риска конфиденциальности для сценария Собственного сбора и использования.

В настоящее время сценарий сбора и использования первой стороной

описывается следующими классами и их подклассами: данные (личные данные, неличные данные), правовая основа, цели деятельности с данными. Подклассы этих классов рассматриваются как категории собираемых данных, которые используются в процедуре оценки (см. Таблицу 4).

Таблица 4 – Категории целей и правовых оснований сбора персональных данных, их важность и вес

Классы и подклассы онтологии	**Категории (класс)**	**Критичность типа ПД**	**Влияние категории**
Purpose	Service Provision	1	0,0
	Security Purpose	1	0,0
	Analytics Personalization	3	2,3
	Adverising Marketing	3	2,3
	Merge Acqisition	2	1,5
	Other	1	1,3
	NotDef*	4	3,0
Legal Basis	Consent	0	0,0
	Legal requirement	0,0	1,7
	Other	1,5	1,7
	NotDef*	0,0	3,0

Критичность каждой категории в таблице 4 присвоена в соответствии с экспертными знаниями. Для целей и юридических оснований десятичный логарифм максимального веса дает 1,77, чтобы увеличить оценку по данным в 1,77 раза, если две категории имеют высокую критичность, и 1,47, если только одна категория имеет высокую критичность.

Алгоритм оценки риска сценария использования первой стороной (FP_RiskSco

Авторы сначала рассчитывают базу оценок рисков конфиденциальности RiskScoreBase на основе критичности личных данных, упомянутых в сценарии использования данных. Этот алгоритм похож для разных сценариев использования.

Пусть PD_classi - это категория личных данных, как определено в таблице 3. Это возвращается функцией getCategory () из алгоритма выше.

Пусть PD_Criticalityi - это критичность i-й категории персональных данных.

Пусть max_criticality - это максимальная критичность среди критичностей категорий данных, не пустых в онтологии. Это возвращается функцией getMaxCritCat () из алгоритма выше. Формализующий алгоритм представлен в приложении A.2.

В дальнейшем мы можем увеличивать или уменьшать эту оценку риска на основе критичности категорий онтологии других классов.

Пусть P_classi - это категория цели использования личных данных, как определено в Таблице 4, а LB_classi - категория правовой основы для использования личных данных. Это возвращается функцией getCategory() из алгоритма выше.

Пусть P_Criticalityi - критичность i-й категории (подкласса) P_classi, LB_Criticalityi - критичность i-й категории LB_classi.

Пусть max_P_Criticality - это максимальная критичность среди критичностей целей использования данных, которые не пусты в онтологии, а max_LB_Critic - максимальная критичность среди критичностей правовой основы для использования данных, которые не пусты в онтологии. Это возвращается функцией getMaxCritCat() из алгоритма выше.

Пусть P_weight - это вес для целевой категории с max_P_Criticality, а LB_weight - вес для юридической категории с max_LB_Criticality. Это возвращается функцией getClassWeight() из алгоритма выше. Формализующий алгоритм представлен в приложении A.3.

Фактически, чтобы снизить FP_RiskScore, необходимо манипулировать весами Цели и Правовой основы.

Для оценки предложенного подхода авторы выбрали четыре политики конфиденциальности, написанные для разных типов сервисов и устройств.

Все политики конфиденциальности, за исключением августовской политики конфиденциальности, были выбраны из набора данных OPP-115 [23], который был создан в 2016 году. Вот почему политики конфиденциальности устарели, и, безусловно, текущие версии политик конфиденциальности компаний содержат всю информацию, необходимую для соответствовать законодательным требованиям GDPR [1], CCPA [23], COPPA [2]. Выбор этих политик объясняется тем, что они представляют различные типы сервисов и имеют аннотации, которые помогают проверять результаты и сущности онтологии.

Таблица 5 – Описание политик конфиденциальности

Company name	**Privacy policy**	**Brief description of company activity**
August Products & Services	24 July 2020	It is a company that produces devices for smart home, such as smart locks, doorbell cameras and other accesso-ries [25]. Their smart lock allows implementing a variety of convenient but privacy risky functions as remotely lock and unlock the door, logging exit/entrance activity of smart lock owners as well as their guests, supports biometrical identi-fication and voice assistant.
Ticketmaster website	July 20, 2012	It is online service of the world entertainment company that organizes different live events for well-known and emerging artists. Their online service allows choosing any entertain-ment event and buying tickets for it, that it is expected that the risks for this policy could be rather high as it should deal with users' financial data.
Cincinnati Museum Center	September 09, 2011	This privacy policy belongs to the cultural center that unites three museums with different exhibits, organizes different events including one for children with possibility to book them in advance. Like in previous case it is ex-pected to have quite high risks due to processing financial data.
Instagram	January 19, 2013	This privacy policy belongs to a popular public social net-work targeted primary for sharing photo and video data. Till 2019 there were no possibilities for online shopping via Instagram.

Авторы сосредоточились на трех аспектах использования данных: сбор и использование данных собственными силами, сбор и обмен данными сторонними организациями и хранение данных для сравнения полученных оценок рисков. Другие аспекты конфиденциальности не были охвачены всеми политиками конфиденциальности. Например, продукты августа не предна-

значены для использования несовершеннолетними до 16 лет, сервисы Instagram и Ticket-master не предназначены для детей младше 13 лет, поэтому риски конфиденциальности для этой конкретной аудитории не рассчитываются. Также следует отметить, что сценарий использования, описывающий уведомление о нарушении конфиденциальности, не был обнаружен в текстах всех политик конфиденциальности. Политика конфиденциальности Cincinatti Museum не содержит информации о хранении Данных. Необходимо отметить, что текущая версия политики конфиденциальности музеев также содержит минимум информации о хранении данных и очень расплывчата.

В большинстве случаев политики конфиденциальности носят общий характер, во многих сценариях использования данных отсутствует информация об определенных аспектах конфиденциальности. Если эта информация не указана явно, она помечается как «Не определено». Оказалось, что эта концепция присутствует практически во всех сценариях использования персональных данных. С одной стороны, его можно использовать для характеристики прозрачности политики - чем больше число соответствующих лиц, тем более расплывчатая и неясная политика. С другой стороны, он имеет самый высокий вес в процедуре оценки риска, и интересно понять, как он влияет на общие оценки риска. Поэтому в экспериментах авторы проводили расчет для обоих случаев, учитывая категорию Not Defined для каждого класса и игнорируя ее.

Таблица 6 содержит оценки, полученные для каждого сценария личного использования. Рассчитанные оценки рисков вполне естественны и легко объяснимы. Политика конфиденциальности Instagram имеет самую низкую RiskScoreBase почти во всех сценариях использования данных, это объясняется тем, что объем собираемых персональных данных минимален, компания не собирает никаких чувствительных и финансовых данных, однако риски довольно высоки. объясняется непрозрачностью целей, обратите внимание, что оценка риска снижается, когда категория «Не определено» не рассматри-

вается. Остальные политики имеют сопоставимую базу данных RiskScoreBase, потому что они собирают почти все типы персональных данных, кроме конфиденциальной. Августовская интеллектуальная блокировка также собирает некоторую служебную информацию, такую как статус блокировки, информацию о пользователях блокировки. Как и в политике Instagram, высокие риски в большинстве случаев объясняются наличием концепции Not Defined. Однако для Музейного центра Цинциннати эти риски сохраняются даже тогда, когда в расчетах не учитывается концепция Неопределенного. При этом поясняется, что Центр делится почти всеми данными с неопределенными или неназванными третьими сторонами. Высокие риски хранения августовского смарт-замка объясняются наличием неограниченного срока хранения данных. Однако, когда авторы этого исследования подробно изучили эту политику, они обнаружили, что эта информация относится только к личным данным в агрегированной форме.

Таблица 6 – Присвоенные ранги P2Onto лицам для разных концепций

Data Usage Scenario	**Risk scores**	**August Products & Services**	**Ticketmaster website**	**Cincinnati Museum Center**	**Instagram**
First Party Collection and Use	RiskScoreBase (with Not Def category)	7,08	6,55	6,55	4,50
	RiskScoreBase (without Not Def category)	6,55	6,55	5,93	3,78
	UsageScenarioRiskScore (with Not Def category)	9,82	9,84	9,65	8,01
	UsageScenarioRiskScore (without Not Def category)	9,67	9,74	9,37	5,96
Third Party Sharing	RiskScoreBase (with Not Def category)	4,24	3,34	6,55	4,50
	RiskScoreBase (without Not Def category)	3,34	3,00	5,93	3,78
	UsageScenarioRiskScore (with Not Def category)	8,55	6,27	9,65	8,67
	UsageScenarioRiskScore (without Not Def category)	5,95	4,96	9,37	6,92

Продолжение таблицы 6

Data Usage Scenario	Risk scores	August Products & Services	Ticketmaster website	Cincinnati Museum Center	Instagram
Data retention	RiskScoreBase (with Not Def category)	6,37	3,34	–	3,40
	RiskScoreBase (without Not Def category)	5,66	3,00	–	1,30
	UsageScenarioRiskScore (with Not Def category)	9,82	5,95	–	6,62
	UsageScenarioRiskScore (without Not Def category)	9,67	4,96	–	2,61

Анализ политики с использованием онтологии выявил много интересных вопросов. Например, смарт-замок собирает данные от гостей, которые посещают своего владельца, это делается для того, чтобы предоставить доступ смарт-замку и организовать процесс приглашения. На рисунке 8 показан этот конкретный сценарий использования данных.

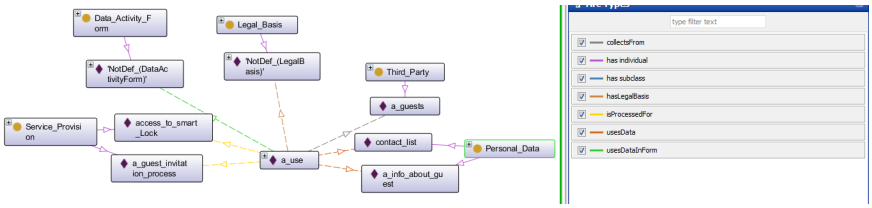

Рисунок 8 – Часть онтологии P2Onto, представляющая практику использования данных для сбора данных первой стороной, обнаруженную в августовской политике конфиденциальности

На рисунке 9 показан сценарий совместного использования данных. Хотя данные передаются третьим лицам в аналитических и маркетинговых целях, формат обмена данными сохраняет конфиденциальность владельца данных.

Это заставило авторов сделать вывод о необходимости учитывать в процедуре расчета риска каждый конкретный сценарий использования данных. И требуется найти баланс между Неопределенными категориями, которые бы

в них присутствовали. Применение онтологии в качестве основы для построения таких правил допускает эти изменения, поскольку все сценарии данных представлены как связанные концепции онтологии. Эта способность онтологии также полезна для объяснения полученных результатов, поскольку ясно, как различные типы персональных данных собираются, обрабатываются и передаются, какие инструменты и опции для доступа, редактирования персональных данных или их удаления доступны конечному пользователю. , так далее.

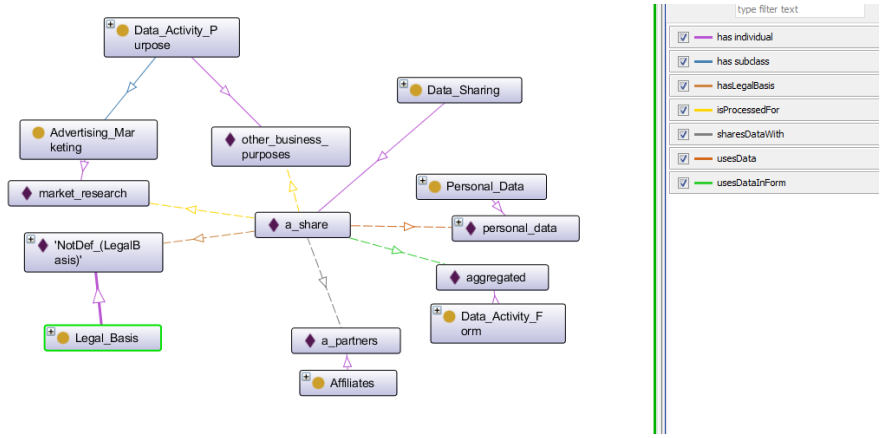

Рисунок 9 – Часть онтологии P2Onto, представляющая практику хранения данных, обнаруженную в августовской политике конфиденциальности

Авторы считают, что эта онтология может служить основой для разработки интерактивных моделей визуализации на основе графов, нацеленных на объяснение рисков конфиденциальности для конечного пользователя в ясной и удобочитаемой форме.

1.4 Заключение

Конфиденциальность в настоящее время вызывает растущую озабоченность. Для защиты конфиденциальности пользователей было разработано множество стандартов. Они обязывают организации четко указывать в политиках конфиденциальности, какие личные данные они используют, для каких целей, кем и как долго. Текст политики конфиденциальности может ввести

пользователей в заблуждение. В этой статье авторы предложили подход к оценке частных рисков на основе онтологии, чтобы сделать понятным и прозрачным для пользователя, какие риски конфиденциальности могут возникать в результате политики конфиденциальности.

Разработанный подход основан на анализе политики конфиденциальности. Он включает в себя анализ текстов политики конфиденциальности, представленных на естественном языке, с использованием сначала методов обработки на естественном языке, генерацию и обработку онтологии политики конфиденциальности для каждой политики, указанной на естественном языке, и вычисление рисков конфиденциальности с использованием созданной онтологии. Предлагаемая онтология политики конфиденциальности, описывающая различные сценарии использования данных, является ядром разработанного подхода. Процесс создания онтологии основан на вопросах компетентности для каждого аспекта конфиденциальности операций по обработке данных, которые позволяют идентифицировать основные сценарии использования персональных данных и устанавливать их характеристики, соответствующие задаче анализа. В статье авторы описали разработанный процесс генерации онтологий, общую онтологию политик конфиденциальности и примеры фрагментов онтологий для конкретных политик.

Также авторы описали разработанный алгоритм расчета рисков конфиденциальности на основе сгенерированной онтологии. Основная идея алгоритма заключается в том, что типы персональных данных и их количество определяют основу оценки риска неприкосновенности частной жизни. Другие аспекты, указанные в политике конфиденциальности, такие как цель, правовая основа и варианты подписки / отказа, могут только увеличивать или уменьшать их. Этот подход демонстрируется в нескольких политиках конфиденциальности. Описаны и проанализированы проведенные эксперименты. Они продемонстрировали применимость подхода для представления основных аспектов использования данных в ясной и удобочитаемой форме и для

объяснения рассчитанной оценки риска.

Кроме того, эксперименты показали, что установление рангов для людей - важный аспект, требующий дополнительных исследований. Еще одно важное направление будущих исследований связано с автоматизацией обнаружения концептов онтологии в тексте политики с использованием методов обработки естественного языка.

1.5 Постановка задачи

Текст…

2 Применение строгих методов анализа текста для формализации политик безопасности

2.1 Статистические модели текстовых документов

Были протестированы две модели векторизованного представления текста – «мешок слов» и модель TF-IDF. Модель «мешок слов» представляет документ в виде матрицы, представленной на рисунке 10. Здесь слова каждого абзаца подсчитываются и сопоставляются с абзацами, в которых они встретились.

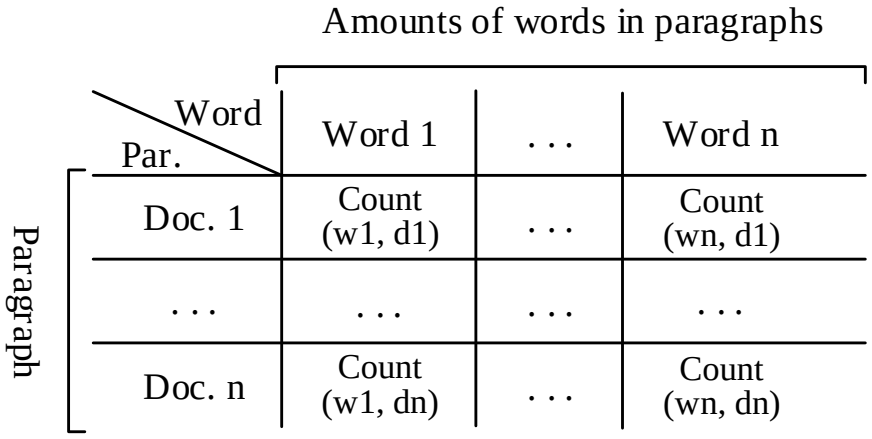

Рисунок 10 – Bag-of-Words матрица

Модель TF-IDF представляет документ в виде матрицы, представленной на рисунке 11. Формула (3) показывает, как можно получить метрику TF-IDF.

$$tfidf(t,d,D) = \frac{n_t}{\sum\limits_k n_k} \times log \frac{\left|D\right|}{\left|\left\{d_i \in D : t \in d_i\right\}\right|}, \qquad (3)$$

где t – термин или слово;

d – конкретный абзац;

D – набор абзацев.

Итак, модель TF-IDF придает больший вес словам которые использова-

ны меньше раз. Это может быть полезно, когда тексты схожи с точки зрения используемых слов, как в нашем случае, для политик безопасности.

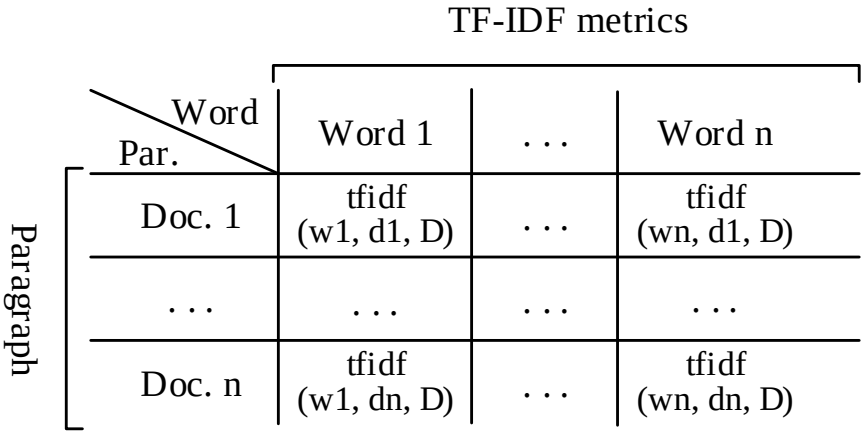

Рисунок 11 – Матрица TF-IDF

2.2 Подход основанный на латентно-семантическом анализе текста

Современные методы кластеризации текстов позволяют определять тематику текстов с высокой точностью. Однако большинство из этих методов принимают тексты с самыми разными темами как вход для алгоритмов. Но тексты со схожими тематиками можно проанализировать с помощью латентно-семантического анализа дважды: группировать тексты по темам один раз, и предоставить еще более детальное разделение их по подтемам во второй раз. Такой подход можно использовать для более точной классификации абзацев с точки зрения их характеристик и аспектов использования персональных данных. Следует отметить, что латентно-семантический поиск сильно зависит от глобального текстового контекста с потерями информации о локальных контекстных отношениях между словами. Были выделены девять тем конфиденциальности, которые следует сопоставить с абзацами согласия пользователя сайта – «сбор личных данных», «сбор данных третьими лицами», «управление личными данными», «механизмы защиты персональных данных» и др. Очевидно, что аспекты обращения с данными состоят из нескольких слов, и в некоторых случаях перекрываются. На основании этих

фактов была выдвинута гипотеза о том, что латентно-семантический поиск способен обнаружить даже незначительную разницу в тексте абзацев при пропуске частых слов. Перед применением латентно-семантического анализа требуется предварительная обработка входных данных. Обычно эта процедура включает очистку данных, удаление гиперссылок, пунктуации и т. д. Также текст политик конфиденциальности был разбит на набор абзацев. Каждый абзац был преобразован в массив слов, которые он содержит. Следующим шагом было удаление наиболее частых, но не столь значимых слов, так называемых стоп-слов. Также была применена операция стемминга, чтобы рассматривать только основную часть всех слов полученных от единого корня.

Пусть A – это матрица абзацев и слов, тогда используя формулу (4)

$$A = U \times S \times V^T, \tag{4}$$

где A – матрица слов и параграфов;

U – ортонормированная матрица U;

V – ортонормированная матрица V;

S – диагональная матрица S, значения которой сингулярны для A.

После того, как матрица была разделена на три компонента, матрица U содержит n-мерные векторы, которые можно интерпретировать как координаты в n-мерном пространстве [30]. Документы могут быть распределены по кластерам по значениям этих координат. Проведенные эксперименты с латентно-семантическим анализом выполнялись с использованием набора данных с открытым исходным кодом, который включает 115 политик безопасности, которые были размечены вручную, и все абзацы присвоены одному или нескольким сценариям использования персональных данных [27]. Результаты экспериментов для модели «мешок слов» представлены в таб-

лице 7, в ней показаны полученные кластеры и соответствующие значения координат.

Таблица 7 – Кластеры политик безопасности для модели Bag-of-Words

№	Координата 1	Координата 2	Координата 3	Координата 4
0	0.634”inform”	0.28”may”	0.276”use”	0.232”servic”
1	0.202“cooki”	0.466“inform”	0.336“site”	0.257“use”
2	0.524“privaci”	0.433“polici”	0.388“cooki”	0.219“site”
3	-0.589“servic”	0.344“site”	0.244“parti”	-0.240“third”
4	-0.504“parti”	0.486 “third”	-0.449“servic”	0.235“advertis”
5	-0.594“site”	0.278“cooki”	0.272“websit”	0.264“privaci”
6	-0.326“may”	0.311“site”	0.307“servic”	-0.293”email”
7	-0.437”may”	-0.369”advertis”	0.345”person”	0.319”cooki”
8	0.501”may”	-0.315”email”	-0.281”use”	-0.264”address”
9	-0.488”user”	-0.384”use”	0.310”provid”	-0.301”websit”

Как видно, результаты противоречивы, поэтому трудно понять, какая из тем каким смыслом обладает. Затем рассчитывалась метрика принадлежности к теме с помощью библиотеки Gensim [31] и результаты снова не были обнадеживающими. Результаты расчета метрики принадлежности кластеру представлены в таблице 8.

Таблица 8 – Принадлежность кластерам

Topic	0	1	2	3	4
Affiliation	2.27	-0.8	0.15	-0.22	-1.2
Topic	5	6	7	8	9
Affiliation	-0.17	-0.15	-0.2	0.22	-0.07

Другие результаты с параграфами, относящимися к другому аспекту обращения с данными, были почти такими же. Результаты представлены в таблице 9.

Таблица 9 – Принадлежность кластерам

Topic	0	1	2	3	4
Affiliation	2.59	-0.76	0.64	0.74	0.13
Topic	5	6	7	8	9
Affiliation	0.14	-0.12	0.23	0.12	0.41

Все протестированные абзацы были сопоставлены с кластером 0, что не может быть верным так как абзацы относились к заведомо разным аспектам обращения с персональными данными.

Результаты экспериментов для модели TF-IDF представлены далее, в таблице 10. Также показывались десять кластеров и значения атрибутов. И, как в первом случае с «мешком слов», по значениям координат невозможно судить о теме кластера.

Таблица 10 – Кластеры политик безопасности для модели TF-IDF

№	Координата 1	Координата 2	Координата 3	Координата 4
0	0.202“cooki”	0.2“may”	0.198“inform”	0.198“site”
1	0.573“cooki”	0.262“browser”	0.195“advertis”	0.182“web”
2	-0.406“media”	0.291“cooki”	0.282“health”	0.279“advertis”
3	-0.453“health”	0.258“email”	-0.204“kaleida”	0.191“address”
4	0.423“health”	0.215“media”	0.205“kaleida”	-0.199“secur”
5	-0.299“advertis”	0.262“health”	-0.252“media”	-0.213“privaci”
6	-0.325“media”	0.263“polici”	0.249“privaci”	0.197”chang”
7	0.280”cooki”	-0.216”device”	-0.183”health”	-0.166”social”
8	-0.223”advertis”	-0.206”teenag”	-0.206”inelig”	0.176”child”
9	-0.263” child”	-0.26”wireless”	0.245”message”	0.239”parent”

Результаты кластеризации снова противоречивы, поэтому трудно сказать, какая конкретная тема описывает какой аспект политики конфиденциальности. В разных темах встречаются одни и те же слова с изменение веса. Для аспектов политики конфиденциальности, которые мы искали нет тем,

которые могли бы их точно описать, поскольку многие из них могут. Затем с помощью библиотеки Gensim был рассчитан показатель принадлежности к теме, и результаты снова не были обнадеживающими. Результаты расчета аффилированности по абзацу одной из политик конфиденциальности представленные в таблице 11.

Таблица 11 – Принадлежность кластерам

Topic	0	1	2	3	4
Affiliation	2.18	-0.97	-0.69	-0.27	0.65
Topic	5	6	7	8	9
Affiliation	0.98	-1.17	0.8	0.27	0.01

Результат для другого абзаца, относящегося к другой политике конфиденциальности, был почти такой же. Результаты представлены в таблице 12.

Таблица 12 – Принадлежность кластерам

Topic	0	1	2	3	4
Affiliation	1.82	0.25	0.49	0.29	-0.04
Topic	5	6	7	8	9
Affiliation	0.74	0.52	-0.04	-0.58	-1.33

Как можно заметить, результаты для модели TF-IDF аналогичны результатам модели «мешка слов», за исключением нескольких незначительных изменений. Все абзацы снова были сопоставлены с кластером 0, что неверно, потому что они на самом деле описывают разные сценарии использования персональных данных. Эти эксперименты позволили сделать вывод, что использование латентно-семантического анализа не дает ценной информации о содержании онлайн-согласия пользователя. Проблема может быть связана с тем, что сценарии использования персональных данных очень похожи между собой, и для того, чтобы различать разные сценарии необходимо

учитывать локальный контекст.

В результате апробации алгоритма латентно-семантического анализа было выяснено что для кластеризации экстремально схожих между собой текстов он подходит не лучшим образом. В связи с этими обстоятельствами было решено обратить внимание на несколько иной подход анализа текста, основанный на контекстно-свободных грамматиках, тегировании по частям речи и синонимическом поиске.

2.3 Подход основанный на латентном размещении Дирихле

Для тестирования подхода авторы использовали два набора данных. Первый набор данных – это OPP-115 с открытым исходным кодом, а второй – это набор данных, созданный авторами и состоящий из политик конфиденциальности только для устройств IoT [12].

Набор данных OPP-115 содержит 115 документов с онлайн-согласиями пользователей веб-сайта. Этот набор данных содержит аннотации сценариев использования личных данных, его авторы обозначили 10 аспектов использования личных данных: “First-party Collection/Use”, “Third-party Sharing/Collection”, “User Choice/Control”, “User Access, Edit and Deletion”, “Data Retention”, “Data Security”, “Policy Change”, “Do Not Track”, “International and Specific Audiences”, “Other”. В большинстве случаев аспекты относятся к абзацам текста, а некоторые абзацы относятся к нескольким категориям одновременно. На рисунке 1 показано распределение абзацев по категориям. Хорошо видно, что есть две основные категории – “Third-party Sharing/Collection” и “First-party Collection and Use”, которые преобладают над остальными.

Чтобы применить LDA к анализу политики конфиденциальности, мы разбили текст политики конфиденциальности на набор абзацев. Каждый абзац был преобразован в массив слов, а затем удалены наиболее частые, но не значащие слова, так называемые «стоп-слова». Мы также выполнили лемматизацию, чтобы обобщить некоторые слова, чтобы добиться более точных результатов.

В ходе экспериментов мы протестировали две модели векторизатора текста – мешок слов и TF-IDF, и оказалось, что метрика TF-IDF предоставляет более подробную информацию о сценариях использования данных, поскольку эта модель векторизатора дает более высокие веса словам, которые реже используются.

Оптимальное количество кластеров, то есть семантических моделей, было определено как 15, поскольку оно соответствует максимальному значению когерентности, рассчитанному с помощью библиотеки Gensim [13]. Важно отметить, что это число отличается от числа категорий, обозначенных создателями набора данных OPP-115.

Результаты экспериментов для модели TF-IDF показаны в таблице 1. В таблице 1 приведен список координат, которые формируют семантические модели темы. Координаты используются для составления гипотезы об использовании личных данных и сценариях его применения/политики конфиденциальности.

Хорошо видно, что большинство извлеченных моделей посвящено сценариям “First-Party Collection and Use” и “Third-Party Sharing/Collection”. Это полностью соответствует распределению категорий в наборе данных. Эти модели различаются характеристиками различных аспектов этих двух сценариев использования. Например, тематическая модель 9 раскрывает варианты согласия / отказа при обмене личными данными в рекламных целях, тематическая модель 6 посвящена использованию файлов cookie первыми и третьими сторонами, некоторые тематические модели предоставляют информацию о типах собираемых личных данных: информация об учетной записи пользователя (тематическая модель 7), финансовые данные (тематическая модель 2), данные отслеживания местоположения и аналитики (тематическая модель 11). Некоторые темы, такие как тематические модели 4 и 10, раскрывают довольно специфические аспекты использования личных данных, такие как безопасность данных, включая случай, когда данные передаются третьим ли-

цам, и уведомление в случае изменения политики. Некоторые тематические модели являются довольно общими, например, модели характеризуют очень общие проблемы, связанные со сбором данных первой стороной и сторонним совместным использованием 0,1 и 3.

Таблица 13 – Тематическое моделирование

№	Координаты семантическогопространства	Возможные сценариииспользования
0	service, friend, story, child, cookie, use, product, email, compromised, card	First-party collection & usage (usage of cookies, e-mail), Special audience (children)
1	schedule, channel, analytic, happy, website, gather, address, mingle, moreover, identifiable	First-party collection (identifiable user data)
2	collect, credit, card, us, address, pursuant, email, service, personal, may	First-party collection: payment credentials
3	state, united, asset, website, policy, personal, privacy, party, third, sm	Third-party sharing
4	security, personal, rating, site, u, disclosure, service, policy, physical, third	Data security (including third-party sharing)
5	party, third, child, service, cookie, personal, personally, site, company, identifiable	Third-party sharing (usage of cookies)
6	service, website, personal, site, cookie, party, third, data, use, us	First-party collection & Third-party sharing (for: services provision, usage of website data and cookies)
7	personal, service, account, information, site, device, u, may, provide, use	First-party collection: user account information
8	device, resume, message, policy, privacy, social, service, site, website, networking	Other
9	opt, collect, site, third, advertising, personal, party, service, u, privacy	First-party collection & Opt-in, opt-out for advertising
10	military, change, policy, time, site, web, page, privacy, cookie, post	Privacy policy change, including notification mechanism
11	navigating, service, google, non, adsense, nielsen, account, collect, device, privacy	First-party collection: device and location information
12	station, feedback, service, consented, java, script, merchant, cookie, child, st	Other
13	cookie, service, third, party, site, website, california, flash, use, technology	Third-party sharing & Special audience: California residents
14	child, forum, trade, age, pii, conversation, chat, branded, personal	Special audience: children

Однако необходимо учитывать, что политики конфиденциальности в

большинстве случаев являются очень общими и неструктурированными, они не содержат четкой спецификации действий по обработке данных. Для некоторых тематических моделей было сложно определить аспекты сценариев использования, мы назвали их “Other”.

Также стоит отметить, что не существует моделей, посвященных хранению данных и аспектам доступа, редактирования и удаления данных. Это могло произойти из-за того, что количество абзацев, содержащих эту информацию, невелико, и они семантически довольно близки к сценарию первичного сбора. В отличие от них мы обнаружили проблемы, посвященные аспектам “International and Special Audience”, “Data Security” и “Privacy Policy Change”, хотя количество вхождений в наборе данных сопоставимо с “Data Retention” и “User Access, Edit and Deletion”.

Второй набор данных состоит из почти 600 политик конфиденциальности поставщиков устройств Интернета вещей [12]. Интересно, что оптимальное количество моделей также было определено как 15. Хотя извлеченные модели были довольно похожи, однако они содержали некоторые дополнительные детали. Как и в предыдущем случае, основная часть политик конфиденциальности посвящена использованию файлов cookie и настроек веб-браузера в сценариях сбора данных и сторонними организациями. Вторая тематическая модель, которая также широко представлена в политиках конфиденциальности, также касается first-party collection но с четко обозначенной основой обработки данных. В отличие от тематических моделей, построенных для набора данных OPP-115, существуют две тематические модели, посвященные праву доступа, редактирования и удаления данных. Этот факт можно объяснить, с одной стороны, большим размером набора данных, а с другой стороны, изменениями в законодательных положениях, которые были приняты недавно и посвящены правам субъекта данных. Данные OPP-115 были созданы в 2016 году до принятия GDPR [1], а набор данных политик конфиденциальности IoT был создан в этом году, и многие компании измени-

ли свои политики конфиденциальности, чтобы соответствовать требованиям GDPR.

Используя извлеченные тематические модели, мы проанализировали содержание политик конфиденциальности и вручную оценили точность индексации абзацев для набора выбранных политик. В общем случае точность, полученная для набора данных IoT, была немного выше, чем для моделей, извлеченных из OPP-115. Например, для политики конфиденциальности Xiaomi [14] мы получили точность 75% и 69% для наборов данных IoT и OPP-115 соответственно. На рисунке 2 показано распределение семантических тематических моделей абзацев в тексте Политики конфиденциальности Xiaomi. Отчетливо видно, что большая часть документа посвящена описанию различных аспектов сбора данных первой стороной – указанию, какие типы данных собираются, есть ли какие-либо варианты выбора/отказа. Полученные результаты также сравнивались с результатами [7] с помощью онлайн-инструмента Pribot [15]. Сравнительный анализ показал, что LDA выявило все основные аспекты использования персональных данных, за исключением одной целевой детской аудитории. Когда мы пересматривали политику, мы посвятили этому аспекту только одно предложение.

2.4 Подход основанный на применении контекстно-свободных грамматик и синонимическом поиске

Другой предложенный подход – подход, основанный на анализе с помощью контекстно-свободных грамматик и синонимического поиска. Синонимический поиск в данном случае – это подмена ключевых слов и их синонимов метками, например «__FP_A__» означает, что это слово и его синонимы считаются акторами первой стороны. Этот метод можно применить ко многим другим словам. Например, сообщения электронной почты, аватары, местоположение также могут быть объектами и синонимами абстрактной метки «__CN__», которая означает существительное сбора или объект сбора. Так все ключевые слова могут быть преобразованы в их смыслы в контексте

предметной области. Маркировка выполняется легко, все слова совпадающие с пулами заменяются метками этих пулов.

Предварительная обработка данных в данном случае состоит из токенизации и лемматизации для более гибкой замены слов на метки их пулов.

При анализе пользовательского согласия сайта недостаточно найти ключевые слова, относящиеся к разным типам персональных данных, например цель и правовую основу распознать гораздо сложнее. Следующий шаг - установить слова отношения в предложениях, чтобы можно было определенно сказать, что ярлыки пулы синонимов связаны друг с другом и формируют логическая цепочку. Один из возможных способов определения отношений слов в тексте на естественном языке – это синтаксический анализ предложения, основанный на частеречной разметке [32]. Имея размеченное по частям речи предложение, парсер грамматики NLTK [33] строит деревья предложений по правилам грамматики. Одно из таких деревьев в обозначениях NLTK можно увидеть на рисунке 12 [33], где «S» – основа предложения, «NP» – именная фраза, «VP» – глагольная фраза, «Adj» – прилагательное, «НОМ» – именное словосочетание, «ПП» – предлог фраза, «Det» – артикль, «V» – глагол, «N» – существительное, «P» – предлог.

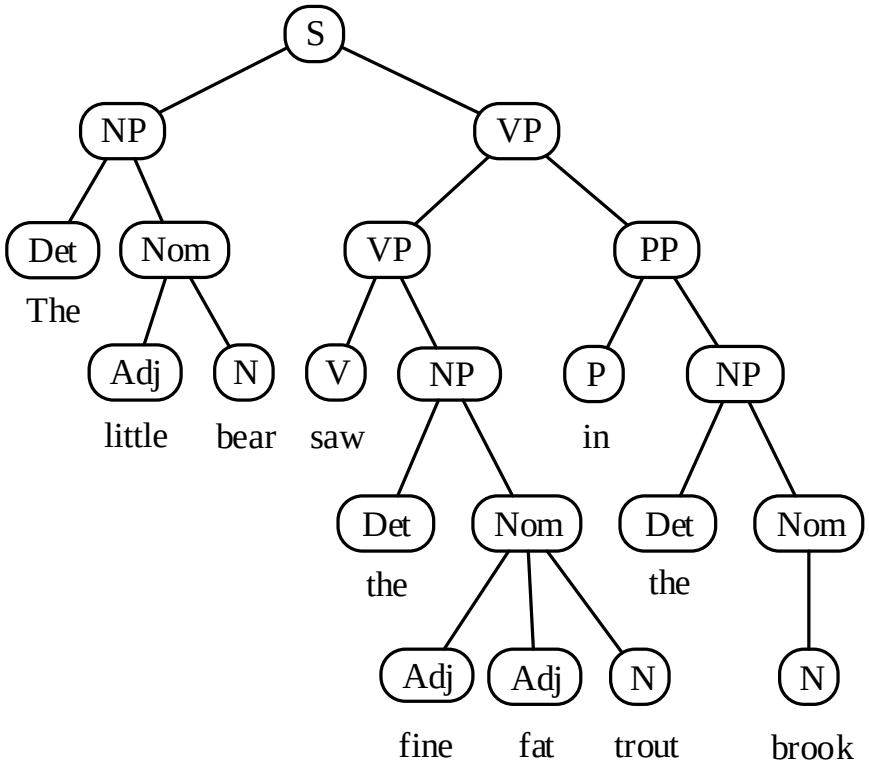

Рисунок 12 – Пример грамматического разбора

В предлагаемом подходе немного другая грамматическая запись. Созданная грамматика представлена в (5).

$$\begin{cases} D \to S \mid S\ D \mid S\ \ U\ D \\ S \to NPG\ \ VBG \\ VPG \to VP \mid VP\ \ VPG \mid VP\ \ U\ \ VPG \\ NPG \to NP \mid NP\ \ NPG \mid NP\ \ U\ \ NPG \\ AJPG \to AJ \mid AJ\ \ APG \mid AJ\ \ U\ \ APG \\ AVPG \to AV \mid AV\ \ APG \mid AV\ \ U\ \ APG \\ VP \to VAPG \mid V\ \ PPG \mid V\ \ PP\ \ APG \\ NP \to NOM \mid DET\ \ NOM \\ NOM \to N \mid AJPG\ \ N \\ PP \to NPG \mid P\ \ NPG \end{cases}, \tag{5}$$

где D – документ,

SB – синтаксическая основа предложения с его зависимостями,

U – союз,

NPG – группа именных фраз,

VPG – группа глагольных фраз,

$AJPG$ – группа однородных прилагательных,

$AVPG$ – группа однородных наречий,

PPG – группа однородных дополнений,

VP – глагольная группа,

NP – именная группа,

NOM – номинальная группа,

P – предлог,

AJ – прилагательное,

AV – наречие,

PP – существительное с предлогом,

N – существительное,

V – глагол,

DET – определяющее слово.

Грамматика из формулы (5) позволяет рекурсивно выделять основу предложения и последовательности глагола, существительного, прилагательного, наречия и т.д. Это все еще не идеальное решение, но попытка найти более сложные предложения в политиках безопасности. Этот подход требует использования пулов синонимов, которые соответствуют различным ключевым словам. Поэтому в грамматику включены метки пулов синонимов, привязанных к части речи. Метки пулов вручную назначены частям речи для преоб-

разования привязок частей речи NLTK, это показано в формуле (6).

$$\begin{cases} U \to NLTK\_CC \\ DET \to NLTK\_DT \\ AJ \to NLTK\_JJ \\ AV \to NLTK\_RB \\ N \to \_\_CN\_\_ \mid \_\_FP\_A\_\_ \mid \_\_TP\_A\_\_ \mid NLTK\_N \\ V \to \_\_CV\_\_ \mid NLTK\_V \end{cases}, \tag{6}$$

где $NLTK\_CC$ – соединение NLTK,

$NLTK\_N$ – все формы существительных NLTK,

$NLTK\_$ – все формы глаголов NLTK,

$NLTK\_DET$ – определители NLTK,

$NLTK\_RB$ – все формы наречий NLTK,

$\_\_FP\_A\_\_$ – метка актора-обладателя персональных данных,

$\_\_TP\_A\_\_$ – третья сторона,

$\_\_CV\_\_$ – глагол сбора,

$\_\_CN\_\_$ – существительное сбора.

Теги, начинающиеся с подчеркивания, являются метками пулов синонимов. Синтаксический анализ выполняет библиотека NLTK. На основе предложенной грамматики, описанной (5) и (6) и разметки лейблами пулов было построено дерево предложения, результат на рисунке 13.

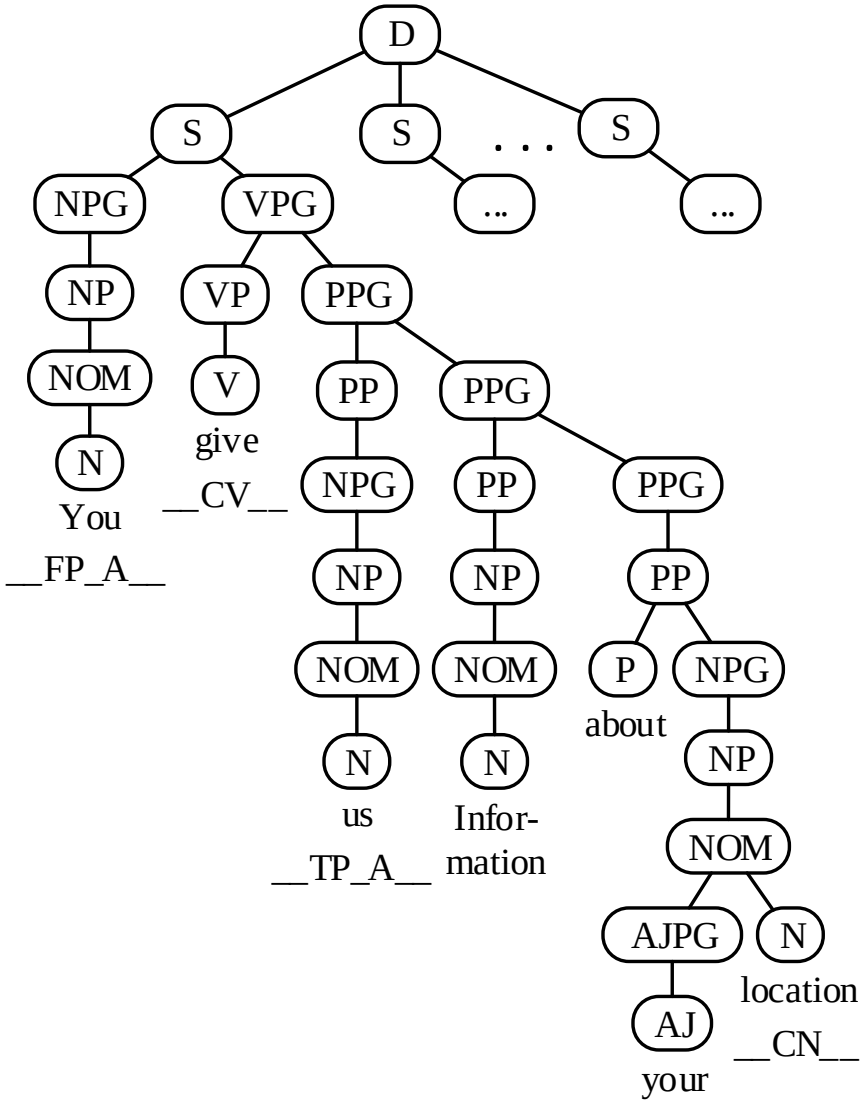

Рисунок 13 – Дерево грамматического разбора

Когда было построено дерево предложений последовательность меток ключевых слов может быть распознана. В этом случае представленная на рисунке 13, последовательность «__FP_A__», «__CV__», «__CN__» хорошо видна. Такие атомарные последовательности, раскрывают значения частей предложения и могут быть объединены в список, после этого весь смысл документов будет описан этим список. Сочетание маркировки ключевых слов и синтаксического анализа дает значения ключевых слов с отношениями между этими словами, определенными в виде древовидных структур. Дерево структура данных более гибкая, чем строка предложения, деревья и особенно поддеревья показывают важные отношения между словами. Запросы к таким

структурам могут дать необходимую информацию для построения логических последовательностей действующих лиц, их действий, субъектов этих действий и, наконец, обстоятельств. Предлагаемый подход определенно имеет такие недостатки, как низкая производительность, вручную определенные пулы синонимов и т.д.

В результате апробации алгоритма латентно-семантического анализа было выяснено что для кластеризации экстремально схожих между собой текстов он подходит не лучшим образом. В связи с этими обстоятельствами было решено обратить внимание на несколько иной подход анализа текста, основанный на контекстно-свободных грамматиках, тегировании по частям речи и синонимическом поиске.

2.5 Выводы по строгим методам текстового анализа

Эксперименты показали, что оба рассмотренных метода имеют как преимущества, так и определенные недостатки. Хотя предложенные подходы, оказались противоречивыми, окончательные результаты заслуживают внимания. Подход с латентно-семантическим поиском оказался не слишком эффективным. Однако, подход основанный на грамматическом анализе предложений и синонимическом поиске дал определенные результаты. Хоть он и не является производительным, с его помощью возможно производить выделение логических цепочек из предложений для получения более формального описания политик безопасности нежели их текстовые варианты.

2.6 Подход основанный на глубоком обучении

Исходя из проведенных исследований стало понятно, что более предпочтительным вариантом решения задачи будет подход с применением моделей с глубоким обучением. Реализация подобного проекта – комплексная задача, ее можно разделить на несколько этапов. Сначала необходимо собрать датасет, потом его разметить для обучения модели, далее обучить модель и получить результаты. Однако сбор датасета тоже является непростой задачей. Для того чтобы осуществить сбор датасета необходим инструмент для

поиска и скачивания вэб-страниц из сети интернет. Затем необходимо произвести очистку данных, удалить все теги со страниц, чтобы можно было передать текст аннотаторам. Все этапы сбора датасета полагаются на базу данных. Она лишена сложного объектно-реляционного моделирования, так как в ней по сути необходимо только хранить промежуточные результаты обработки текстовых файлов.

3 Проектирование инструментария

3.1 Техническое задание «Инструментарий для сбора датасета»

3.1.1 Основные положения технического задания

3.1.2 Скрейпер вэб-страниц

Скачивание веб-страниц будет производиться инструментом написанным на языке Python, с помощью библиотек можно скачивать страницы анализировать данные с них, переходить по гиперссылкам и много другое. Такой инструмент позволит просматривать и сохранять содержимое страниц в автоматическом режиме без вмешательства пользователя. Таким образом в автоматическом режиме можно сохранить и проанализировать огромное количество текстовой информации.

3.1.3 Очистка скачанных страниц политик

Для очистки страниц от кода разметки планируется использовать библиотеку «html sanitizer». Очистка кода необходима для того, чтобы аннотаторы могли максимально сфокусироваться на анализе текста, таким образом получая чистый текст они не будут отвлекаться на не имеющие значения в контексте задачи фрагменты.

3.1.4 Инструмент разметки датасета

Инструмент разметки датасета планировалось реализовать с помощью веб-технологий. Серверная часть будет полагаться на приложение, написанное на PHP, которое будет регулировать порядок выдачи текста на аннотирование. Процесс разметки высокодинамичен, поэтому невозможно избежать написания качественной клиентской части приложения на языке javascript. Это позволит сделать работу аннотаторов максимально производительной, в «одну сессию», так как страница не будет перезагружаться, однако все изменения, которые будут вносится, сохранятся.

3.1.5 Фреймворк глубокого обучения

Для создания и тренировки модели анализа текста планируется использовать фреймворк машинного обучения «Keras». Он позволяет быстро создавать классификаторы с самыми разными конфигурациями и любых типов.

После того как классификатор будет сконфигурирован останется лишь обучить его на датасете, полученном ранее.

Обученный классификатор будет в состоянии определять различные характеристики политики безопасности и аспекты обращения с данными, что позволит в автоматическом режиме формировать краткие отчеты о безопасности предоставляемого соглашения.

3.2 Методика сбора

Планируя решение появившейся задачи важно уделить внимание источникам дынных для сбора, потому что без них невозможно будет продолжать работу. Это важно еще и потому что необходимо будет адоптировать инструмент сбора данных под конкретные веб-ресурсы, так как на каждом из них реализована собственная html-разметка.

Исходя из ориентированности дата сета на умные устройства, логичным выглядит обращение к крупным торговым площадкам, так как они занимаются дистрибьюцией подобных устройств. На сайтах торговых площадок можно осуществлять поиск продукции и получать данные о ней в том числе и производителя продукции. Типовая разметка веб-страниц располагает для получения такой информации, так как существует лишь несколько вариантов наполнения страницы продукции.

Торговые площадки не предоставляют ссылки на официальные сайты производителей. Поэтому необходимо организовать поиск официальных сайтов производителей. Поисковые движки предоставляют API для поиска, однако некоторые из них являются платными, другие выдают совершенно неприемлемые результаты. С другой использование поисковых движков, пред-

назначенных для реальных пользователей, дает наилучшие результаты из возможных, скорее всего это связано с клиентоориентированностью, то есть получая запрос близкий к наименованию бренда с большей вероятностью будет выдана официальная страница производителя в Интернете.

Далее важной задачей является определение какая из ссылок в результате запроса наиболее четко соответствует искомому производителю. Получение официальных веб-сайтов производителей задача на первый взгляд сложная, однако результаты ручной проверки показали, что лучшим вариантом является поисковый запрос с названием производителя «как есть». В таком случае вебсайт производителя оказывается на первой странице результата поискового запроса, а если не оказывается, значит у этой компании его с очень большой вероятностью нет.

Получив ссылки предполагаемых официальных сайтов, мы получаем доступ к страницам, на которые они ведут. Поиск политики безопасности на уже обнаруженном сайте производителя является тривиальной задачей. Сейчас на абсолютном большинстве сайтов в футере имеется ссылка, названная как “Privacy” или “Privacy Policy”. Футер доступен на любой странице сайта и является частью глобальной навигационной системы сайта, в него вынесена информация, которая пригождается не так часто как, например, информация из верхних баров и меню, однако тем не менее эта информация важна, и помимо ссылок на политику безопасности зачастую содержит контактные данные и прочую организационную информацию.

Таким образом можно получить ссылки на политики безопасности производителей умной продукции. Далее необходимо произвести обработку скачанных политик безопасности.

3.3 Методика очистки

Очистка политик безопасности является комплексной задачей. Получив политику безопасности, необходимо вырезать все теги, которые несут в себе динамику, то есть все элементы управления. Такие элементы как всплы-

вающие модальные диалоговые окна тоже не могут содержать текст политики безопасности. Изображения, помещенные на странице, так же не относятся к политике безопасности. Таким образом получается, что большое количество тегов необходимо агрессивно удалять еще до начала анализа страницы, так как они точно не содержат полезной информации.

Далее необходимо применить обработку, которая включала бы в себя преобразование разметки: недопустимые теги должны быть развернуты, определенные комбинации вложенных тегов должны быть заменены на более тривиальные. Также необходимо очистить теги от атрибутов, так как в них не содержится полезной информации или чего-либо способного положительно сказаться на структуре очищенного документа. Затем по всему дереву DOM осуществляется рекурсивный обход с целью слияния тегов, где это возможно, или оборачивания сырых текстов. В ходе данного этапа также производится нормализация пунктуации и настройки отступов текстов, чтобы привести их к читабельному виду.

После указанных двух этапов очистки, следует заключительный, на котором из тегов извлекается текст, то есть параграфы, представленные в виде одной длинной строки. Это делается, потому что расставленные определенным образом переводы на новую строку могут по тем или иным причинам не подходить, и это будет более гибким решением, потому что где требуется можно применить лайн-врапинг.

3.4 Методика разметки

Текст…

3.5 Потенциальные проблемы

Еще до решения задачи были выделены потенциальные проблемы, способные замедлить процесс разработки и сбора дата сета. Потенциально возможные проблемы при реализации приложений по добного типа следующие:

1) блокировка из-за подозрительных заголовков браузера,

2) блокировка из-за слишком частого обращения с запросами,

3) как следствие 2-х предыдущих пунктов требование подтвердить, что это не попытка автоматического доступа (ввод капчи).

4) Невидимые элементы разметки,

5) динамически формируемые страницы торговых площадок и политик безопасности,

6) промахи при сборе данных из-за частично некорректных результатов поиска на торговых площадках и в поисковых движках.

Проблемы 1, 2, 3 решаются использованием разных заголовков браузера попеременно. Также отправка запросов ограничена по частоте от 2 до 6 секунд, ограничение выбирается случайным образом. Такие решения позволяют крайне редко попадать под подозрения, потому что в таком случае поведение максимально похоже на поведение реального пользователя, соответственно процент успеха при попытке получить данные с веб-страницы значительно повышается. Стоит отметить, что данные ограничения очень эффективно обходятся за счет использования прокси-серверов, которые позволяют менять ip-адреса. Еще одним важным и эффективным инструментом для является профиль браузера. Он позволяет запускать безголовый браузер с определенной историей использования будь то куки-файлы, история запросов или аутентификация на различных сервисах. Наличие такой предыстории у браузера для некоторых сайтов является доказательством, что он не автоматизирован.

Проблема 4 решается следующим образом. Попав на страницу политики безопасности, можно исполнить код на javascript, который загрузит на страницу библиотеку для работы с деревом DOM и удалит невидимые элементы разметки.

Проблема 5 решается использованием безголового браузера, который полнофункционален с точки зрения воспроизведения контента, так как поддерживает исполнение javascript кода на странице. Таким образом страница будет загружена и динамические элементы будут созданы, после чего можно

будет их обработать. Однако на некоторых веб-сайтах для того, чтобы получить ту или иную информацию необходимо заполнить форму. С такими обстоятельствами сложно бороться – разметка всегда различается, но таких случаев крайне мало, поэтому исключение их из рассмотрения будет оправданным.

Проблема 6 может отчасти решиться конкретизацией поискового запроса путем прибавления к названию производителя ключевых слов и продукции, которая им производится. Хотя этот вариант и показал гораздо более качественные результаты нежели чем поиск производителя «как есть», иногда все же попадается шум.

3.6 Приложение вэб-скрейпер

3.6.1 Первичная декомпозиция и планирование

Начальным этапом решения задачи является первичная декомпозиция, в ее результате выделяются подзадачи различной важности, которые должны быть решены для доведения цикла разработки до конца. В данном случае можно выделить следующие подзадачи:

1) определение источника информации о различной IoT-продукции,

2) отправка поискового запроса,

3) получение результатов запроса (список IoT-продуктов),

4) определение производителей IoT-продукции,

5) поиск официальных сайтов производителей в сети интернет,

6) поиск раздела «политика безопасности» на сайтах производителей,

7) скачивание политик безопасности,

8) очистка скачанных веб-документов от лишних элементов разметки,

9) слияние тегов и оборачивание сырого текста,

10) нормализация пунктуации и отступов,

11) извлечение текста из тегов.

Получение списка производителей возможно на электронных торговых площадках, типовая разметка веб-страниц располагает для получения та-

кой информации, так как существует лишь несколько вариантов наполнения страницы продукции.

Получение официальных веб-сайтов производителей задача на первый взгляд сложная, однако результаты ручной проверки показали, что лучшим вариантом является поисковый запрос с названием производителя «как есть». В таком случае веб-сайт производителя оказывается на первой строчке результата поискового запроса, а если не оказывается, значит у этой компании его с очень большой вероятностью нет.

3.6.2 Структура приложения вэб-скрейпера

Исходя из результатов декомпозиции, эффективным подходом выглядит представление приложения в виде последовательно выполняющихся подпрограмм так, что входом модуля является результат работы предыдущего модуля, то есть в виде конвейера. Схема организации приложения представлена на рисунке 14.

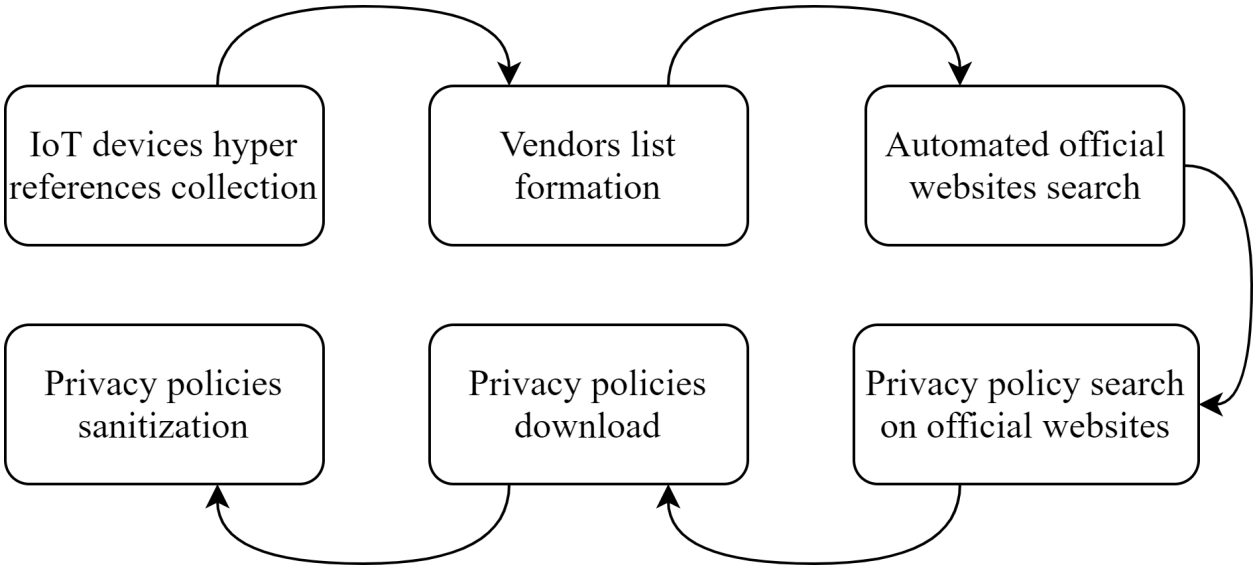

Рисунок 14 – Схема организации приложения

Далее была разработана композиционная модель приложения, на ней присутствуют все необходимые для решения задач модули. Схема представлена на рисунке 15.

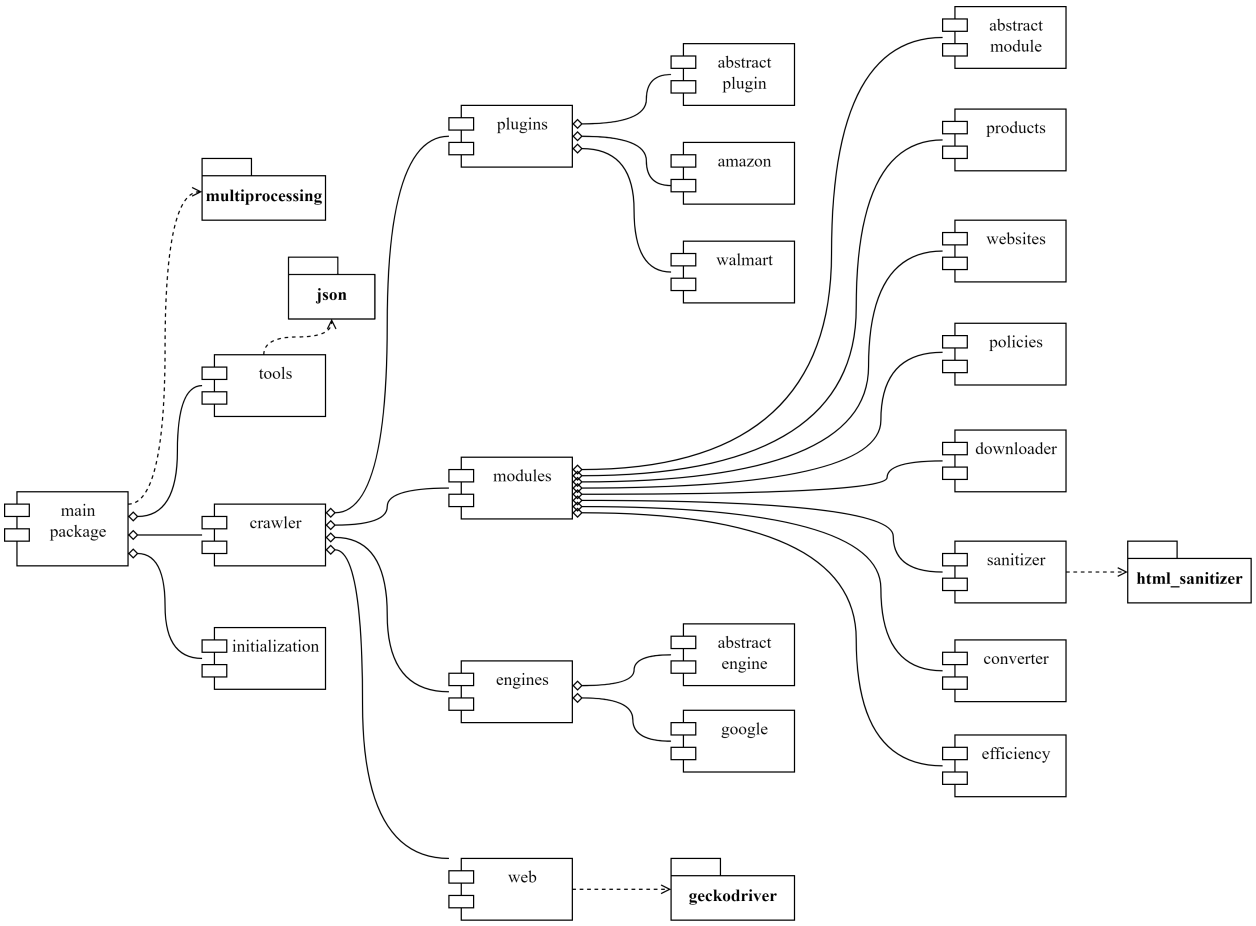

Рисунок 15 – Композиционная модель приложения

3.6.3 Средства разработки вэб-скрейпера

Для реализации приложения были выбраны следующие средства:

1) python 3.9,

2) «безголовый» браузер Firefox,

3) драйвер для управления «безголовым» браузером «geckodriver»,

4) библиотека html-sanitizer для очистки скачанных веб-документов.

Выбор «безголового» браузера обусловлен потребностью в отрисовке страниц, так как на некоторых веб-страницах разметка генерируется с помощью javascript. Это делает невозможным использование простого скачивания, не обходима страница именно с исполненными скриптами, в противном случае будет невозможно получить требуемую информацию. В то же время браузер лишен графического интерфейса, чем снижается потребление вычислитель ных ресурсов.

Таким образом приложение построено на 4 основных абстракциях.

1) Концепция модуля – одна из основополагающих, так как модулем в данном случае выступает любая подпрограмма, участвующая в сборе данных, принимающая входные данные в виде json-файла, и на выходе дающая так же json-файл чтобы следующий в очереди модуль мог отработать. Модули могут быть написаны с нуля, а могут расширять возможности уже существующих посредством механизма наследования. Таким образом можно не переписывая существующий код, а только добавляя новый изменять поведение программы и адаптировать ее под разные задачи сбора данных.

2) Концепция конвейера – этот элемент поочередно вызывает модули и передает данные из одного модуля в другой. В результате отработки всех модулей поэтапно решается поставленная задача, то есть сбор данных из интернет-источников. Конвейер может быть сконфигурирован, в него могут быть помещены любые модули, реализующие соответствующий интерфейс. Также может быть сконфигурирована последовательность запуска модулей сбора данных.

3) Концепция поискового движка – данная концепция порождена в связи с необходимостью сделать приложение как можно более гибким. Такой абстрактный элемент позволяет менять используемые поисковые движки, применять к результатам поиска алгоритмы для определения какие результаты удовлетворяют условиям поиска, а какие нет.

4) Концепция плагина – плагин обеспечивает сбор данных с какой-либо конкретной торговой площадки. Данная концепция использована так же для обеспечения гибкости приложения – для устранения привязки к набору конкретных торговых площадок. Использую механизм наследования можно переопределить поведение плагина для работы с любой другой торговой площадкой.

На рисунке 2 модуль «main» отвечает за запуск программы, разверты-

вание основных ее частей. Там же происходит инициализация пула процессов для мультипроцессинга затратных задач таких как, например, взаимодействие с «безголовым» браузером. Он так же отвечает за последовательное исполнение подпрограмм элементов конвейера. Он осуществляет прием выходных и передачу входных данных модулей.

Модуль «initialization» производит проверку файловой системы и создает необходимые директории в папке ресурсов.

Модуль «tools» содержит вспомогательные функции, в частности для ввода и вывода данных в формате json.

Модуль «crawler» отвечает за получение данных с веб-страниц, в нем агрегированы все инструменты для сбора и очистки данных.

Модуль «plugins» включает в себя набор плагинов, каждый из которых адаптирован для получения требуемой информации с определенного шаблона веб-страничной разметки. Некоторое поведение инкапсулировано в абстрактном плагине для увеличения «reusability» кода. Получая адрес на вход, данный плагин производит скачивание страницы и с помощью набора шаблонов пытается извлечь информацию. Данный модуль записывает полученную с помощью плагинов информацию в json-файл для большей прозрачности и возможности сохранения результатов между запусками приложения, например, для пропуска данного этапа и использования его сохраненных результатов работы.

Данные полученные с помощью модулей «products», «websites», «policies», «downloader», «sanitizer», «converter» и «efficiency» записывается в json-файлы для большей прозрачности и возможности сохранения результатов между запусками приложения, например, при пропуске какого-либо из этапов и использования его сохраненных результатов работы. Модуль «products» получение производителей IoT-продуктов. Модуль «websites» получение официальных сайтов производителей. Модуль «policies» получение веб-ссылок на политики безопасности. Модуль «downloader» отвечает за скачивание стра-

ниц и их сохранение в отведенную для этого директорию. Модуль «sanitizer» отвечает за очистку скачанных веб-страниц от не нужных тегов и ссылок. Модуль «converter» производит перевод политик безопасности из веб-страничного вида в текстовое представление. Модуль «efficiency» производит расчет статистики по дата сету.

Модуль «web» отвечает за взаимодействие с вебсайтами будь то торговые площадки или сайты производителей IoT-продуктов. В нем используется geckodriver для управления «безголовым» браузером.

Модуль «proxy» содержит инструменты для скачивания и автоматического применения бесплатных прокси-серверов. Однако ввиду ненадежности бесплатных, есть так же возможность задать список выделенных прокси-серверов.

Для обеспечения наиболее гибкой настройки как можно больше настроек выведено в отдельный конфигурационный файл. В нем задаются:

1) параметры для библиотеки html-sanitizer, в частности набор допустимых тегов и допустимых атрибутов;

2) параметры безголового браузера, в том числе количество повторных попыток при сбоях, появлении каптчи и так далее, набор юзерагентов для перебора, флаги использования кэширования, флаг запуска браузера в режиме без графического интерфейса, флаг использования прокси, пути для логов, а также путь до профиля браузера;

3) список директорий и файлов, в которые происходит сохранение результатов сбора данных;

4) количество процессов для одновременного сбора данных на многоядерных конфигурациях.

Для настройки работы заменяемых элементов таких как поисковые движки плагины и модули, предусмотрены отдельные файлы, в которых создаются те или иные конфигурируемые объекты.

Учитывая конвейерную организацию и передачу результатов из модуля в модуль посредством json-файлов, структура дата сета следующая: каждый модуль имеет свой json-файл для записи результатов. По сути результаты – это массив из python-словарей, каждый словарь является своего рода кортежем, эти кортежи обладают избыточностью данных, однако, таким образом достигается максимальная простота формализации данных. Каждый элемент – IoT устройство, обладающее набором информационных полей: идентификатор; ссылка на страницу на торговой площадке; наименование производителя; ключевое слово, по которому было найдено устройство; ссылка на сайт производителя; ссылка на политику безопасности; путь к сохраненной оригинальной страницы политики безопасности; путь к очищенной политике безопасности; путь к текстовой версии политики безопасности; хэш, сгенерированный по тексту политики; блок статистики по структурным элементам, таким как нумерованные и ненумерованные списки, элементы списков, таблицы, параграфы, длина политики в символах. Пример такой разметки можно увидеть на рисунке 16.

```
{
  "id": 1,
  "url": "https://www.walmart.com/ip/GreaterGoods-Smart-Scale-BT-Connected-Body-Weight-Bathroom-Scale-BMI-Body-Fat-Muscle-Mass-Water-Weight-FSA-HSA-Approved/696264102",
  "manufacturer": "greater goods",
  "keyword": "smart scale",
  "website": "http://greatergoods.com",
  "policy": "http://greatergoods.com/legal/privacy-policy",
  "original_policy": "D:\\source\\repos\\iot-dataset\\original_policies\\greatergoods.com-legal-privacy-policy.html",
  "processed_policy": "D:\\source\\repos\\iot-dataset\\processed_policies\\greatergoods.com-legal-privacy-policy.html",
  "plain_policy": "D:\\source\\repos\\iot-dataset\\plain_policies\\greatergoods.com-legal-privacy-policy.html.txt",
  "policy_hash": "9d63c3eeb2a4ef4ad0b4428ad56d4be5",
  "statistics": {
    "length": 25888,
    "table": 0,
    "ol": 0,
    "ul": 7,
    "li": 27,
    "p": 39,
    "br": 5
  }
}
```


Рисунок 16 – Пример кортежа дата сета

В веб-краулере также предусмотрена возможность явного указания адресов для скачивания политик безопасности, для чего предусмотрен отдельный json-файл, содержащий элементы со схожей структурой. В нем можно указывать любые из полей – они будут заполнены соответствующе, а незаполненные поля останутся равными «null». Явно заданные для скачивания политики считываются непосредственно на этапе скачивания, таким образом данные о названии производителя и другие данные которые участвуют в более ранних стадиях сбора несут сугубо справочный характер. Статистические показатели политик безопасности рассчитываются на последнем этапе работы приложения, что означает их перезапись после каждого запуска, при

условии, что модуль расчета статистики активен.

3.7 Инструмент разметки датасета

Инструмент разметки датасета планировалось реализовать с помощью веб-технологий. Серверная часть будет полагаться на приложение, написанное на PHP, которое будет регулировать порядок выдачи текста на аннотирование. Процесс разметки высокодинамичен, поэтому невозможно избежать написания качественной клиентской части приложения на языке javascript. Это позволит сделать работу аннотаторов максимально производительной, в «одну сессию», так как страница не будет перезагружаться, однако все изменения, которые будут вносится, сохранятся.

3.7.1 Объектное моделирование приложения

Объектная модель инструмента представлена на рисунке 17.

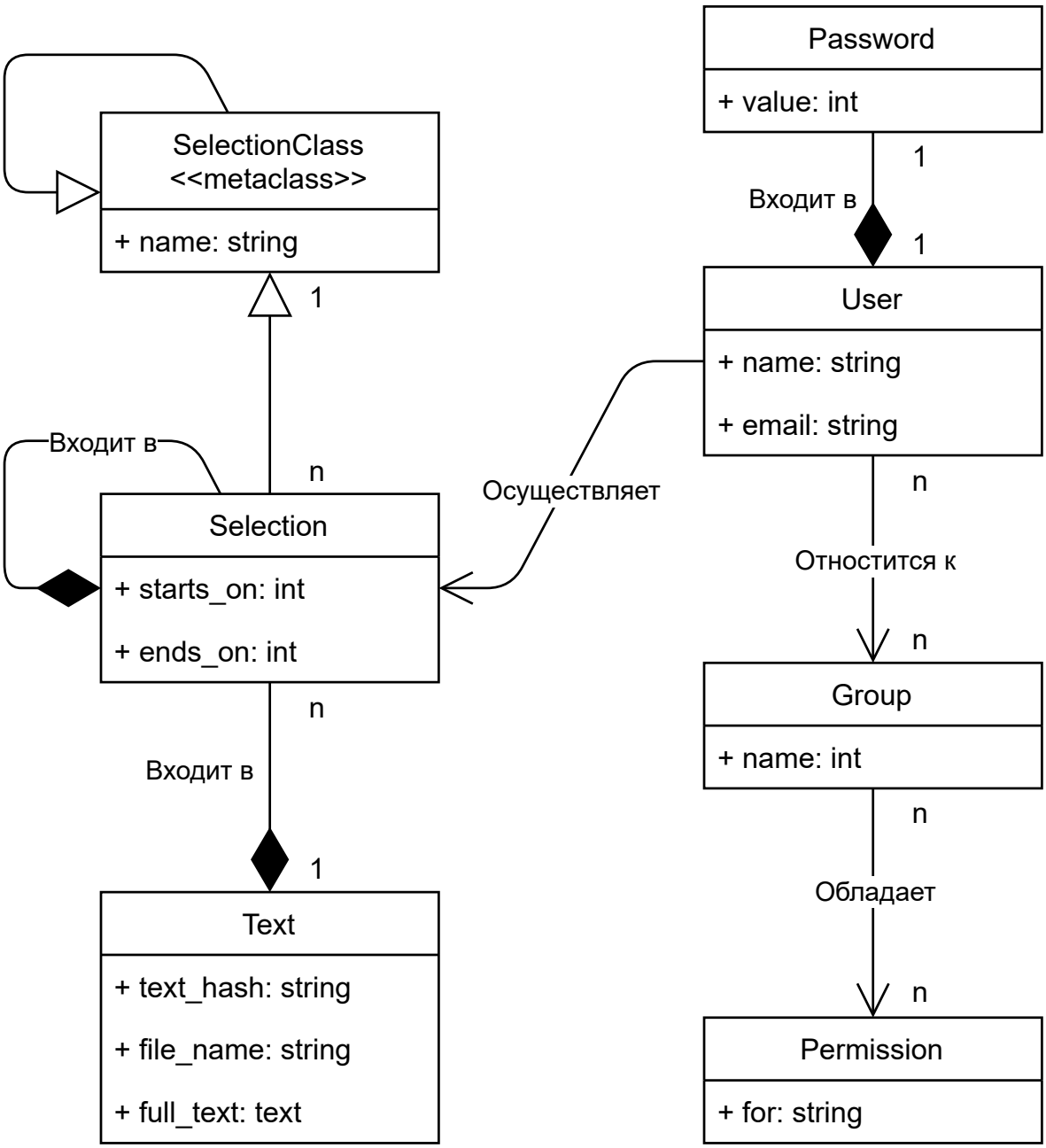

Рисунок 17 – Объектная модель

3.7.2 Реляционная модель приложения

Реляционная модель инструмента представлена на рисунке 18.

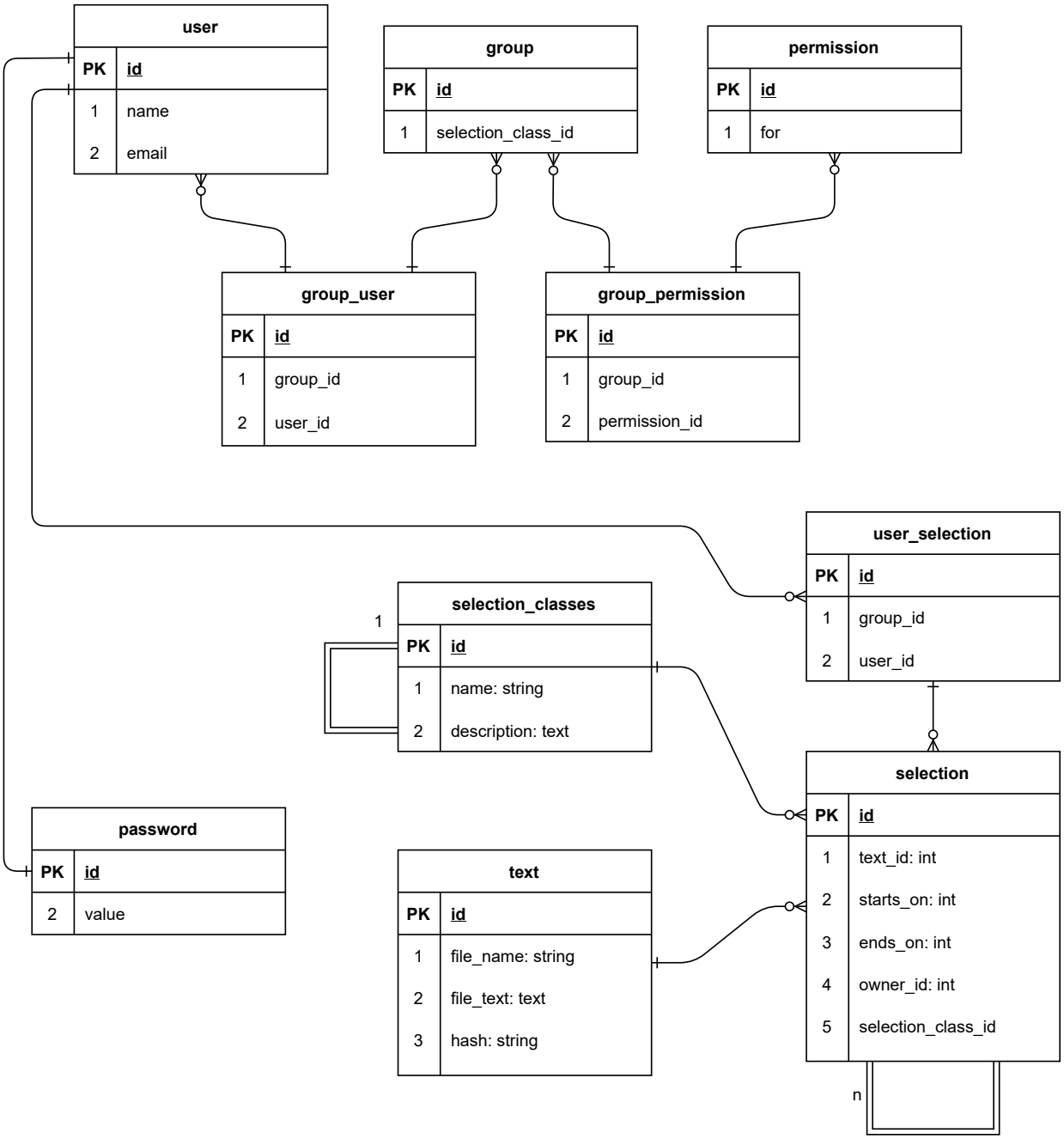

Рисунок 18 – Реляционная модель

3.7.3 Проектирование пользовательского интерфейса

Презентационный прототип интерфейса инструмента представлен на рисунке 19.

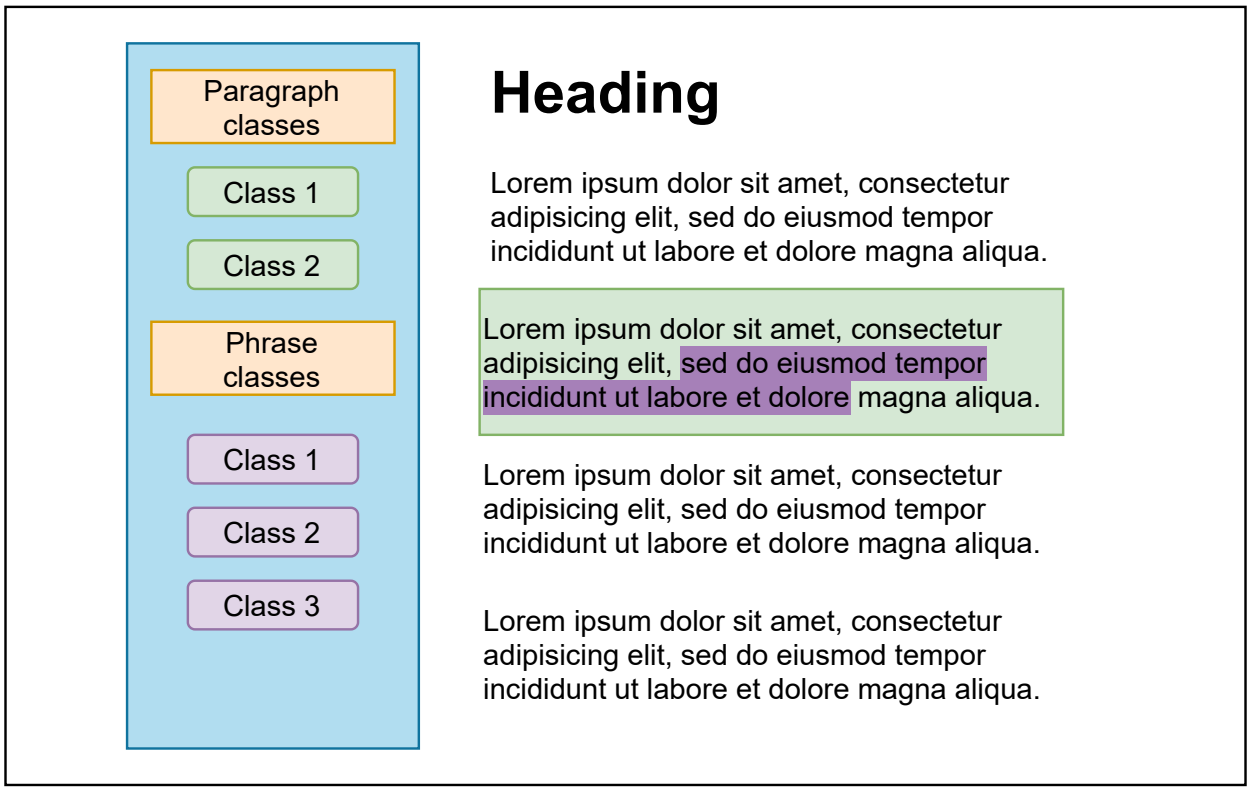

Рисунок 19 – Презентационный прототип интерфейса

3.7.4 Средства разработки инструмента разметки

4 Технические детали реализации инструментария

4.1 Полученные в результате реализации исходные коды

В соответствии с результатами декомпозиции, выбора средств и проектирования приложение было реализовано. Характеристики полученных классов и функций приведены далее, в таблицах ?-?. Исходные коды представлены в приложениях Б и В.

4.2 Полученный в результате сбора данных дата сет

Поиск осуществлялся на торговых площадках amazon и walmart, брались результаты поискового запроса по первым 30-ти страницам, по категориям «smart scale», «smart watch», «smart bracelet», «smart lock», «smart bulb», «smart navigation system», «smart alarm clock», «smart thermostat», «smart plug», «smart light switch», «smart tv», «smart speaker», «smart thermometer», «smart air conditioner», «smart video doorbell», «robot vacuum cleaner», «smart air purifier», «gps tracking device», «tracking sensor», «tracking device», «indoor camera», «outdoor camera», «voice controller». Всего производителей было найдено приблизительно 160. Стоит отметить, что результат является приемлемым, так как многие производители на данной торговой площадке не имеют вы деленного вебсайта, а пользуются услугами amazon, то есть на таких страницах действует политика безопасности amazon, а не производителя. Также стоит отметить, что у некоторых продуктов явно не указан производитель, что сократило количественно результат поиска.

Всего было проанализировано 57150 моделей умной продукции, из них для 51727 (90,5%) были определены производители. Всего уникальных производителей было найдено 6161, из них 1419 (23%) имеют официальную вебстраницу. Проанализировав найденные веб-сайты были собраны 798 политик безопасности, разумеется, среди них имеется определенный процент промахов, если производитель имеет сходство с каким-либо другим более крупным. Из дата сета были исключены политики безопасности, длина которых

в символах не превышала 1000. Это объясняется тем, что некоторые производители имеют на своем сайте страницу с политикой безопасности, но по каким-то причинам эта страница не наполнена. Примеры таких случаев приведены на рисунках 20 и 21. Таким образом полноценных уникальных политик безопасности осталось 592.

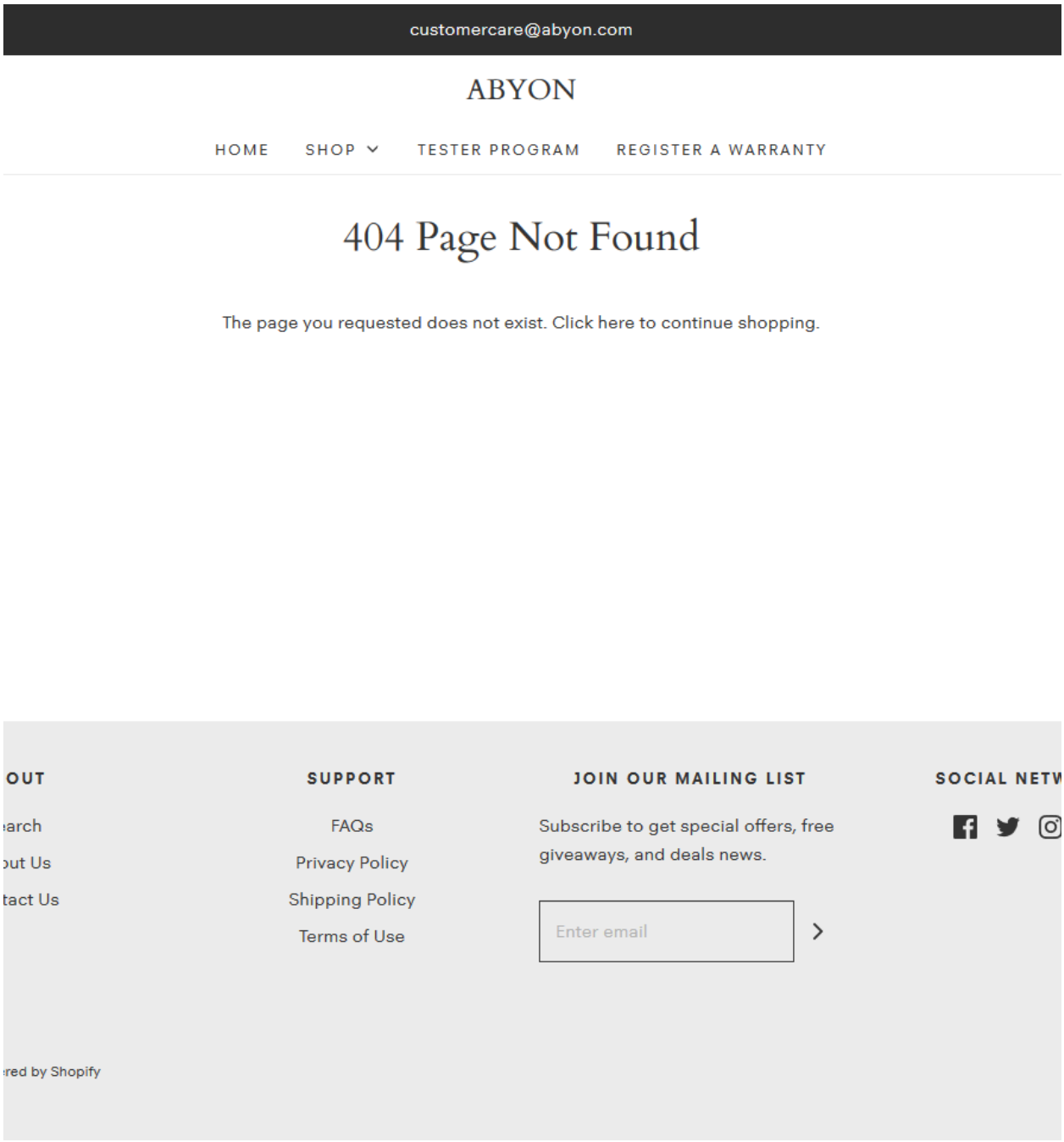

Рисунок 20 – Пример отсутствующей политики

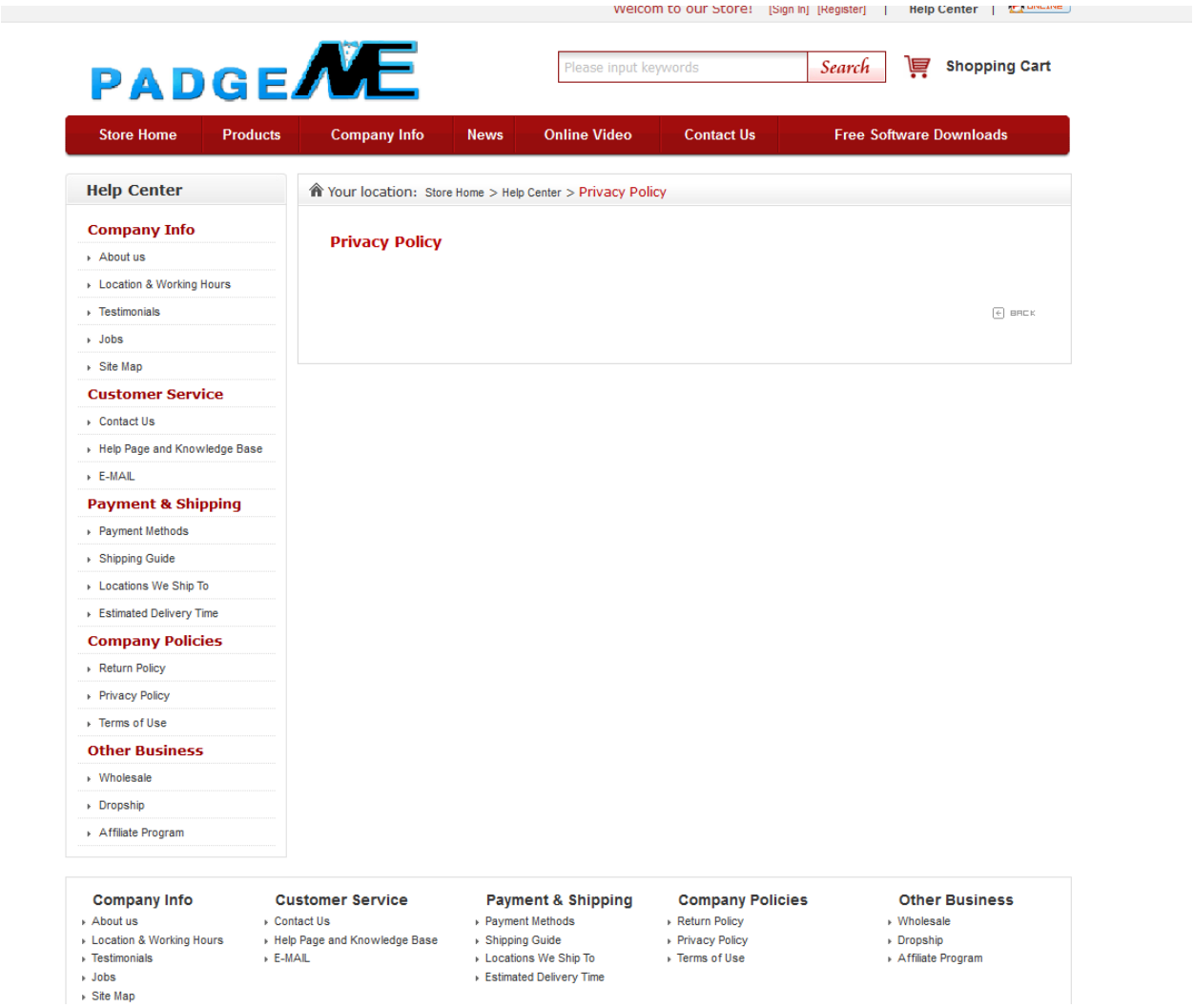

Рисунок 21 – Пример отсутствующей политики

Некоторые из производителей, которые не имеют собственного веб-сайта и политика безопасности которых не была найдена, пользуются услугами хостинга интернет-магазина непосредственно на amazon. В таком случае, будучи частью интернет-магазина на них распространяется политика безопасности площадки, на которой они размещают свои предложения, причем политики могут различаться для разных стран. Случаи с использованием отдельных политик безопасности под различные типы устройств не были зафиксированы, хотя такие случаи и существуют, проще прибегнуть к явному заданию адресов политик, нежели чем к попытке автоматизировать процесс сбора, так как остаются непрозрачными способы выявления подобных ситуаций.

На рисунках 22 и 23 приведены статистические данные по объемам абзацев политик и самих докумнетов соответственно.

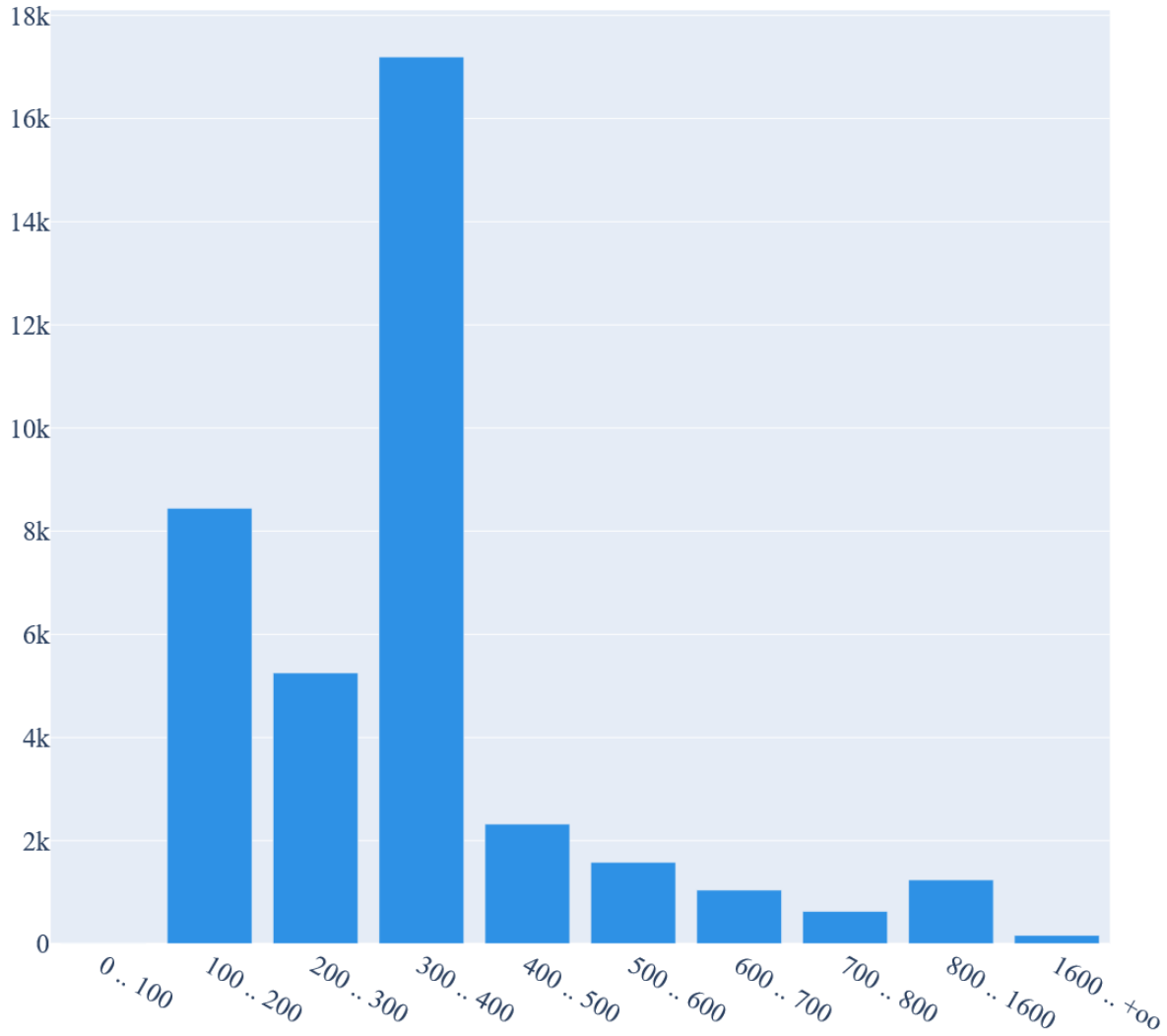

Рисунок 22 – Распределение политик по объему параграфа

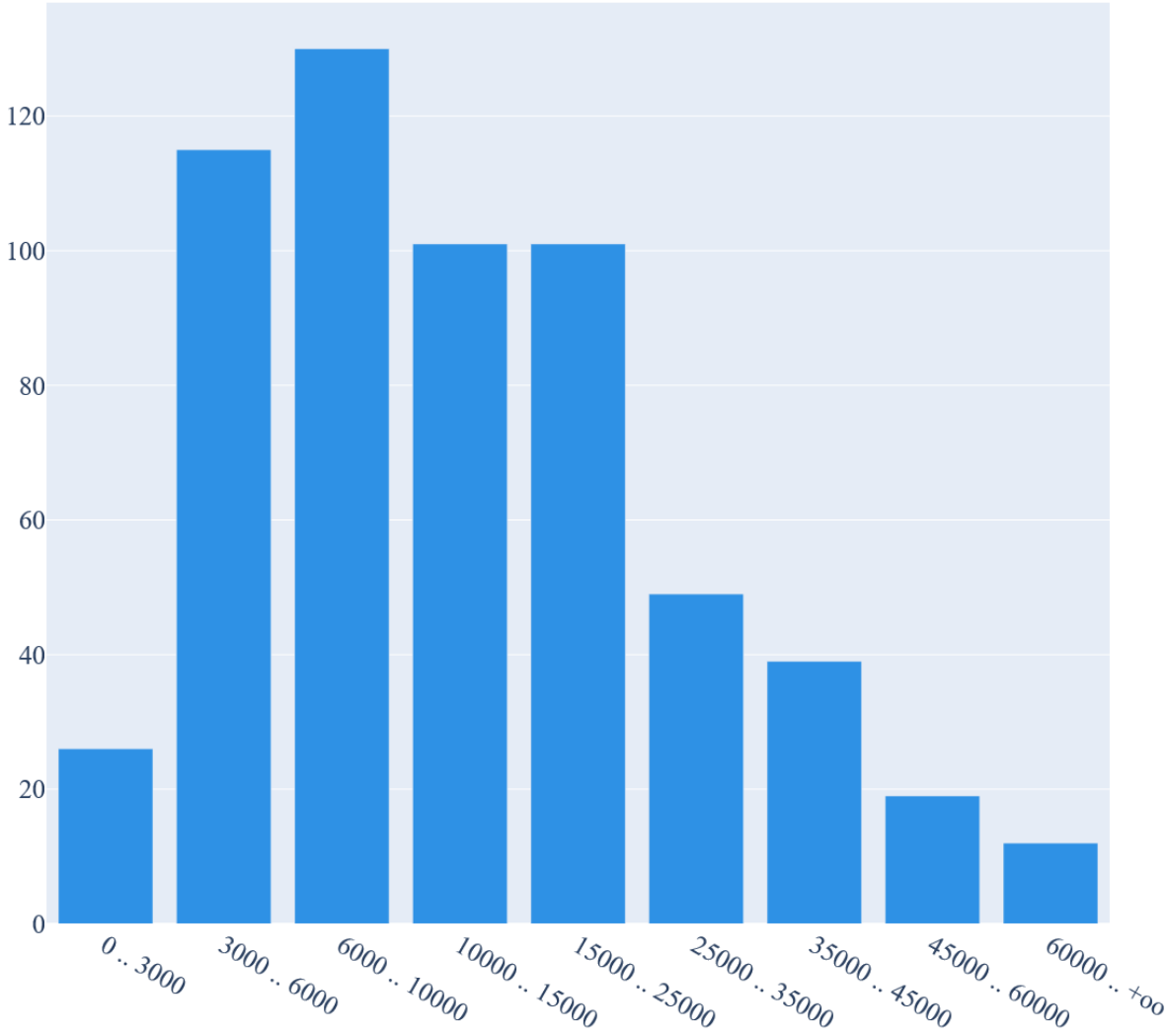

Рисунок 23 – Распределение политик по объему документа

Подсчет количества заголовков сложно организовать автоматизированно в связи с большим разнообразием html-разметки. На каждом сайте своя разметка, своя конвенция по нумерованию секций, заголовков, списков. На некоторых сайтах списки и заголовки нумеруются средствами html, на других нумерация проставлена вручную. Все это порождает разношерстность данных, и их обработка становится сложной с точки зрения учета всех возможных вариантов. Поэтому авторы решили прибегнуть к простому подсчету количества строк длиной меньше 100 символов и не содержащих при этом маркеров «list item». Такой подход не даст очень точных показателей, но может дать приблизительные значения. На рисунках 24 и 25 приведена стати-

стика по структурным элементам политик безопасности в двух частях. Здесь изображены детальные распределения структурных элементов для каждой из найденных политик безопасности.

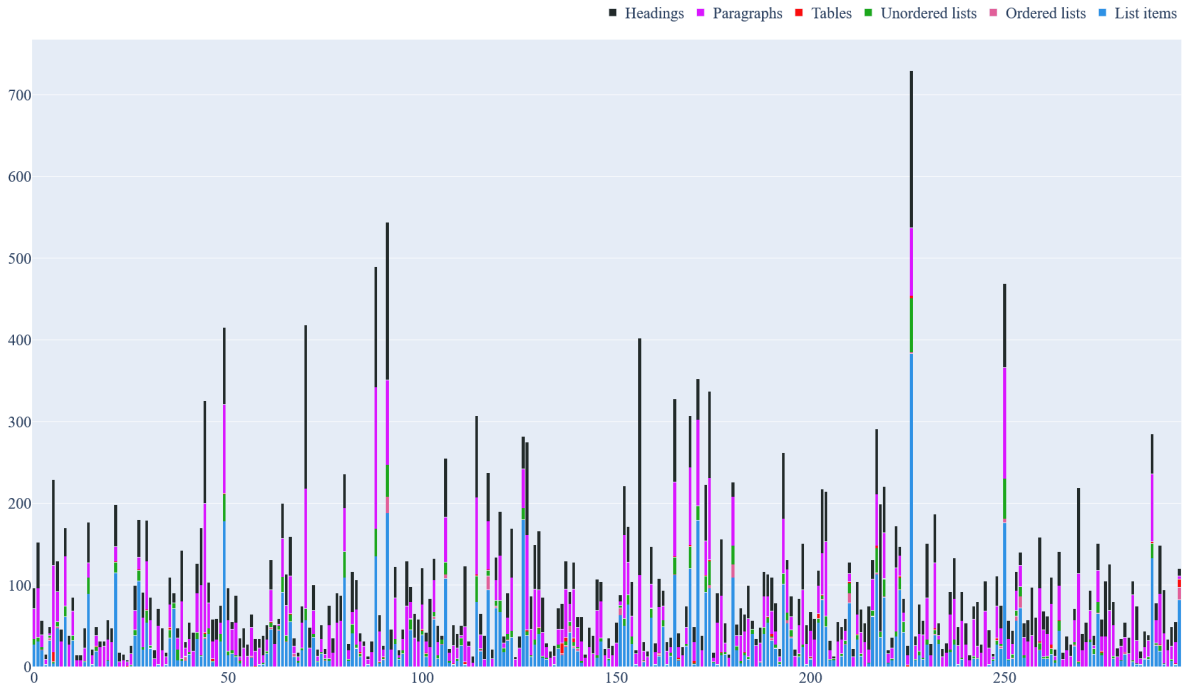

Рисунок 24 – Статистика первых 246 политик в IoT дата сете по структурным элементам

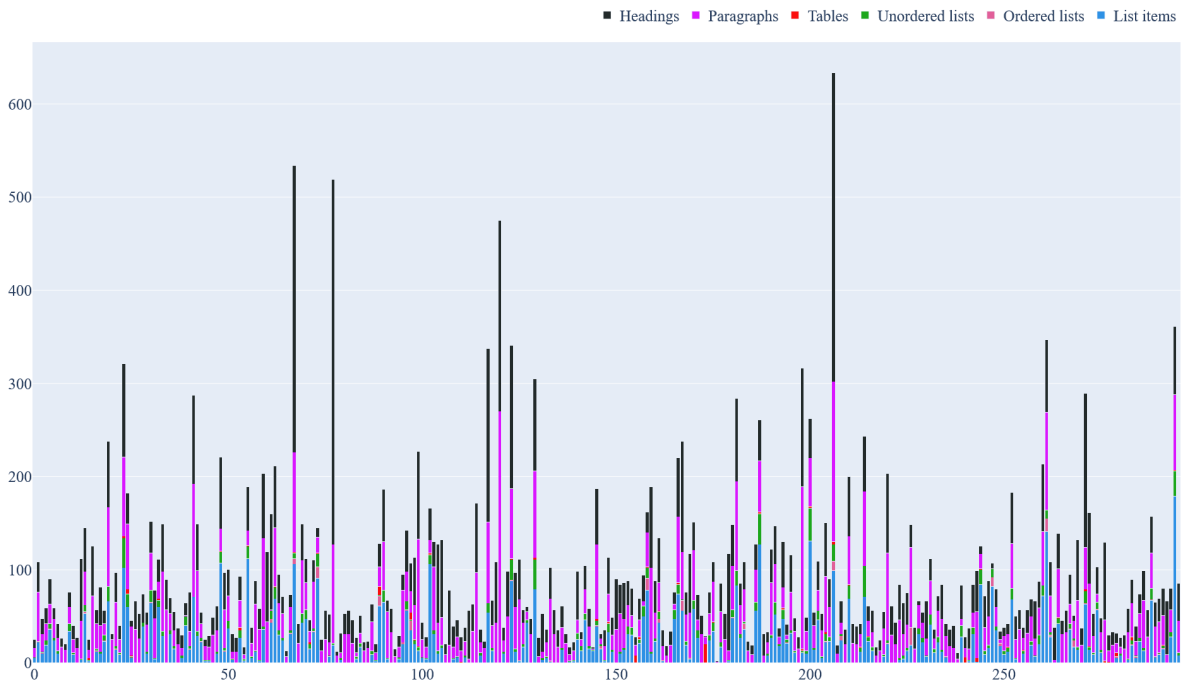

Рисунок 25 – Статистика последних 246 политик в IoT дата сете по структурным элементам

Таким образом можно описать среднестатистическую политику безопасности, которая состоит из 31.5 абзацев, 33 заголовков, 23.6 элементов перечислений, 0.7 нумерованных списков, 4.4 ненумерованных списка, 0.5 таблиц.

Для дополнительного статистического анализа дата сета, он был кластеризован с помощью латентного размещения Дирихле. Как и в [-] для кластеризации политики безопасности были разбиты на абзацы, после чего была проведена предобработка, состоящая из лемматизации и удаления пунктуации и так называемых «стоп слов». В таблице 14 приведены результаты моделирования тем в IoT дата сете. В [-] уже была исследована точность латентного размещения Дирихле, его преимущества и недостатки, на основании чего IoT дата сет был проанализирован именно таким способом. По ним видно, что с помощью такой кластеризации можно выделить различные аспекты политик безопасности.

Таблица 14 – Тематическое моделирование

№	Координаты семантическогопространства	Возможные сценариииспользования
0	email, send, promotional, communication, marketing, opt, product, service, message, list	First-party collection, Opt-in, opt-out messages and notifications to end user
1	party, third, service, information, privacy, website, share, policy, site, advertising	Third parties sharing for marketing purposes
2	removed, href, hyperref, question, contact, privacy, us, please, policy, comment	Contact information: company
3	cookie, device, browser, service, address, website, site, collect, information, use	First-party collection: browser and device information
4	child, age, entering, detection, year, fill, redirected, show, knowingly	Special audience: children
5	sensor, educational, suggestion, top, acquirer, mailing, employment, job, taking, clickstream	First-party collection: device and service specific information
6	corporate, automated, storefrontdigest, indefinite, personalization, direction, administrator, token, shop, employed	Other
7	data, personal, right, request, processing, information, necessary, legal, purpose, law	First-party collection: right to edit, access, with specified (legal) basis of data processing
8	sponsor, push, reply, default, swiss, desire, becoming, correspondence, calling, representative	Other
9	asset, service, product, merger, company, item, list, business, another, referral	Third-party sharing in case of company acquisution and merging
10	erasure, unaffiliated, input, approximate, format, appliance, pref, persistent, canadian, short	Right to erase
11	address, name, information, account, email, promotion, password, u, collect, contact	First-party collection: personal and account information
12	security, protect, safety, hosted, secure, violate, property, others, technical, law	Data security
13	california, state, resident, institution, law, united, ccpa, right, request, country	Special audience: California residents
14	change, policy, privacy, statement, time, notice, pci, payment, ds, update	Privacy policy changes

На рисунке 7 приведены результаты кластеризации дата сета. При кластеризации порог аффилиации абзаца политики безопасности был установлен в 0.3, параграф относился к нескольким кластерам, если вероятность аффилиации с ним была больше указанного порога. По графику на рисунке 26 можно судить, какую часть от общего объема текстов занимают те или иные аспекты политик безопасности.

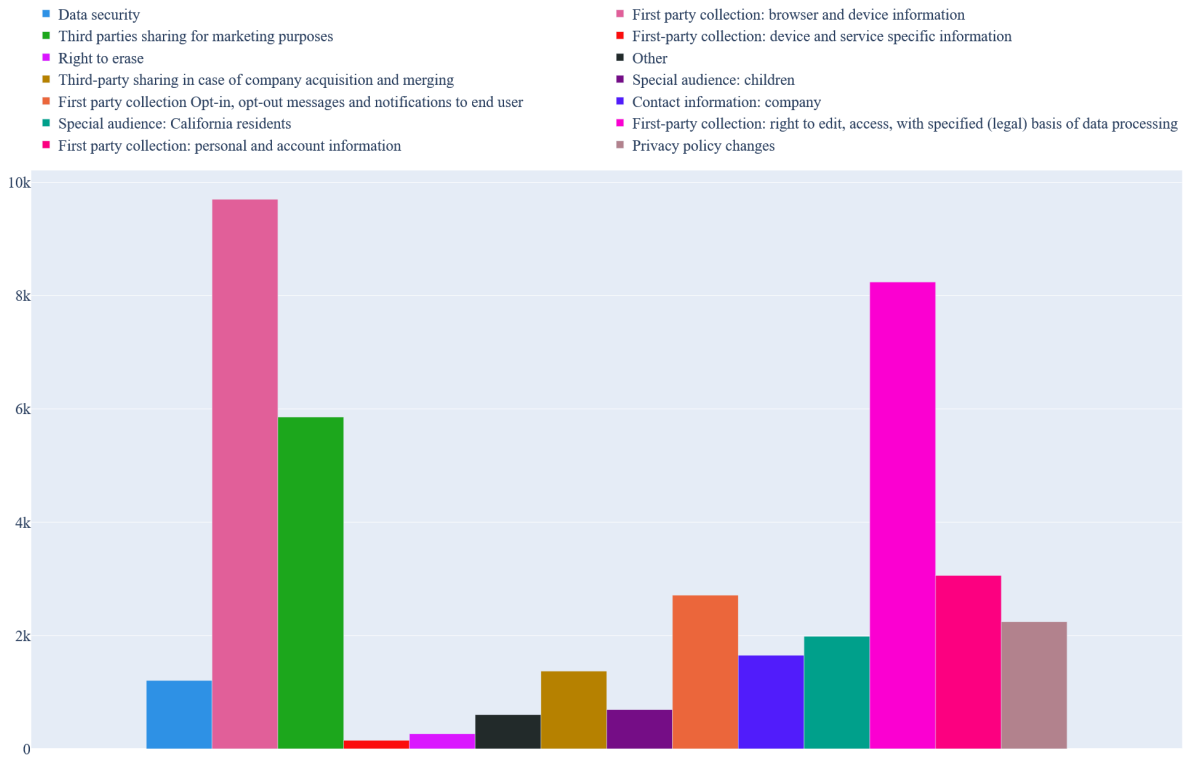

Рисунок 26 – Статистика аспектов в IoT дата сете

Как заключение статистического обзора сформированного дата сета на рисунке 27 и 28 приведено детальное распределение аспектов политик безопасности по каждой конкретной политике. Здесь в виде гистограммы представлены распределения всех 15 аспектов, выделенных алгоритмом LDA. Каждый абзац может относиться к нескольким аспектам с порогом аффилиации 0.3.

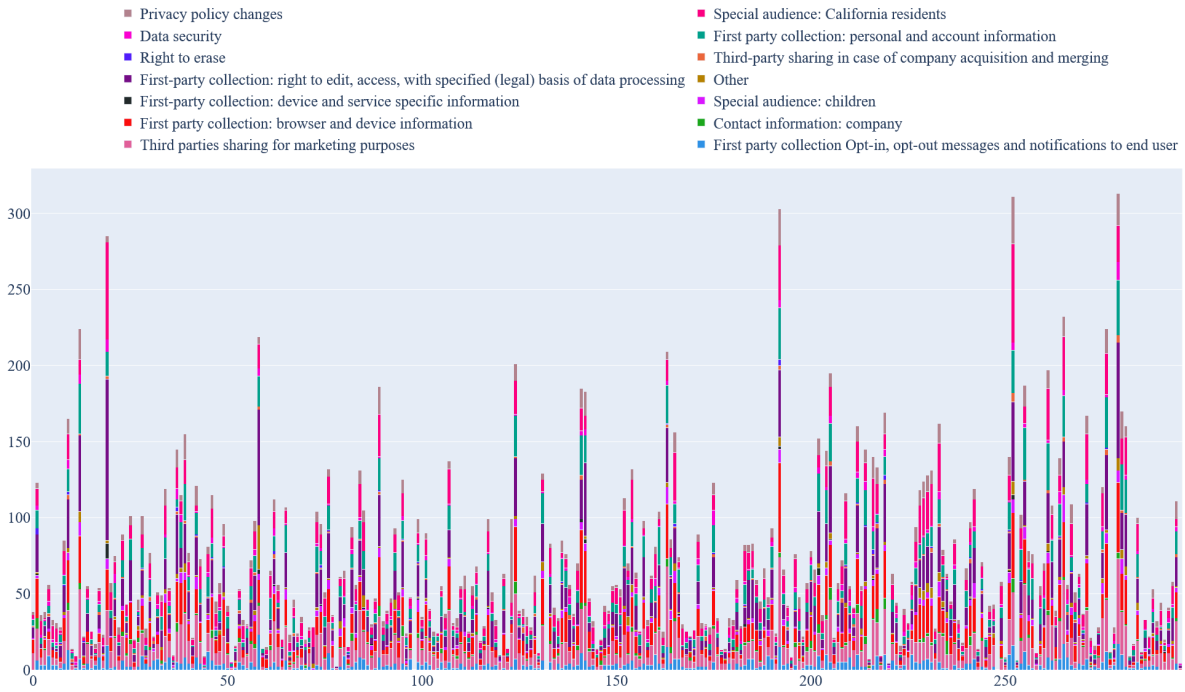

Рисунок 27 – Статистика первых 246 политик в IoT дата сете по аспектам

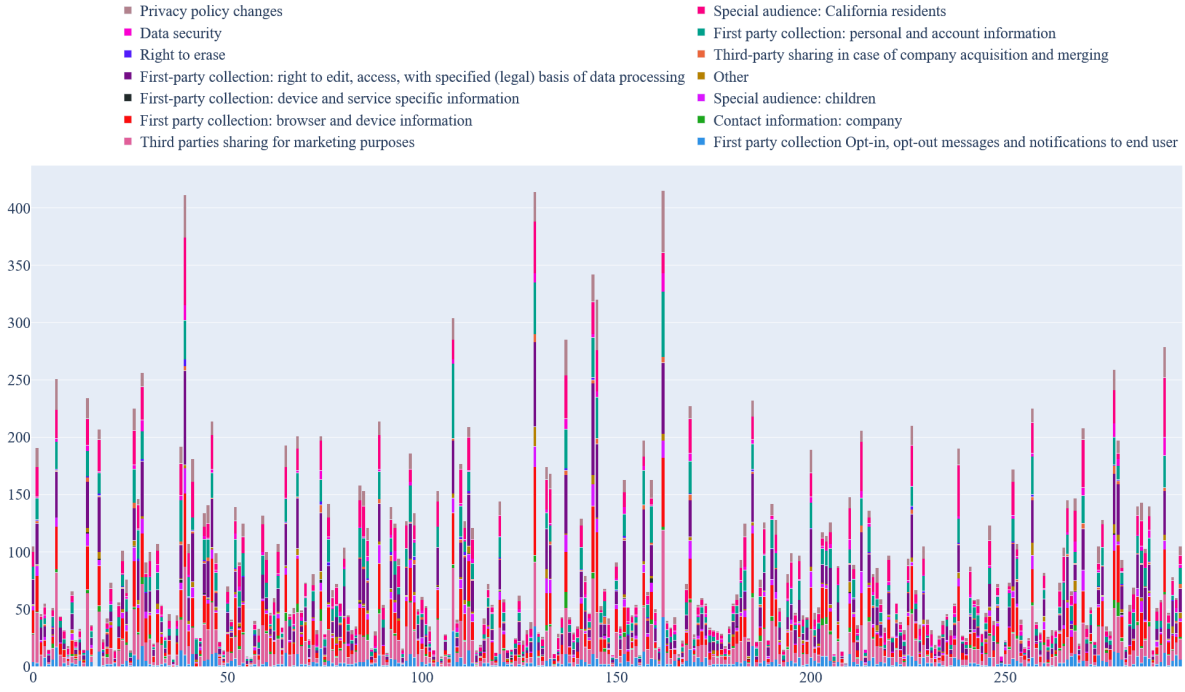

Рисунок 28 – Статистика последних 246 политик в IoT дата сете по аспектам

4.3 Полученный в результате реализации инструмент разметки

Текст…

4.4 Результаты решения поставленной задачи с помощью разработанного инструментария

Текст…

5 Составление бизнес-плана по коммерциализации результатов научно-исследовательской работы магистранта

Текст…

ЗАКЛЮЧЕНИЕ

Исходя из анализа методов формализации политик безопасности, было принято решение продолжать движение в сторону создания инструментов разметки датасетов, и моделей глубокого обучения. Таким образом было проведено первичное планирование процесса выполнения выпускной квалификационной работы магистра.

В результате выполнения работы было спроектировано и реализовано требуемое программное средство для сбора датасета, ориентированного на политики безопасности, и позволяющего создавать, обучающие выборки, ориентированные на формирование онтологического представления предметной области.

В ходе выпускной квалификационной работы были успешно проделаны следующие шаги:

– провести анализ предметной области;

– разработать методики сбора, очистки и разметки обучающей выборки;

– спроектировать инструментарий для построения обучающей выборки, обеспечивающей обучение классификатора с учетом онтологического представления предметной области;

– реализовать инструментарий для построения обучающей выборки, обеспечивающей обучение классификатора с учетом онтологического представления предметной области.

Все задачи, поставленные в выпускной квалификационной работе, были успешно выполнены. Файлы исходных кодов приложения приведены в приложениях Б и В. Электронная версия данной пояснительной записки к выпускной квалификационной работе представлена в приложении Г.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 General Data Protection Regulation, GDPR homepage. URL: https://gdpr .eu (дата обращения 14.02.2021).

2 Children's Online Privacy Protection Rule ("COPPA"). Available online: https://www.ftc.gov/enforcement/rules/rulemaking-regulatory-reform-proceedings/childrens-online-privacy-protection-rule (accessed on 30 March 2021).

3 Health Information Privacy. Available online: https://www.hhs.gov/hipaa/index.html (accessed on 30 March 2021).

4 Novikova, E., Doynikova, E., Kotenko, I.: P2Onto: Making Privacy Policies Transparent. In Proceedings of The 3rd Interna-tional Workshop on Attacks and Defenses for Internet-of-Things (ADIoT 2020), In Conjunction with ESORICS 2020. 4-6 No-vember 2020, Paris, France. Computer Security, Lecture Notes in Computer Science (LNCS), Springer, 2020; vol. 12501; pp. 235-252. DOI: https://doi.org/10.1007/978-3-030-64330-0_15

5 Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J.R., Russell N.C., Sadeh, N.: MAPS: Scaling Pri-vacy Compliance Analysis to a Million Apps. In: Proceedings on Privacy Enhancing Technologies, 66, 2019, https://ir.lawnet.fordham.edu/faculty_scholarship/1040.

6 Oltramari, A., Piraviperumal, D., Schaub, F., Wilson, S., Cherivirala, S., Norton, T., Russell, N., Story, P., Reidenberg, J., Sadeh, N.: PrivOnto: A semantic framework for the analysis of privacy policies. Semantic Web, 9(2), 2018; pp. 185-203.

7 Palmirani, M., Martoni, M., Rossi, A., Bartolini, C., Robaldo, L.: Legal ontology for modelling GDPR concepts and norms. Legal Knowledge and Information Systems. IOS Press, 2018. doi: https://doi.org/10.3233/978-1-61499-935-5-91.

8 Pandit, H. J., O'Sullivan D., Lewis D. An Ontology Design Pattern for Describing Personal Data in Privacy Policies. WOP@ISWC, 2018.

9 Kumar V.B., Iyengar R., Nisal N., Feng Y., Habib H., Story P., Cherivirala S., Hagan M., Cranor L., Wilson C., Schaub F., and Sadeh N.: Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In: Pro-ceedings of The Web Conference 2020 (WWW '20), pp. 1943-1954. New York, NY, USA, Association for Computing Machin-ery, 2020.

10 Sathyendra, K. M., Schaub, F., Wilson, S., Sadeh., N.: Automatic extraction of opt-out choices from privacy policies. In Proc. AAAI Symposium on Privacy-Enhancing Technologies, AAAI Fall Symposium - Technical Report, 2016.

11 Ashley, P., Hada, S., Karjoth, G., Schunter, M.: E-p3p privacy policies and privacy authorization. In: Proc. of the ACM work-shop on Privacy in the Electronic Society (WPES 2002), Washington, DC, USA, 2002.

12 Karjoth, G., Schunter, M.: Privacy policy model for enterprises. In: Proc. of the 15th IEEE Computer Security Foundations Workshop, Cape Breton, Nova Scotia, Canada, 2002.

13 Ardagna, C.A., De Capitani di Vimercati, S., Samarati, P.: Enhancing User Privacy Through Data Handling Policies. In: Damiani E., Liu P. (eds) Data and Applications Security XX. DBSec 2006. Lecture Notes in Computer Science, vol. 4127. Springer, Berlin, Heidelberg, 2006.

14 Pardo, R., Le Métayer, D.: Analysis of Privacy Policies to Enhance Informed Consent. In: Foley S. (eds) Data and Applica-tions Security and Privacy XXXIII. DBSec 2019. Lecture Notes in Computer Science, vol. 11559. Springer, Cham, 2019.

15 Gerl, A., Bennani, N., Kosch, H., Brunie, L.: LPL, Towards a GDPR-Compliant Privacy Language: Formal Definition and Us-age. Trans. Large-Scale Data- and Knowledge-Centered Systems 2018, 37, pp. 41-80.

16 NIST Privacy Risk Assessment Methodology (PRAM). Available online: https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/resources (accessed on 30 March 2021).

17 De, S.J., Le Metayer, D.: Privacy Risk Analysis to Enable Informed Privacy Settings. In: 2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), London, pp. 95-102, 2018.

18 The Usable Privacy Policy Project. Available online: https://usableprivacy.org/ (accessed on 30 March 2021).

19 IoT Security Compliance Framework. Available online: https://www.iotsecurityfoundation.org/best-practice-guidelines/ (accessed on 30 March 2021).

20 GSMA IoT Security Guidelines and Assessment. Available online: gsma.com/iot/iot-security/iot-security-guidelines/ (ac-cessed on 30 March 2021).

21 Data privacy vocabularies and controls community group. Available online: https://www.w3.org/community/dpvcg/ (ac-cessed on 30 March 2021).

22 PROV_O: The PROV Ontology Homapage. Available online: https://www.w3.org/TR/prov-o/#Agent (accessed on 30 March 2021).

23 The Usable Privacy Policy Project Data and Tools. Available online: https://usableprivacy.org/data (accessed on 30 March 2021).

24 California Consumer Privacy Act 2018. Available online: https://oag.ca.gov/privacy/ccpa (accessed on 30 March 2021).

25 August Device and Service Privacy Policy Homepage. Available online: https://august.com/pages/privacy-policy#product (accessed on 30 March 2021).

26 Electronic Passes in Moscow during lockdown. URL: https://www.cnews.ru/news/top/2020-05-25_moskovskij_sajt_s_propuskami (дата обращения 14.02.2021).

27 Harshvardhan J. Pandit, Declan O'Sullivan, and Dave Lewis. Personalised Privacy Policies. 2018.

28 Hamza Harkous, Kassem Fawaz, Remi Lebret, Florian Schaub3, Kang G. Shin, and Karl Aberer. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. 2018. arXiv:1802.02561v2.

29 Evgenia Novikova, Elena Doynikova, and Igor Kotenko. P2Onto: Making Privacy Policies Transparent. Springer, 2020.

30 Landauer, T. K., Foltz, P. W., and Laham, D. An Introduction to Latent Semantic Analysis. Discourse Processes, 25, 1998, pp. 259-284. DOI: https://doi.org/10.1080/01638539809545028.

31 Gensim topic modeling library, Gensim homepage. URL: https://radimrehurek.com/gensim (дата обращения 14.02.2021).

32 Sachini Weerawardhana, Subhojeet Mukherjee, Indrajit Ray, and Adele Howe. Automated Extraction of Vulnerability Information for Home Computer Security, pages 356-366. Springer, 2015. DOI: https://doi.org/10.1007/978-3-319-17040-4_24.

33 Natural Language ToolKit, Analyzing Sentence Structure, NLTK homepage. URL: https://www.nltk.org/book/ch08.html (дата обращения 14.02.2021).

ПРИЛОЖЕНИЕ А

А.1 Алгоритм

```
coeff = getSumPdWeight(PD_class)
if coeff < 1:
    coeff = 10 + coeff
RiskScoreBase = max_criticality * (1 + lg(coeff))
```

А.2 Алгоритм

```
if (LB_weight + P_weight) < 1:
    coeff = 1
else:
    coeff = 1 + lg(LB_weight + P_weight)
FP_RiskScore = RiskScoreBase * coeff
if FP_RiskScore > 10:
    FP_RiskScore = RiskScoreBase + (10 - RiskScoreBase) * \
        (coeff / (1 + lg(6))
Return FP_RiskScore
```

А.3 Алгоритм

```
UsageScenarioRiskScore = RiskScoreBase * coeff,

where UsageScenarioRiskScore – privacy risk score for the data usage scenario;
       RiskScoreBase – the base of privacy risk score;
       coeff – coefficient that increase or decrease the risk depending on the
           aspects specified in the privacy policy.

The 'algorithms pseudocode is provided below.

rC = getRoots(C) //C is the set of ontology classes
for rci from rC:  //rC is the set of root ontology classes corresponding
                                                    to the usage scenarios
    coeff = 0    //coefficient that depends on the aspects specified
                                                    in the privacy policy
    rcc = getChilds(rci)       //get childs of the selected root class
    for rcck from rcc:
        scc = getSubclasses(rcck)  //get subclasses of the selected
```

```
                                                                class
        for sccr from scc:
            catr = getCategory(sccr)  //determine category of the
                                                                subclass
            add catr to cat  //forming the set of categories for the
                                                                subclass
        maxCritCat = getMaxCritCat(cat)  //define category with max
                                                            criticality
        classWeight = getClassWeight(maxCritCat)  //define weight of
                                        the category with max criticality
        if rcck is PersonalData:
            pd_coeff = getSumPdWeight(cat)  //calculate sum of weights
                                            for personal data subclasses
            if pd_coeff < 1:
                pd_coeff = 10 + pd_coeff
            RiskScoreBase = maxCritCat(1 + lg(pd_coeff))  //calculate
                                        the base of privacy risk score
        else:
            coeff += classWeight  //calculate sum of max weights for
                        other categories subclasses (not personal data)
        if coeff < 1:
            risk_coeff = 1
        else:
            risk_coeff = 1 + lg(coeff)
    UsageScenarioRiskScore = RiskScoreBase * risk_coeff
    if UsageScenarioRiskScore > 10:        //scaling to 0 to 10 scale
        UsageScenarioRiskScore = RiskScoreBase + (10 - RiskScoreBase) *
            risk_coeff / (1 + lg(coeff))
add UsageScenarioRiskScore to RiskScores  //set of privacy risk scores
                                                        for usage scenarios
return RiskScores
```

ПРИЛОЖЕНИЕ Б

Архив с исходными кодами вэб-скрейпера.

ПРИЛОЖЕНИЕ В

Архив с исходными кодами инструмента для разметки датасета.

ПРИЛОЖЕНИЕ Г

Электронная версия пояснительной записки.