«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 _____ Кузнецов М.Д.
подпись

Руководитель к.т.н., доцент _____ Новикова Е.С.
(Уч. степень, уч. звание) *подпись*

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 _____ Кузнецов М.Д.

подпись

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

(Уч. степень, уч. звание) *подпись*

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*

РЕФЕРАТ

Поясн. зап. 58 стр., 19 рис., 8 табл., 9 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 Обзор текущего состояния предметной области

Текст…

1.2 Методика формализации политик безопасности с применением онтологического представления предметной области

Текст…

1.3 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 1. Здесь слова каждого абзаца подсчитываются и сопоставляются с абзацами, в которых они встретились.

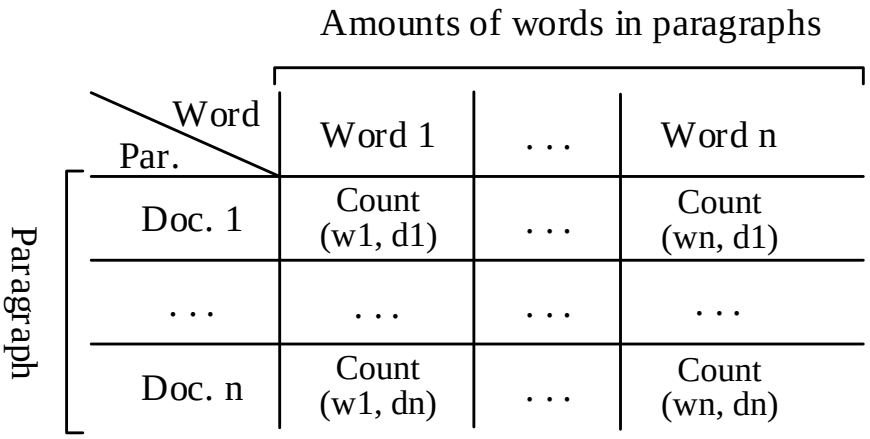

Рисунок 1 – Bag-of-Words матрица

Модель TF-IDF представляет документ в виде матрицы, представленной на рисунке 2. Формула (1) показывает, как можно получить метрику TF-IDF.

$$tfidf(t,d,D) = \frac{n_t}{\sum\limits_k n_k} \times log \frac{\left|D\right|}{\left|\left\{d_i \in D : t \in d_i\right\}\right|}, \qquad (1)$$

где t – термин или слово;

d – конкретный абзац;

D – набор абзацев.

Итак, модель TF-IDF придает больший вес словам которые использованы меньше раз. Это может быть полезно, когда тексты схожи с точки зрения используемых слов, как в нашем случае, для политик безопасности.

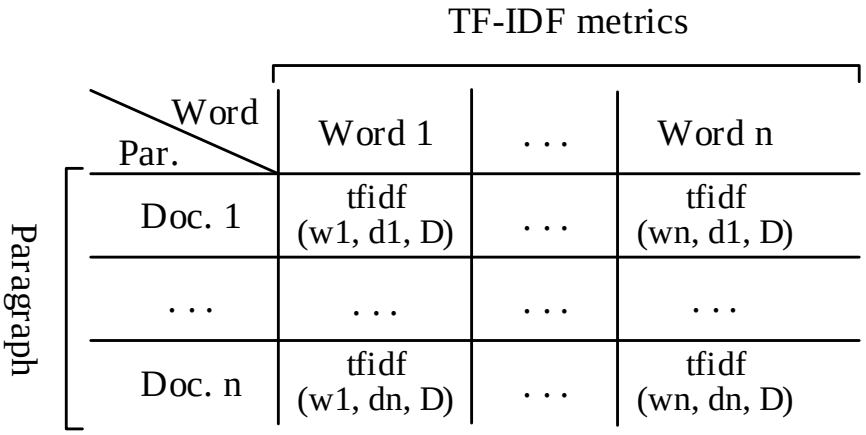

Рисунок 2 – Матрица TF-IDF

1.4 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 согла-

сия 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 (2)

$$A = U \times S \times V^T, \tag{2}$$

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 [6]. 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 [3]. Результаты экспериментов для модели «мешок слов» представлены в таблице 1, в ней показаны полученные кластеры и соответствующие значения координат.

Таблица 1 – Кластеры политик безопасности для модели Bag-of-Words

№	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 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 [7] и результаты снова не были обнадеживающими. Результаты расчета метрики принадлежности кластеру представлены в таблице 2.

Таблица 2 – 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 3.

Таблица 3 – 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 4. Также показывались десять кластеров и значения атрибутов. И, как в первом случае с «мешком слов», по значениям координат невозможно судить о теме кластера.

Таблица 4 – Кластеры политик безопасности для модели Bag-of-Words

№	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 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”

Продолжение таблицы 4

№	Coordinate 1	Coordinate 2	Coordinate 3	Coordinate 4
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 5.

Таблица 5 – 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 7.

Таблица 6 – 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.

1.5 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 конфи-

денциальности.

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 лицам, 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 7 – Тематическое моделирование

№	Координаты семантического пространства	Возможные сценарии использования
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)

Продолжение таблицы 7

№	Координаты семантического пространства	Возможные сценарии использования
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

Продолжение таблицы 7

№	Координаты семантического пространства	Возможные сценарии использования
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. Вто-

рая 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 изменили 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.

1.6 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 [8]. Имея размеченное по частям речи предложение, парсер грамматики NLTK [9] строит деревья предложений по правилам грамматики. Одно из таких деревьев в обозначениях NLTK можно увидеть на рисунке 3 [9], 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.

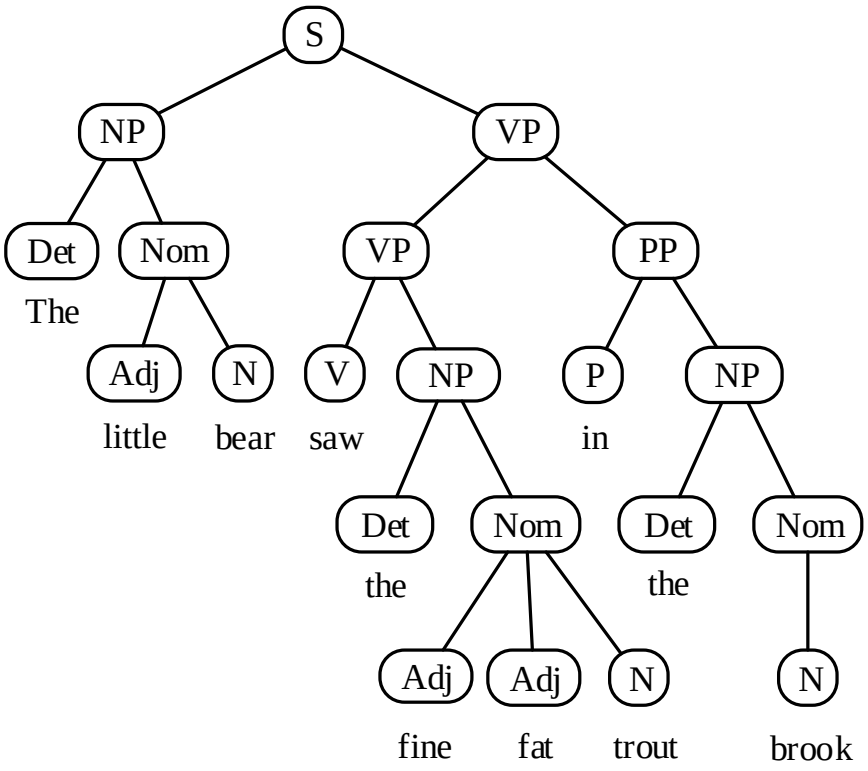

Рисунок 3 – Пример грамматического разбора

В предлагаемом подходе немного другая грамматическая запись. Созданная грамматика представлена в (3).

$$\begin{cases} D \to S \mid S\ D \mid S\ \ U\ D \\ S \to NPG\ \ VBG \\ VPG \to VP \mid VP\ \ VPG \mid VP\ \ U\ \ VPG \\ NPG \to NP \mid NP\ \ NPG \mid NP\ \ U\ \ NPG \\ AJPG \to AJ \mid AJ\ \ APG \mid AJ\ \ U\ \ APG \\ AVPG \to AV \mid AV\ \ APG \mid AV\ \ U\ \ APG \\ VP \to VAPG \mid V\ \ PPG \mid V\ \ PP\ \ APG \\ NP \to NOM \mid DET\ \ NOM \\ NOM \to N \mid AJPG\ \ N \\ PP \to NPG \mid P\ \ NPG \end{cases}, \tag{3}$$

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 (3) 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 (4).

$$\begin{cases} U \to NLTK\_CC \\ DET \to NLTK\_DT \\ AJ \to NLTK\_JJ \\ AV \to NLTK\_RB \\ N \to \_\_CN\_\_ \mid \_\_FP\_A\_\_ \mid \_\_TP\_A\_\_ \mid NLTK\_N \\ V \to \_\_CV\_\_ \mid NLTK\_V \end{cases}, \tag{4}$$

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 (3) и (4) и разметки лейблами пулов было построено дерево предложения, результат на рисунке 4.

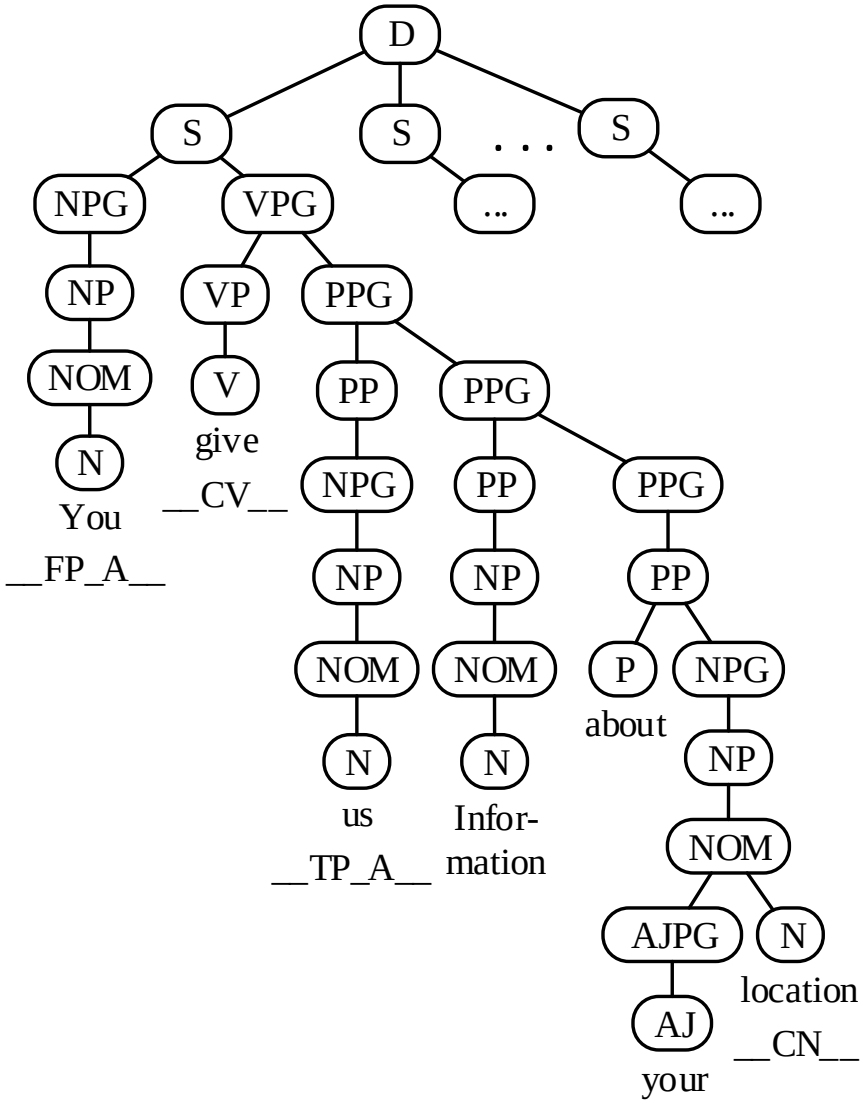

Рисунок 4 – Дерево грамматического разбора

Когда было построено дерево предложений последовательность меток ключевых слов может быть распознана. В этом случае представленная на рисунке 4, 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.

1.7 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.

1.8 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.

2 Проектирование инструментария

2.1 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, предназначенных 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.

2.2 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 всплывающие 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.

2.3 Методика разметки

Текст…

2.4 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», ино-

гда все же попадается шум.

2.5 Техническое задание «Инструментарий для сбора датасета»

2.5.1 Основные положения технического задания

2.5.2 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.

2.5.3 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.

2.5.4 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.

2.5.5 Фреймворк глубокого обучения

Для создания и тренировки модели анализа текста планируется использовать фреймворк машинного обучения «Keras». Он позволяет быстро созда-

вать 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.

2.6 Приложение вэб-скрейпер

2.6.1 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 такой 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.

2.6.2 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 5.

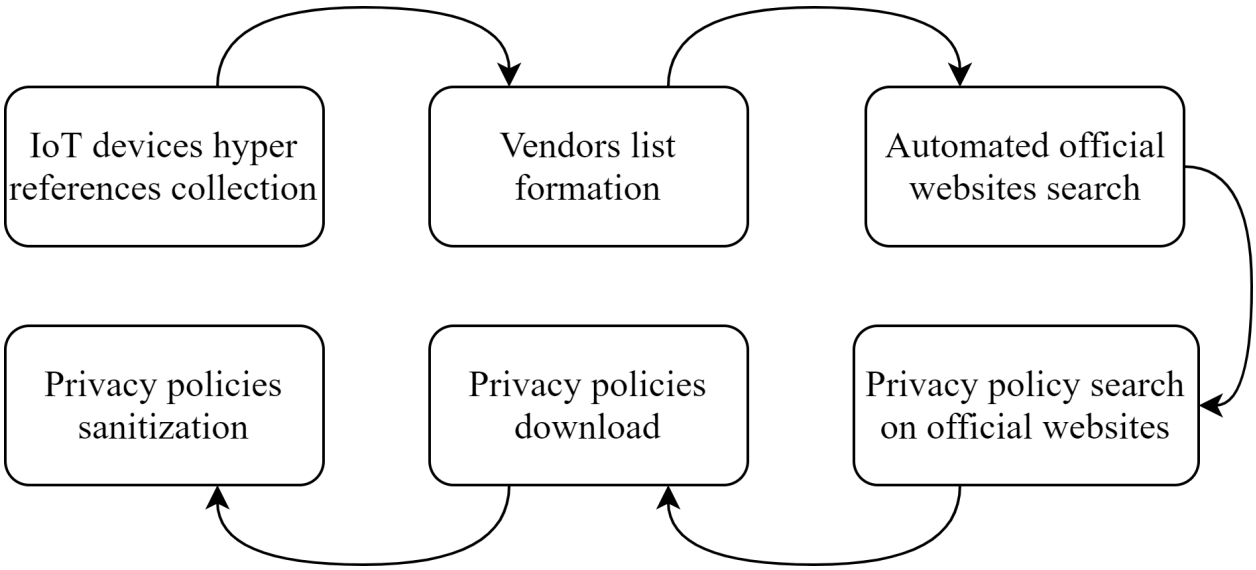

Рисунок 5 – Схема организации приложения

Далее была разработана композиционная модель приложения, на ней присутствуют все необходимые для решения задач модули. Схема представлена на рисунке 6.

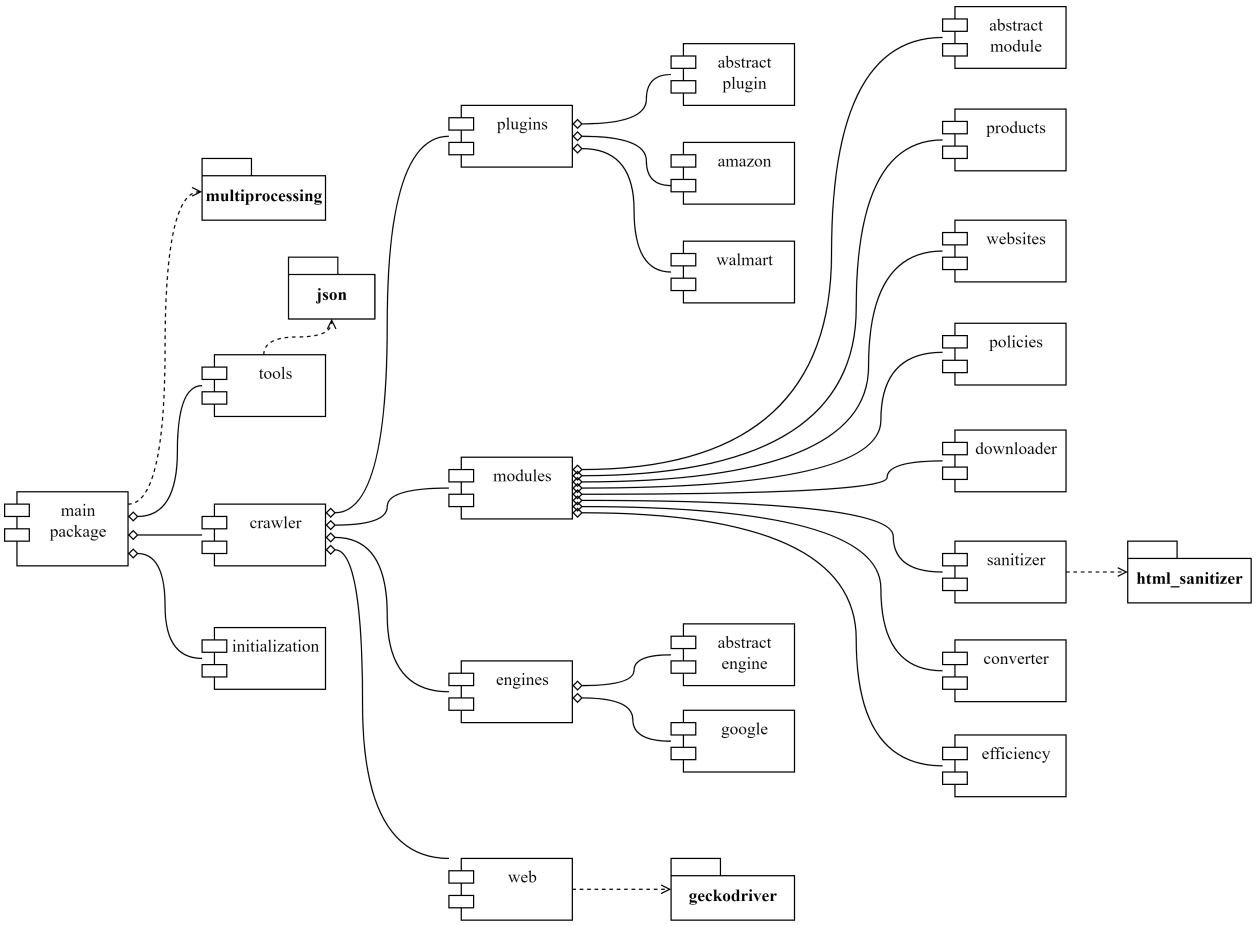

Рисунок 6 – Композиционная модель приложения

2.6.3 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 7.

```
{
  "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 7 – 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.

2.7 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.

2.7.1 Объектное моделирование приложения

Объектная модель инструмента представлена на рисунке 8.

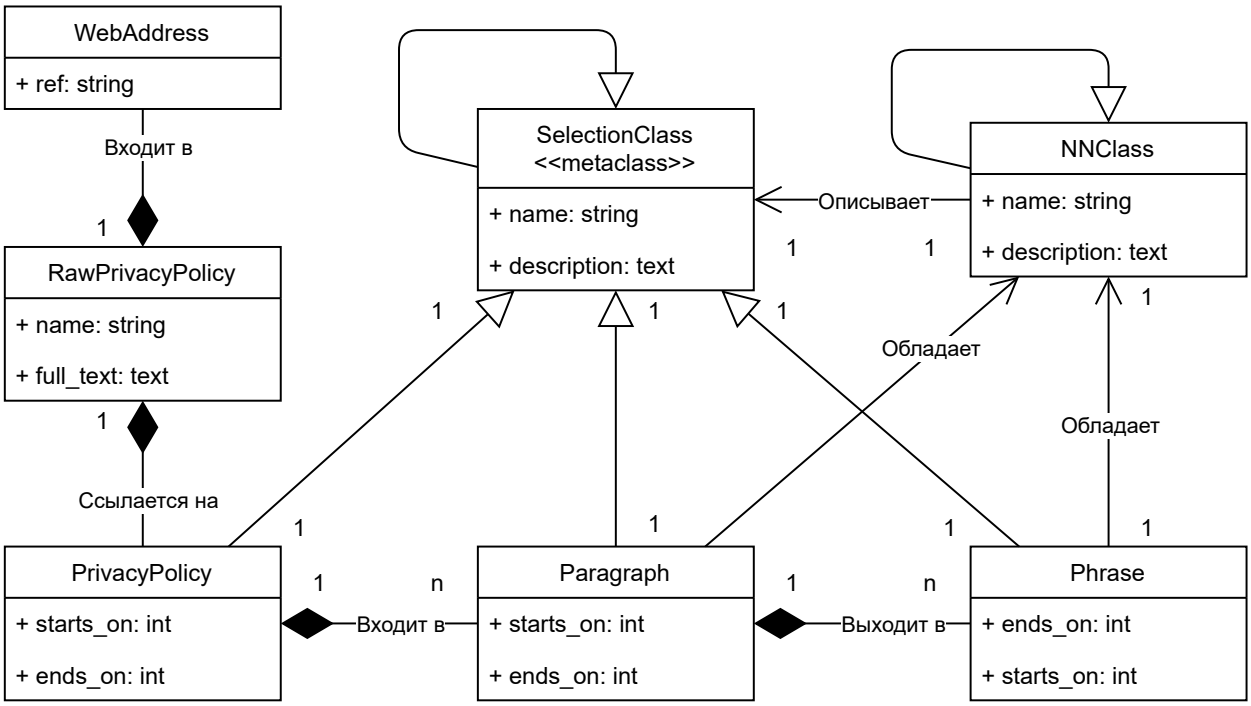

Рисунок 8 – Объектная модель

2.7.2 Реляционная модель приложения

Реляционная модель инструмента представлена на рисунке 9.

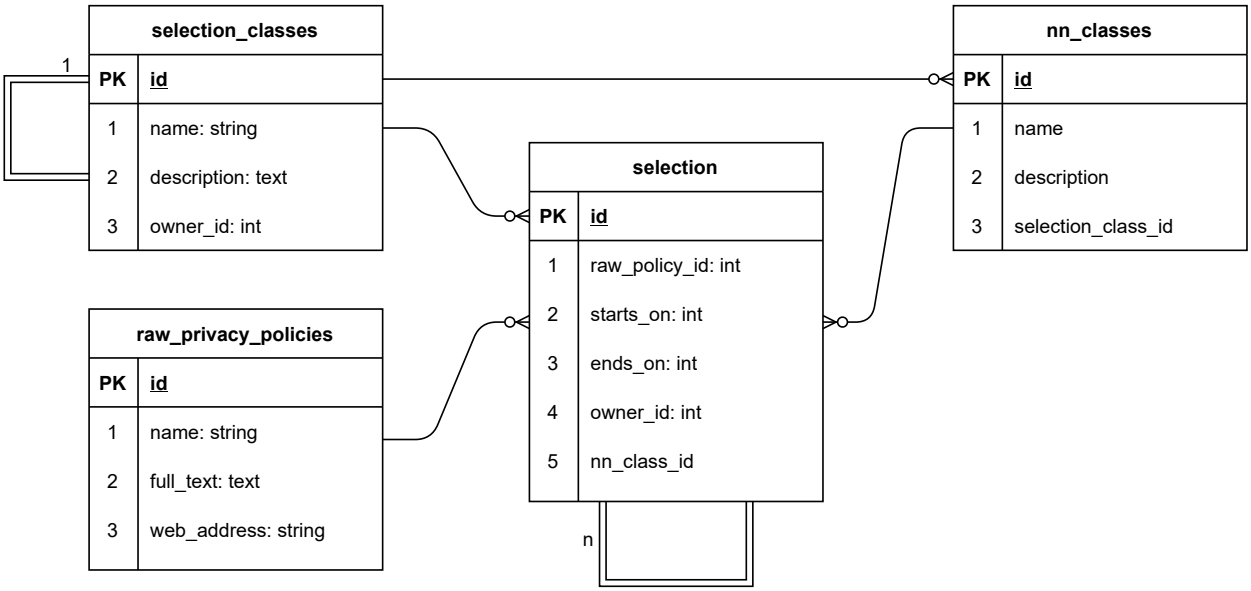

Рисунок 9 – Реляционная модель

2.7.3 Проектирование пользовательского интерфейса

Презентационный прототип интерфейса инструмента представлен на рисунке 10.

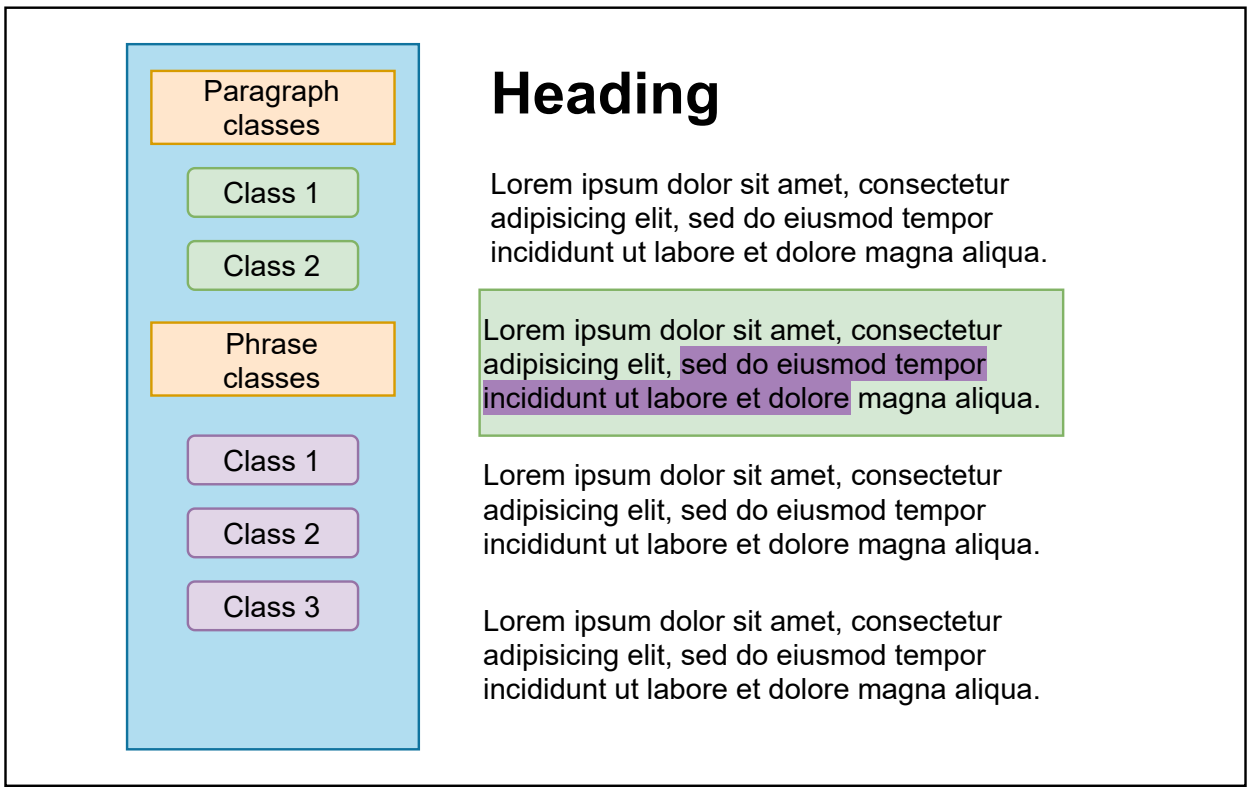

Рисунок 10 – Презентационный прототип интерфейса

2.7.4 Средства разработки инструмента разметки

3 Реализация инструментария

3.1 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 приложении А.

3.2 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 11 и 12. Таким образом полноценных уникальных политик безопасности осталось 592.

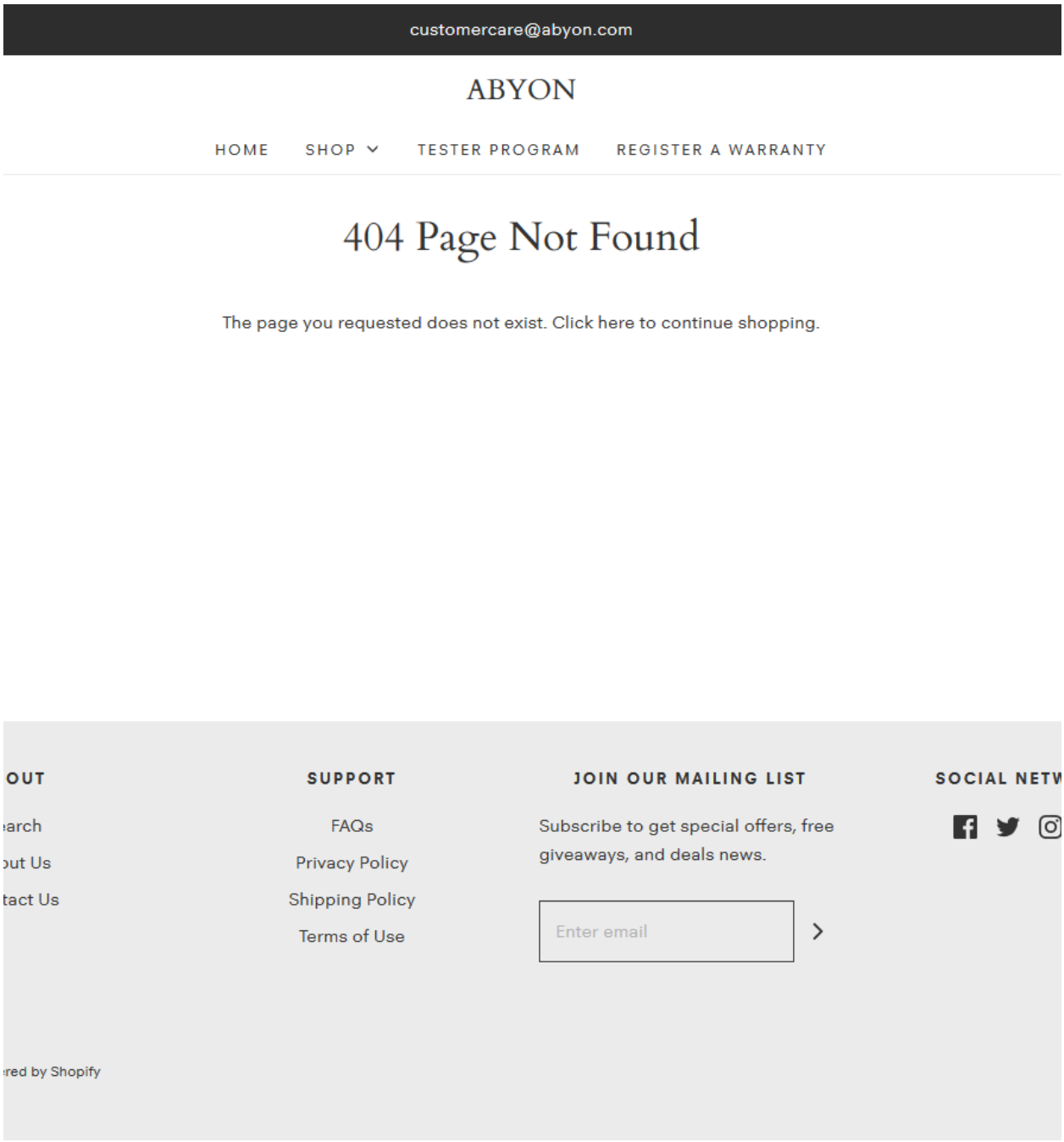

Рисунок 11 – Пример отсутствующей политики

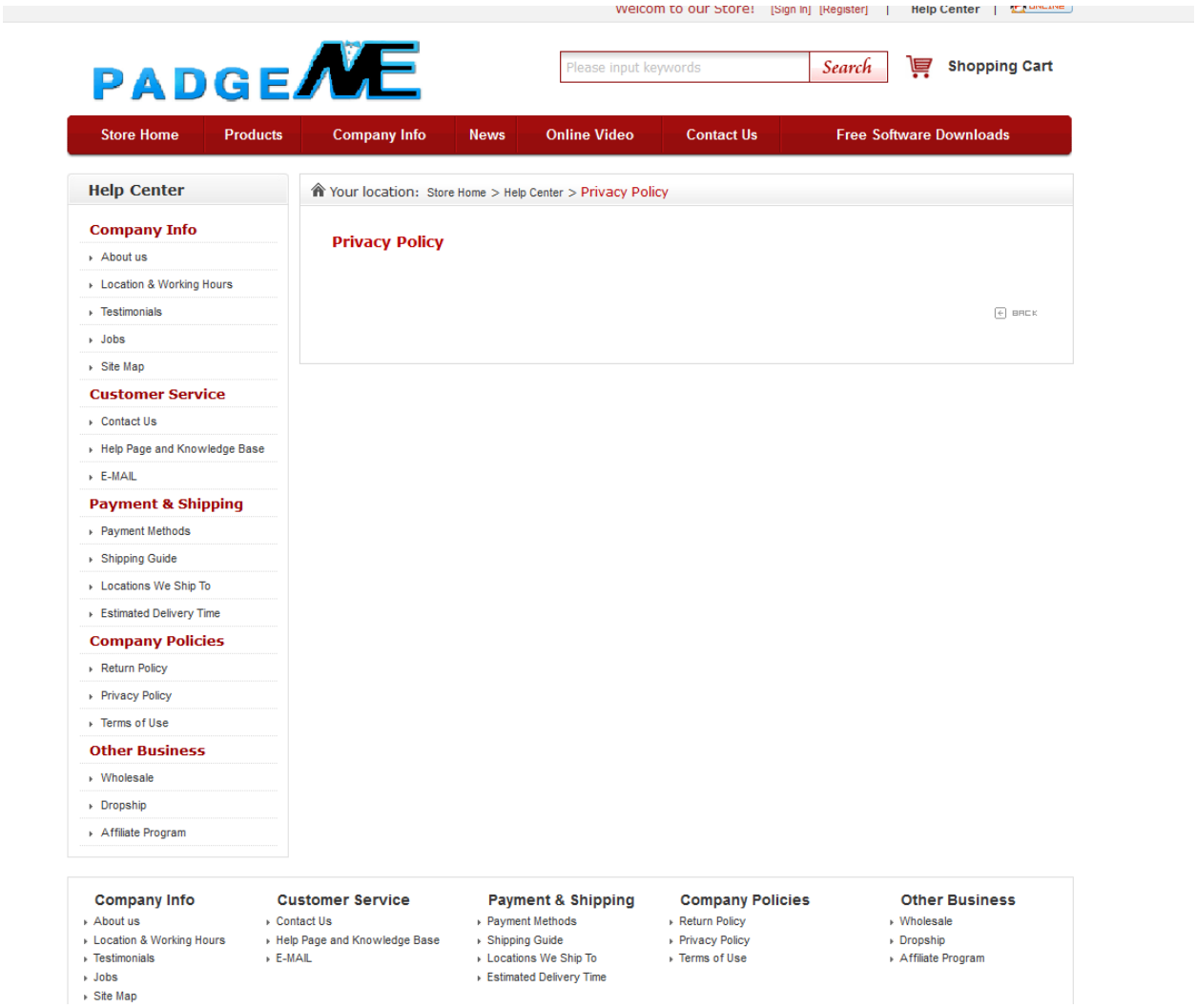

Рисунок 12 – Пример отсутствующей политики

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 13 и 14 приведены статистические данные по объемам абзацев политик и самих докумнетов соответственно.

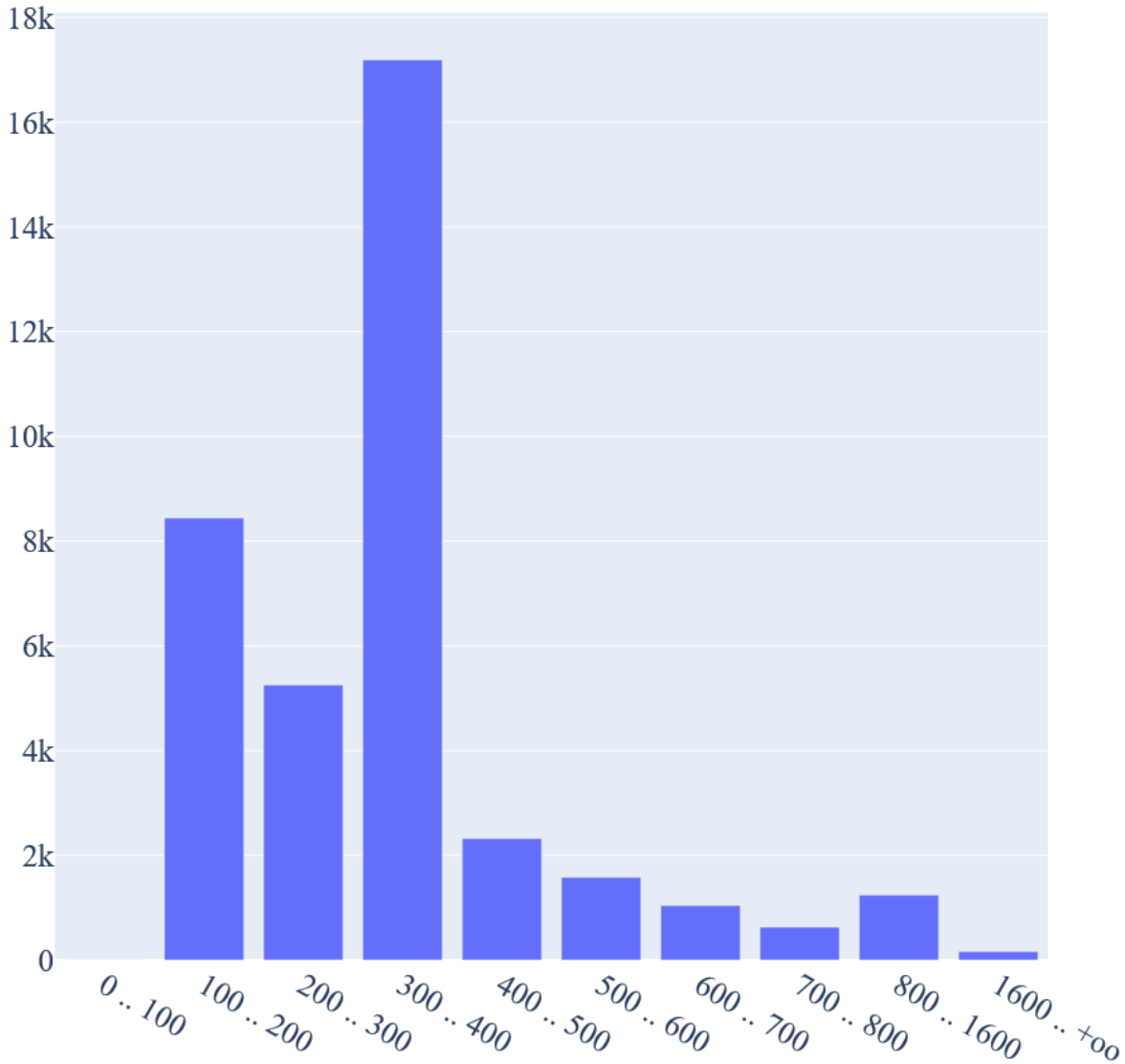

Рисунок 13 – Распределение политик по объему параграфа

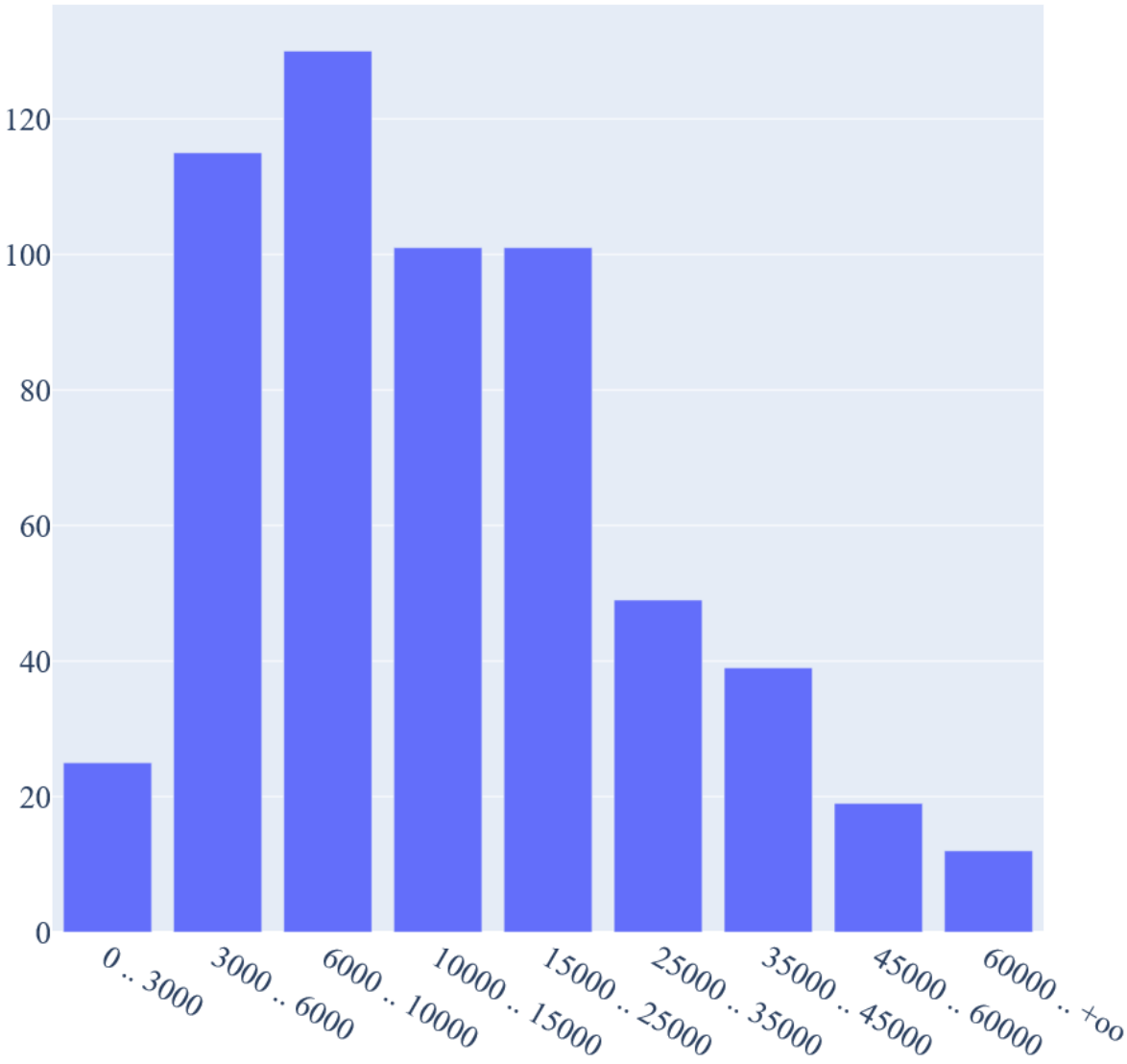

Рисунок 14 – Распределение политик по объему документа

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 15 и 16 приведена статистика по структурным элементам политик безопасности в двух частях. Здесь изображены детальные распределения структурных элементов для каждой из найденных политик безопасности.

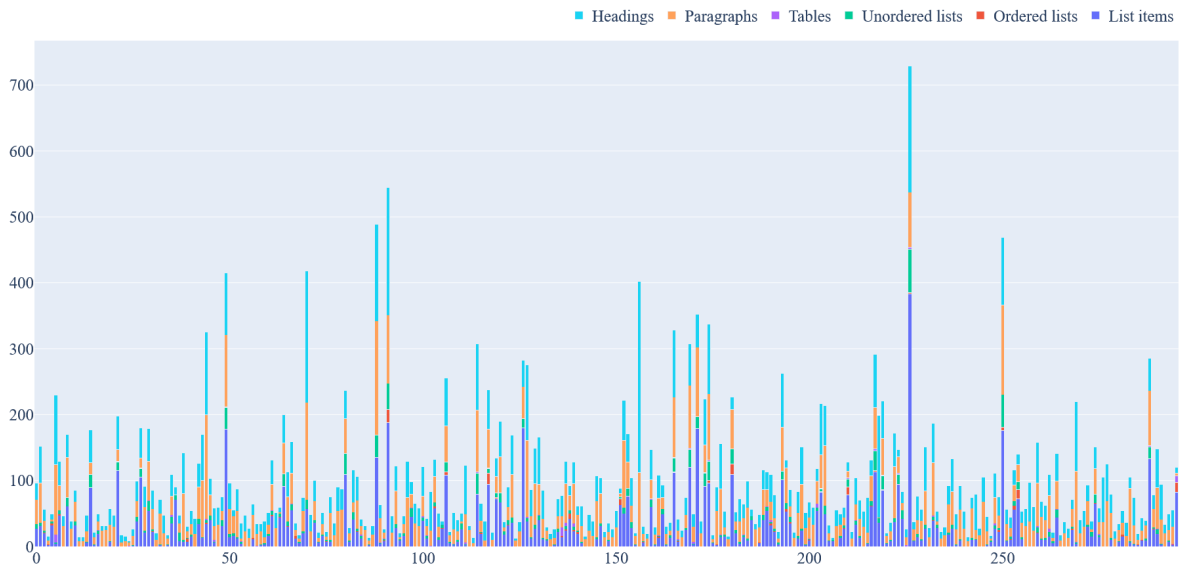

Рисунок 15 – Статистика первых 246 политик в IoT дата сете по структурным элементам

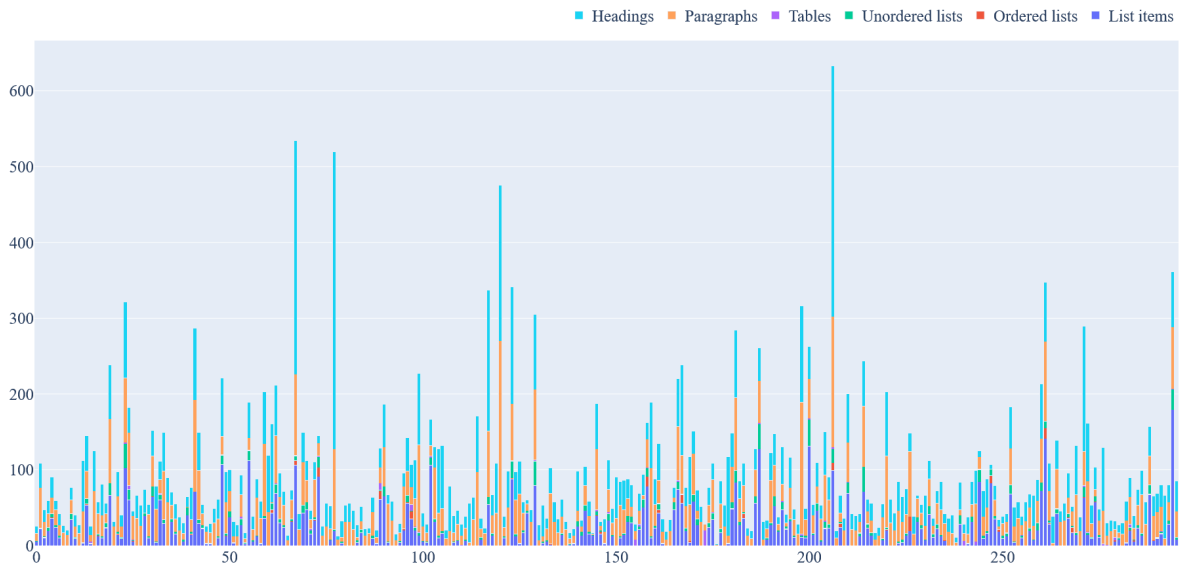

Рисунок 16 – Статистика последних 246 политик в IoT дата сете по структурным элементам

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 8 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 8 – Тематическое моделирование

№	Координаты семантического пространства	Возможные сценарии использования
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

Продолжение таблицы 8

№	Координаты семантического пространства	Возможные сценарии использования
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 17 можно судить, какую часть от общего объема текстов занимают те или иные аспекты политик безопасности.

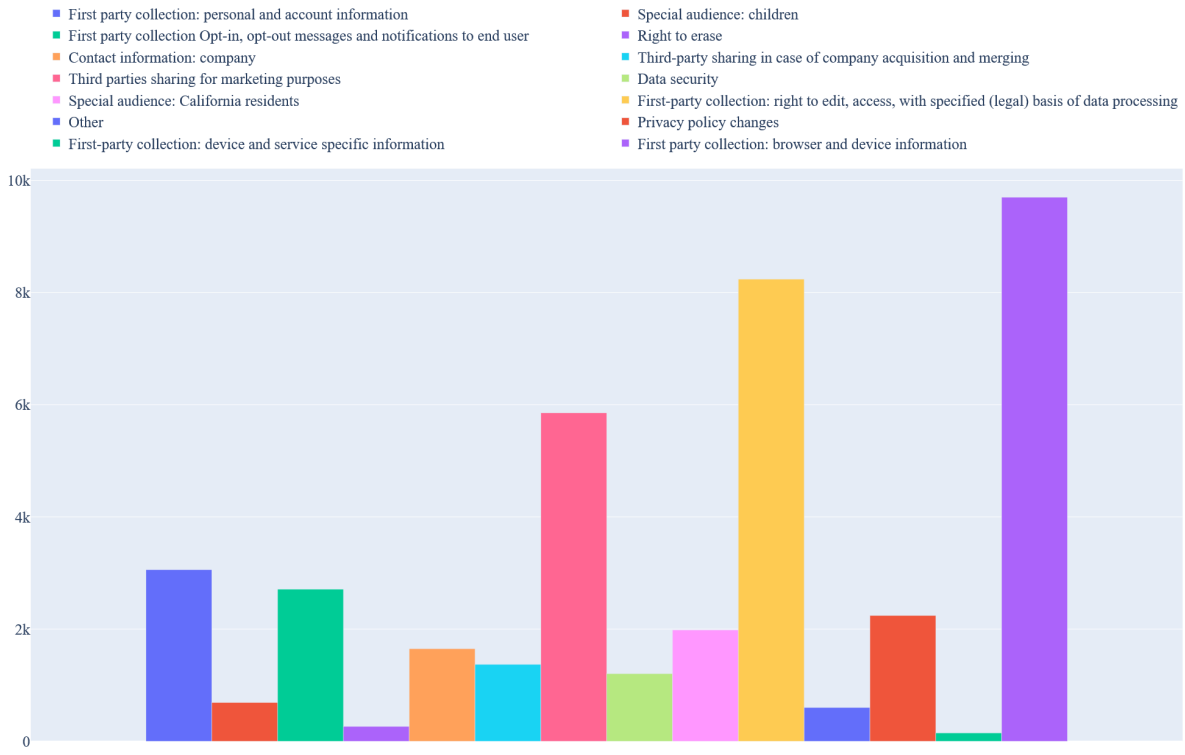

Рисунок 17 – Статистика аспектов в IoT дата сете

Как заключение статистического обзора сформированного дата сета на рисунке 18 и 19 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.

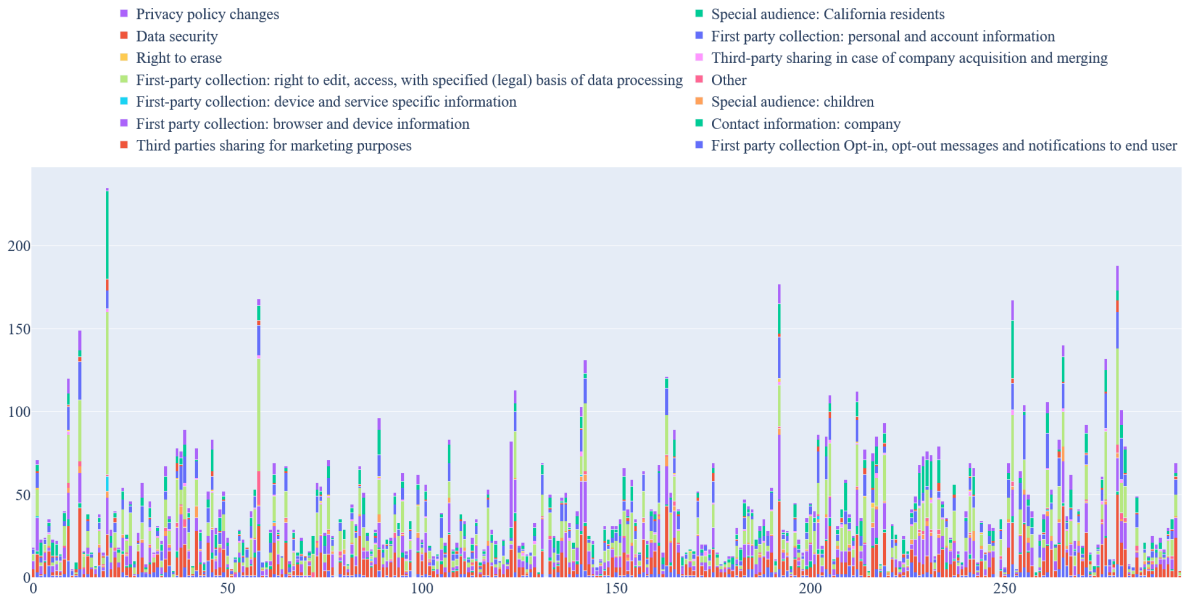

Рисунок 18 – Статистика первых 246 политик в IoT дата сете по аспектам

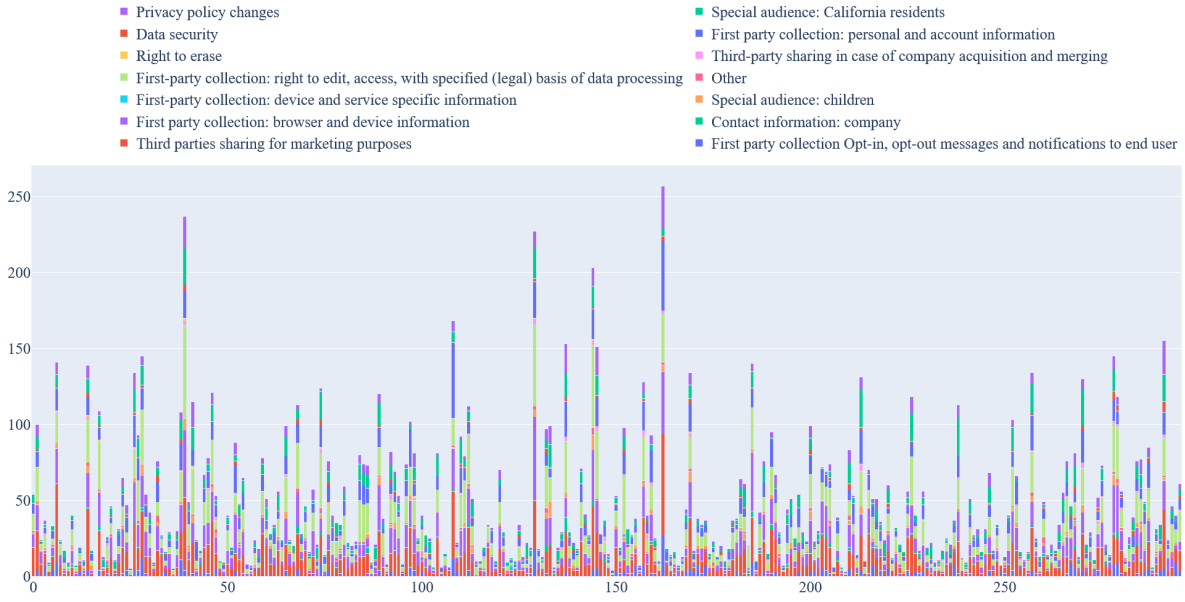

Рисунок 19 – Статистика последних 246 политик в IoT дата сете по аспектам

3.3 Полученный в результате реализации пользовательский интерфейс инструмента разметки

Текст…

3.4 Результаты решения поставленной задачи с помощью разработанного инструментария

В результате работы программы были найдены 65 политик безопасности, разумеется среди них имеется определенный процент промахов, если производитель имеет сходство с каким либо другим более крупным. Поиск осуществлялся по торговой площадке amazon, брались результаты поискового запроса по первым 10-ти страницам, по категориям «smart scales», «smart watches», «smart locks» и «smart bulbs». Всего производителей было найдено приблизительно 160. Стоит отметить, что результат является приемлемым, так как многие производители на данной торговой площадке не имеют выделенного веб-сайта, 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.

4 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 приложении А. Электронная версия данной пояснительной записки к выпускной квалификационной работе представлена в приложении А.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

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

2 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).

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

4 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.

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

6 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.

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

8 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.

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

ПРИЛОЖЕНИЕ А

Архив с исходными кодами вэб-скрейпера и инструмента для разметки датасета.