TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
Отделение школы (НОЦ) – Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Интеллектуальный анализ текстовых данных с применением методов машинного обучения

УДК: 004.912:004.422.6:004.85

Студент

Группа	ФИО	Подпись	Дата
8ПМ7И	Кульневич Алексей Дмитриевич		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Губин Е.И.	к.ф.-м.н.		

КОНСУЛЬТАНТЫ ПО РАЗДЕЛАМ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
старший преподаватель	Потехина Н.В.			

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Горбенко М.В.	к.т.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Губин Е.И.	к.ф.-м.н.		

Томск – 2019 г.

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОБУЧЕНИЯ ПО ООП

Код результата	Результат обучения
Общие по направлению подготовки 09.04.04 «Программная инженерия»	
Р1	Способность проводить научные исследования, связанные с объектами профессиональной деятельности
Р2	Способность разрабатывать новые и улучшать существующие методы и алгоритмы обработки данных в информационно-вычислительных системах
Р3	Способность составлять отчеты о проведенной научно-исследовательской работе и публиковать научные результаты
Р4	Способность проектировать системы с параллельной обработкой данных и высокопроизводительные системы
Р5	Способность осуществлять программную реализацию информационно-вычислительных систем, в том числе распределенных
Р6	Способность осуществлять программную реализацию систем с параллельной обработкой данных и высокопроизводительных систем
Р7	Способность организовывать промышленное тестирование создаваемого программного обеспечения
Профиль «Технологии больших данных»/ «Big data solutions»	
Р8	Способность исследовать и анализировать большие данные, создавать их модели и интерпретировать структуры данных в таких моделях
Р9	Способность понимать принципы создания, хранения, управления, передачи и анализа больших данных с использованием новейших технологий, инструментов и систем обработки данных в высокопроизводительных сетях
Р10	Способность применять теорию распределенной системы управления базами данных к традиционным распределенным системам реляционных баз данных, облачным базам данных, крупномасштабным системам машинного обучения и хранилищам данных

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
Отделение школы (НОЦ) – Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП

_____ _____ _____
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

Магистерской диссертации

Студенту:

Группа	ФИО
8ПМ7И	Кульневич Алексей Дмитриевич

Тема работы:

Подготовка исходных данных для построения кредитного скоринга	
Утверждена приказом директора	№1436/с от 25.02.2019

Срок сдачи студентом выполненной работы:	06.06.2019

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе	Данные из открытых источников: – Cyberleninka; – Статьи из Академии Google; – Новостные статьи Reuters.
Перечень подлежащих исследованию, проектированию и разработке вопросов *(аналитический обзор по литературным источникам с целью выяснения достижений мировой науки техники в рассматриваемой области; постановка задачи исследования, проектирования, конструирования; содержание процедуры исследования, проектирования, конструирования; обсуждение результатов выполненной работы; наименование дополнительных разделов, подлежащих разработке; заключение по работе).*	1. Аналитический обзор подходов к решению задач: – Извлечение именованных сущностей; – Разрешение кореференции; – Извлечение ключевых слов; – Автореферирование; – Тематическое моделирование. 2. Реализация методов решения данных задач

	3. Реализация системы анализа документов, включающую в себя: – Backend часть; – Frontend часть; – База данных (elasticsearch); – Обработка данных в фоновом режиме. 4. Финансовый менеджмент; 5. Социальная ответственность; 6. Application of Russian named entity recognition and coreference resolution in the oil industry 7. Заключение.
Перечень графического материала	Рисунки и таблицы для описания исследования предметной области, реализованного сервиса, составления финансовой стороны и уточнения социальной ответственности.

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Социальная ответственность	**Горбенко Михаил Владимирович, доцент ООД ШБИП, к.т.н.**
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	**Потехина Нина Васильевна, старший преподаватель ОСГН**
Обязательное приложение на английском языке	**Диденко Анастасия Владимировна, доцент ОИЯ ШБИП, к.ф.н.**

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

1 Аналитический раздел

2 Сервис

3 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение

4 Социальная ответственность

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Кульневич Алексей Дмитриевич		

TOMSK POLYTECHNIC UNIVERSITY ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

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

Инженерная школа информационных технологий и робототехники
Направление подготовки – 09.04.04 Программная инженерия
Уровень образования магистратура
Отделение школы (НОЦ) – Отделение информационных технологий
Период выполнения: весенний семестр 2018 /2019 учебного года

Форма представления работы:

Магистерская диссертация

КАЛЕНДАРНЫЙ РЕЙТИНГ-ПЛАН
выполнения выпускной квалификационной работы

Срок сдачи студентом выполненной работы:	06.06.2019

Дата контроля	Название раздела (модуля) / вид работы (исследования)	Максимальный балл раздела (модуля)
09.02.2019	*1 Аналитический модуль / Описание используемых подходов и моделирование системы*	*30*
18.04.2019	*2 Сервис / Описание требований к сервису, обзор технического решения и реализованного веб-сервиса*	*30*
11.05.2019	*3 Финансовый менеджмент, ресурсоэффективность и ресурсосбережение*	*20*
27.05.2019	*4 Социальная ответственность*	*20*

СОСТАВИЛ:
Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		

СОГЛАСОВАНО:

Руководитель ООП

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент ОИТ ИШИТР	Губин Е.И.	к.ф.-м.н.		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ПМ7И	Кульневич Алексей Дмитриевич

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	Магистратура	Направление/специальность	Программная инженерия

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:	
1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*- Оклад инженера – 21760;* *- Оклад руководителя от организации – 35000;* *- Оклад научного руководителя – 33664;*
2. *Нормы и нормативы расходования ресурсов*	*- Годовая норма амортизации составляет 33.3 %.*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*- Размер страховых взносов равный 30%;* *- Районный коэффициент г. Томск.*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*- Анализ потенциальных потребителей;* *- Диаграмма Исикавы;* *- Анализ конкурентных технических решений.*
2. *Разработка устава научно-технического проекта*	*Постановка целей проекта, определение ожидаемого результата и критериев приемки проекта.*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*- Построение диаграммы Ганта;* *- Планирование работ;* *- Формирование бюджета затрат;* *- Анализ рисков проекта.*
4. *Определение ресурсной, финансовой, экономической эффективности*	*Расчет финансового показателя эффективности. Описание общих выводов.*
Перечень графического материала ***(с точным указанием обязательных чертежей):***	
1. *Сегментация рынка* 2. *Результаты оценки конкурентных систем анализа документов* 3. *Диаграмма Исикавы* 4. *Календарный график проекта* 5. *Бюджет затрат* 6. *Реестр рисков*	

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Ст. преподаватель ОСГН ШБИП	Потехина Н.В.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Кульневич Алексей Дмитриевич		

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа	ФИО
8ПМ7И	Кульневич Алексей Дмитриевич

Школа		Отделение школы (НОЦ)	
Уровень образования	Магистр	Направление/специальность	Программная инженерия

Исходные данные к разделу «Социальная ответственность»:	
1. Описание рабочего места (рабочей зоны, технологического процесса, механического оборудования)	*В соответствии с ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования».*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности.	*- ГОСТ 12.2.032-78 ССБТ - СанПиН 2.2.4.548–96 - СанПиН 2.2.4/2.1.8.562–96 - СанПиН 2.2.2/2.4.1340–03 - ГОСТ 12.1.009-2009 - ГОСТ 12.1.038-82 ССБТ - ГОСТ Р 22.3.03-94*
2. Анализ выявленных вредных факторов проектируемой производственной среды	*- Освещение - Микроклимат - Шум - Психофизиологические факторы: нервно-психические перегрузки*
3. Анализ выявленных опасных факторов проектируемой произведённой среды.	*- Электрический ток (источник – ПК) - Короткое замыкание - Статическое заземление (источник – ПК)*
4. Охрана окружающей среды.	*Воздействие объекта на атмосферу, гидросферу отсутствует. Воздействие на литосферу происходит при утилизации ПК, используемого для разработки, а также утилизации люминесцентных ламп освещения.*
5. Защита в чрезвычайных ситуациях.	*Возможной чрезвычайной ситуацией при разработке алгоритма является возникновение пожара на рабочем месте.*

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Горбенко М. В.			

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ПМ7И	Кульневич Алексей Дмитриевич		

РЕФЕРАТ

Выпускная квалификационная работа 114 с., 34 рис., 22 табл., 26 источников.

Ключевые слова: анализ данных, обработка естественного языка, извлечение именованных сущностей, разрешение кореференции, извлечение ключевых слов, автореферирование текста, тематическое моделирование.

Объект исследования – работа с текстовой информацией, в частности, в нефтегазовой сфере.

Предмет исследования – использование методов интеллектуального анализа при анализе документов.

Цель работы – исследование методов обработки естественного языка и реализация система анализа текстов данных с применением методов машинного обучения.

В процессе исследования проводилось изучение, анализ, тестирование и сравнение различных существующих методов анализа и обработки естественного языка реализация алгоритмов извлечения именованных сущностей, разрешения кореференции, извлечения ключевых слов, автореферирования текста и тематического моделирования.

В результате исследования была реализована система, состоящая из основных компонентов: backend, frontend, data processing worker частей.

Для backend части был использован фреймворк Flask, для аналитики – фреймворки Keras и PyTorch, frontend части был использован Vue.Js фреймворк

Область применения: разработанная система может быть использована в административных и научно-технических организациях, где документооборот (поиск, анализ, работа с документами) требует большого количества человеко-часов.

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ, СОКРАЩЕНИЯ

Обработка естественных языков – направление искусственного языка и математической лингвистики. В рамках направления изучаются проблемы компьютерного анализа и синтеза естественных языков, также означает понимание языка, а синтез означает генерацию грамотного текста.

LSTM-сети – разновидность архитектуры рекуррентных нейронных сетей. В отличие от традиционных сетей, LSTM-сеть хорошо приспособлена к обучению на задачах классификации, обработки и прогнозирования временных рядов, когда важные события разделены временными лагами с неопределенной продолжительностью и границами.

Извлечение именованных сущностей – задача, состоящая в распознавании в тексте именованных сущностей (которыми являются слова и словосочетания) и их классификация по предопределенным категориям, например, личности (person), организации (organization), продукты (product) и другие. Даная задача является подзадачей общей задачи извлечения информации, которая состоит в извлечении структурированных данных из источников неструктурированной или слабоструктурированной информации.

Машинное обучение – семейство алгоритмов, способных обучаться. Различают два основных типа: обучение по прецедентам, которое основано на выявлении общих закономерностей и дедуктивное обучение, позволяющее формализовать знания экспертов в виде базы знаний.

Глубинное обучение – совокупность методов машинного обучения, основанных на обучении представлениям, а не специализированным алгоритмам под конкретные задачи.

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Основным направлением данной работы является исследование современных методов анализа текстовой информации.

Обработка естественных языков включает в себя различные техники интерпретации человеческого языка, начиная от статистических подходов, машинного обучения до подходов, основанных на лингвистических правилах или классических алгоритмических подходов.

В области обработки естественных языков существует определенное количество библиотек, таких как:

– Spacy; [1]
– NLTK; [2]
– StanfordNLP; [3]
– PyMorphy. [4]

Каждая из библиотек имеет различную встроенную логику, позволяющую «из коробки» использовать различные инструменты, как например, извлечение именованных сущностей, получение морфологических признаков слов и т. д.

Большинство инструментов для обработки естественных языков реализованы именно для английского языка, что является не удобным при необходимости работы с другими языками. Стоит отметить, что существуют библиотеки для русского языка, такие как:

– Zamgi; [5]
– Abby Compreno. [6]

Zamgi является одной из первых систем для русского языка, позволяющей автоматически извлекать именованные сущности. В ней имеются как свои достоинства, так и недостатки. Данная система имеет обобщающий подход, то есть основана на стилистике, а не на правилах. Но имеются ограничения – невозможно извлечь сущности, которые начинаются не с заглавной буквы.

Что касается Abby Compreno – это коммерческий продукт для ограниченного числа пользователей, проверка качества и функционала которого

достаточна затруднительна в силу данных причин. Основной интерес в данной работе представляет два аспекта при работе с текстами:

– анализ текста;

– поиск по текстам.

В рамках работы над анализом текста были проанализированы и реализованы для русского языка следующие задачи:

– извлечение именованных сущностей;

– извлечение кореференции;

– извлечение ключевых слов;

– автореферирование текста;

– тематическое моделирование текстов.

Каждая из задач будет подробно разобрана в дальнейшем. Конечным результатом данной работы является веб-сервис, который также будет в дальнейшем полностью описан.

1 АНАЛИТИЧЕСКИЙ МОДУЛЬ

Основную часть разрабатываемого сервиса представляет аналитика, которая представлена рядом алгоритмов, реализующих предобработку текста, методы анализа последовательностей, извлечения ключевой информации и тематическое моделирование.

Предобработка текста включает в себя:

– лемматизацию;

– удаление стоп-слов;

– представление слов текста в векторном формате;

– обогащение дополнительными признаками (морфология, тэги).

Методы анализа последовательностей включают в себя:

– распознавание именованных сущностей;

– разрешение кореференции.

Методы извлечения ключевой информации включают в себя:

– извлечение ключевых фраз (в том числе слов);

– автореферирование текста.

Далее, в данном разделе будут подробно разобрано подходы к решению каждой из задач.

1.1 Распознавание именованных сущностей

Распознавание именованных сущностей — это процесс, когда алгоритм принимает на вход строку (предложение, параграф или корпус текстов) и определяет релевантные объекты (такие, например, как люди, местоположения или организации) в данных. Задача распознавания именованных сущностей является подзадачей извлечения информации. [7]

В общем случае задача относится к разметке последовательностей (sequence labeling), где задачей является пометка категориальной меткой членов последовательности. Помимо распознавания именованных сущностей, также существует частеречная разметка и задача разрешения кореференции.

Пример распознавания именованных сущностей изображен на рисунке 1.

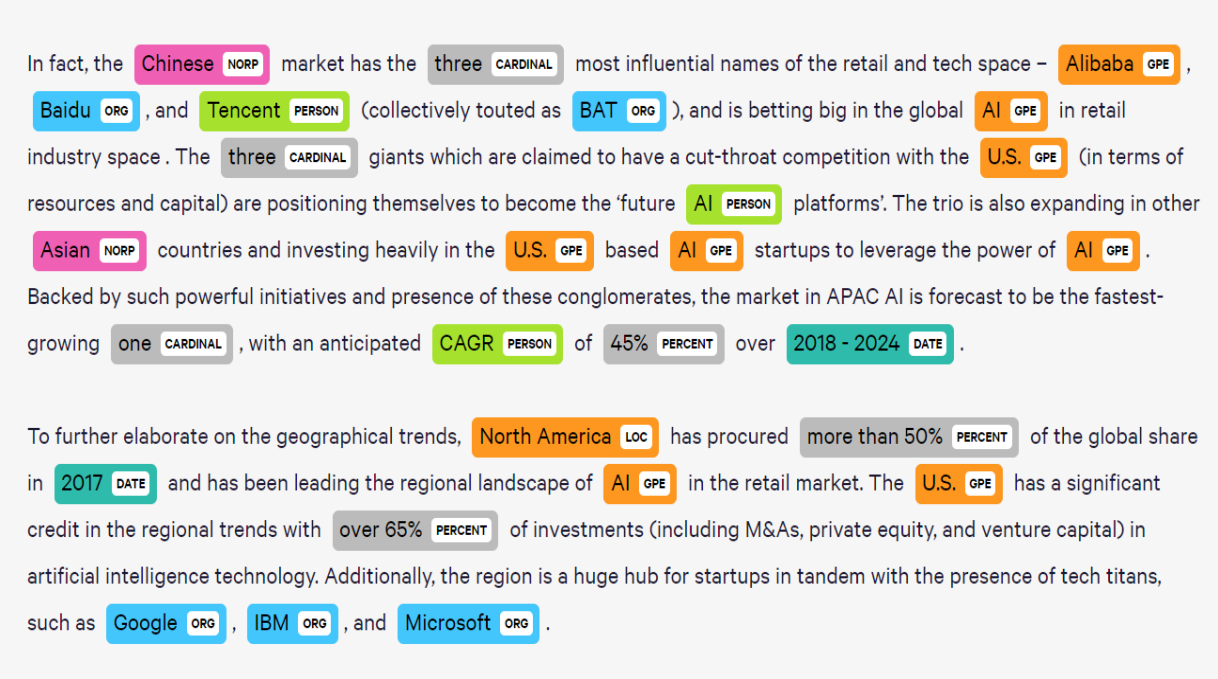

Рисунок 1 – Распознавание именованных сущностей

Существует несколько подходов к решению данной задачи:

– статистические (пример с использованием Conditional Random Fields);

– основанные на лингвистических правилах (пример Natasha) [8];

– основанные на машинном обучении (двунаправленные нейронные сети долгосрочной памяти);

– для проверки качества реализованной системы распознавания именованных сущностей существуют следующие метрики:

1) Precision — это качество предсказанных сущностей, покрываемых в соответствии с золотым стандартом (например, разметка эксперта-лингвиста);

2) Recall — это степень покрытия алгоритма в соответствии с золотым стандартом;

3) F1-Score — это гармоническое среднее precision и recall.

Золотой стандарт (ground truth) относится к тренировочным данным для задач классификации в задачах обучения с учителем. Используется в статистических моделях (а также моделях машинного обучения) для доказательства или опровержения исследовательской гипотезы.

F1 Score представлен на рисунке 2.

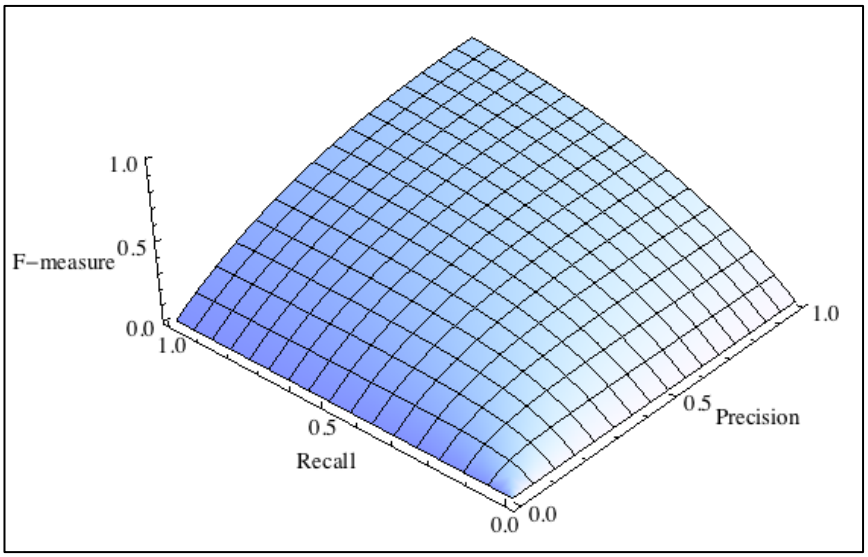

Рисунок 2 – F1-Score

1.1.1 Подходы

Традиционно, одним из первых являлся подход, основанный на рукописных правилах, где использовались знания лингвистики, словари понятий. Недостатком данного подхода является отсутствие возможности обобщения подхода для распознавания новых сущностей, а также сложность составления правил.

Визуально, алгоритм решения представлен на рисунке 3.

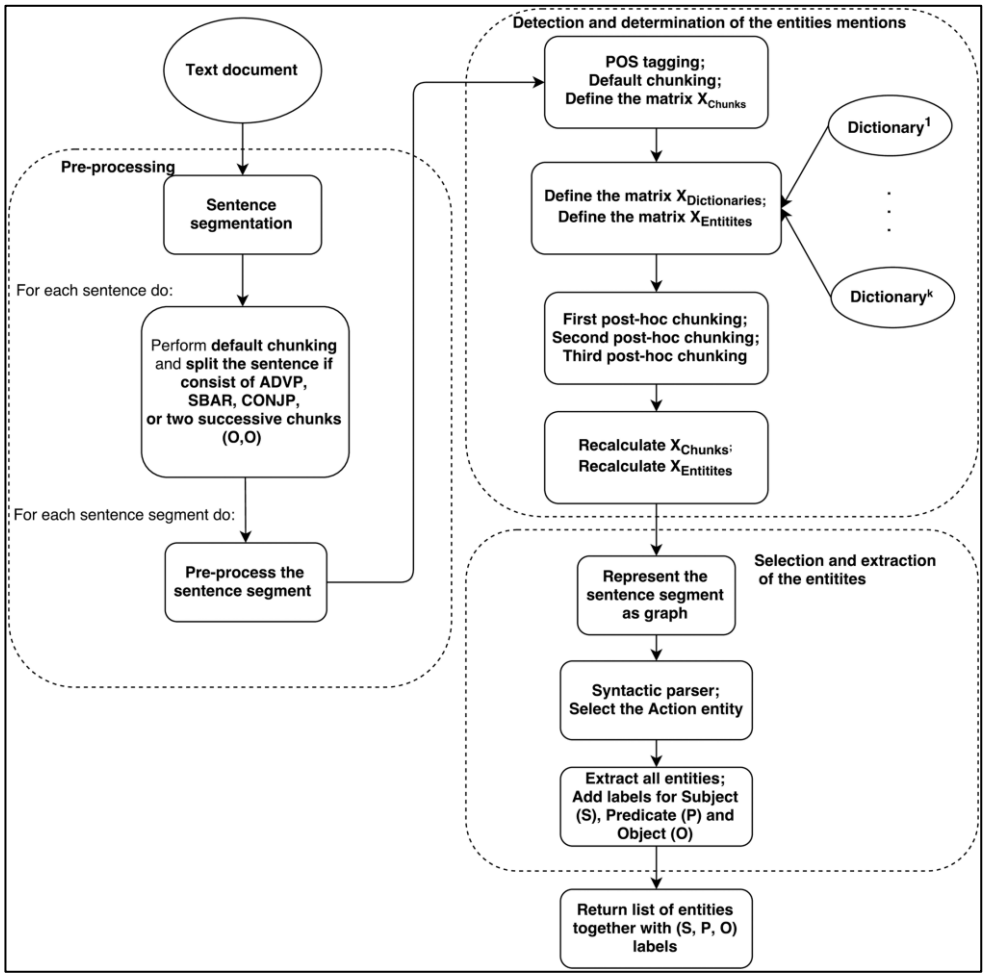

Рисунок 3 – Подход основанный на правилах

Статистические подходы являются более универсальными в случае задачи распознавания именованных сущностей. Методы, как правило, основываются на биграммах, триграммах, а также n-граммах.

Существует несколько статистических моделей, используемых для распознавания именованных сущностей, как например скрытые марковские модели, модель максимальной энтропии, а также условные случайные поля.

Скрытая марковская модель (HMM) является графовой моделью, позволяющей выражать условную вероятность распределений, основываясь на ограниченной истории (свойство Маркова).

Существует два типа условных контекстов в HMM:

– Наблюдаемый контекст: контексты, непосредственно относящиеся к наблюдению;

– Скрытый контекст: контексты относящиеся к скрытым (латентным) состояниям.

В n-грамм моделировании мы предполагаем, что наблюдение слова зависит от наблюдаемых предыдущих слов. Но, мы знаем, что имеет смысл не только слово, но и контекст его использования. Во многих случаях важно знать синтаксическую категорию (часть речь) для предсказания следующего слова.

В скрытых Марковских моделях состояние явно невидимо, но выход зависит от состояния, которое видимо. Каждое состояние имеет вероятностное распределение на множестве выходных слов. Таким образом, последовательность слов, генерируемая HMM моделью, дает некоторую информацию о последовательности состояний.

Стоит заметить, что термин «скрытая» относится к состоянию, через которое модель проходит, а не к параметрам модели. Даже, если все параметры заданы точно, модель остается скрытой.

Принцип максимума энтропии утверждает, что наиболее характерными распределениями вероятностей состояний неопределенной среды являются такие распределения, которые максимизируют выбранную меру неопределенности при заданной информации о «поведении» среды.

Модель максимальной энтропии подходит для извлечения именованных сущностей, поскольку она сочетает различные формы контекстной информации и не накладывает каких-либо предположений о распределении тренировочных данных.

Основная идея модели – это выбор вероятностного распределения, имеющего наибольшую энтропию, и удовлетворяет определенным условиям.

Определенные условия ограничивают модель к поведению согласно множеству статистик, собранных из тренировочных данных. В частности, ограничения зависят от признаков.

К примеру, частеречная разметка с помощью данной модели, реализованная Адвэйтом Ратнапархи (Adwait Ratnaparkhi) из университета Пенсильвании дала точность 96,6% на тренировке. [9]

Метод условных случайных полей рассматривает условное распределение (y ∣ x) последовательности меток y∈Y, а вектор x∈X состоит из наблюдаемых элементов. Из наблюдаемых и выходных элементов конструируется набор бинарных функций-признаков, которые могут включать в себя любое количество элементов.

Приведем пример:

$$F_i(x, y) = 1, \tag{1}$$

если y = <GEO>, y начинается с большой буквы, x = «улица», иначе = 0.

Формально, модель условных случайных полей рассматривается как обобщение Модели максимальной энтропии и Скрытых Марковских моделей и может быть выражена следующим образом:

$$\text{p (y ∣ x)} = \frac{1}{z(x)} \exp\Big(\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x)\Big), \tag{2}$$

где λ_i представляет веса, назначенные различным признакам на этапе тренировки;

Z(x) это степень нормализации, которая может быть выражена:

$$Z(x) = \sum_{y \in \text{Y}} \exp\Big(\sum_t \sum_k \lambda_k f_k(y_{t-1}, y_t, x)\Big) \tag{3}$$

При использовании условных случайных полей, важно определить пространство признаков. В подходе при сравнении Принципа максимума энтропии и условных случайных полей, были использованы следующие признаки:

– тэги частей речи слов;

– газеттиры.

Газеттиры представляют собой фиксированные словари сущностей, используемые в системах обработки естественных языков. Примеры газеттиров: списки локаций, персон или организаций, которые можно скачать с внешних источников для использования.

Газеттиры имеют следующие недостатки:

– список сущностей ограничен;

– список не зависит от контекста (изменив порядок слов, с помощью газеттира будет проблематично это учесть);

– построение списков – это монотонная работая, требующая много времени.

Таким образом, данный подход является ограниченным, в случае появления новых сущностей возникают проблемы необходимости постоянного добавления новых объектов в список.

Пример промышленных реализаций для русского языка – это система распознавания именованных сущностей для русского языка Zamgi-NER.

Данная система способна распознавать сущности следующих типов:

– персоны;

– географические названия;

– продукты;

– события.

Особенностью данной системы является то, что типы определяются не словарем, а на основе статистических алгоритмов. С одной стороны, это может привести к ошибкам в определении типа сущности (например, "Красная Москва – когда-то это были самые замечательные духи" может определиться как география), но с другой стороны система способна корректно определить новый, ранее не встречавшийся тип.

Количество типов и описание их классов задается на этапе обучения (получения статистической модели).

Основным недостатком данной системы является чувствительность к регистру и проблема с учетом контекста, то есть, система способна распознать лишь сущности с заглавной буквы, а также, если слово записано с ошибкой, с незаглавной буквы либо имеет омонимы, то система не сможет корректно распознать тип сущности.

Одним из наиболее универсальных подходов сегодня являются подходы, основанные на машинном обучении.

Наиболее подходящими являются рекуррентные сети, которые имеют основное предназначение в обработке последовательностей.

Однако, простые рекуррентные сети слабо справляются с сохранением информации о контексте длинного окна.

Пример для обработки естественных языков: «Я родился во Франции, поэтому, мой родной язык французский.». Данный пример является достаточно простым и в случае, если мы хотим предсказать слово «французский», достаточно учесть контекст данного предложения.

Куда более сложным является пример, подобный этому: «Я родился во Франции в 1949 году. Родители работали на лесопилке, поэтому денег было мало. Моим родным языком с детства был французский.».

В случае, представленном выше, для предсказания родного языка, нам необходимо учесть контекст 3 предложений.

Рекуррентные сети быстро теряют информацию о прошлом, что в англоязычной литературе называется «Vanishing Gradient Problem», суть которой заключается в том, что в глубоких архитектурах нейронных сетей, более ранние слои в сети медленнее обучаются по сравнению с последними. [10] [11]

Данная проблема решается при использовании рекуррентных сетей долгосрочной памяти (Long Short-Term Memory Recurrent Neural Networks - LSTM), благодаря механизму долгосрочной памяти, имеется возможность сохранения информации о более длинных контекстах.

Каждая ячейка LSTM состоит из четырех компонентов:

– входные ворота;

– выходные ворота;

– механизм «забывания»;

– механизм сохранения информации (memory block).

Формальное описание данных компонентов представлено ниже:

$$i_t = \sigma\left(W_{ix}\, x_t + W_{ih}\, h_{t-1} + b_i\right), \quad (4)$$

$$f_t = \sigma(W_{fx}\, x_t + W_{fh}\, h_{t-1} + b_f), \quad (5)$$

$$c_n = g(W_{cx}\, x_t + W_{ch}\, h_{t-1} + b_c), \quad (6)$$

$$c_t = f_t \circ c_{t-1} + i_t \circ c_n, \quad (7)$$

$$h_t = o_t \circ g(c_t), \quad (8)$$

$$o_t = \sigma(W_{ox}\, x_t + W_{oh}\, h_{t-1} + b_o), \quad (9)$$

где σ, g – сигмоидная и функция гиперболического тангенса;

$\circ$ – поэлементное умножение;

W – матрица весов;

B – bias;

i, f, o, c – входные, выходные ворота, механизмы забывания и сохранения информации соответственно.

Корректное распознавание именованных сущностей в предложениях зависит от контекста слова. Более того, для корректного распознавания, более полезно, если мы будем учитывать, как прошлый контекст, так и будущий.

1.1.2 Используемый подход

Для решения данной задачи подходящим является использование двунаправленных рекуррентных сетей долгосрочной памяти (Bi-Directional LSTM). Вычисления для двунаправленных сетей состоят из двух шагов:

– слой прямого распространения вычисляет вектор левого контекста;

– слой обратного распространения слой вычисляет вектор правого контекста. [12]

Выходы данных слоев конкатенируются для представления полного вектора последовательности. На выходе сети, вместо стандартного для задачи классификации, softmax-слоя добавляется слой условных случайных полей для более быстрого и точного решения задачи классификации последовательности.

Данный подход продемонстрирован на рисунке 4:

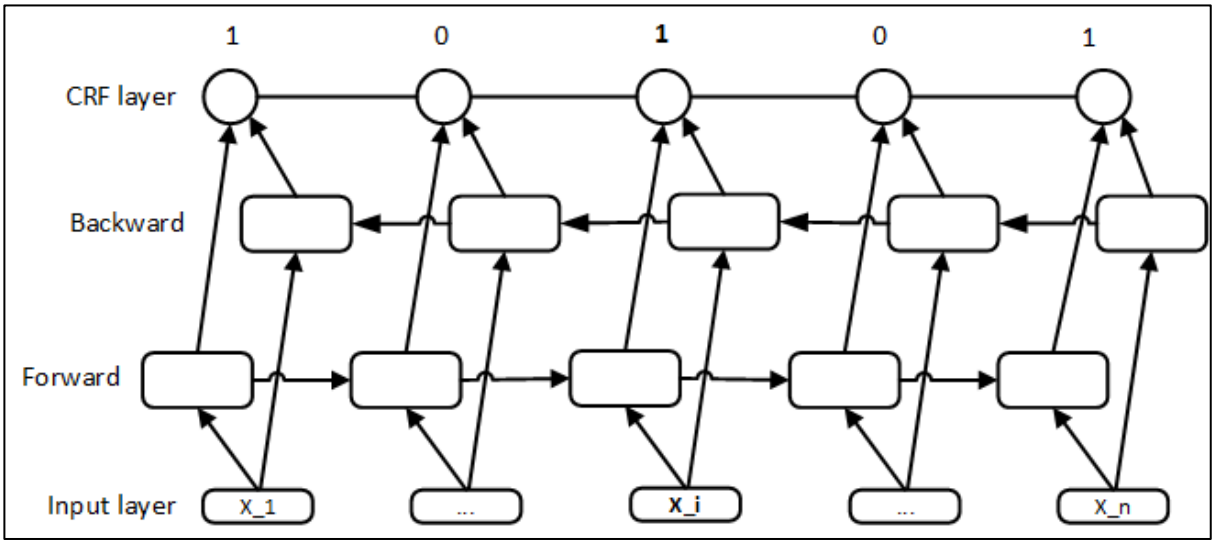

Рисунок 4 – Двунаправленная нейронная сеть с случайными условными полями

Для построения вектора нами были использованы следующие признаки:

– векторное представление слов;

– векторное представление символов;

– тэги частей речи;

– дополнительные тэги: география и другие.

Использование данного подхода в совокупности с используемыми признаками позволило достичь следующих результатов:

– организации: precision – 81%, recall – 80%, f1-score – 81%;

– локации: precision – 75%, recall – 78%, f1-score – 76%;

– персоны: precision – 89%, recall – 87%, f1-score – 88%;

– даты: precision – 97%, recall – 98%, f1-score – 97%;

– продукты нефтегазовой отрасли precision – 73%, recall – 87%, f1-score – 0.6351.

Данное решение было реализовано на языке Python с использованием фреймворков машинного обучения: tensorflow, keras, sklearn.

1.2 Разрешение кореференции

Кореферентность – отношение между компонентами текста, в котором они ссылаются на один и тот же референт.

Алгоритмы распознавания сущностей как правило определяют в тексте упоминания некоторых сущностей и относят их к тому или иному классу. Разрешение кореференции позволяет объединить одинаковые (относящиеся к одной и той же сущности) упоминания, связать их с объектом действительности.

Примеры:

– анафора: «*Книга* лежит на столе. *Она* тяжелая.»;

– метонимия: «*Писать* заставил публику читать *себя*» (себя – труды писателя);

– катафора: «Если *они* разозлятся из-за шума, то *соседи* вызовут полицию»;

– употребление слов, объединяющих несколько сущностей: «*Кэрол* и *Боб* посетят вечеринку. *Они* приедут вместе.»;

– различные формы употребления одной и той же сущности: «*Москва* занимает первое место в России по численности населения. Кремль находится в центральной части *города*.».

Применение алгоритмов разрешения кореференции:

– семантические поисковые системы;

– рекомендательные системы;

– анализ текстовых документов;

– вопрос-ответные системы;

– извлечение ключевых фраз;

– автоматизированное построение онтологии предметной области. [13]

1.2.1 Подходы

Существуют эвристические подходы к разрешению кореференции, основанные на различных правилах (грамматиках). Экспертная система разрешения кореференции на основе правил может иметь высокую точность распознавания кореференции (низкую долю ложных срабатываний), однако полнота разрешения кореференции не очень высока. Большая полнота распознавания требует создания большого числа правил – покрыть все виды кореференции для каждой сущности становится невозможным.

Для разрешения кореференции могут применяться алгоритмы машинного обучения, в частности, алгоритмы обучения с учителем и без учителя. Алгоритмы обучения с учителем обладают хорошей обобщающей способностью и работают на примерах, которых не было в обучающей выборке, что позволяет добиться повышения полноты результатов. К недостаткам относится зависимость качества распознавания от качества и количества размеченных данных. Алгоритмы обучения без учителя не требуют размеченных данных, но и не обладают высоким качеством работы.

В качестве алгоритмов обучения с учителем для разрешения кореференции часто используются нейронные сети, обученные на достаточно большой обученной выборке. На рисунке 5 изображена визуализация результатов разрешения кореференции.

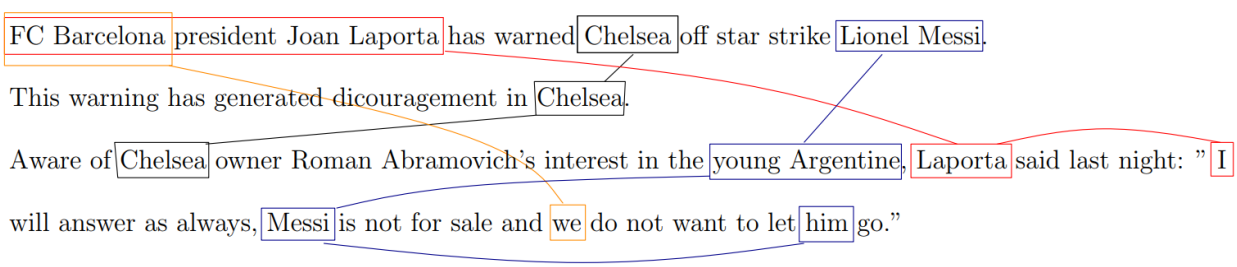

Рисунок 5 – Визуализация кореференции

Некоторые виды моделей для разрешения кореференции:

– Mention-pair - модель обучения с учителем. Она основана на бинарной классификации – кореферируют две сущности или нет. Применение таких моделей затруднено на практике: во-первых, свойство транзитивности кореференции может не соблюдаться (упоминание может кореферировать с двумя другими упоминаниями, которые не кореферируют между собой). Во-вторых, большинство сущностей не кореферируют друг с другом, что затрудняет обучение – классы сильно не сбалансированы;

– Mention-ranking – такие модели для каждой сущности ранжируют другие сущности по вероятности кореференции. Если упоминания сущностей не определены другой моделью, то такие зачастую структура таких моделей подразумевает расчет вероятностей, являются ли данные части текста упоминаниями сущностей и кореферируют ли они.

Комбинация алгоритмов распознавания сущностей, разрешения кореференции и связывания именованных сущностей позволяет определять связь упоминаний некоторой сущности в тексте с сущностями в базе данных.

1.2.2 Используемый подход

Поскольку подходы на основе бинарной классификации по упомянутым выше причинам не позволяют добиться высокого качества разрешения кореференции, используется подход на основе ранжирования.

Одна нейронная сеть на основе вектора упоминания получает оценку (не обязательно вероятность) наличия у упоминания предшественника – другого упоминаний той же сущности. Вторая нейронная сеть на основе векторов пары упоминаний получает оценку наличия кореференции. Определенные значения затем могут быть отсортированы для определения имеет ли упоминание сущности предшественников и если имеет, то определить, какое упоминание является предшественником. Оценки на выходе нейронных сетей необязательно должны означать вероятность, для ранжирования это не требуется.

Вектора упоминаний сущностей получаются следующим образом: берутся вектора слов внутри и вне (в пределах некоторого окна, размер которого зависит в том числе от количества доступных для обучения данных) упоминания сущности. Данные вектора берутся из модели FastText (аналог Word2vec).

На рисунке 6 изображена пример структуры на основе ранжирования.

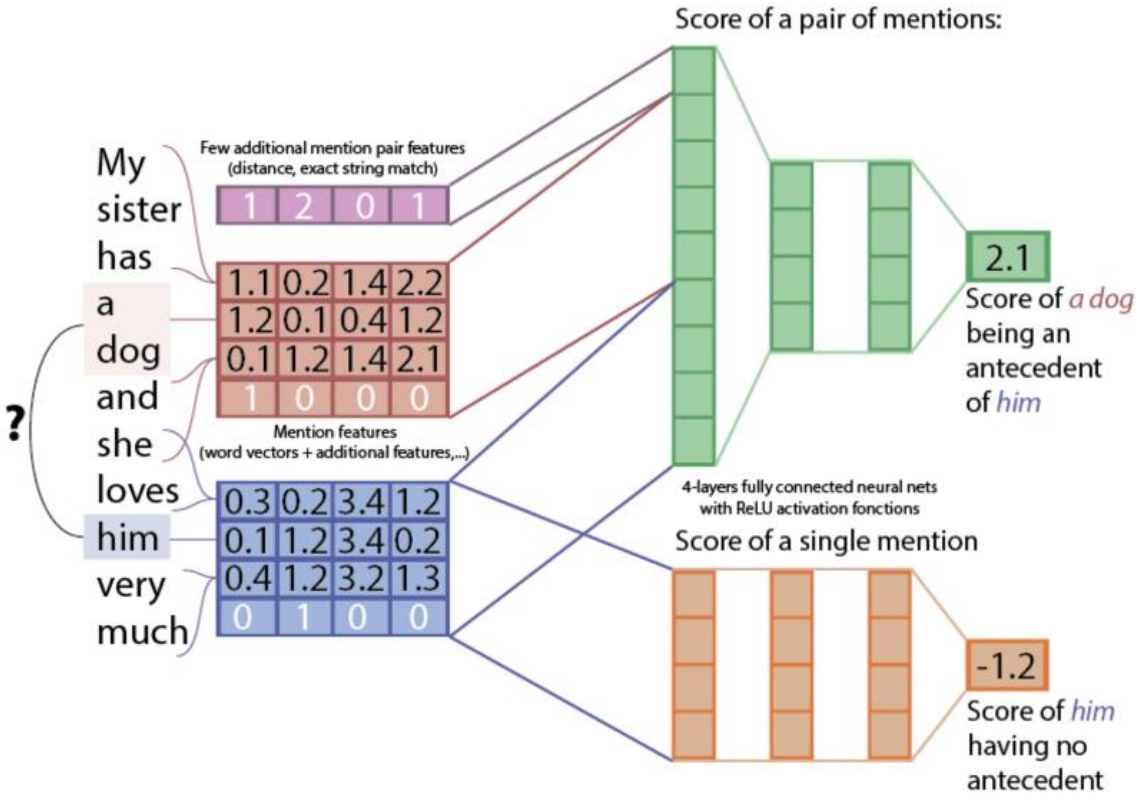

Рисунок 6 – Пример структуры на основе ранжирования

1.3 Автореферирование текста

В силу большого количества текстовых данных, генерируемых каждый день, встает вопрос о времени получения информации и необходимости получать основную информацию в более сжатом виде. Данная проблема адресована к решению проблемы извлечения автореферата текста.

Автореферирование текста – это задача создания короткого, точно отражающего суть всего документа.

1.3.1 Подходы

Существует два основных подхода к суммаризации текста:

– экстрактивный подход;

– абстрактивный подход.

Экстрактивный подход подразумевает под собой извлечение ключевых предложений из документа. Данный подход включает техники ранжирования ключевых предложений по важности (которая оценивается через различные метрики), после чего выбирается N наиболее релевантных предложений.

На рисунке 7 изображено визуальное представление экстрактивного подхода:

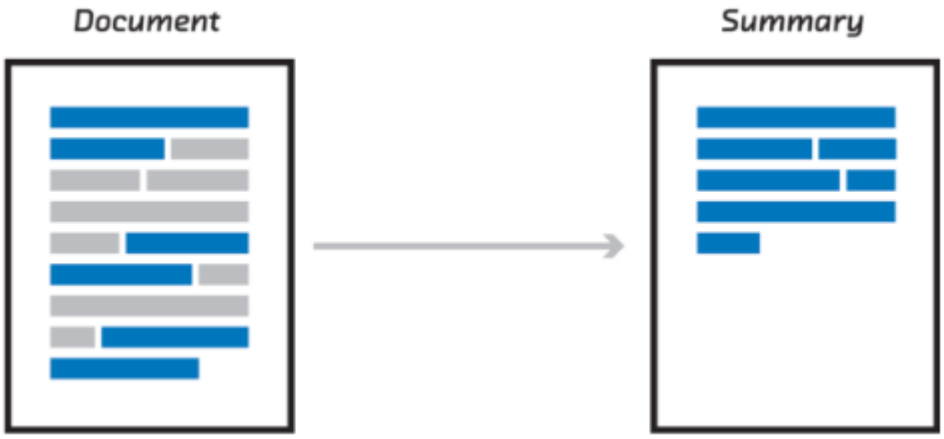

Рисунок 7 – Экстрактивный подход к автореферированию текста

Абстрактивный подход подразумевает под собой генерацию нового текста, захватывая смысл исходного текста. Это более сложный подход в силу того, что для данного подхода, как правило, используются нейронные сети архитектуры Encoder-Decoder и требуется большое количество тренировочных данных формата: Исходные текст и автореферат текста.

Одна из архитектур нейронной сети для абстрактивного подхода имеет следующие блоки: из эмбеддингов слов и топиков исходного текста извлекаются признаки с помощью свёрточных слоёв, затем вычисляется скалярное произведение признаков кодировщика и декодировщика. После этого

производится генерация предложения. На рисунке 8 изображено визуальное представление абстрактивного подхода:

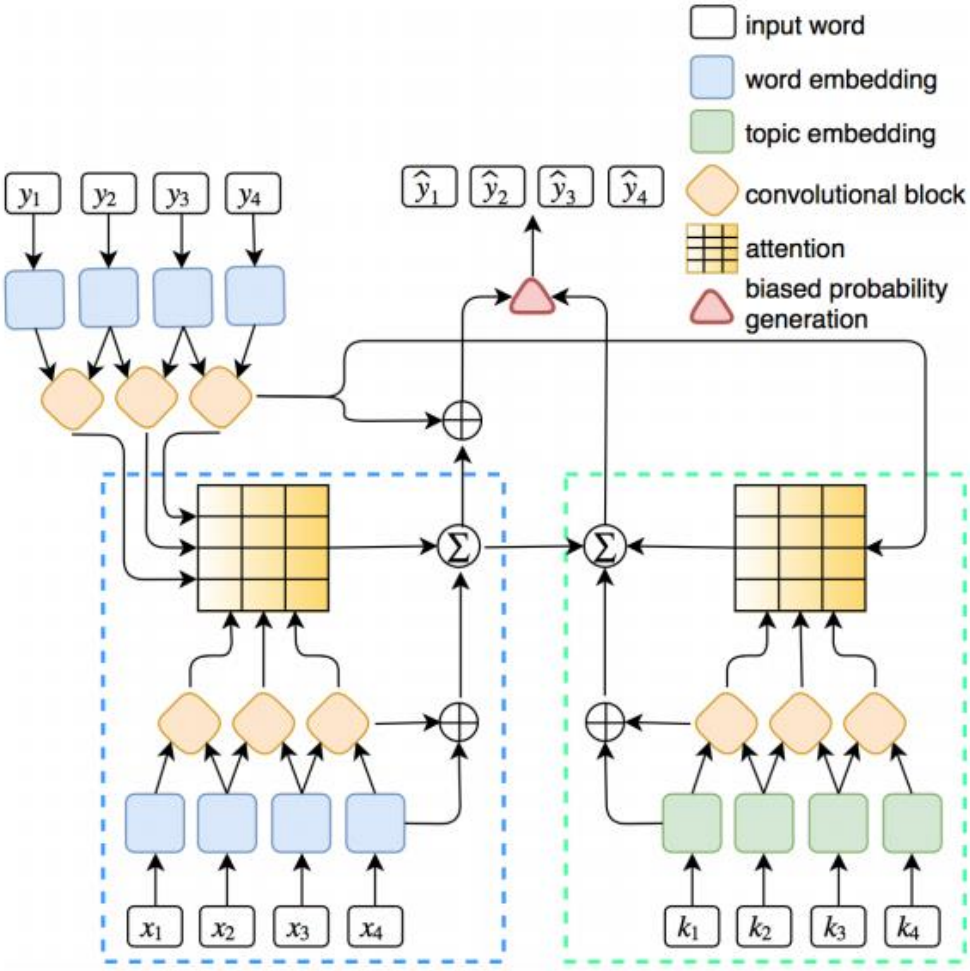

Рисунок 8 – Абстрактивный подход к автореферированию текста

В данной системе реализован экстрактивный подход в силу отсутствия выборки для обучения нейросети задаче генерации текста.

1.3.2 Используемый подход

TextRank – это алгоритм, основанный на графе, позволяющий извлекать наиболее релевантные предложения из текста, что и будет являться авторефератом текста. Основная идея заключается в следующем: чем больше у каждой вершины связей, тем выше ее ранг. [14]

Допустим, у нас есть предложение: «Kylo Ren confronts Finn and Rey in the snowy forest». На первом шаге мы разбиваем текст на предложения. Имея

предложения, мы можем построить граф, где каждое предложение являются вершинами и ребрами – это связи предложений между друг другом или k наиболее схожих предложений по определенным оценкам.

Основной интерес представляет расчет схожести между предложениями, где мы можем оценить отношение количества общих слов к общему количеству всех слов или какими-либо другими способами.

После этого необходимо использовать алгоритм PageRank, который имеет следующий смысл: выполнение случайных переходов по графу, где каждый случайный переход начинается из произвольной точки, далее на основе веса ребра выбирается соседняя вершина и так далее. К примеру, в вершине A (предложении) мы имеем соседей B (0.65), C (0.04), D (0.27), тогда мы можем вычислить вероятность перехода от A к каждой из соседних вершин следующим образом:

– A => B = 0.65 / (0.65 + 0.04 + 0.27);
– A => C = 0.04 / (0.65 + 0.04 + 0.27);
– A => D = 0.27 / (0.65 + 0.04 + 0.27).

В данном случае, при наибольшей вероятности перехода A => B, можно заключить, что переход в вершину B является наиболее приоритетным.

Следующая задача в данном алгоритме – это выбрать оптимальную длину переходов. Существует элегантное решение данной проблемы: берется один большой маршрут, на каждом шаге имеющий вероятность X, что он посетит соседа и вероятностью (1-X), что случайно совершит переход на случайную вершину, которая станет началом новой цепочки переходов. Обычно, X выбирается в диапазоне 0.8-0.85.

Каждый раз выполняя случайный переход, мы прибавляем 1 к коэффициенту предложения, переход на которое мы совершили.

В конце мы сортируем предложения по количеству посещений – это и будет оценкой нашего ранжирования. Выбирая топ-N слов по порядку, мы имеем автореферат (или саммари) нашего текста.

Таким образом, у алгоритма имеются следующие параметры:

– текст (из которого мы хотим получить автореферат);

– сколько предложений мы хотим получить в автореферате;

– функция оценки схожести предложений;

– длина пути переходов;

– X (вероятность перехода к соседу и 1-X перехода к случайной вершине).

Визуально, концепцию алгоритм можно представить, как на схеме, изображенной на рисунке 9.

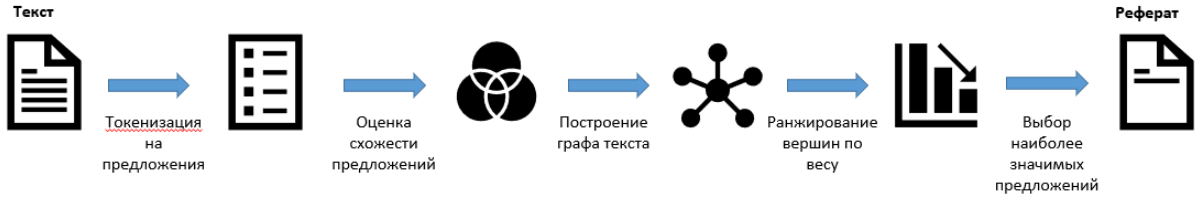

Рисунок 9 – Концепция алгоритма TextRank

1.4 Извлечение ключевых слов

Извлечение ключевых слов определяется как задача автоматического определения терминов, наилучшим образом описывающих содержание документа. Необходимость извлечения ключевых слов появляется в задачах интеллектуального анализа текстов, информационного поиска и обработки естественного языка. Ключевые слова применяются для автоматического индексирования коллекций документов или для категоризации и классификации текстов, которые используются в построении систем, предназначенных для управления документами, автореферирования, высокоуровневого семантического анализа, поиска по категориям, классификации и кластеризации веб-сайтов и документов, создания тематических словарей и многого другого. Автоматическое извлечение ключевых слов является обязательным этапом обработки документов в системах такого рода.

Все существующие системы автоматического извлечения ключевых слов основаны на одном общем базовом алгоритме: формирование множества фраз-кандидатов с последующей селекцией из этого множества конечных ключевых фраз на основе оценивающего алгоритма. Несмотря на наличие большого разнообразия методик разработка системы автоматического извлечения ключевых слов встречает определенные трудности. Наиболее значительные проблемы, требующие настройки фундаментальных параметров системы, обусловлены лексическими особенностями и грамматическими нормами языка, на котором представлен текст. К простым в обработке естественным языкам относятся английский и французский языки, большую грамматическую сложность имеют немецкий и русский языки, а хинди считается одним из языков, наиболее трудно поддающихся компьютерной обработке. В качестве одной из трудностей, присущей задаче извлечения ключевых слов, можно указать сложность извлечения определенных лексических групп, например, именных групп. Еще одна сложность – определение списка стоп-слов, которые удаляются при первичной обработке документа, к ним, как правило, относятся предлоги, союзы, междометия, артикли и другие слова, не несущие смысловую нагрузку.

На выбор метода и его реализации в системе влияют область применения системы и требования по ее работе, например, наличие корпуса документов и возможность обучения системы в процессе ее работы позволяют построить статистическую модель корпуса и дополнять ее при добавлении новых документов, в то время как для получения ключевых слов для одиночного документа преимущественно используются методы, основанные построении графовых или векторных образов текста. Кроме того, при разработке системы семантического анализа, к задачам которого относится определение ключевых слов, рекомендуется учитывать тематику и специфику структуры обрабатываемых документов.

1.4.1 Подходы

Методы извлечения ключевых слов могут быть разделены на следующие группы:

1) простые статистические подходы;
2) лингвистические подходы;
3) подходы с использованием машинного обучения;
4) гибридные подходы.

Лингвистические подходы используют лингвистические особенности слов, предложений и документа. Лексический, синтаксический, семантический и дискурсивный анализ являются одними из наиболее распространенных, но вместе с тем сложных анализов.

Подходы контролируемого машинного обучения создают модель, которая обучается на основе набора ключевых слов. Они требуют ручной разметки набора обучающих данных, которая является чрезвычайно утомительной и непоследовательной (иногда запрашивает предварительно определенную таксономию). Таким образом, для извлечения ключевых слов из нового документа применяется индуцированная модель. Этот подход включает в себя метод наивного Байеса, метод опорных векторов, пакетирование и т. д. Таким образом, методы требуют обучающих данных и часто зависят от предметной области. Системе необходимо заново обучать и настраивать модель каждый раз при изменении предметной области. Индукция модели может быть очень сложной и длительной при работе с массивными наборами данных.

Гибридные подходы для извлечения ключевых слов объединяют все методы, упомянутые выше. Одним из таких подходов является модель векторного пространства.

Простейшим подходом, используемым для выделения ключевых слов, является критерий частотности, позволяющий выделять важные слова в документе. Однако, этот метод показывает довольно слабые результаты.

Стоит перечислить наиболее популярные методы на сегодняшний день:

– TF-IDF;

– TextRank;

– Topical PageRank;

– Themed PageRank.

TF-IDF статистика – это общий инструмент для извлечения не только векторов слов, но и извлечения ключевых слов. В данном подходе ключевые слова извлекаются не из конкретного документа, а из набора. Также, если произвести предварительную кластеризацию, то имеется возможность извлекать документы из кластера, которые должны являться однородными по смыслу.

Визуально данный подход представлен на рисунке 10.

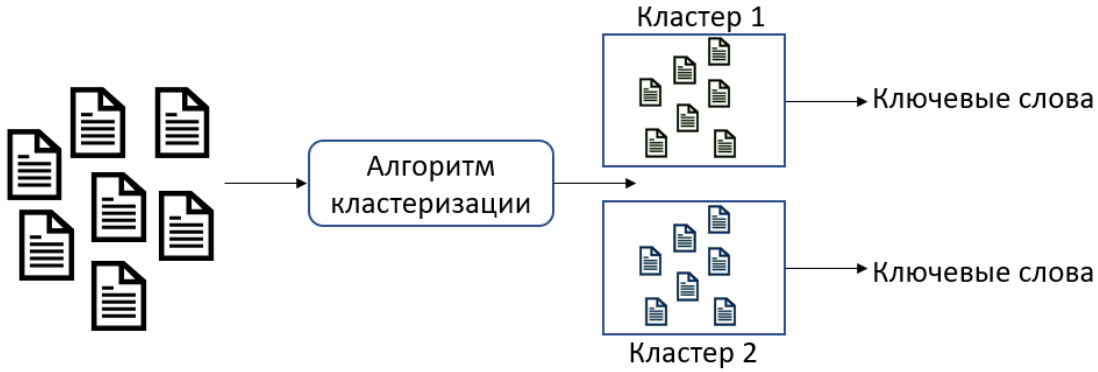

Рисунок 10 – Извлечение ключевых слов кластера с помощью TF-IDF

TF-IDF извлекает ключевые слова из документа методом выбора N слов с наибольшим значением TF-IDF документа. Формула TF – отношение числа вхождений некоторого слова к общему числу слов документа. Таким образом, оценивается важность слова в пределах отдельного документа. IDF – это инверсия частоты, с которой некоторое слово встречается в коллекции документов. Учёт IDF позволяет уменьшить вес широкоупотребительных слов. Для каждого отдельного уникального слова в пределах конкретной коллекции документов существует только одно значение IDF.

Таким образом получается, что наибольший вес будут иметь слова, часто употребляемые в конкретном документе и реже встречаемые в других, что позволяет в рамках алгоритма отфильтровать стоп-слова.

Пример расчетной таблицы алгоритма TF-IDF представлен на рисунке 11.

	Document Keyword	tf						idf	wdt=tf*idf				
		d	d_1	d_2	d_3	d_4	df	log(n/df)	d	d_1	d_2	d_3	d_4
k_1	google	3	0	0	2	0	2	2.7	8.1	0	0	5.4	0
k_2	*pasar* (market)	1	2	1	1	0	4	2.4	2.4	4.8	2.4	2.4	0
k_3	browsing	2	0	0	3	0	2	2.7	5.4	0	0	8.1	0
k_4	online	1	0	0	4	0	2	2.7	2.7	0	0	10.8	0
k_5	*iklan* (ad)	3	1	0	0	0	2	2.7	8.1	2.7	0	0	0
k_6	network	1	0	0	1	2	3	2.52	2.52	0	0	2.52	5.04
k_7	user	1	0	0	0	0	1	2.7	2.7	0	0	0	0
k_8	facebook	2	0	0	1	0	2	2.7	5.4	0	0	2.7	0
k_9	mobile	4	0	0	3	0	2	2.7	10.8	0	0	8.1	0
k_{10}	sosial	1	1	0	0	1	3	3	3	3	0	0	3

Рисунок 11 – Пример расчета TF-IDF

При использовании других алгоритмов (не TF-IDF) стоит отметить, что очень важно произвести фильтрацию стоп-слов, поскольку, наиболее релевантными словами станут предлоги, союзы и другие слова, которые не несут большой смысловой нагрузки.

Topical PageRank (TRP) алгоритм – это вариация TextRank с добавлением веса наиболее важных слов из выбранной темы и распределений слов-тем, сгенерированной моделью тематического моделирования. Данный подход хорошо подходит для очень больших корпусов текстов, в тоже время он становится вычислительно сложнее предыдущих алгоритмов.

Themed PageRank – это еще одно улучшение TextRank и PageRank алгоритмов. Сначала он извлекает пространство Latent Dirichlet Allocation (LDA), которое представляет коллекцию документов в виде множества тем, описывающих конкретные области знаний. Затем рассчитывается модель для каждого топика, основываясь на модифицированном Personalised PageRank. В зависимости от поисковых нужд вход (одно или несколько ключевых слов или документов конвертируются в распределение по темам, которые затем используются как линейная комбинация множества тематико-специализированных оценок модели в уникальную оценку, представляющую

соответствие между входными данными поиска и документа. LDA – это байесовская генеративная вероятностная модель для коллекции дискретных данных, которая стала популярной для тематического моделирования в научных корпусах текстов. В LDA документ корпуса моделируется с помощью явного представления конечного набора тем, где каждая тема моделируется конечным множеством слов в корпусе. В Themed PageRank LDA используется для воспроизведения тем, которые позволяют связать документы и термины в корпусе текста. Основой модели являются не слова, а термины, поскольку, именно термины позволяют извлекать факты (информацию) из технических текстов.

Это также позволяет сделать модель масштабируемой на больших корпусах текстов. Метод, который используется для нахождения технических терминов довольно прост и не требует больших вычислительных возможностей, однако не является наилучшим по качеству определения терминов.

1.4.2 Используемый подход

EmbedRank – это один из наиболее современных методов извлечения ключевых слов. Данный метод был использован в конечном приложении.

Методы извлечения ключевых слов на основе обучения с учителем требуют большого количества размеченных данных и слабо обобщаются вне специфики тренировочных данных.

В отличии от большинства алгоритмов обучения без учителя, которые имеют низкую точность и слабо обобщаются, EmbedRank использует вектор предложения (sentence embeddings). EmbedRank достигает высокого F-score по сравнению с алгоритмами основанными на графах и хорошо подходит для обработки текстов в реальном времени, что хорошо подходит для системы семантического анализа текстов, где время имеет важное значение.

Важно отметить, что для извлечения ключевых слов, для алгоритма требуется только документ, из которого непосредственно предполагается

извлечение ключевых слов, в отличие от алгоритмов, которые требуют целые корпуса, связанные с данными документом по смыслу.

Метод основывается на современных разработках, где представление текста произвольной длины подразумевает под собой эмбеддинг предложения в векторном пространстве. Это позволяет рассчитывать семантическую близость среди фрагментов текста, используя метрики схожести эмбеддингов, как например, косинусное расстояние.

EmbedRank гарантирует соблюдение двух основных свойств ключевых фраз – информативность и разнообразие. Информативность, получается из расстояния между эмбеддингом фразы-кандидата и текста всего документа, а разнообразие через вычисление расстояний между эмбеддингами фраз.

Эмбеддинги слов имеют важное значение в представлении слов как векторов в непрерывном векторном пространстве, решая основные проблемы: отсутствие смысловой связи между словами (в случае использования эмбеддингов, смысловая связь рассчитывается на основе схожести контекстов слов), а также размерности (методы, вроде one-hot encoding, имеют размер вектора равный размеру словаря, что достаточно неэффективно по использованию памяти).

Формально, алгоритм EmbedRank состоит из трех основных шагов:

– извлечение фраз-кандидатов из текста;

– использование эмбеддингов предложений и текстов для сравнения (размеры векторов используются равными);

– на основе ранжирования кандидатов (по векторному расстоянию между векторами ключевых фраз и вектором документа) выбирается набор N ключевых фраз. [15]

Алгоритм EmbedRank представлен на рисунке 12.

Рисунок 12 – Алгоритм EmbedRank

В векторном пространстве можно представить, как объекты, расположенные в разных областях, где центральным звеном является вектор документа и объекты, расположенные наиболее близко к нему, имеют большую оценку при ранжировании кандидатов в ключевые фразы. Визуально это представлено на рисунке 13.

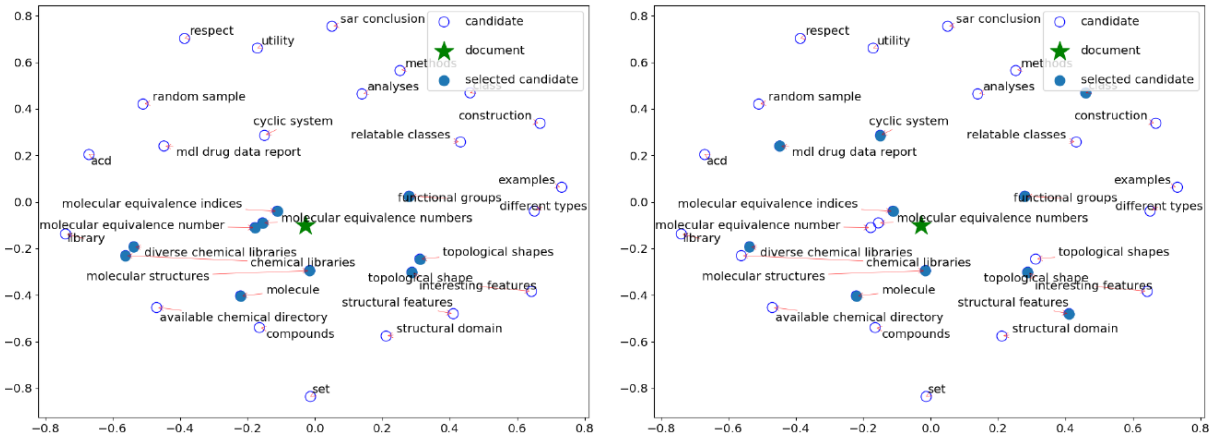

Рисунок 13 – Извлечение ключевых слов с использованием EmbedRank

1.5 Тематическое моделирование

Тематическое моделирование – это технология статистического анализа текстов для автоматического выявления тематики в больших коллекциях документов. Тематическая модель определяет, к каким темам относится каждый документ, и какими словами описывается каждая тема. Для этого не требуется никакой ручной разметки текстов, обучение модели происходит без учителя.

Похоже на кластеризацию, но тематическая кластеризация является «мягкой» и допускает, чтобы документ относился к нескольким кластерам-темам. Тематическое моделирование не претендует на понимание смысла текста, однако оно способно отвечать на вопросы «о чём этот текст» или «какие общие темы имеет эта пара текстов».

Для чего используется тематическое моделирование:

- разведочный информационный поиск (exploratory search) в электронных библиотеках;
- поиск по смыслу, а не ключевым словам;
- обнаружение и отслеживание событий в новостных потоках;
- выявление тематических сообществ в социальных сетях;
- построение профилей интересов пользователей в рекомендательных системах;
- категоризация индентов в системах разговорного интеллекта;
- поиск мотивов в нуклеотидных и аминокислотных последовательностях;
- аннотирование изображений;
- поиск аномального поведения объектов в видеопотоке;
- выявление паттернов поведения клиентов по транзакционным данным.

Тематическая модель формирует сжатое векторное представление текста, которое помогает классифицировать, рубрицировать, аннотировать, сегментировать тексты. В отличие от известных векторных представлений семейства x2vec (vord2vec, paragraph2vec, graph2vec и т. д.), в тематических векторах каждая координата соответствует теме и имеет содержательную интерпретацию. Модель привязывает к каждой теме список ключевых слов или фраз, который описывает семантику этой темы.

LDA, латентное размещение Дирихле – самая известная и часто используемая тематическая модель. Проблема данной модели в том, что задача тематического моделирования имеет очень много (бесконечно много) решений,

и LDA выбирает одно из них, не предоставляя никаких средств для выбора лучшего решения под конкретную задачу.

Регуляризация служит для задания желаемых свойств тематической модели в виде оптимизационных критериев. Например, есть регуляризаторы, которые улучшают качество классификации текстов, повышают точность и полноту поиска, повышают различность тем, обеспечивают максимально возможную разреженность решения, учитывают дополнительные нетекстовые данные, и т. д. [16]

1.5.1 Инструмент BigARTM

Аддитивная регуляризация (ARTM) позволяет задать сразу несколько критериев-регуляризаторов. Например, чтобы построить тематическую модель новостного потока, необходимо учесть несколько верхних уровней уже имеющегося рубрикатора, научить модель учитывать время документов, разделять темы на подтемы и создавать новые темы по необходимости. ARTM позволяет складывать регуляризаторы от разных моделей, создавая комбинации моделей с заданными свойствами под конкретные приложения. Это приводит к модульной технологии тематического моделирования с высокой степенью повторного использования кода.

BigARTM реализует несколько механизмов, которые снимают многие ограничения простых моделей типа PLSA или LDA и расширяют спектр приложений тематического моделирования:

– Regularization. Регуляризаторы, которые можно комбинировать в любых сочетаниях;

– Modality. Модальности, которыми можно описывать нетекстовые объекты внутри документов;

– Hierarchy. Тематические иерархии, в которых темы разделяются на подтемы;

– Intratext. Обработка текста как последовательности тематических векторов слов;

– Co-occurrence. Использование данных о совместной встречаемости слов;

– Hypergraph. Тематизация сложно структурированных транзакционных данных;

Теперь немного подробнее об этих механизмах:

– мультимодальные тематические модели позволяют обрабатывать документы, содержащие не только слова, но и токены других модальностей. Это могут быть метаданные документа – авторы, время, источник, рубрики, и т.д. Это могут быть также токены, находящиеся внутри текста – ссылки, теги, словосочетания, именованные сущности, объекты на изображениях, записи о действиях пользователей, и т.д. Модальности помогают строить темы с учётом дополнительной информации. С другой стороны, темы помогают выявлять семантику нетекстовых модальностей, предсказывать или рекомендовать значения пропущенных токенов;

– мультиязычные тематические модели реализуются как частный случай мультимодальных. Модальностями являются языки. В системах кроссязычного и мультиязычного тематического поиска запрос даётся на одном языке, а ответ может быть получен на других языках. Например, пользователь имеет текст патента на русском языке, и хочет найти близкие патенты на английском;

– иерархические тематические модели используются для автоматической рубрикации текстов. В BigARTM тематическая иерархия строится сверху вниз по уровням. Каждая дочерняя тема связывается с одной или несколькими родительскими. Каждая родительская тема может разделиться на несколько подтем, либо перейти на следующий уровень целиком;

– внутритекстовые регуляризаторы позволяют учитывать порядок слов, синтаксические связи, деление текста по предложениям и абзацам и другую внутритекстовую информацию. Он используется для тематической сегментации текстов, при этом сегментация влияет на темы. Это позволяет отойти от гипотезы

«мешка слов» – самого критикуемого допущения в тематическом моделировании;

– тематические модели со-встречаемости используют данные о совместной встречаемости слов в локальных контекстах, например, в предложениях. Они основаны на дистрибутивной гипотезе – предположении, что смысл слова в языке определяется совокупностью всех слов, встречающихся в его локальных контекстах. Получаемые векторные представления слов имеют те же свойства, что и в моделях дистрибутивной семантики семейства x2vec. Они лучше инкапсулируют смыслы слов и точнее решают задачи семантической близости. При этом тематические векторные представления, в отличие от векторов x2vec, имеют интерпретируемые координаты;

– гиперграфовые тематические модели используются для описания транзакционных данных. В обычном тексте транзакция – это запись о том, что слово встретилось в документе. В более сложных приложениях данные не сводятся к парным транзакциям и описывают взаимодействия трёх и более объектов. Например, транзакция (u, b, p) в рекламной сети – «пользователь u кликнул баннер b, расположенный на странице p»; финансовая транзакция (b,s,g) – «покупатель b купил товар g у продавца s». Транзакциями могут быть любые наборы объектов. Предложение в тексте – это тоже транзакция, состоящая из слов. Транзакции могут быть пересекающимися или вложенными. Модель строит тематические векторные представления для всех объектов, участвующих в транзакциях, независимо от их природы. Это наиболее общий вид тематических моделей, которые можно строить с использованием BigARTM.

Следующие регуляризаторы реализованы в библиотеке BigARTM:

– сглаживание распределений терминов в темах. Используется для выделения фоновых тем, собирающих общую лексику языка или общую лексику данной коллекции;

– сглаживание распределений тем в документах. Используется для выделения фоновых слов в каждом документах;

– разреживание распределений терминов в темах. Используется для выделения лексических ядер предметных тем как относительно небольшой доли слов словаря;

– разреживание распределений тем в документах. Используется для выделения относительно небольшой доли предметных тем в каждом документах;

– декоррелирование распределений терминов в темах. Используется для повышения различности лексических ядер предметных тем;

– отбор тем путём обнуления вероятности темы во всех документах. Используется для выведения из модели незначимых тем. Позволяет оптимизировать число тем, начиная с заведомо избыточного числа тем и постепенно удаляя ненужные.

Следующие метрики качества реализованы в библиотеке BigARTM:

– перплексия;

– разреженность;

– средняя чистота тем;

– средняя контрастность тем;

– средний размер лексического ядра тем;

– доля фоновых слов во всей коллекции. [17]

1.5.2 Автоматическая регуляризация тематических моделей с помощью библиотеки bigARTM

В основе рассматриваемой тематической модели лежит вероятностная тематическая модель, представленная на рисунке 14.

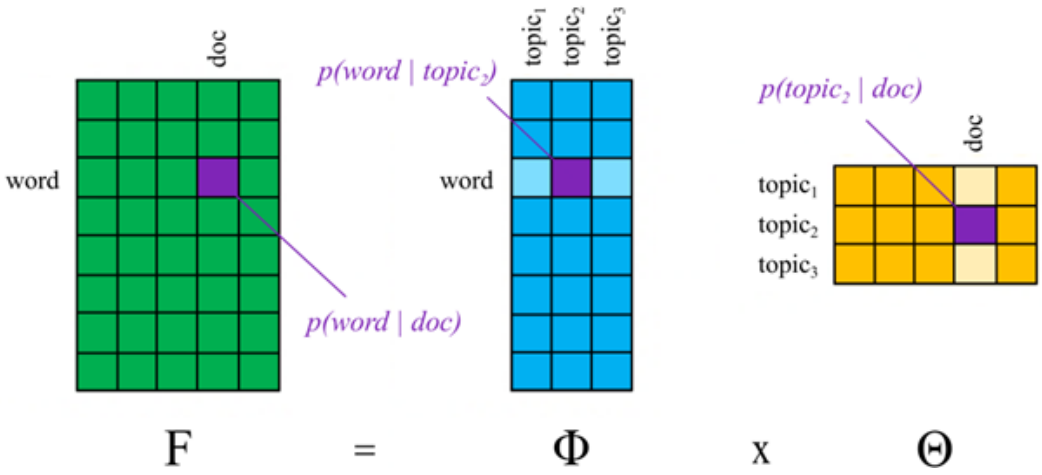

Рисунок 14 – Вероятностная тематическая модель

Так как задача поиска матриц распределения слов в темах (Фи) и распределения тем в документах (Тета) имеет неединственное решение, решается задача оптимизации.

Дополнительные условия, используемые для того, чтобы придать модели некоторые свойства или как-то улучшить её, называются регуляризаторами.

1.5.3 Реализация работы с моделью ARTM

Предложенная реализация использует библиотеку BigARTM при этом осуществляется автоматический подбор регуляризаторов.

Возможны несколько режимов работы:

1. загрузка готовой модели;
2. создание модели с нуля;
3. тестирование модели на разное количество тем.

1.5.4 Предобработка текста

Для корректной работы алгоритмов обработки естественного языка чаще всего используют те или иные методы предобработки текста, описанные, в том числе, раннее.

Как правило, данные подходы проверяются экспериментально и заранее нет методического подхода, определяющего, какие методы покажут себя лучше

всего. Для получения наиболее осмысленных результатов в задаче тематического моделирования были экспериментально определены следующие методы предобработки текста:

1. приведение всех слов к нижнему регистру;
2. удаление всех знаков препинания;
3. удаление всех цифр;
4. приведение всех слов к начальной форме (с помощью библиотеки pymorphy2);
5. удаление стоп-слов и слов с высокой частотностью (nltk.corpus import stopwords и слов из подготовленного текстового документа);
6. приведение датасета к формату vowpal wabbit.

Рассмотрим пример.

Исходный текст: «Профильные комитеты Совета Федерации рекомендуют палате одобрить законопроект об изменении границ между Москвой и Московской областью. Как отмечается в отзывах комитетов на данный законопроект, изменения границ между Москвой и Московской областью позволят «повысить инвестиционную привлекательность как Москвы, так и области, что крайне важно для экономического и градостроительного развития». Соглашение об изменении границ подписано на днях мэром Москвы Сергеем Собяниным и губернатором Московской области Борисом Громовым. «Изменение границы между Москвой и областью носит характер уточнения, цель которого придать юридический статус фактически сложившейся ситуации и границе в целом. Земельные участки Москве передаются общей площадью 723,46 гектара, а в область отойдут земли площадью 328,45 гектара», - сказал «Интерфаксу» глава комиссии Совета Федерации по жилищной политике и жилищно-коммунальному хозяйству Валерий Парфенов.».

Строка в формате vowpal wabbit: «business199005.txt |text профильный комитет федерация палата одобрить законопроект москва московский область отмечаться отзыв комитет законопроект москва московский область повысить инвестиционный привлекательность москва область экономический

градостроительный развитие соглашение мэр москва сергей собянин губернатор московский область борис громов москва область уточнение придать юридический земельный участок москва передаваться площадь гектар область отойти земля площадь гектар интерфакс комиссия федерация жилищный политика жилищно коммунальный хозяйство валерий парфен.».

1.5.5 Создание новой модели

Для создания модели с нуля необходим датасет в формате vowpal wabbit и путь куда модель будет сохраняться.

Алгоритм обучения модели включает в себя следующие этапы:

1. автоматизированный выбор оптимального количества тем на основе когерентности;
2. поиск оптимальных параметров регуляризации для заданного набора регуляризаторов;

После того как модель была создана:

1. модель можно сохранить (для дальнейшего использования уже готовой);
2. выгрузить топ-слова;
3. сохранить словарь. Во время обучения artm создаёт свой словарь для обучения, в котором рассчитаны частотные характеристики слов, на которых обучена модель. Необходимо для расчёта когеренции;
4. выгрузка матрицы theta. Матрица theta представляет собой таблицу;
5. предсказание топиков для документа, не участвовавшего в обучении;

1.5.6 Регуляризаторы и метрики

В модели используются следующие регуляризаторы:

- декорреляция тем;
- разреживание распределения тем в документах;
- сглаживание распределения слов в темах.

Каждый из регуляризаторов добавляется последовательно в указанном порядке. Для каждого регуляризатора выбирается оптимальное значение параметра регуляризации (из нескольких рассматриваемых). После того, как первый регуляризатор настроен, происходит выбор второго, без изменений для первого и тд.

Выбор регуляризаторов осуществляется на основе трёх метрик:

1. перплексия;
2. разреживание матрицы распределения тем в документах;
3. разреживание матрицы распределения слов в темах.

Выбирается тот параметр регуляризации, который дал лучший результат с точки зрения разреживания матриц, незначительно ухудшив при этом перлексию.

1.5.7 Предсказание топиков нового документа

Для предсказания топикой документа необходимо иметь готовую модель. Будет она получена путём загрузки или обучена с нуля не имеет значения для предсказания. Для этого необходимо:

– объект модели;
– ассоциированный с ней список стоп-слов;
– 3 - текст, для которого предсказываем топики.

Пример:

Исходный текст: «Как сообщает РИА Новости, в ночном клубе Reina, по данным СМИ, неизвестные в костюмах Санта-Клауса открыли стрельбу из автоматов Калашникова. Правда, у властей города другая версия: они утверждают, что действовал одиночка.

BREAKING | Suspect wearing Santa Claus is allegedly hiding in the night club where many killed. #IstanbulAttack pic.twitter.com/1KrPdYdYBj

— Vocal Europe (@thevocaleurope) December 31, 2016

Под пулями погибли 39 человек, 69 получили ранения. К данной минуте опознан 21 человек, почти все из них – иностранцы. По предварительным данным, среди жертв теракта нет россиян. Что касается раненых, то информация еще уточняется.».

Результат: «text_id |text риа новость ночное клуб reina данные неизвестный костюм сантаклаус стрельба автомат калашников правда власть одиночка breaking suspect wearing santa claus is allegedly hiding in the night club where many killed istanbulattack pictwittercomkrpdydybj vocal europe thevocaleurope december пуля погибнуть человек ранение минута опознать человек иностранец данные жертва теракт россиянин раненый уточняться».

['topic_0', 'topic_12', 'topic_19'].

Расшифровка топиков (ключевые слова):

– «topic_0»: [«фильм», «театр», «русский», «интернет», «человек», «тег», «культура», «режиссёр», «язык», «кино», «церковь», «сеть», «актёр», «премия», «музей»];

– «topic_12»: [«сообщить», «ребёнок», «агентство», «дом», «летний», «мужчина», «район», «орган», «человек», «обнаружить», «полиция», «правоохранительный», «убийство», «данные», «источник»];

– «topic_19»: [«человек», «местный», «телеканал», «американский», «посольство», «сообщить», «власть», «взрыв», «жертва», «погибший», «погибнуть», «штат», «теракт», «данные», «страна»].

Матрица theta представляется собой таблицу, в которой представлено распределение текстов по топикам. Для всех текстов, на которых обучалась модель, написаны вероятности их отнесения к каждому топику.

Результат обработки матрицы theta. В примере представлен фрагмент результирующего словаря:

– ‘russia237311.txt’: [(‘topic_10’, 0.19455838), (‘topic_12’, 0.8054416)];

– ‘russia237312.txt’: [(‘topic_3’, 0.27257246), (‘topic_6’, 0.60759944), (‘topic_11’, 0.11982813)];

– ‘russia237313.txt’: [(‘topic_7’, 0.37931034), (‘topic_13’, 0.47485793), (‘topic_17’, 0.14583172)];

– ‘russia237386.txt’: [(‘topic_7’, 0.45833334), (‘topic_17’, 0.5416667)];

– ‘russia237387.txt’: [(‘topic_1’, 0.11175502), (‘topic_8’, 0.8270205)];

– ‘russia237392.txt’: [(‘topic_6’, 0.8125971)];

– ‘russia237394.txt’: [(‘topic_3’, 0.99999994)].

2 СЕРВИС

Для реализации сервиса необходимо прежде всего определить несколько важных факторов:

– функциональные требования сервиса;

– нефункциональные требования;

– определение схемы решения.

Ответив на данные вопросы, подход к разработке будет более систематический и позволит совершить как можно меньше ошибок в последующем.

2.3 Описание требований к сервису

NLP Service предназначен для упрощения анализа и поиска документов, представленных в различных форматах. Для этого предполагается использовать современные методы выделения ключевых слов и смысловых частей из текста.

Основные функции:

– извлечение из загруженных документов текста и дополнительной информации, включая: семантические атрибуты, темы, краткое содержание;

– отображение извлеченной информации для каждого документа;

– хранение исходных документов и всей извлеченной информации;

– поиск по извлеченной из документов информации;

– управление доступом к загруженным документам и извлеченной из них информации.

Дополнительно система должна обладать следующими свойствами:

– позволять одновременно работать нескольким пользователям;

– допускать хранение не менее 100.000 документов без существенной деградации производительности;

– обладать web-интерфейсом.

2.3.1 Функциональные требования

К функциональным требованиям относится обработка документов, а также отображение результатов обработки.

При обработке документов необходимо реализовать следующие функции:

– загрузка документа на сервер;

– извлечение текста:

1) поддерживаем следующие форматы: pdf, doc, docx, txt;

2) для изображений или сканов извлекаем текст с помощью OCR.

– извлечение семантических атрибутов (named entity recognition, coreference resolution, keyword extraction);

– тематическое моделирование:

1) необходимо перестраивать модель каждые N документов.

– summary extraction:

1) экстрактивный подход.

Также система должна хранить исходный документ, результаты извлечение текста и обработки каждого документа, построенные модели.

При отображении результатов обработки загруженных документов необходимо реализовать:

– поиск документов по атрибутам;

– отображение результатов обработки конкретных документов:

1) отображение извлеченного текста;

2) выделение подсветкой именованных сущностей;

3) кластеров упоминаний;

4) отношение к тематикам по %;

5) извлеченных ключевых слов из документа.

2.3.2 Администрирование и управление доступом

Система должна позволять управлять доступом к документам и результатам их обработки.

Для этого не необходимо реализовать:

– регистрацию пользователей;

– создание групп пользователей;

– назначение/сопоставление пользователей группам;

– назначение/сопоставление документов группам;

– назначение/управление списком администраторов.

2.3.3 Нефункциональные требования

Развертывание:

– требования к клиенту:

1) браузер Chrome, Edge;

– требований к серверу:

1) Ubuntu / CentOS + nvidia-docker + docker-compose;
2) память> 24GB, GPU – Cuda.

Производительность:

– не менее 2-х пользователей одновременно. Не ожидаем большого количества пользователей одновременно, но система не должна быть однопользовательской.

Время обработки документа:

– среднее ~ 10 мин;

– максимальное – 30 мин.

2.4 Обзор технического решения

Компоненты:

– frontend – реализует пользовательский интерфейс в браузере;

– backend – все необходимое api для frontend;

– Elasticsearch – система индексации документов; [18]

– data processing worker – фоновый процесс(ы) в котором выполняются алгоритмы, занимающие длительное время.

2.4.1 Frontend

Содержит следующие страницы (группы страниц/диалоги):

– страница обработанного документа;

– страница поиска;

– страница регистрации;

– страница входа в систему;

– страница администрирования.

Реализация: в браузере с использованием vue.js

2.4.2 Backend

Реализуем следующие группы функций:

– авторизация и управление доступом;

– загрузка документов в Elasticsearch;

– выдачу результатов обработки документа (из Elasticsearch);

– проектирование (с учётом прав доступа) запросов к данным в Elasticsearch.

Реализация: python, flask, Gunicorn/gevent.

2.4.3 Elasticsearch

Для хранения документов и других данных системы используется ElasticSearch:

– Users Index – данные о пользователях;

– Workspaces index – данные о группах документах/отделах;

– Documents indices – индексы для хранения документов и результатов их обработки;

– Data processing worker – отдельный модуль по обработке данных.

Процесс (пока предполагаем, что процесс будет один) обработки данных реализует следующие алгоритмы:

1. Обработка документов – по мере загрузки документов в Elasticsearch;
2. Обновление моделей.

Ограничение по памяти: реализуется с помощью настроек docker.

2.5 Описание веб-сервиса

Для удобства работы с системой был реализован прототип системы, представленный в виде веб-сервиса, который состоит из трех компонентов:

1. Frontend – реализует пользовательский интерфейс в браузере. Реализован при помощи Vue.js, Webpack, Vuetify. Содержит следующие страницы: [19]

– страница регистрации пользователей;
– страница входа в систему;
– панель администрирования пользователей;
– панель администрирования рабочих пространств;
– страница загрузки документов;
– страница поиска;
– страница просмотра информации о документе.

2. Backend – все необходимое api для frontend, созданное при помощи Python, Flask, WSGI HTTP Server. Были реализованы следующие группы функций: [20]

– авторизация и управление доступом;
– реализация CRUD операций над сущностями базы данных (пользователи, документы, рабочие пространства);
– вызов функций анализа документов.

3. Поисковая система, созданная при помощи ElasticSearch, позволяет осуществлять поиск по документам, а также хранит сущности, необходимые для работы системы:

– пользователи;
– рабочие пространства;
– документы.

2.5.1 Аутентификация и авторизация

Система авторизации построена JSON Web Token [21].

При аутентификации, когда пользователь успешно входит в систему, используя свои учетные данные, возвращается веб-токен JSON. Всякий раз, когда пользователь хочет получить доступ к защищенному маршруту или ресурсу, пользователь должен отправить JWT, обычно в заголовке авторизации, используя схему Bearer. Защищенные маршруты сервера будут проверять допустимый JWT в Authorization заголовке, и, если он присутствует, пользователю будет разрешен доступ к защищенным ресурсам.

В системе предусмотрено два типа токенов: токен доступа и токен обновления. Токен доступа используется для доступа к защищенным ресурсам. Его время жизни – 15 минут. После истечения времени жизни токена доступа необходимо отправить на сервер запрос содержащий токен обновления, в ответ на который придет новый токен доступа.

POST
/api/refresh
Headers:
Authorization: Bearer <refresh_token>
Response json:
{‘access_token’: <access_token>}

Время жизни токена обновления 60 дней, по истечении которых пользователю необходимо заново войти в систему и получить новый токен обновления.

На случай потери доступа к аккаунту администратора предусмотрено аварийное восстановление прав и получение доступа к учетной записи администратора по умолчанию, с помощью следующего запроса:

POST
/api/restore_default_admin
Response json:
{}

Для администратора по умолчанию восстанавливаются права системного администратора, токен обновления с ограниченным временем действия сохраняется в логи сервера. После это токен обновления можно использовать для получения доступа к учетной записи администратора по умолчанию через web-интерфейс.

Для входа в систему используется следующий метод:

POST
/api/login
Request json:
{'username': <username>,'password': <password>}
Response json:
{'access_token': <access_token>,'refresh_token': <refresh_token>}

В ответ на запрос пользователь получает токен доступа и токен обновления для дальнейшей работы с системой.

Вид пользовательского интерфейса для входа в систему представлен на рисунке 15.

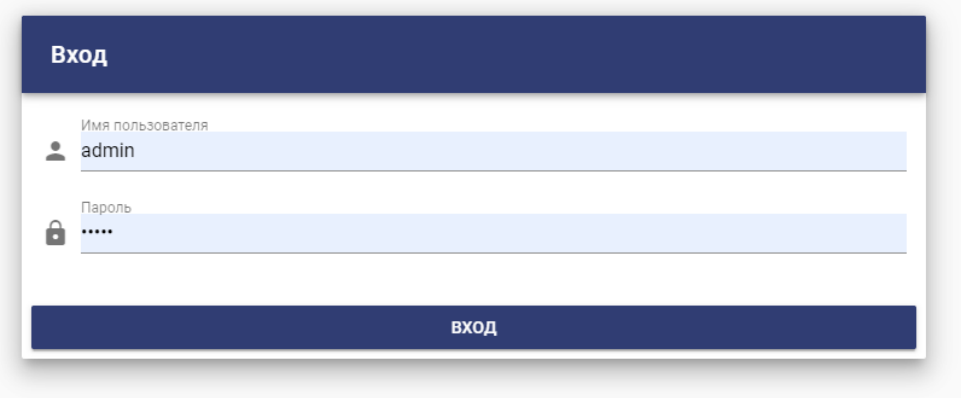

Рисунок 15 – Вид формы входа в систему

Регистрация пользователя происходит с помощью следующего метода:

POST
/api/user

Request json:

{'username': <username>, 'password': <password>}

Response json в формате Elastic Index Api:

{"_id": "4c1fb0f2166cdb9cd3eaa247662bea7450397e02",
"_index": "e12735c585a54b5f9a1571959cc9a41c",
"_source": {"username": "aas"},
"_type": "user",
"_version": 1,
"found": true}

Для регистрации требуется отправить логин пользователя и пароль. После этого в базе ElasticSearch создается запись о новом пользователе.

Вид пользовательского интерфейса для регистрации представлен на рисунке 16.

Рисунок 16 – Вид формы регистрации

После входа в систему предусмотрена возможность смены пароля пользователя самим пользователем или системным администратором.

PUT
/api/user/[user_id]/password
Headers
Authorization: Bearer <access_token>
Параметры
user_id - идентификатор пользователя.
Request json
Для пользователя: {"old_password": <old_password>, "new_password": <new_password>}
Для администратора: {"new_password": <new_password>}
Response json:

```
{"_id": "4c1fb0f2166cdb9cd3eaa247662bea7450397e02",
 "_index": "93c509ec64104d4797079cd992c3bddb",
 "_primary_term": 1,
 "_seq_no": 1,
 "_shards": {"failed": 0,
             "successful": 1,
              "total": 2},
 "_type": "user",
 "_version": 2,
 "forced_refresh": true,
 "result": "updated"}
```

Вид пользовательского интерфейса для смены пароля пользователем представлен на рисунке 17.

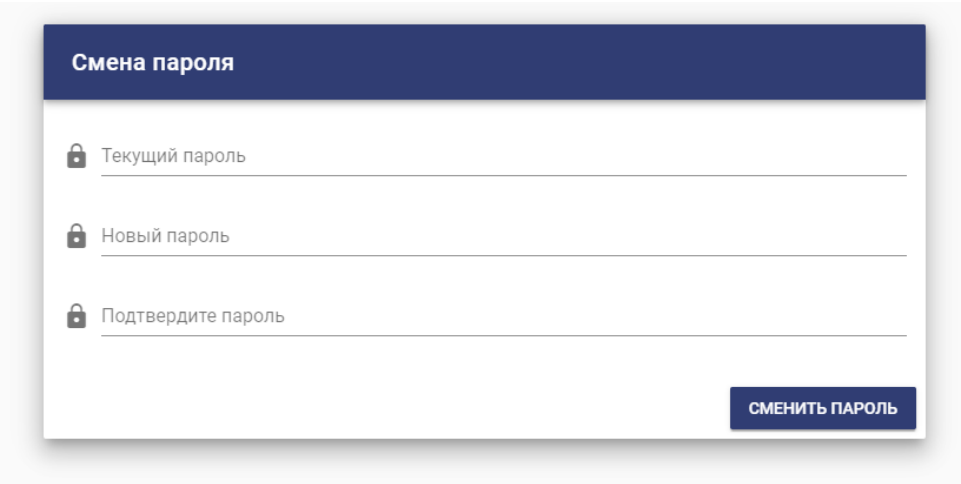

Рисунок 17 – Вид формы смены пароля

Для системного администратора предусмотрена страница управления пользователями, представленная на рисунке 18.

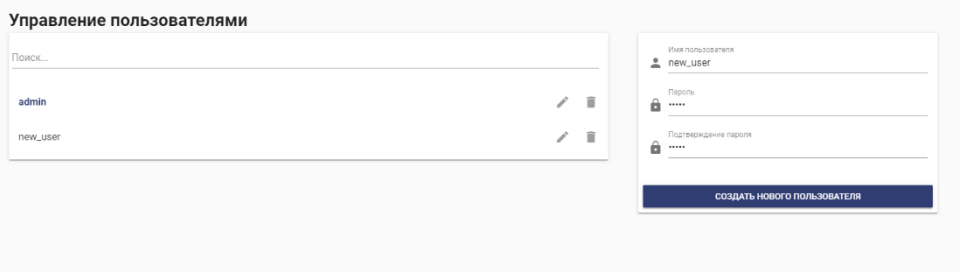

Рисунок 18 – Страница управления пользователями

При переходе на данную страницу выполняется запрос всех существующих в системе пользователей:

GET
/api/user
Headers
Authorization: Bearer <access_token>
Response json:
[{"_id": "6bbf7c81860fe7f355612c7ee02dac6c4ca0ba9d",
"_index": "173d760a83614ff9b865f7c908cfeaf7",
"_score": 1.0,
"_source": {"username": "49d9d9481b124445b773dc2a323c7a10"},
"_type": "user"}]

Также можно получить информацию о конкретном пользователе по его id:

GET
/api/user/[user_id]
Headers
Authorization: Bearer <access_token>
Параметры
user_id - идентификатор пользователя.
Response json:
{"_id": "4c1fb0f2166cdb9cd3eaa247662bea7450397e02",
"_index": "83bcef91c6a74f608b128501647d8bde",
"_source": {"username": "aas"},
"_type": "user",
"_version": 1,
"found": true}

Или по его логину:

GET
/api/user/[user_name]/name
Headers
Authorization: Bearer <access_token>
Параметры
user_name - имя пользователя.
Response json:
{"_id": "4c1fb0f2166cdb9cd3eaa247662bea7450397e02",
"_index": "83bcef91c6a74f608b128501647d8bde",
"_source": {"username": "aas"},
"_type": "user",
"_version": 1,
"found": true}

Системный администратор может удалять пользователей из системы:

DELETE
/api/user/[user_id]
Headers
Authorization: Bearer <access_token>
Параметры
user_id - идентификатор пользователя.
Response json:
{"_id": "da7f78c93cf6f022dbc39cae94fcbd8fd5c1c23a",
"_index": "ad0b44aa3168417d8ee546878e3aef3c",
"_primary_term": 1,
"_seq_no": 3,
"_shards": {"failed": 0,
"successful": 1,
"total": 2 },
"_type": "user",
"_version": 4,
"forced_refresh": true,
"result": "deleted"}

А также изменять информацию о пользователях и их уровень доступа. Могут быть изменены следующие поля:

– system_admin - true/false - выдача пользователю прав системного администратора.

– workspace_access - уровни доступа пользователя к различным рабочим пространствам. Допустимые уровни доступа и их возможности представлены в таблице 2.1.

Таблица 2.1 - Уровни доступа пользователей к рабочим пространствам

Тип пользователя	Возможности
User	Поиск по документам рабочей области, просмотр документов.
Power user	То же что и User, загрузка новых документов в рабочую область.
Admin	То же что и Power user, удаление документов из рабочей области, выдача прав другим пользователям.
PUT /api/user/[user_id] *Headers* Authorization: Bearer <access_token> *Параметры* user_id - идентификатор пользователя. *Request json:* Могут быть изменены следующие поля: – system_admin - true/false изменяется только пользователем с правами системного администратора. Пример: – { "system_admin": true } – workspace_access - пользователем с правами системного администратора или администратором соответсвующего workspace. Допустимые уровни доступа к workspace user, power_user, admin Пример: – { "workspace_access": { "workspace1_id>": "admin", "workspace2_id>": "user") } *Response json:* {"_id": "4c1fb0f2166cdb9cd3eaa247662bea7450397e02", "_index": "93c509ec64104d4797079cd992c3bddb", "_primary_term": 1, "_seq_no": 1, "_shards": {"failed": 0, "successful": 1, "total": 2}, "_type": "user", "_version": 2, "forced_refresh": true, "result": "updated"}	

Вид пользовательского интерфейса изменения информации о пользователе представлен на рисунке 19.

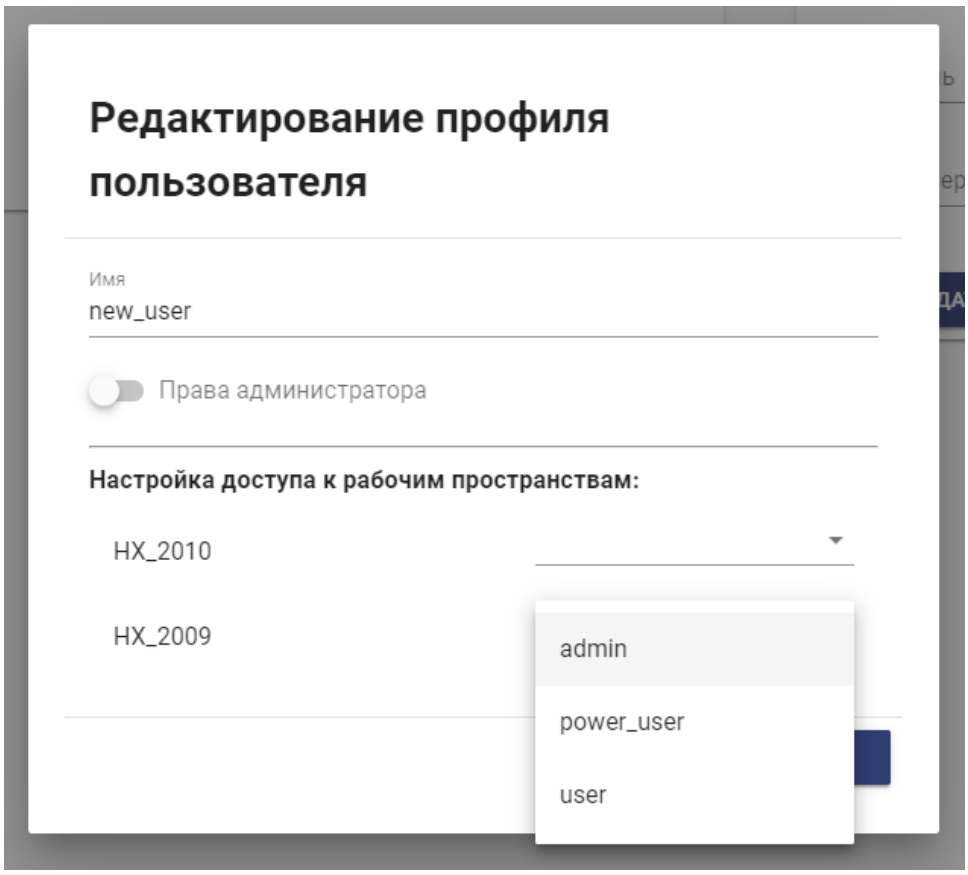

Рисунок 19 – Вид формы изменения информации о пользователе

2.5.2 Работа с рабочими пространствами

Для создания, удаления и изменения рабочих пространств требуются права системного администратора. Администратор может управлять рабочими пространствами, с помощью интерфейса управления рабочими пространствами, представленного на рисунке 20. Получение списка всех уже существующих пространств возможно с помощью следующего метода:

```
GET
/api/workspace
        Headers
        Authorization: Bearer <access_token>
        Response json:
        [{"_id": "ab95ec02d0a44694833283d5df1ce187",
          "_index": "949d6525bfea404a9e4a2e778b88f098",
          "_source": {"name": "some name", "description": "some description"},
          "_type": "workspace",
          "_version": 1,
          "found": true}]
```

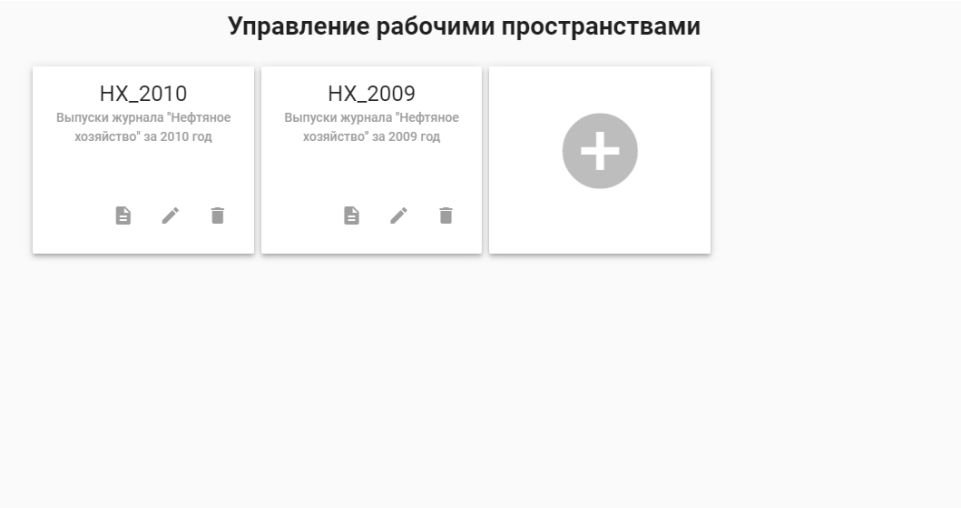

Рисунок 20 – Вид интерфейса управления рабочими пространствами

Для создания нового рабочего пространства предусмотрен следующий метод:

POST
/api/workspace
Headers
Authorization: Bearer <access_token>
Request json:
{'name': <workspace name>, 'description': <workspace description>}
Response json:
{"_id": "ab95ec02d0a44694833283d5df1ce187",
"_index": "949d6525bfea404a9e4a2e778b88f098",
"_primary_term": 1,
"_seq_no": 0,
"_shards": {"failed": 0,
"successful": 1,
"total": 2},
"_type": "workspace",
"_version": 1,
"forced_refresh": true,
"result": "created"}

Вид формы создания нового рабочего пространства представлен на рисунке 21.

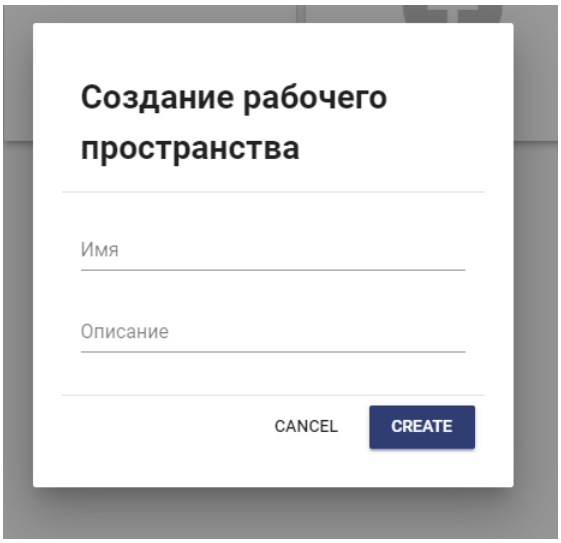

Рисунок 21 – Вид формы создания нового рабочего пространства

Предусмотрена возможность изменения информации о ранее существующем рабочем пространстве с помощью следующего метода:

PUT
/api/workspace/[workspace_id]

Headers

Authorization: Bearer <access_token>

Параметры

workspace_id - идентификатор workspace.

Request json:

Все поля не обязательные.

{'name': <workspace name> 'description': <workspace description> }

Response json:

```
{"_id": "ab95ec02d0a44694833283d5df1ce187",
 "_index": "949d6525bfea404a9e4a2e778b88f098",
 "_primary_term": 1,
 "_seq_no": 1,
 "_shards": {"failed": 0,
             "successful": 1,
             "total": 2},
 "_type": "workspace",
 "_version": 2,
 "forced_refresh": true,
 "result": "updated"}
```

Форма редактирования рабочего пространства представлена на рисунке 22.

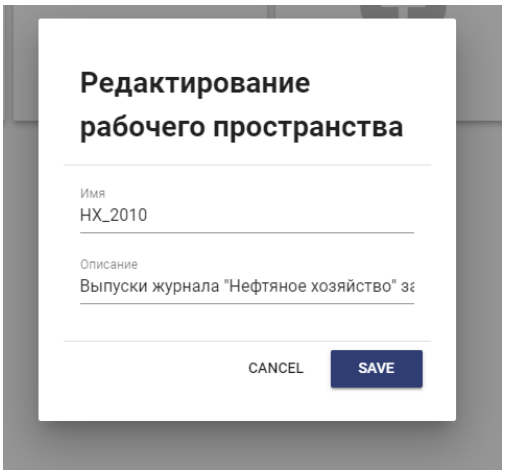

Рисунок 22 – Вид формы редактирования существующего рабочего пространства

Метод для удаления рабочего пространства:

DELETE
/api/workspace/[workspace_id]

Headers

Authorization: Bearer <access_token>

Параметры

workspace_id - идентификатор workspace.

Response json:

{"_id": "ab95ec02d0a44694833283d5df1ce187",
"_index": "949d6525bfea404a9e4a2e778b88f098",
"_primary_term": 1,
"_seq_no": 2,
"_shards": {"failed": 0,
"successful": 1,
"total": 2},
"_type": "workspace",
"_version": 3,
"forced_refresh": true,
"result": "deleted"}

2.5.3 Работа с документами

Перед началом работы необходимо загрузить документы в рабочее пространство. Загрузка документов происходит с помощью следующего метода:

POST
/api/document/[workspace_id]

Headers

Authorization: Bearer <access_token>

Параметры

workspace_id - идентификатор workspace.

Request:

user_filename - не обзательное имя, используемое для отображения и скачивания

multipart/form-data name='file' user_filename=''

Response json:


```
{"_id": "1",
"_index": "workspace_4021429bc0a4495f8f40df28e0500bbf",
"_primary_term": 1,
"_seq_no": 0,
"_shards": {"failed": 0,
            "successful": 1,
            "total": 2},
"_type": "doc",
"_version": 1,
"forced_refresh": true,
"result": "created"}
```


Вид страницы загрузки документов представлен на рисунке 23. На данной странице можно добавить загружаемые документы путем перетаскивания в область загрузки или выбрав из списка документов, находящихся на устройстве пользователя. Для каждого документа можно добавить название, и выбрать рабочую область, в которую будет загружен документ.

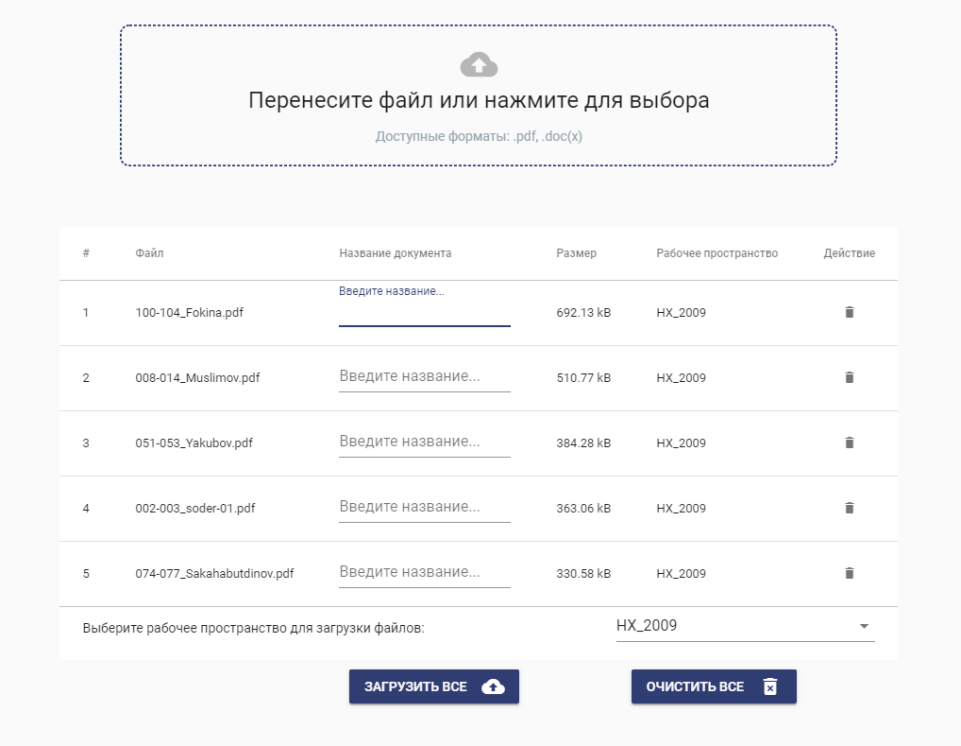

Рисунок 23 – Вид страницы загрузки документов в рабочее пространство

Для поиска документов в системе предусмотрена страница поиска, изображенная на рисунке 24.

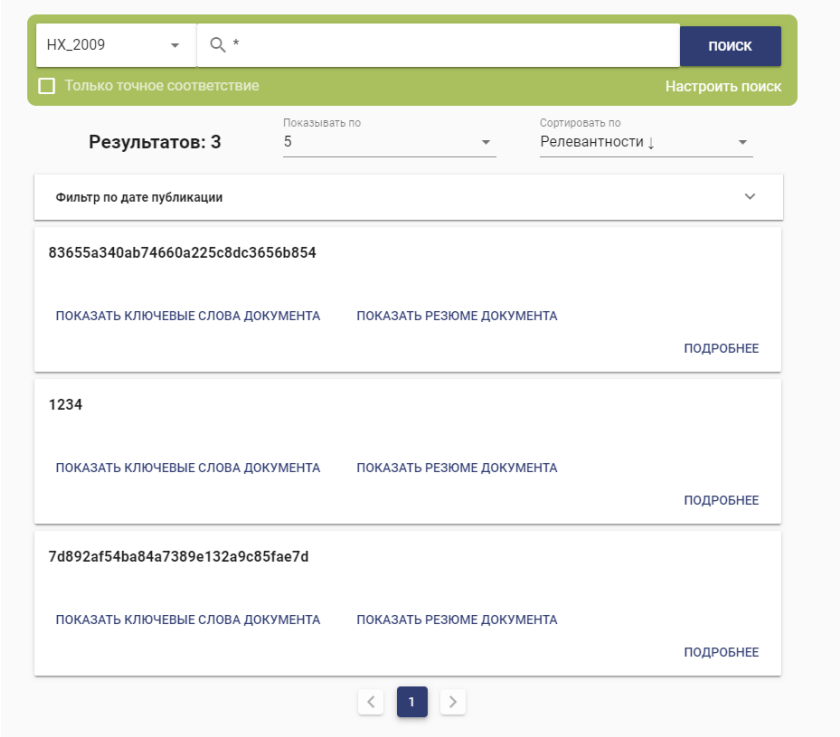

Рисунок 24 – Общий вид страницы поиска

Для осуществления поиска формируется поисковый запрос, поля которого представлены в таблице 2.2.

Таблица 2.2 - Поля поискового запроса

Поле	Значение
query	Строка поискового запроса
sort	Параметр сортировки
highlight	Параметры подсветки результатов
source_fields	Поля, по которым производится поиск

Результаты возвращаются в поле в source, в зависимости от передаваемых параметров. Количество возвращаемых документов ограничено 100.

POST
/api/document/[workspace_id]/query

Headers

Authorization: Bearer <access_token>

Параметры

workspace_id - идентификатор workspace.

Request json:

```
{"query" : <query>,
 "sort" : <sort criteria>
 "highlight" : <highlight criteria>
 "source_fields" : <source_fields>}
```

Response json:

```
[{"_id": "5f5ffe92e3fa48c79accb9a126bf1471",
  "_index": "workspace_25d96cc598cd49b2be466985ba5ccd13",
  "_score": 0.45798382,
  "_source": {...},
  "_type": "doc"}]
```

Для выбора полей документов, по которым производится поиск, в системе предусмотрены расширенные настройки поиска, представленные на рисунке 25.

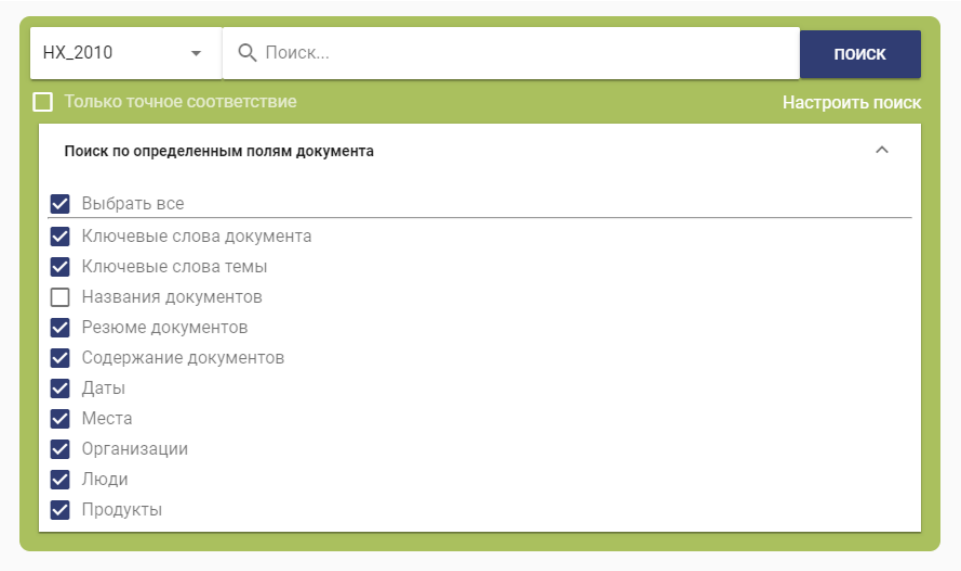

Рисунок 25 – Расширенные настройки поиска

Возможна фильтрация результатов поиска по дате документов (Рисунок 26).

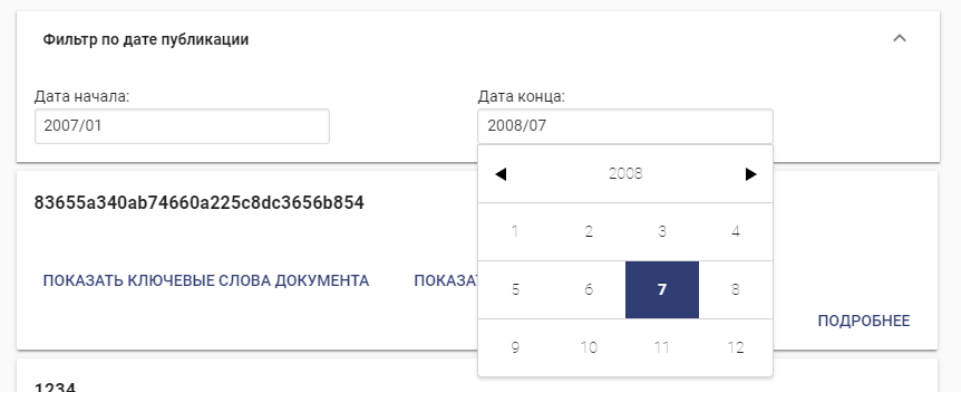

Рисунок 26 – Фильтрация документов по дате

Каждый документ в поисковой выдаче представлен в следующем виде, представленном на рисунке 27. Для каждого документа указываются места совпадения поискового запроса с текстом документа, ключевые слова и краткое содержание текста.

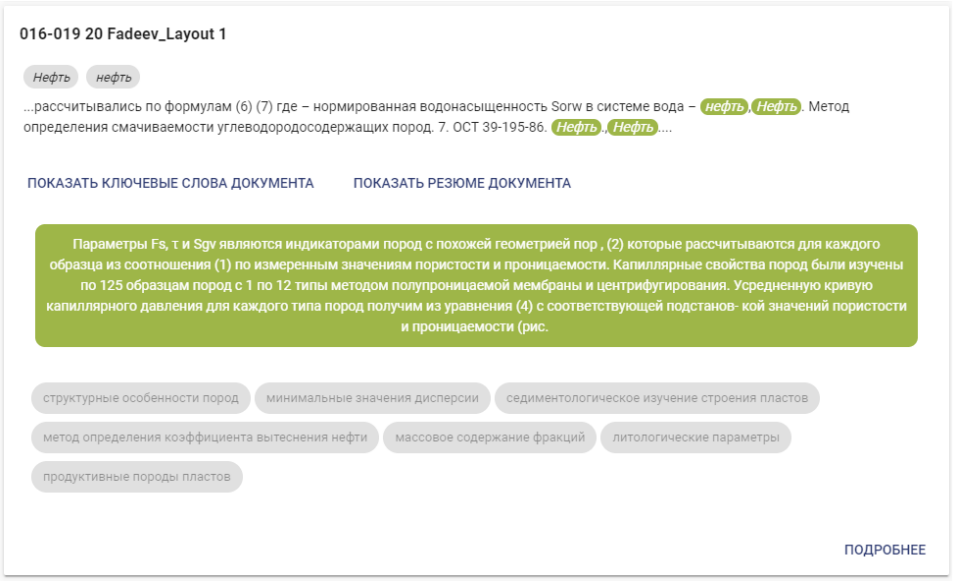

Рисунок 27 – Представление документа в поисковой выдаче

Для получения подробной информации о конкретном документе используется следующий метод:

GET
/api/document/[workspace_id]/[document_id]

Headers

Authorization: Bearer <access_token>

Параметры

workspace_id - идентификатор workspace. document_id - идентификатор документа.

Response json:

```
{"_id": "1",
"_index": "workspace_4021429bc0a4495f8f40df28e0500bbf",
"_source": {
    "attachment":
       {"content": <текст извлеченный ingest plugin>,
        "content_length": 2498,
        "content_type": "application/pdf",
        "date": "2019-02-07T07:10:25Z",
        "language": "ru"},
    "data": <исходный файл в base64>,
    "original_filename": <имя исходного файла>,
    "user_filename" : "my_file.pdf"
    "text" : <текст документа>,
    "vw" : <нормализованный текст документа в формате Vompal Wabbit>,
    "named_entities" : <результат работы Named Entity Recognition>,
    "coreferences" : <результат работы Coreference Resolution>,
    "keywords" : <список ключевых слов, выделенных Topic Modelling>},
"_type": "doc",
"_version": 1,
"found": true}
```

На рисунке 28 представлен вид страницы обзора подробной информации о документе.

Рисунок 28 – Вид страницы обзора документа

Ниже можно увидеть результаты анализа документа, а именно: резюме, ключевые слова, именованные сущности и кореференции, а также текст исходного документа. На рисунке 29 изображены ключевые слова.

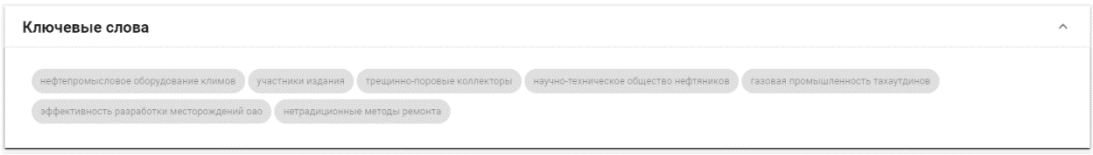

Рисунок 29 – Ключевые слова

На рисунке 30 представлены именованные сущности.

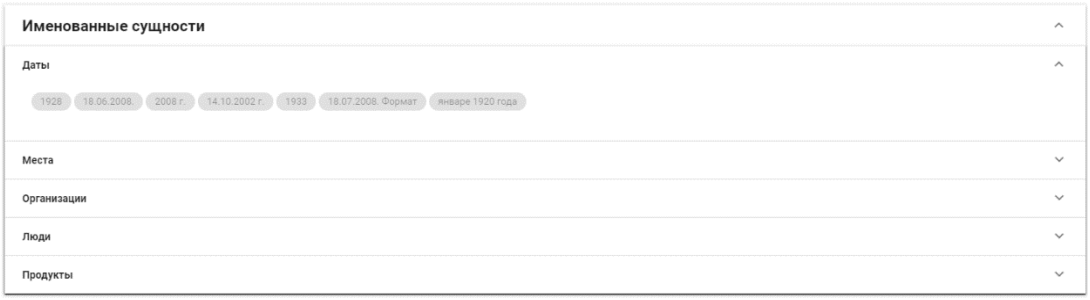

Рисунок 30 – Именованные сущности

На рисунке 31 изображены кореференции.

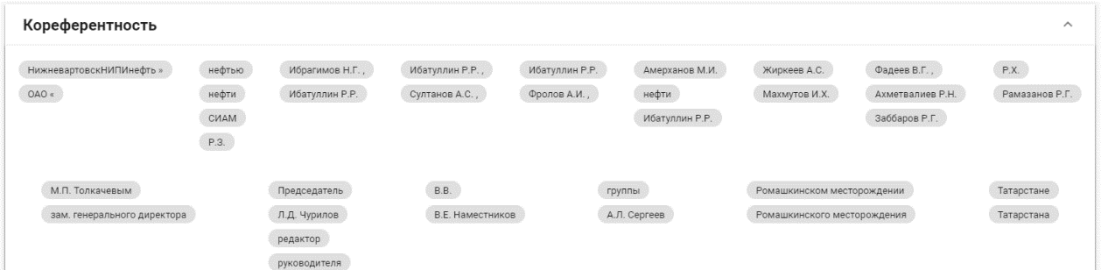

Рисунок 31 – Кореференции

На рисунке 32 представлен исходный документ.

Рисунок 32 – Исходный документ

Также существует рекомендательный блок, в котором предлагаются статьи того же кластера тематического моделирования. Рекомендательный блок изображен на рисунке 33.

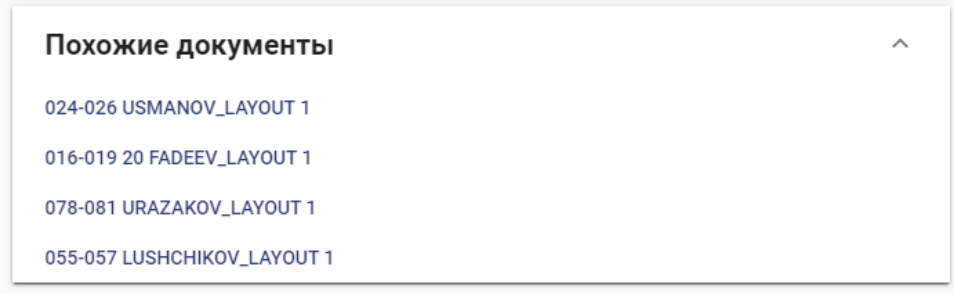

Рисунок 33 – Окно рекомендации похожих документов

Также предусмотрена возможность скачать исходный документ, с помощью кнопки “Скачать документ” и следующего метода:

GET
/api/document/[workspace_id]/[document_id]/download

Headers

Authorization: Bearer <access_token>

Параметры

workspace_id - идентификатор workspace. document_id - идентификатор документа.

Response:

Файл в двоичном представлении.

3 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

В рамках данной работы была спроектирована и разработана система NLP Service, предназначенная для анализа, поиска и обработки документов, представленных в различных форматах. Алгоритм предполагает использование современных методов выделения ключевых слов и смысла из текста.

Целью данного раздела является определение перспективности и успешности разрабатываемой системы обработки документов.

Для достижения цели были поставлены следующие задачи:

– определить потенциальных потребителей системы;

– провести анализ конкурентных решений;

– проанализировать факторы и определить связанные с ними проблемы, используя диаграмму Исикавы;

– определить цель и результат проекта, составить рабочую группу проекта, а также зафиксировать финансовые ресурсы;

– спланировать работу, распределить задачи между участниками проекта и определить трудоемкость работ для каждого исполнителя;

– сформировать бюджет проекта;

– провести анализ рисков.

3.1 Предпроектный анализ

3.1.1 Потенциальные потребители результатов исследования

Для определения потенциальных потребителей требуется определить целевой рынок и произвести его сегментирование. Для того, чтобы определить организации, которым необходима данная разработка, было проведено сегментирование целевого рынка. Сегментация проводилась по следующим критериям: размер организации и вид обработки документов. Карта сегментирования представлена в таблице 3.1.

Таблица 3.1 – Сегментация рынка

		Вид обработки документов	
		Семантический поиск	Полнотекстовый
Размер организации	Крупные	Econophysica Ltd (крупная)	
	Средние	Организация B (средняя)	
	Мелкие		Организация C (мелкая)

Таким образом, конечными потребителями являются крупные и средние предприятия, для которых характерен высокий документооборот и наличие больших массивов информации.

Примером могут являться научно-технические центры, где собраны сотни тысяч документов и поиск определенного документа может занимать до получаса времени. Если в компании работает более 100 сотрудников, имеющих дело с документами, подобная система может сократить десятки человеко-часов в неделю.

3.1.2 Анализ конкурентных технических решений

Проведем анализ конкурентного технического решения с помощью оценочной карты. В качестве конкурентных систем рассматривались решения StanfordNLP ($Б_{к1}$) и Abby Compreno ($Б_{к2}$).

StanfordNLP представляет собой библиотеку инструментов для обработки естественных языков. Данная библиотека включает в себя методы предобработки, разметки последовательностей, извлечения информации. Также, достоинством является поддержка многих языков, но в тоже время, инструментарий для русского языка очень ограничен.

Abby Compreno представляет собой решение, выполняющее комплексный анализ, включающий в себя:

– Синтаксический анализ;

– Семантический анализ.

Данная система является конечной, то есть, включает в себя пользовательский интерфейс с различным функционалом. Недостатком, в свою очередь, является ограниченный список компаний, кто может использовать

данное решение, а также более низкое качество работы алгоритмов для русского языка, по сравнению с качеством для английского языка.

Анализ представлен в таблице 3.2.

Позиция разработки и конкурентов оценивается по каждому показателю экспертным путем по пятибалльной шкале, где 1 – наиболее слабая позиция, а 5 – наиболее сильная. Веса показателей, определяемые экспертным путем, в сумме должны составлять 1.

Стоит отметить, что Abby Compreno имеет ряд преимуществ:

– Богатый функционал;

– Поддержка нескольких языков.

Недостатки:

– Закрытость для рынка (только отдельным юридическим лицам);

– Слабая поддержка русского языка.

StanfordNLP в свою очередь представляет не конечную систему, а скорее фреймворк, имеющий следующие достоинства:

– Богатый функционал;

– Часть решений является бесплатной.

Недостатки:

– Отсутствие русского языка;

– Отсутствие удобного визуального интерфейса.

Прежде, чем представить результаты анализа, необходимо описать определения для каждого из критериев:

– Удобство в использовании определяет, насколько приложение отвечает естественности взаимодействия пользователя с системой;

– Практическая польза алгоритмов для конечного пользователя определяет, насколько алгоритмы полезны в решении проблем, которые испытывает пользователь без использования системы;

– Потребность в ресурсах и памяти определяет требования к ресурсам сервера, в частности ОЗУ, ЦПУ и другое;

– Потенциальные возможности функционального расширения системы определяются подходами, используемыми в системе. Приведем конкретный пример: в случае использования rule-based подходов есть вероятность постепенной деградации расширяемости, в то время как использование машинного обучения решает данную проблему;

– Функциональная мощность определяет, на сколько тот или иной алгоритм хорошо работает. В случае NLP алгоритмов, это возможно оценить с помощью различных метрик качества;

– Удобство графического интерфейса определяет, насколько удобно воспринимать человеку информацию, получаемую из системы;

– Актуальность на рынке определяет популярность направления и технологий, используемых в системе;

– Послепродажное обслуживание определяет возможности трудностей в использовании и, соответственно, последующим обслуживанием системы;

– Открытость на рынке определяет возможности покупки системы любому лицу.

Таблица 3.2 – Результаты оценки конкурентных систем анализа документов

Критерии оценки	**Вес крите-рия**	**Баллы**			**Конкурентоспособность**		
		$Б_ф$	**$Б_{к1}$**	**$Б_{к2}$**	**$К_ф$**	**$К_{к1}$**	**$К_{к2}$**
1	2	3	4	5	6	7	8
Технические критерии оценки ресурсоэффективности							
1. Удобство в использовании	0,15	5	3	2	0,75	0,45	0,3
2. Практическая польза алгоритмов для конечного пользователя	0,15	5	3	3	0,75	0,45	0,45
3. Потребность в ресурсах памяти	0,03	5	4	5	0,15	0,12	0,15

Продолжение таблицы 3.2

4. Потенциальные возможности функционального расширения системы	0,18	4	5	4	0,72	0,9	0,72
5. Функциональная мощность	0,09	5	3	3	0,45	0,27	0,27
6. Удобство графического интерфейса	0,05	5	2	3	0,25	0,1	0,15
Экономические критерии оценки ресурсоэффективности							
1. Актуальность на рынке	0,04	2	3	4	0,08	0,12	0,16
2. Цена	0,2	4	4	3	0,8	0,8	0,6
3. Послепродажное обслуживание	0,04	5	4	4	0,2	0,16	0,16
4. Открытость на рынке	0,02	4	4	4	0,08	0,08	0,08
5. Требование сертификации разработки	0,05	5	5	5	0,25	0,25	0,25
Итого	**1**				**4,48**	**3,7**	**3,29**

Из таблицы сравнения собственного сервиса обработки документов, с сервисами StanfordNLP и Abby Compreno, можно сделать вывод, что уязвимость конкурентных решений связана неудобством эксплуатации, ограниченным функционалом и пониженной производительностью труда, с точки зрения анализа и обработки документов.

Таким образом, реализация собственного сервиса является более предпочтительнее, чем использование стороннего существующего решения.

Разработанный сервис, в отличии от представленных для сравнения, является бесплатной полноценной системой с поддержкой русского языка и пользовательским интерфейсом. При ее проектировании и создании применялись современные методы для анализа документов, с применением

технологии машинного обучения. Кроме того, система имеет возможность самообучения при ее использовании, что повышает качество результата обрабатываемого документа.

3.1.3 Диаграмма Исикавы

Диаграмма причины-следствия Исикавы (Cause-and-Effect-Diagram) – это графический метод анализа, формирования причинно-следственных связей между факторами и проблемами в исследуемой проблеме или ситуации. Диаграмма представлена на рисунке 23.

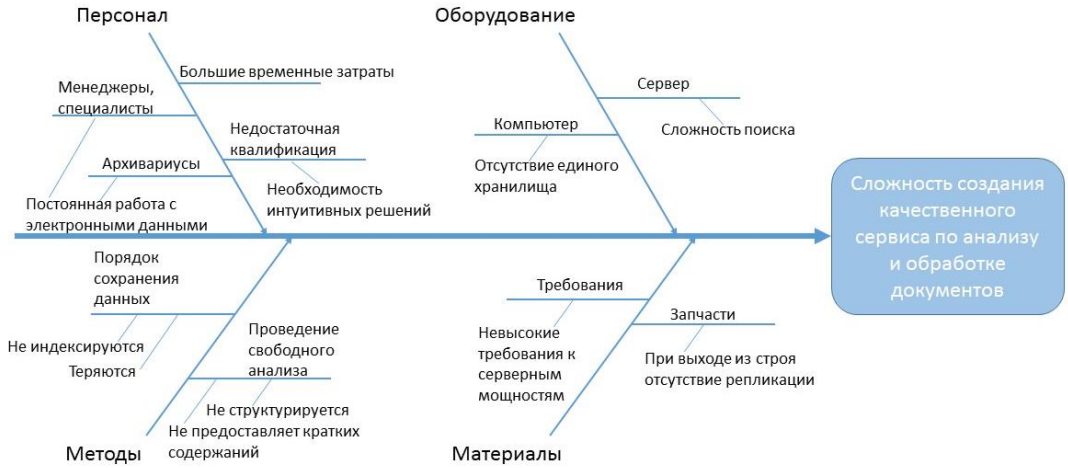

Рисунок 34 – Диаграмма Исикавы

В диаграмме Исикавы, представленной выше представлены основные проблемы, касающиеся четырех основных факторов: персонал, оборудование, методы, материалы. Стоит отметить, что сервис анализа и обработки документов на сегодняшний день не является чем-то популярным в силу того, что технологии в данной области еще не являются совершенными, поэтому обучение персонала и объяснение принципов систем подобного рода является нетривиальной задачей.

3.2 Инициация проекта

В процессе инициации проекта определяются начальные цели и содержание, а также фиксируются финансовые ресурсы. Определяются внутренние и внешние заинтересованные стороны проекта, которые будут взаимодействовать и влиять на общий результат научного проекта.

3.2.1 Цели и результат проекта

В данном разделе приведена информация о заинтересованных сторонах проекта, целях проекта и критериях их достижения. Информация по заинтересованным сторонам представлена в таблице 3.3.

Таблица 3.3 – Заинтересованные стороны проекта

Заинтересованные стороны	**Ожидание сторон**
Организация-заказчик	Готовое решение по анализу и поиску документов
Научный руководитель, студент	Готовая магистерская диссертация
Пользователь	Повышение удобства и снижение времени работы с документами

В таблице 3.4 представлена иерархия целей проекта и критерии достижения данных целей.

Таблица 3.4 – Цели и результат проекта

Цели проекта:	Разработка системы семантического поиска и анализа документов
Ожидаемые результаты:	Разработка и внедрение Document Search and Analysis System внутри организации-заказчика
Критерии приемки результата проекта:	Сервис реализован и работает Алгоритмы работают для русского языка и показывают адекватный результат: – По метрикам качества (precision, recall, f1) при наличии ground truth; – По человеческой оценке.

Продолжение таблицы 3.4

Требования к результату проекта:	1. Выполнены все пункты технического задания; 2. Разработанный функционал полностью соответствует проектным решениям.

3.2.2 Организационная структура проекта

В таблице 3.5 представлена информация о рабочей группе проекта.

Таблица 3.5 – Рабочая группа проекта

№	ФИО, место работы, должность	Роль в проекте	Функции
1	Кульневич Алексей Дмитриевич, Econophysica Ltd, инженер по анализу данных	Инженер	1. Проектирование 2. Реализация 3. Внедрение
2	Зюбин Сергей Александрович, Econophysica Ltd, Head Of Practice	Руководитель проекта от организации	1. Проверка разработки 2. Помощь во внедрении
3	Губин Евгений Иванович, ТПУ, кандидат физико-математических наук	Научный руководитель	1. Составление научных целей и задач 2. Проверка документации

3.2.3 Ограничения и допущения проекта

Ограничения проекта – это все факторы, которые могут послужить ограничением степени свободы участников команды проекта, а также «границы проекта» - параметры проекта или его продукта, которые не будут реализованы в рамках данного проекта. Ограничения проекта представлены в таблице 3.6.

Таблица 3.6 – Ограничения проекта

Фактор	Ограничения
Бюджет проекта	250000 рублей
Источник финансирования	Econophysica Ltd
Сроки проекта	29.01.2019-01.06.2019
Дата утверждения плана управления проектом	29.01.2019
Дата завершения проекта	01.06.2019

В данном разделе были рассмотрены цели и ожидаемый результат, организационная структура, а также ограничения и допущения проекта. К ограничениям относятся бюджет и сроки.

3.3 Планирование проекта

3.3.1 Структура работ в рамках проекта

В данном разделе производится декомпозиция задач и разграничение зоны ответственности между участниками проекта.

В таблице 3.7 представлено распределение исполнителей по видам работ.

Таблица 3.7 – Распределение задач

Основные этапы	**№**	**Содержание работ**	**Должность исполнителя**
Постановка задач	1	Описание требований	Руководитель от организации, научный руководитель
	2	Анализ предметной области	Руководитель от организации, научный руководитель, инженер
	3	Разработка технического задания	Руководитель от организации, инженер
Проектирование	4	Разработка архитектуры сервиса	Руководитель от организации, инженер
Разработка	5	Разработка аналитического модуля системы	Инженер
	6	Разработка веб-сервиса	Инженер
	7	Разработка модуля обработки входных данных	Инженер
Отладка решения и внедрение	8	Развертывание системы	Инженер
	9	Внедрение разработки	Руководитель от организации, инженер
Оформление документации	10	Написание отчетной документации	Инженер
	11	Проверка работы	Научный руководитель

3.3.2 Определение трудоемкости выполнения работ

Для определения ожидаемых сроков выполнения проекта необходимо оценить его трудоемкость. Воспользуемся формулой:

$$t_{\text{ож}i} = \frac{3 * t_{\min i} + 2 * t_{\max i}}{5} \quad (10)$$

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

t_{mini} – минимально возможная трудоемкость выполнения заданной *i*-ой работы (оптимистическая оценка: в предположении наиболее благоприятного стечения обстоятельств), чел.-дн.;

t_{maxi}– максимально возможная трудоемкость выполнения заданной *i*-ой работы (пессимистическая оценка: в предположении наиболее неблагоприятного стечения обстоятельств), чел.-дн.

Исходя из ожидаемой трудоемкости работ, определяется продолжительность каждой работы в рабочих днях Tpi, учитывающая параллельность выполнения работ несколькими исполнителями:

$$t_{pi} = \frac{t_{\text{ож}i}}{\text{Ч}_i} \quad (11)$$

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

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

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

Для удобства составления календарного плана и графика работ необходимо перевести длительность каждого из этапов из рабочих дней в календарные дни. Для этого воспользуемся следующей формулой:

$$T_{\text{к}i} = T_{pi} * k_{\text{кал}} \quad (12)$$

где $t_{кi}$ – продолжительность выполнения i-й работы в календарных днях;

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

$k_{кал}$ –коэффициент календарности.

Коэффициент календарности определяется по следующей формуле:

$$\text{Ткал} = \frac{\text{Т}_{\text{кал}}}{\text{Т}_{\text{кал}} - \text{Т}_{\text{вых}} - \text{Т}_{\text{пр}}} = \frac{365}{365 - 66} = 1{,}22 \quad (13)$$

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

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

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

В соответствии с производственным календарем (для 6-дневной рабочей недели) в 2019 году 365 календарных дней, 299 рабочих дней, 66 выходных/праздничных дней. В таблице 3.8 представлены подробные временные расчеты этапов отдельных видов работ.

Таблица 3.8 – Временные показатели проведения научного исследования

Наименование работы	**Исполнители работы**	**Трудоемкость работ, чел-дни**			**Длительность работ, дни**	
		t_{min}	t_{max}	$t_{ож}$	$T_р$	$T_к$
Описание требований	Руководитель от организации	1	2	1,4	2	3
	Научный руководитель	1	2	1,4	2	3
Анализ предметной области	Руководитель от организации	2	3	2,4	3	4
	Научный руководитель	1	2	1,4	2	3
	Инженер	5	8	6,2	7	9
Разработка технического задания	Руководитель от организации	1	2	1,4	2	3
	Инженер	3	5	3,8	4	5
Разработка архитектуры сервиса	Руководитель от организации	2	4	2,8	3	4
	Инженер	12	15	13,2	14	18
Разработка аналитического модуля системы	Инженер	25	28	26,2	27	33
Разработка веб-сервиса	Инженер	7	9	7,8	8	10
Разработка модуля обработки входных данных	Инженер	7	9	7,8	8	10
Развертывание системы	Инженер	2	3	2,4	3	4
Внедрение разработки	Руководитель от организации	1	2	1,4	2	3
	Инженер	2	3	2,4	3	4

Продолжение таблицы 3.8

Написание отчетной документации	Инженер	45	55	49	49	60
Проверка работы	Научный руководитель	1	2	1,4	2	3

Для иллюстрирования календарного плана-графика используется диаграмма Ганта, где представлено наглядное отображение графика и распределения работ между участниками проекта. Диаграмма представлена в таблице 3.9.

Таблица 3.9 – Календарный график проекта

№	Содержание работ	Должность исполнителя	Продолжительность выполнения работ				
			Февр.	Март	Апрель	Май	Июнь
1	Описание требований	РО, НР					
2	Анализ предметной области	РО, НР, И					
3	Разработка технического задания	РО, И					
4	Разработка архитектуры сервиса	РО, И					
5	Разработка аналитического модуля системы	И (50%)					
6	Разработка веб-сервиса	И (50%)					
7	Разработка модуля обработки входных данных	И (50%)					
8	Развертывание системы	И (50%)					
9	Внедрение разработки	РО, И (50%)					
10	Написание пояснительной документации	И (50%)					
11	Проверка работы	НР					

3.4 Бюджет проекта

Бюджет научно-технического исследования включает в себя стоимость всех расходов, необходимых для выполнения работ по магистерской диссертации. По окончании формирования бюджета, были выявлены следующие группы затрат:

- основная заработная плата исполнителей;
- дополнительная заработная плата исполнителей;
- отчисления во внебюджетные фонды (страховые отчисления);
- амортизация;
- накладные расходы.

3.4.1 Расчет амортизации

При расчете материальных затрат учитывались вся стоимость материалов, которая понадобилась на выполнение данной магистерской работы. В ходе выполнения работы был использован ПК организации. Срок полезного пользования для офисных машин составляет 2-3 года, компьютер взят на 3 года. Стоимость ПК составляет 40000 рублей. Написание работы составляет 4 месяца. Тогда норма амортизации рассчитывается следующим образом:

$$A_n = \frac{1}{3} \times 100\% = 33.33\% \quad (14)$$

Годовые амортизационные отчисления составляют:

$$A_r = 40000 \times 0.33 = 13200 \text{ рублей} \quad (15)$$

Ежемесячные амортизационные отчисления составляют:

$$A_m = \frac{13200}{12} = 1100 \text{ рублей} \quad (16)$$

Итоговая сумма амортизации основных средств составляет:

$$A = 1100 \times 4 = 4400 \text{ рублей} \quad (17)$$

3.4.2 Расчет основной заработной платы исполнителей

Данный раздел включает в себя заработную плату научного руководителя, руководителя от организации и инженера. Расчет выполняется на основе трудоемкости выполнения каждого этапа и величины месячного оклада исполнителя.

Трудоемкость исполнителей на разных стадиях выполнения магистерской диссертации была просуммирована и представлена в виде количества дней. Таким образом, если согласно плану, работа над проектом должна вестись в течение 127 дней, то реально затраченное время каждого исполнителя в днях отличается от данной цифры, а также от того, что можно увидеть из диаграммы Ганта (так как на ней декомпозиция происходит с точностью до дней, а не часов).

Для расчета основной заработной платы необходимо рассчитать среднедневную заработную плату:

$$З_{дн} = \frac{З_м \times М}{F_д} \quad (18)$$

где $З_м$ – оклад работника за месяц, руб.;

М – количество месяцев работы без отпуска в течение года, 24 дня;

$F_д$ – действительный годовой фонд рабочего времени персонала, 243 дня.

Основная заработная плата рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \times Т_р \times К_р \quad (19)$$

Расчет основной заработной платы представлен в таблице 3.10.

Таблица 3.10 – Расчет основной заработной платы

Исполнитель	**Оклад, руб./мес.**	**Средняя дневная ставка, руб.**	**$Т_р$**	**$К_р$**	**Основная заработная плата, руб.**
Научный руководитель	33664	1440,8	6	1,3	11238

Продолжение таблицы 3.10

Руководитель от организации	35000	1497,9	12		23368
Инженер	21760	931,3	74		89590,3
				Итого $С_{осн}$	124196,2

3.4.3 Расчет дополнительной заработной платы исполнителей

В данную статью включается сумма выплат, предусмотренных законодательством о труде. Например, оплата очередных и дополнительных отпусков; оплата времени, связанного с выполнением государственных и общественных обязанностей; выплата вознаграждения за выслугу лет и т. п. (в среднем — 12% от суммы основной заработной платы).

Расчеты дополнительной заработной платы представлены в таблице 3.11.

Таблица 3.11 – Расчет дополнительной заработной платы

Исполнитель	**Основная з/п, руб.**	**К.**	**Дополнительная з/п, руб.**
Научный руководитель	11238	0,12	1348,6
Руководитель от организации	23368		2804,2
Инженер	89590,3		10750,9
		Итого $С_{доп}$	14903,6

3.4.4 Расчет итоговой заработной платы исполнителей

Исходя из расчетов, обозначенных в таблицах 3.10 и 3.11, была рассчитана итоговая заработная плата исполнителей, которая представлена в таблице 3.12.

Таблица 3.12 – Итоговая заработная плата

Исполнитель	**Основная з/п, руб**	**Дополнительная з/п, руб**	**Итоговая з/п, руб**
Научный руководитель	11238	1348,6	12586,5
Руководитель от организации	23368	2804,2	26172,1

Продолжение таблицы 3.13

Инженер	89590,3	10750,9	100341,1
		Итого $C_{зп}$	139009,7

3.4.5 Расчет отчислений во внебюджетные фонды

При составлении расходов на проект, необходимо учитывать обязательные отчисления по установленным законодательством Российской Федерации нормам органам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС) от затрат на оплату труда работников, что в сумме составляет 30%.

Отчисления представлены в таблице 3.13.

Таблица 3.13 – Отчисления во внебюджетные фонды

Исполнитель	**Зарплата, руб**	**Сумма отчислений, руб**
Научный руководитель	12586,5	3776
Руководитель от организации	26172,1	7851,6
Инженер	100341,1	30102,4
	Итого $C_{внеб}$	41729,9

3.4.6 Расчет накладных расходов

В данном разделе производится расчет тех затрат, не вошедших в расчеты предыдущих расходов: канцелярия, печать, оплата электроэнергии, пользование услугой интернет. Такие расходы требуют низких затрат денежных средств относительно заработной платы исполнителей, поэтому величина коэффициента накладных расходов $k_{накл}$ была взята в размере 16%.

Расчет накладных расходов производится по формуле:

$$C_{накл} = k_{накл} \times C_{пр} = 0{,}16 \times 139009{,}7 = 29636{,}76 \quad (20)$$

где $k_{накл}$ – величина коэффициента накладных расходов;

$C_{зп}$ – сумма предыдущих затрат.

3.4.7 Формирование бюджета проекта

Исходя из произведенных расчетов, сумма всех расходов была рассчитана и представлена в таблице 3.14.

Таблица 3.14 – Бюджет затрат

Расходы	Сумма, руб	%
Основная заработная плата исполнителей	124196,2	2,05
Дополнительная заработная плата исполнителей	14903,6	57,80
Отчисления во внебюджетные фонды	41729,9	6,94
Амортизация	4400	19,42
Накладные расходы	29636,76	13,79
Итого	**214866,5**	**100**

Рассчитанный бюджет не превышает бюджета в 250000 рублей, указанного в ограничениях проекта (таблица 3.7, раздел 3.4.1).

3.5 Реестр рисков проекта

Идентифицированные риски проекта включают в себя возможные неопределенные события, которые могут возникнуть в проекте и вызвать последствия, которые повлекут за собой нежелательные эффекты. Информация о потенциальных рисках приведена в таблице 3.15.

Таблица 3.15 – Реестр рисков

№	Риск	Потенциальное воздействие	Вероятность наступления	Влияние риска	Уровень риска	Способы смягчения	Условия наступления
1	Технологическое отставание	Снижение качества работы алгоритмов по сравнению с конкурентами	1	3	высокий	Правильная архитектура системы для возможности итеративного изменения	Недостаточный анализ предметной области

Продолжение таблицы 3.15

2	Недостаточная производительность системы	Снижение конкурентоспособности по сравнению с аналогами	1	2	средний	Использование алгоритмов машинного обучения сводит риск к минимуму	Технологическое отставание алгоритмов системы
3	Узкий сегмент потребителей	Высокая конкуренция	2	3	средний	Более детальный анализ конкурентов	Реализация широко потребляемой системы
4	Увеличение требований со стороны заказчика	Появление дополнительных трудозатрат	2	1	низкий	Обсуждение проекта с заказчиком на каждом этапе сотрудничества	Ошибки в составлении технического задания

В результате идентификации рисков были рассмотрены потенциальные риски, часть которых могут привести к неконкурентоспособности разработанной системы. Однако, воздействие данных рисков возможно свести к минимуму, если следовать представленным способам смягчения.

3.6 Определение экономической эффективности исследования

По описанным критериям, разобранных в данном разделе, можно сформировать итоговый результат эффективности исследования.

Потенциальными потребителями данной работы являются организации, имеющие в своих бизнес-процессах временные затраты на документооборот: научно-технические центры, вузы.

Анализ конкурентоспособности показал, что разработанный сервис в сравнении с конкурентными решениями характеризуется бесплатной полноценной системой с поддержкой русского языка и пользовательским интерфейсом, а также возможностью самообучения при ее использовании.

Для повышения качества управления проектом работ была произведена декомпозиция задач и разграничение зоны ответственности между участниками проекта. Общая длительность исследования составила около четырех месяцев.

В результате идентификации рисков были рассмотрены потенциальные риски, часть которых могут привести к неконкурентоспособности разработанной системы. Для их устранения предложены методы смягчения рисков.

Рассчитываемая величина затрат составила 207 485 рублей. Исходя из ограничений, накладываемых на проект, максимальный бюджет не должен превышать 250000 рублей.

Для определения экономической эффективности разработки был рассчитан интегральный финансовый показатель, который определяется по следующей формуле:

$$I_{\text{фин}} = \frac{\Phi_{\text{р}}}{\Phi_{max}}, \quad (21)$$

где $I_{\text{фин}}$ – интегральный финансовый показатель разработки;

$\Phi_{\text{р}}$ – стоимость исполнения работ;

Φ_{max} – максимально допустимая стоимость исполнения проекта.

$$I_{\text{фин}} = \frac{214866{,}5}{250000} = 0{,}86 \quad (22)$$

Таким образом значения финансового показателя составляет 0,83, что свидетельствует об эффективном использовании финансовых ресурсов.

Одним из основных положительных эффектов исследования является систематический анализ, проработка методов обработки естественных языков и сбор их в сервис, что позволит в перспективе получить пользовательскую обратную связь и доработать сервис до промышленных масштабов.

4 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

4.1 Введение

Алгоритм, разработанный в процессе написания магистерской диссертации, предусматривается сервисом для анализа, обработки и поиска документов. Разработанный алгоритм позволит уменьшить время на поиск, анализ и обработку сотен тысяч документов. Предназначен для крупных предприятий, для которых характерна обработка большого количества документов и наличие больших массивов информации.

Выполняемая работа заключалась в проектировании и разработке сервиса для поиска и обработки документов. Таким образом, работу можно классифицировать как работу разработчика программного обеспечения.

Разработка осуществлялась в офисе типа Open Space, на территории работодателя, за настольным персональным компьютером. Работодателем является компания “Econophysica Ltd”, специализирующаяся в разработке программного обеспечения на заказ.

4.2 Правовые и организационные вопросы обеспечения безопасности

В процессе написания магистерской диссертации, разработка алгоритма проводилась в офисном помещении за персональным компьютером. Для комфортного времяпрепровождения на территории рабочего места, организации необходимо соблюдать ряд правил для офисного помещения и персональных компьютеров, изложенных в трудовом праве российской федерации.

Работодатели при установке ПК обязаны выполнить следующий перечень требований:

- к помещению;
- к освещению;
- к организации медицинского обследования пользователей;
- и других.

Также немаловажным фактором при выборе компьютеров для сотрудников является возможность конструкции компьютера изменять положение ПК в различных плоскостях (горизонтальные или вертикальные), с возможной устойчивой фиксацией в положении, которая удобна пользователю. Цвет корпуса ПК должен быть спокойным без блестящих деталей, которые бы вызывали повышенную утомляемость глаз. Экран монитора должен содержать регулировку яркости и контрастности, что каждый работник мог установить нужный режим, которые будет соответствовать чувствительности глаз.

Устройство рабочего стола должна быть использовано для оптимального размещения используемого оборудования. Кроме того, форма рабочего стола должна быть удобна для поддержания рациональной позы пользователя, так, чтобы он мог менять положения своего тела для предупреждения утомления. В соответствии с ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования» [20] к рабочему месту предъявляются следующие основные требования:

– при организации рабочего места следует учитывать антропометрические показатели женщин (если работают только женщины) и мужчин (если работают только мужчины); если работают и женщины, и мужчины – общие средние показатели женщин и мужчин;

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

Продолжительность рабочего дня не должна превышать 40 часов в неделю. Вид трудовой деятельности за компьютерным устройством (компьютер, мобильное устройство), в рамках выполнения выпускной квалификационной работы, соответствует группе «В» — творческая работа в режиме диалога с компьютерным устройством. Категория данной трудовой деятельности соответствует III (до 6 часов непосредственной работы за компьютером).

4.3 Профессиональная социальная ответственность

4.3.1 Анализ вредных и опасных факторов, которые может создать объект исследования

В процессе рассмотрения безопасности в рабочей зоне, необходимо было выявить вредные и опасные факторы, которые могут возникнуть на рабочем месте. Факторы считаются вредными, если его воздействие на человека может привести его к заболеванию. Опасный фактор может привести человека к травме.

Также необходимо было описать мероприятия по защите исследователя и пользователей конечного продукта от действия данных факторов.

Факторы, влияющие на работу с компьютером представлена в таблице 4.1.

Таблица № 4.1 — Негативные факторы при работе с компьютером

Наименование видов работ и параметров производственного процесса	**Факторы (ГОСТ 12.0.003-74 ССБТ)**	**Нормативные документы**
Вредные факторы		
Работа с компьютером	Повышенная или пониженная температура воздуха рабочей зоны	СанПиН 2.2.4.548-96
	Повышенная или пониженная влажность воздуха рабочей зоны	СанПиН 2.2.4.548-96
	Повышенный уровень шума на рабочем месте	СанПиН 2.2.4/2.1.8.562-96
	Недостаточная освещенность рабочей зоны	СанПиН 2.2.1/2.1.1.1278-03
	Повышенный уровень электромагнитных излучений	СанПиН 2.2.2/2.4.1340-03
Опасные факторы		
Работа с компьютером	Опасность поражения электрическим током	ГОСТ 12.1.038-82
	Пожаровзрывоопасность	ГОСТ 12.1.041-83

4.3.2 Микроклимат

Одним из необходимых условий труда является обеспечение нормального микроклимата в рабочей зоне, оказывающие значительное влияние на самочувствие человека.

Так как работа с использованием персонального компьютера является основной, (диспетчерские, операторские, расчетные, кабины и посты управления, залы вычислительной техники и др.) и связана с нервно-эмоциональным напряжением, должны быть обеспечены оптимальные параметры микроклимата для категории работ 1а и 1б в соответствии с действующими санитарно-эпидемиологическими нормативами микроклимата производственных помещений. На других рабочих местах следует поддерживать параметры микроклимата на допустимом уровне, соответствующем требованиям указанных выше нормативов.

Содержание вредных химических веществ в таких рабочих зонах не должно превышать предельно допустимых концентраций загрязняющих веществ в атмосферном воздухе населенных мест в соответствии с действующими гигиеническими нормативами. Также, в помещениях с персональными компьютерами должна проводится ежедневная влажная уборка. Нормативным документов, отвечающим за гигиенические требования к микроклимату производственных помещений, является СанПиН 2.2.4.548-96 [21]. В документе указаны все нормативные требования к микроклимату на рабочих местах, у всех видов производственных помещений.

В таблице 4.2 представлены фактические, оптимальные и допустимые показатели микроклимата рабочей зоны.

Таблица 4.2 — Параметры микроклимата на рабочем месте

Период года	**Кат. раб.**	**Температура воздуха, 0С**			**Относительная влажность воздуха, %**			**Скорость движения воздуха, м/с**		
		Факт.	Опт.	Доп.	Факт.	Опт.	Доп.	Факт.	Опт.	Доп.

Продолжение таблицы 4.2

Холодный	Ia	23	22-24	20-25	55	40-60	15-75	0,1	0,1	0,1
Теплый	Ia	23	23-25	21-28	50	40-60	15-75	0,1	0,1	0,1-0,2

Оптимальный микроклимат является необходимым на рабочих местах, так как создает комфортное нахождение человека в рабочей зоне, а также обеспечивает его высокий уровень работоспособности. Такие микроклиматические условия обеспечивают благоприятное состояние организму человека и не вызывают отклонений в состоянии его здоровья.

4.3.3 Шум

Превышение уровня шума является распространенным вредным фактором на рабочем месте. Его нарушение влечет за собой негативное воздействие не только на органы слуха, но и на весь организм человека в целом, через центральную нервную систему.

Среди источников шума выделяют принтеры, систему охлаждения, множительная техника, осветительные приборы дневного света, а также шумы, проникающие извне.

Уровень шума не должен превышать значений, установленных СанПиН 2.2.4/2.1.8.562-96 [22] и составлять не более 50 дБА. Допустимые значения уровней звукового давления представлены в таблице 4.3.

Таблица 4.3 — Допустимые значения уровней звукового давления компьютера

Уровни звукового давления в октавных полосах со среднегеометрическими частотами								**Уровни звука в дБА**
31,5 Гц	63 Гц	125 Гц	250 Гц	500 Гц	1000 Гц	2000 Гц	4000 Гц	8000 Гц	
86 дБ	71 дБ	61 ДБ	54 дБ	49 дБ	45 дБ	42 дБ	40 дБ	38 дБ	50

Снижению уровня шума способствует установка звукопоглощающих материалов (плиты, панели), подвесных акустических потолков, а также установка малошумного оборудования.

4.3.4 Освещенность

Недостаточная освещенность рабочей зоны оказывает негативное влияние на зрительную систему человека. Происходит снижение концентрации внимания, усталость центральной нервной системы, что приводит к снижению производительности труда.

Уровень освещения на поверхности рабочего стола в зоне размещения документа, согласно СанПиН 2.2.2/2.4.1340-03 [23], должен быть в диапазоне от 300 до 500 лк. Уровень освещенности экрана не должен превышать 300 лк. Яркость осветительных приборов, находящихся в поле зрения, не должна превышать 200 кд/м2.

Приведем расчет искусственного освещения для прямоугольного помещения, размерами: длина А = 5 м, ширина В = 6 м, высота Н = 4 м, количество ламп N = 12 шт.

Определим расчетную высоту подвеса светильников над рабочей поверхностью (h) по формуле:

$$h = H - h_p - h_c, \tag{23}$$

где Н – высота потолка в помещении, м;

h_p – расстояние от пола до рабочей поверхности стола, м;

h_c – расстояние от потолка до светильника, м.

Вычислим расчетную высоту подвеса светильников над рабочей поверхностью по формуле 4.1 для компьютерной аудитории кафедры программной инженерии:

$$h = 4 - 0.8 - 0.01 = 3.19 \text{ м}, \tag{24}$$

Индекс помещения определяется по формуле:

$$i = \frac{S}{h(A+B)} \tag{25}$$

где S – площадь помещения, м^2;

A – длина комнаты, м;

B – ширина комнаты, м;

h – высота подвеса светильников, м.

Индекс помещения для компьютерной аудитории кафедры программной инженерии:

$$i = \frac{30}{3.19(5+6)} = 0.83 \tag{26}$$

Исходя из того, что потолок в помещении чистый бетонный, а также свежепобеленные стены без окон, согласно методическим указаниям, примем коэффициенты отражения от стен ρ_c=*70%* и потолка $\rho_п$=*50%*. По таблице коэффициентов использования светового потока для соответствующих значений i, ρ_c, $\rho_п$, примем η=0,29.

Освещенность помещения рассчитывается по формуле:

$$E_{\text{ф}} = \frac{n\eta\Phi}{Sk_z z} \tag{27}$$

где Φ – световой поток светильника, лм;

S – площадь помещения, м^2;

$k_з$ – коэффициент неравномерности освещения;

n – число светильников;

η – коэффициент использования светового потока.

Коэффициент запаса k учитывает запыленность светильников и их износ. Для помещений с малым выделением пыли k = 1,5. Поправочный коэффициент z – это коэффициент неравномерности освещения. Для люминесцентных ламп z = 1,1. В помещении находятся светильники ЛВО 4×18 CSVT, с

люминесцентными лампами типа L 18W/640 с потоком F = 1200 лм. Учитывая все параметры, рассмотренные выше, найдем освещенность:

$$E_{\text{ф}} = \frac{48 * 0.29 * 1200}{30 * 1.5 * 1.1} = 337 \text{ лк} \quad (28)$$

В рассматриваемом помещении освещенность должна составлять 300 лк [23]. В данном помещении освещенность находится в пределах нормы, следовательно, дополнительные источники света не нужны.

Вся электротехника, и компьютеры в том числе, производят электромагнитное излучение. В таблице 4.4 представлены временные допустимые уровни электромагнитных полей, создаваемые компьютерами на рабочих местах, согласно СанПиНу 2.2.2/2.4.1340-03 [23].

Таблица 4.4 — Временные допустимые уровни электромагнитных полей

Наименование параметров		**Временные допустимые уровни электромагнитных полей**
Напряженность электрического поля	в диапазоне частот 5 Гц-2 кГц	25 В/м
	в диапазоне частот 2 кГц-400 кГц	2,5 В/м
Плотность магнитного потока	в диапазоне частот 5 Гц-2 кГц	250 нТл
	в диапазоне частот 2 кГц-400 кГц	25 нТл
Поверхностный видеомонитора	электростатический потенциал экрана	500В

4.3.5 Психофизиологические факторы

Продолжительность непрерывной работы за компьютерным устройством, без регламентированного перерыва, не должна превышать 2 часа. Длительность регламентированных перерывов составляет 20 минут (после 1,5-2,0 часа от начала рабочей смены и обеденного перерыва). Также, необходимо уделять

время нерегламентированным перерывам (микропаузы), длительность которых составляет 1-3 минуты.

Для снижения воздействия вредных факторов, устанавливаются перерывы в работе для отдыха сотрудников. Суммарное время регламентированных перерывов при работе с персональным компьютером зависит от категории трудовой деятельности и уровня нагрузки за рабочую смену. В таблице 4.5 приведено суммарное время отдыха для каждой категории работ.

Таблица 4.5 — Суммарное время перерывов

Категория работы с ПК	**Уровень нагрузки за рабочую смену при видах работ с ПК**			**Суммарное время регламентированных перерывов при 8-часовой смене, мин.**
	Группа А, количество знаков	Группа Б, количество знаков	Группа В, количество знаков	
I	до 20000	до 15000	до 2	50
II	до 40000	до 30000	до 4	70
III	до 60000	до 40000	до 6	90

4.3.6 Статическое электричество

В помещениях, оборудованных ПЭВМ, токи статического электричества чаще всего возникают при прикосновении персонала к любому из элементов ПЭВМ. Такие разряды опасности для человека не представляют, однако кроме неприятных ощущений могут привести к выходу оборудования из строя.

Для предотвращения образования и защиты от статического электричества в помещении используются нейтрализаторы и увлажнители, а полы имеют антистатическое покрытие в виде поливинилхлоридного антистатического линолеума.

Также в СанПиН 2.2.2/2.4.1340-03 установлен максимальный допустимый электростатический потенциал экрана видеомонитора – 500 В.

В качестве мер уменьшения влияния вредных факторов на пользователя используются защитные фильтры для мониторов, увлажнители воздуха. Должны

использоваться розетки с заземлением. Требуется проводить регулярную влажную уборку.

4.3.7 Электрический ток

Среди распространенных опасностей в рабочей зоне находится и поражение электрическим током. Опасность поражения определяется величиной тока проходящего через тело человека I или напряжением прикосновения U. Напряжение считается безопасным при напряжении прикосновения U < 42 В.

При получении человеком разряда электрического тока могут быть получены электротравмы, электрические удары и даже летальный исход (согласно ГОСТ 12.1.009-2009 [24]).

Для защиты от поражения электрическим током следует выполнить следующие пункты:

– обеспечить недоступность токоведущих частей от случайных прикосновений;

– электрическое разделение цепей;

– устранить опасность поражения при появлении напряжения на разных частях.

В таблице 4.6 представлены предельно допустимые значения напряжения прикосновения и тока на рабочем месте (согласно ГОСТ 12.1.038-82 [25]).

Таблица 4.6 — Допустимые значения напряжения прикосновения и тока

Род тока	**Напряжения прикосновения, В**	**Ток, мА**
	Не более	
Переменный, 50 Гц	2,0	0,3
Постоянный	8,0	1,0

Согласно электробезопасности, рабочее место относится к помещениям без повышенной опасности поражения электрическим током. Данный фактор

характеризуется отсутствием условий, создающих повышенную или особую безопасность от разряда электрическим током.

4.4 Экологическая безопасность

Экологическая безопасность и охрана окружающей среды являются одними из важнейших факторов при выполнении работ любого характера. При работе в офисном помещении за персональным компьютером отсутствуют выбросы в окружающую среду и нет влияния на жилищную зону.

Так как при разработке данной магистерской диссертации использовался персональный компьютер, необходимо описать правильную утилизацию компьютерного лома после его выхода из строя. В соответствии с постановлением правительства юридическим лицам запрещено самостоятельно утилизировать компьютерную технику. Для этого необходимо найти специальную компанию, которая занимается утилизацией в частном порядке.

В нормативном документе СанПиН 2.2.2/2.4.1340-03 [23], даются следующие общие рекомендации по снижению опасности для окружающей среды, исходящей от компьютерной техники:

- применять оборудование, соответствующее санитарным нормам и стандартам экологической безопасности;
- применять расходные материалы с высоким коэффициентом использования и возможностью их полной или частичной регенерации;
- отходы в виде компьютерного лома утилизировать;
- использовать экономичные режимы работы оборудования.

4.5 Безопасность при чрезвычайных ситуациях

4.5.1 Анализ вероятных чрезвычайных ситуаций

Самым распространенным чрезвычайным обстоятельством в офисе является пожар. Такое рабочее место относится к категории “В”

(пожароопасные), так как в данном помещении присутствует пыль, вещества и материалы, способные при взаимодействии с воздухом гореть. Возникновение пожара может произойти по нескольким факторам:

- возникновением короткого замыкания в электропроводке вследствие неисправности самой проводки или электросоединений и электрораспределительных щитов;
- возгоранием устройств вычислительной аппаратуры вследствие нарушения изоляции или неисправности самой аппаратуры;
- возгоранием мебели или пола по причине нарушения правил пожарной безопасности, а также неправильного использования дополнительных бытовых электроприборов и электроустановок;
- возгоранием устройств искусственного освещения.

4.5.2 Мероприятия по предотвращению чрезвычайных ситуаций и порядок действия в случае возникновения чрезвычайных ситуаций.

Для устранения возможных причин возгорания следует проводить следующие мероприятия:

1. Организация мероприятий:
   - противопожарный инструктаж обслуживающего персонала;
   - обучение персонала техники безопасности;
   - разработка инструкций, планов эвакуаций и пр.
2. Эксплуатационные мероприятия:
   - соблюдение эксплуатационных норм оборудования;
   - выбор и использование современных автоматических средств пожаротушения.
3. Технические мероприятия:
   - профилактический осмотр и ремонт оборудования;
   - соблюдение противопожарных мероприятий при устройстве электропроводок, оборудования, систем отопления и пр. [7]

ЗАКЛЮЧЕНИЕ

Началом работы стало изучение современных решений в области обработки текстовых данных с применением современных подходов, в частности, машинного обучения. После было определено, что данная отрасль на данный момент находится в достаточно зачаточном состоянии по сравнению, например, с использованием машинного обучения в задачах компьютерного зрения.

Далее были изучены основные направления в обработке естественных языков, в частности:

– разметка последовательностей: извлечение именованных сущностей, разрешение кореференции, частеречная разметка;

– извлечение ключевой информации: извлечение ключевых фраз и автореферирование текста;

– тематическое моделирование.

После проведенного исследования в области существующих задач каждого из направлений, был выделен список задач, требующих решения для реализации сервиса, позволяющего проводить анализ и поиск текстовой информации с учетом семантики содержимого.

Качество алгоритмов, где была в наличии ground truth выборка, было протестировано, как пример, извлечение именованных сущностей было проверено с помощью f1-оценки.

На основе разработки аналитических алгоритмов для работы с русским языком, был реализован веб сервис, включающий в себя основные компоненты:

– база данных;

– модуль обработки данных;

– frontend-часть;

– backend-часть.

Следует отметить, что это исследование является стартовой точкой и требует существенных доработок. В частности, требуется более глубокое исследование семантического поиска.

Одной из идей по его улучшению является апробирование варианта векторизации запросов и ответов и возможности их ранжирования на основе векторной близости. Трудностью по получению устойчивого результата является невозможность количественно оценить качество алгоритма на данном этапе развития, поскольку хороший результат в данном случае – это субъективная оценка.

СПИСОК ПУБЛИКАЦИЙ СТУДЕНТА

1. Закиров А.Р., Кирьянов Е.Л., Буханов Н.В., Белозеров Б.В. (ООО «Газпромнефть НТЦ»), Кульневич А.Д., Чугунов Р.А. (Компания «Эконофизика»), Сливкин С.С. (Томский политехнический университет). Когнитивные технологии исследования информационных массивов для восстановления неявных знаний и данных.

2. Кульневич А.Д., Радишевский В.Л. Machine learning for natural language processing tasks // "Distributed Computing and Grid-Technologies in Science and Education" GRID 2018, Book of Abstracts, Дубна, 10-14 сентября 2018 - С. 139.

3. Радишевский В.Л., Кульневич А.Д. Botnet in PyPy to speed up the work of the Earley parser // "Distributed Computing and Grid-Technologies in Science and Education" GRID 2018, Book of Abstracts, Дубна, 10-14 сентября 2018 - С. 139.

4. Кульневич А. Д., Сергеева Н. Д., Чугунов Р. А. Система раннего детектирования пневмонии на основе методов глубокого обучения // Вестник Амурского государственного университета. Серия: Естественные и экономические науки. - 2018. - Вып. 83. - С. 35-40.

5. Kulnevich A. D., Radishevsky V. L., Chugunov R. A., Shevchuk A. A. Application of russian named entity recognition and coreference resolution in the oil industry // CEUR Workshop Proceedings. - 2018 - Vol. 2267. - p. 378-382.

6. Radishevsky V. L., Kulnevich A. D., Chugunov R. A., Shevchuk A. A. Distributed GLR-parser for Natural Languag Processing // CEUR Workshop Proceedings. - 2018 - Vol. 2267. - p. 374-377.

7. Кульневич А. Д. Введение в нейронные сети //Молодой ученый. – 2017. – №. 8. – С. 31-36.

8. Сергеева Н. Д., Кульневич А. Д., Чугунов Р. А. Автоматизированная система учета рабочего времени // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно- практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 417-418.

9. Чугунов Р. А., Кульневич А. Д. Разработка информационной системы поддержки междисциплинарных курсовых проектов // Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов, аспирантов и молодых ученых, Томск, 3-7 Декабря 2018. - Томск: ТПУ, 2019 - С. 385-386.

10. Кульневич А. Д., Радишевский В. Л. Интеллектуальный анализ аэрофотоснимков // Молодежь и современные информационные технологии: сборник трудов XV Международной научно- практической конференции студентов, аспирантов и молодых ученых, Томск, 4-7 Декабря 2017. - Томск: ТПУ, 2018 - С. 77-78.

11. Радишевский В. Л., Кульневич А. Д. Распределенный брокер сообщений KAFKA для высокоскоростной передачи и агрегации данных // Молодежь и современные информационные технологии: сборник трудов XV Международной научно- практической конференции студентов, аспирантов и молодых ученых, Томск, 4-7 Декабря 2017. - Томск: ТПУ, 2018 - С. 284-285.

12. Кульневич А. Д., Радишевский В. Л. Интеллектуальный анализ аэрофотоснимков // Молодёжь и современные информационные технологии. Сборник трудов XV Международной научно-практической конференции студентов, аспирантов и молодых учёных (4-7 декабря 2017 г). – С. 77-78.

13. Радишевский В. Л., Кульневич А. Д. Распределенный брокер сообщений KAFKA для высокоскоростной передачи и агрегации данных // Молодёжь и современные информационные технологии. Сборник трудов XV Международной научно-практической конференции студентов, аспирантов и молодых учёных (4-7 декабря 2017 г). – С. 286-287.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Industrial-Strength Natural Language Processing in Python. Режим доступа: https://spacy.io/ [Электронный ресурс].
2. Natural Language Toolkit documentation. Режим доступа: https://www.nltk.org/ [Электронный ресурс].
3. StanfordNLP 0.2.0 - Python NLP Library for Many Human Languages. Режим доступа: https://stanfordnlp.github.io/stanfordnlp/ [Электронный ресурс].
4. Морфологический анализатор pymorphy2. Режим доступа: https://pymorphy2.readthedocs.io/en/latest/ [Электронный ресурс].
5. Zamgi. Режим доступа: https://github.com/zamgi [Электронный ресурс].
6. ABBYY Compreno. Режим доступа: https://www.abbyy.com/ru-ru/science/technologies/compreno/ [Электронный ресурс].
7. Konkol M. Named Entity Recognition //PhD Study Report/ – 2012. – P. 29.
8. Rule-based named entity recognition library for Russian language. Режим доступа: https://github.com/natasha/natasha [Электронный ресурс].
9. Ratnaparkhi A. A Maximum Entropy Model for Part-Of-Speech Tagging //Conference on Empirical Methods in Natural Language Processing/ – 1996. – P. 133-142.
10. Hochreiter S. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems/ – 1998. – P. 10.
11. Lee K., He L., Lewis M., Zettlemoyer L. End-to-end Neural Coreference Resolution //Empirical Methods in Natural Language Processing/ – 2017. – P. 188-197.
12. Mihalcea R., Tarau P. TextRank: Bringing Order into Texts // Association for Computational Linguistics/ – 2004. – P. 404-411.

13. Bennani-Smires K., Musat C., Hossmann A., Baerswyl M., Jaggi M. Simple Unsupervised Keyphrase Extraction using Sentence Embeddings // Association for Computational Linguistics/ – 2018. – P. 221-229.

14. Воронцов К. В. Обзор вероятностных тематических моделей. Режим доступа: https://is.gd/almt5W [Электронный ресурс].

15. Vorontsov K., Frei O., Apishev M., Romov P., Dudarenko M. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections //Communications in Computer and Information Science/ – 2015. – Vol. 542. – P. 370-381

16. Gormley C., Tong Z. Elasticsearch: The Definitive Guide // O'Reilly Media/ – 2015. – P. 665.

17. Vue.JS. Режим доступа: https://ru.vuejs.org/index.html [Электронный ресурс].

18. Flask. Режим доступа: http://flask.pocoo.org/ [Электронный ресурс].

19. Introduction to JSON Web Tokens. Режим доступа: https://jwt.io/introduction/ [Электронный ресурс].

20. ГОСТ 12.2.032-78 ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования.

21. СанПиН 2.2.4.548–96. Гигиенические требования к микроклимату производственных помещений.

22. СанПиН 2.2.4/2.1.8.562–96. Шум на рабочих местах, в помещениях жилых, общественных зданий и на территории застройки.

23. СанПиН 2.2.2/2.4.1340–03. Санитарно-эпидемиологические правила и нормативы «Гигиенические требования к персональным электронновычислительным машинам и организации работы».

24. ГОСТ 12.1.009-2009. Система стандартов безопасности труда. Электробезопасность.

25. ГОСТ 12.1.038-82 ССБТ. Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов.

26. ГОСТ Р 22.3.03-94. Безопасность в ЧС. Защита населения. Основные положения.