Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Школа Инженерная информационных технологий и робототехники
Направление подготовки 09.04.01 Информатика и вычислительная техника
Отделение школы (НОЦ) Отделение информационных технологий

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

Тема работы
Разработка системы умного поиска с генерацией обзорных научных статей

УДК 004.89:001.814.2

Студент

Группа	ФИО	Подпись	Дата
8ВМ03	Хайров Марк Альбертович		

Руководитель ВКР

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОИТ	Иванова Ю.А.	к.т.н., доцент		

КОНСУЛЬТАНТЫ:

По разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент ОСГН	Былкова Т.В.	к.э.н.		

По разделу «Социальная ответственность»

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ООД ШБИП	Федоренко О.Ю.	д.м.н.		

ДОПУСТИТЬ К ЗАЩИТЕ:

Руководитель ООП	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ОИТ	Спицын В.Г.	д.т.н, профессор		

Томск – 2022 г

ПЛАНИРУЕМЫЕ РЕЗУЛЬТАТЫ ОСВОЕНИЯ ООП

Код компетенции	Наименование компетенции
Универсальные компетенции	
УК(У)-1	УК(У)-1. Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий
УК(У)-2	УК(У)-2. Способен управлять проектом на всех этапах его жизненного цикла
УК(У)-3	УК(У)-3. Способен организовывать и руководить работой команды, вырабатывая командную стратегию для достижения поставленной цели
УК(У)-4	УК(У)-4. Способен применять современные коммуникативные технологии, в том числе на иностранном (-ых)языке (-ах), для академического и профессионального взаимодействия
УК(У)-5	УК(У)-5. Способен анализировать и учитывать разнообразие культур в процессе межкультурного взаимодействия
УК(У)-6	Способен определить и реализовать приоритеты собственной деятельности и способы ее совершенствования на основе самооценки
Общепрофессиональные компетенции	
ОПК(У)-1	Способен самостоятельно приобретать, развивать и применять математические, естественно-научные, социально-экономические и профессиональные знания для решения нестандартных задач, в том числе в новой или незнакомой среде и в междисциплинарном контексте
ОПК(У)-2	Способен разрабатывать оригинальные алгоритмы и программные средства, в том числе с использованием современных интеллектуальных технологий, для решения профессиональных задач
ОПК(У)-3	Способен анализировать профессиональную информацию, выделять в ней главное, структурировать, оформлять и представлять в виде аналитических обзоров с обоснованными выводами и рекомендациями
ОПК(У)-4	Способен применять на практике новые научные принципы и методы исследований
ОПК(У)-5	Способен разрабатывать и модернизировать программное и аппаратное обеспечение информационных и автоматизированных систем

ОПК(У)-6	Способен разрабатывать компоненты программно-аппаратных комплексов обработки информации и автоматизированного проектирования
ОПК(У)-7	Способен адаптировать зарубежные комплексы обработки информации и автоматизированного проектирования к нуждам отечественных предприятий
ОПК(У)-8	Способен осуществлять эффективное управление разработкой программных средств и проектов
Профессиональные компетенции	
ПК(У)-1	Способен к созданию программного обеспечения для анализа, распознавания и обработки информации, систем цифровой обработки сигналов
ПК(У)-2	Способен проектировать сложные пользовательские интерфейсы
ПК (У)-3	Способен управлять процессами и проектами по созданию (модификации) информационных ресурсов
ПК (У)-4	Способен осуществлять руководство разработкой комплексных
ПК(У)-5	Способен проектировать и организовывать учебный процесс по образовательным программам с использованием современных образовательных технологий

Министерство образования и науки Российской Федерации
федеральное государственное автономное образовательное учреждение
высшего образования
**«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ»**

Инженерная школа информационных технологий и робототехники
Направление подготовки (специальность) 09.04.01 «Информатика и вычислительная техника»
Отделение информационных технологий

УТВЕРЖДАЮ:
Руководитель ООП
_____ _____ Спицын В.Г.
(Подпись) (Дата) (Ф.И.О.)

ЗАДАНИЕ
на выполнение выпускной квалификационной работы

В форме:

магистерской диссертации

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

Студенту:

Группа	ФИО
8ВМ03	Хайров Марк Альбертович

Тема работы:

Получение коллективного концентрата редкоземельных металлов из монацита	
Утверждена приказом директора (дата, номер)	03.02.2022 № 34-63/с

Срок сдачи студентом выполненной работы:	

ТЕХНИЧЕСКОЕ ЗАДАНИЕ:

Исходные данные к работе *(наименование объекта исследования или проектирования; производительность или нагрузка; режим работы (непрерывный, периодический, циклический и т. д.); вид сырья или материал изделия; требования к продукту, изделию или процессу; особые требования к особенностям функционирования (эксплуатации) объекта или изделия в плане безопасности эксплуатации, влияния на окружающую среду, энергозатратам; экономический анализ и т. д.).*	Набор статей из журнала «Journal of Membrane Science» в формате pdf с наличием текстового слоя
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).*	Реферат; Введение; 1. Обзор литературы; 2. Извлечение данных из научных статей; 3. Разработка системы поиска; 4. Финансовый менеджмент, ресурсоэффективность и ресурсосбережение; 5. Социальная ответственность; Выводы; Список использованных источников.

Перечень графического материала *(с точным указанием обязательных чертежей)*	

Консультанты по разделам выпускной квалификационной работы
(с указанием разделов)

Раздел	Консультант
Финансовый менеджмент, ресурсоэффективность и ресурсосбережение	Былкова Татьяна Васильевна
Социальная ответственность	Федоренко Ольга Юрьевна

Дата выдачи задания на выполнение выпускной квалификационной работы по линейному графику	24.01.2022

Задание выдал руководитель:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Доцент	Иванова Ю.А.	к.т.н., доцент		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ03	Хайров Марк Альбертович		

РЕФЕРАТ

Выпускная квалификационная работа содержит 123 страниц, 23 рисунков, 37 таблиц, 90 литературных источников, 1 приложение.

Ключевые слова: автоматическая суммаризация, научные статьи, word2vec, кластеризация, Longformer, поиск по тексту.

Объектом исследования является система автоматизированного научного поиска. Предметами исследования выступают алгоритмы, лежащие в основе системы поиска, в том числе система автоматической суммаризации научных статей.

Цель работы – разработка поисковой системы для автоматизации научного поиска.

В исследовании представлен обзор существующих технологий автоматической генерации рефератов по научным статьям, предложен способ извлечения данных из файлов научных статей алгоритм поиска с генерацией обзорного текста по запросу.

Область применения: научно-исследовательские и конструкторские работы.

Экономическая эффективность проекта достигается за счёт выбора наиболее простых для реализации и эффективных алгоритмов.

Разработаны меры по снижению влияния вредных и предотвращения воздействия опасных факторов. Разработка оказывает вред окружающей среде только с точки зрения образования твёрдых отходов из люминесцентных ламп и комплектующих оргтехники. Разработаны меры предотвращения пожара и правил поведения во время пожара.

В дальнейшем планируется расширение базы знаний, оптимизация разработанных алгоритмов и интеграция табличных данных и рисунков для генерации обзорной научной статьи по теме.

Оглавление

ВВЕДЕНИЕ

Исследовании научной литературы является достаточно долгим и трудоёмким процессом, а объём научных работ не перестаёт возрастать. Автоматизация поиска и анализа научной литературы позволит несколько ускорить НИОКР и упростить задачи, связанные с анализом научной литературы.

Для решения данной проблемы в работе предлагается использовать поиск среди множества документов по контексту на основе векторного представления слов и биграмм с получением реферата по заданной теме на основе найденных релевантных документов.

Технологии контекстного поиска и нашли широкое применение на сегодняшний день. Использования контекста запроса для поиска информации используется в вопросно-ответных системах, суммаризация используется в некоторых агрегаторах новостей. Существуют системы для поиска и/или суммаризации научных статей, но в основном для работ из областей компьютерной лингвистики или медицины. В данной работе предлагается экстраполяция различных методов обработки и понимания текста на слабо изученную предметную область – научные работы по темам химии и химической технологии.

Целью данного проекта является создание системы научного поиска с функцией генерации обзорной научной статьи.

Для достижения поставленной цели были выработаны следующие задачи:

1. анализ существующих решений автоматической суммаризации научных статей;
2. извлечение данных из научных статей для формирования базы знаний;
3. построение алгоритма поиска текстов;
4. реализация алгоритма суммаризации научных статей.

1 ОБЗОР ЛИТЕРАТУРЫ

1.1 Суммаризация

Задача автоматической суммаризации текста состоит в том, чтобы уменьшить объём исходного текста, при этом представив ключевую информацию [14].

1.1.1 Виды суммаризации

С точки зрения задачи суммаризации научных статей важно выделить два признака по которым классифицируют типы суммаризации: метод суммаризации и количество документов, используемых для суммаризации.

По методу суммаризация делится на:

- экстрактивную;
- абстрактивную;
- гибридную.

Экстрактивный подход основывается на извлечении предложений напрямую из резюмируемого текста с использованием отбора предложений. Данный подход подразумевает нахождение важных с точки зрения информации предложений с последующим их объединением для получения сжатой версии текста [1] .

Преимущество экстрактивного подхода заключаются в возможности извлечь важную информацию с корректными фактами.

Недостаток заключается в том, что зачастую полученный реферат может быть непоследовательным, предложения могут быть несвязны между собой, в результате чего значительно ухудшается читаемость сгенерированного текста.

Абстрактивный подход подразумевает интерпретацию исходного текста при помощи методов обработки естественного языка для генерации сжатого текста, при чём часть этого текста может и не фигурировать в исходном тексте за счёт перефразирований, что схоже с тем, как резюмирует текст человек [1] .

Преимущество абстрактивного подхода заключаются в возможности получить последовательный, связный и читаемый текст.

Среди недостатков можно отметить то, что в результате абстрактивной суммаризации часть фактов может быть потеряно, а часть может быть искажена.

Гибридные методы комбинируют экстрактивный и абстрактивный методы. Как правило в таких методах применяются две модели: экстрактор и абстрактор [1].

Данный подход позволяет пользоваться преимуществами двух методов суммаризации. Из недостатков имеется возрастающая сложность решения.

По количеству документов суммаризация бывает:

- монодокументная;
- многодокументная.

В соответствии с названием монодокументная суммаризация производится по одному источнику, а во многодокументной их сразу несколько.

1.1.2 Метрики качества суммаризации

Существуют три группы методов оценки качества автоматической суммаризации: экспертная оценка, автоматические методы и полуавтоматические методы.

1.1.2.1 Экспертная оценка

Экспертная оценка автоматически сгенерированного реферата заключается в том, что группа экспертов по заданным критериям оценивает качество полученного текста. В том числе возможно сравнение автоматически сгенерированного текста с текстом, написанным экспертом. Не исключены такие случаи, что автоматически сгенерированный реферат может быть в чём-то лучше текста эксперта [2], и ручная оценка позволяет выявить такие прецеденты.

1.1.2.2 ROUGE (автоматические методы)

Среди методов автоматической оценки качества суммаризации применяются различные разновидности метрики ROUGE [16] [3], одна из наиболее популярных разновидностей - ROUGE-N.

ROUGE-N – это доля n-грамм из эталонного резюме-примера, которая оказалась в автоматически сгенерированном резюме.

ROUGE-N recall – представляет долю совпадающих n-грамм для сгенерированного реферата и реферата-примера в общем числе n-грамм примера.

$$recall = \frac{count_{match}(gram_n)}{count_{ref.summ.}(gram_n)} \quad (1)$$

ROUGE-N precision – представляет долю совпадающих n-грамм для сгенерированного реферата и реферата-примера в общем числе n-грамм сгенерированного реферата.

$$precision = \frac{count_{match}(gram_n)}{count_{gen.summ.}(gram_n)} \quad (2)$$

ROUGE-N F1 метрика – это среднее гармоническое между предыдущими двумя метриками:

$$F1_{score} = \frac{precision \cdot recall}{precision + recall} \quad (3)$$

Также существуют такие метрики как: ROUGE-L – наиболее длинная общая последовательность, ROUGE-W – взвешенная наиболее длинная общая последовательность и ROUGE-S – статистика совпадений по скип-граммам.

1.1.2.3 PYRAMID (полуавтоматические методы)

Метод оценки качества суммаризации PYRAMID [4] основывается на гипотезе о том, что существует эталонный стандарт реферата текста.

Хотя люди склонны резюмировать тексты по-разному, выделяя разную информацию из исходного текста, предполагается, что можно собрать эталонный реферат из множества рефератов, составленных людьми используя

информацию о частоте появления той или иной информации. По частоте можно судить о значимости информации и оценить качество суммаризации.

1.1.2.4 Автоматические методы оценки читаемости

Обозначенные выше автоматические и полуавтоматические методы позволяют понять в том числе, какое количество важной информации было извлечено, но есть методы, которые также позволяют оценить читаемость текста. Существуют следующие методы автоматической оценки читаемости текста:

- Индекс удобочитаемости Флеша-Кинкейда [5]:

$$FKGL = 0{,}39 \cdot \left(\frac{\text{число слов}}{\text{число предложений}}\right) + 11{,}8 \cdot \left(\frac{\text{число слогов}}{\text{число слов}}\right) - 15{,}59, \quad (4)$$

- Индекс туманности Ганнинга [6]:

$$GFI = 0{,}4 \cdot \left[\left(\frac{\text{число слов}}{\text{число предложений}}\right) + 100 \cdot \left(\frac{\text{число сложных слов}}{\text{число слов}}\right)\right], \quad (5)$$

- Индекс Колман — Лиау [7]:

$$CLI = 0.0588L - 0.296S - 15.8, \quad (6)$$

где L – среднее количество букв на 100 слов; S – среднее количество предложений на 100 слов.

1.2 Суммаризация научных статей

Автоматическая суммаризация множества научных статей несколько выделяется на фоне общей суммаризации набора текстов.

Научные статьи имеют следующие особенности [8]:

1) чёткая общая структура;
2) значительный объём текста;
3) цель суммаризации никогда не уникальна;
4) наличие рисунков, таблиц, формул, схем и алгоритмы;
5) специфичная для предметной области лексика.

Во-первых, общая структура любой научной статьи:

- аннотация;
- вступление (мотивация);
- лит. обзор;
- методология;
- экспериментальная секция;
- результаты и обсуждение;
- список источников.

Во-вторых, научные статьи, как правило, имеют большое количество текста для автоматической суммаризации, даже если речь идёт о монодокументной суммаризации.

В-третьих, информация, которую необходимо получить из научного исследования может быть различной, то может быть метод, результаты, ограничения или какие-либо ещё аспекты работы.

В-четвёртых, часть информации может подаваться в виде таблиц, рисунков, формул или псевдокода, что является уже отдельной задачей [13] [8].

В-пятых, суммаризация научных статей требует умение выделить общее, для чего суммаризирующая модель должна обладать некоторым пониманием текстов и, возможно, предметной области. Например, в случае исследования одной и той же темы в разных статьях может приводиться разная аргументация и даже разные выводы. Задаче же суммаризации новостей предполагает освещение одного и того же события с разных точек зрения.

На данный момент есть два основных направления связанных с суммаризацией научных статей: автоматическая генерация аннотации и суммаризация на основе цитирований. Рассмотрим достижения в одном и другом направлении.

1.2.1 Автоматическая генерация аннотаций

В работе Льорета (2011) [18] [9] были предложены два подхода для автоматической генерации аннотации: один – экстрактивный под названием COMPENDIUM$_E$, другой основанный на гибридном подходе - COMPENDIUM$_{E\,A}$.

В основе экстрактивного алгоритма четыре этапа:

1) предварительная обработка (т. е. токенизация, сегментация предложений, удаление стоп-слов и частеречная разметка);

2) удаление избыточности [19] [10];

3) идентификация релевантности предложений, при которой каждому предложению присваивается оценка, отражающая его важность на основе двух признаков — принципа количества кода [11, 20] и частоты употребления (term frequency или TF) [12, 21] — а затем ранжируется в соответствии с их баллам; и

4) генерация аннотации, при которой выбираются предложения с наивысшим рангом для создания окончательного варианта аннотации в том же порядке, в котором предложения появляются в исходном документе. Таким образом, сгенерированный реферат является экстрактивным.

Гибридный метод в качестве основы и добавляет этапы сжатия и объединения текста между третьим и четвертым этапами экстрактивного алгоритма выше. Новые предложения генерируются либо путем объединения информации из двух предложений, либо путем сокращения длинного предложения и разбития его на более короткие.

Ян и др. (2016) [13, 22] предложили систему аннотирования, которая описывает наиболее важные аспекты научной статьи на основе весов данных. Алгоритм состоит из двух этапов: вычисление весов предложений и выбор значимых предложений. На первом этапе рассматривается семантическая информация из цитирующего текста. Авторы рассматривают аннотацию целевой статьи, как объект, который содержит основные аспекты работы и цитирует значимые предложения тела статьи. Сеть цитирований также позволяет выделить значимые предложения тела статьи. На основе последних вычисляются

веса для предложений и оценивается их значимость. Из значимых предложений составляется конечная аннотация.

Сламет и др. (2018) [14] предложили простую систему, которая автоматически генерирует реферат статьи для индонезийского языка. Алгоритм состоит из четырёх шагов. Во-первых, этап предварительной обработки (состоящий из извлечения предложения, понижение регистра символов, маркировки, фильтрации и стемминга [15]) используется для подготовки входного текста к следующему этапу. Затем вычисляется TF-IDF [16] для каждого слова в предварительно обработанном тексте. Используя косинусное расстояние [17] и моделирование векторного пространства (VSM) [16], вычисляется сходство между текстом и 20 ключевыми словами выходных данных TF-IDF, а предложения ранжируются на основе их показателей сходства. Конечная аннотация составляется из десяти лучших предложений.

1.2.2 Суммаризация на основе цитирований

Цитирующий текст научных статей часто содержит наиболее важную информацию о цитируемой работе. При помощи цитирований можно оценить вклад той или иной работы и получить представления о её достоинствах и недостатках с точек зрения разных авторов.

В ранних работах по автоматической суммаризации на основе цитирований [18-20] статьи суммаризировались по средствам извлечения набора предложений из множества цитирующего текста. Более поздние работы указали [21-24] указали на некоторые проблемы с использованием предложений цитирования. При цитировании предложений обсуждение целевой статьи обычно пересекается с обсуждением других цитируемых статей или с содержанием нерелевантной информации цитирующей статьи. В качестве решения этой проблемы было предложено использование различных частей текста из цитируемой статьи.

Далее будет представлен обзор методов автоматической суммаризации научных статей на основе только цитирований и также с использованием текста оригинальной статьи.

1.2.2.1 Суммаризация на основе только цитирований

Казвинян и Радев (2008) предложили систему суммаризации на основе цитирующих предложений под названием C-LexRank [19]. В данной модели цитирующие предложения представлены в виде вершин графа связанные рёбрами с весами, которые представляют меры подобия между предложениями. На основе меры подобия предложения разбиваются на кластеры. При генерации реферата по предложению выбирается из каждого кластера.

В 2011 году Абу-Джбара и Радев [20] отметили некоторые проблемы вышеупомянутой модели, связанные с тем, что в цитирующих предложениях может содержаться нерелевантная информация, что приводит к снижению читаемости, согласованности, увеличению размера реферата и потери важной информации во время ранжирования предложений. Авторами был улучшен данный алгоритм за счёт маркировки ссылок, введения классификатора для них (удалить, сохранить или заменить на местоимение) и введения системы для группировки цитирующих предложений по отношению к разделу (введение, постановка задачи, метод, результаты и ограничения).

1.2.2.2 Суммаризация на основе цитирований с использованием текста оригинальной статьи

В 2015 году Галгани и др. [25] предложили использовать как цитирования, так и полный текст цитируемой статьи, и реферат цитирующей статьи при наличие последнего. Рефераты генерируются либо (1) ранжированием извлеченного текста цитирований, с целью нахождения общих понятий между нескольким цитатами и генерацией реферата из них, либо (2) путём измерения сходства между каждым предложением в целевой статье и цитированием с последующим ранжированием предложений в порядке убывания. Во втором случае резюмирующими предложениями будут предложения, имеющие

наибольшее сходства с цитирующими предложениями. Основная идея заключается в том, что цитаты представляют основные аспекты целевой статьи и, следовательно, могут использоваться для выбора сегментов, которые представляют эти аспекты.

В цитирующих предложениях не достаёт контекста о результатах работы [26]. Для решения этой проблемы Cohan and Goharian (2015) [21] предложили использовать текстовых срезов из цитируемой статьи. Авторы используют в качестве представления цитирующих предложений вектор из их n-грамм для нахождения наиболее релевантных текстовых срезов, которые затем группируются по темам; предложения далее ранжируются по информативности. Реферат составляется либо путём итеративного поиска предложений с наивысшим рейтингом, либо с использованием жадной стратегии.

В работе Ronzano and Saggion (2016) [27] было исследовано влияние контекста цитирования и то, к какой части статьи он относится, на качество суммаризации. Было выявлено, что использование контекста цитирования повышает качество суммаризации. Наилучшие результаты (по среднему ROUGE-2) были получены, когда в качестве примера реферата использовалась аннотация и предложения для реферата брались из тела статьи и контекста цитирований

В 2018 году Коханом и Гохаряном был представлен фреймворк, решающий проблему неточности текста цитирования [22]. В основе решения также был эмбединг и классификатор, позволяющий найти подходящий контекст для каждого цитирующего предложения.

В 2017 была создана система суммаризации CitationAS [28]. В алгоритме используется набор правил для идентификации предложений цитирования, состоит он из трёх этапов. Первый — это кластеризация, в которой используются три алгоритма: кластеризация суффиксного дерева (STC) [29], Lingo [30] и разделение K-средним [31]. На этом этапе цитирующие предложения сначала получают представление в виде вектором VSM [32], а TF-IDF [33] используется

для вычисления весов признаков; похожие предложения группируются в кластеры. Затем используется комбинация Word2Vec [34] и WordNet [35] для генерации меток кластеров, после чего кластеры с похожими метками объединяются. В конце, кластеры сортируются по размеру, а предложения извлекаются для формирования конечного реферата при помощи ранжирования. Одним из преимуществ CitationAS является то, что генерируемое резюме является исчерпывающим и репрезентативным по теме, хотя и содержит некоторое избыточное содержание.

В 2018 году было предложено решение по суммаризации научных статей [36] с использованием метода максимизации признаков [37]. Предлагаемые системы суммаризации являются статистическими, не имеют параметров, не зависят от языка и не нуждаются в дополнительных корпусах. Общая структура предлагаемой системы состоит из пяти основных этапов. Во-первых, входной текст подвергается предварительной обработке (т. е. удалению стоп-слов и стеммингу), а вес слова вычисляется для набора ключевых слов в заголовке статьи, подзаголовках и аннотации. Затем вычисляются веса предложений, используя среднее значение весов его слов. Размер итогового реферата определяется на третьем этапе на основании распределения весов. В конце может быть применена обработка текста для удаления избыточного содержания.

В одной из последних работ [38] была предложена система суммаризации научных статей, которая основывалась на гипотезе, что вклад любой статьи лучше всего описывается ее целью, используемым методом и конечным результатом. Базируясь на этой гипотезе, авторы использовали классификатор на основе k-ближайших соседей [39] и бутстреп [40] для извлечения трёх понятий, обозначенных выше, названия целевой статьи, аннотации и контекста цитирования (т. е. цитируемого текста). На основе полученной информации строится граф знаний для графического представления связи извлечённых понятий и цитирований.

1.2.3 Автоматическая суммаризация научных статей при помощи трансформеров

Из методов глубокого обучения наиболее популярным решением для многодокументной суммаризации являются модели на основе трансформеров. Они имеют ряд преимуществ над другими методами как то, что они могут находить зависимости между документами, с чем есть проблемы при использовании свёрточных нейронных сетей в виду их ограниченного поля зрения, и возможность распараллеливать вычисления, чего не позволяют рекуррентные нейронные сети [12].

Для задач многодокументной суммаризации успешно применяются предобученные трансформеры, при этом такие модели могут быть обучены на задаче монодокументной суммаризации или вообще на датасете для другой задачи, что позволяет избежать нехватки данных в области многодокументной суммаризации [12].

Трансформеры успешно применялись для задач суммаризации коротких текстов в сравнении с текстами из научных статей [41]. Относительно недавно появились разработки трансформеров для длинных последовательностей [42, 43] и также адаптации трансформеров для задач, связанных с суммаризацией научных статей [44].

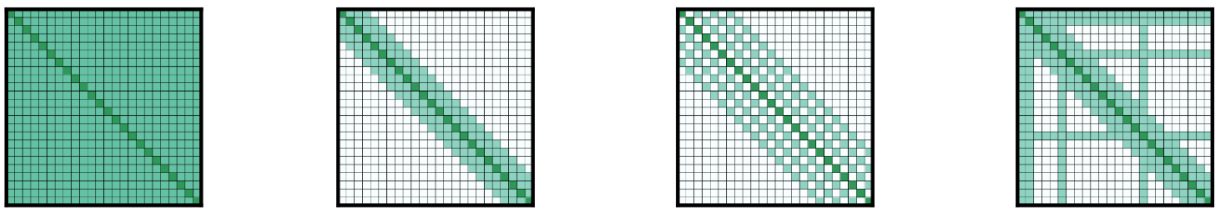
1.2.3.1 Решения на основе трансформеров для работы с длинными последовательностями

Обходы ограничений трансформеров по работе с длинными последовательностями варьируются в зависимости от задач [45]. Например, для задачи классификации является вполне приемлемым решением усечение документа [46]. Также существует метод, при котором документ разбивается на чанки, которые обрабатываются моделью отдельно [47]. В другом методе используется подход из двух стадий, где на первой стадии происходит поиск релевантного документа, на второй – его обработка [48].

Недостатком всех вышеперечисленных методов является то, что они приводят к потере информации. Предпочтительнее реализация, при котором вся последовательность обрабатывается за один проход.

Одной из последних тенденций в построении трансформеров для работы с длиными последовательностями является одновременное использование механизмов глобальной и локальной памяти (Рисунок 1 г).

В трансформерах с полным механизмом внимание увеличение зоны внимания сопровождается увеличением используемой памяти согласно квадратичному закону n^2 (Рисунок 1 а). Трансформеры ETC [48], BigBird [42] и Longformer [43] используют совмещённый механизм внимания. На данный момент лучшие результаты из перечисленных моделей показывают модели семейства Longformer [43].



а б в г

Рисунок 1 – Виды «внимания», используемые в Longformer [42]:

а – полное внимание; б – внимание со скользящим окном (локальное); в – внимание с расширенным скользящим окном; г – глобальное и локальное внимание

В моделях семейства Longformer в зависимости от задач используются различные подходы, которые позволяет снизить использование памяти.

В случае авторегрессии наилучшие результаты получаются при использовании расширенное скользящее окно. Здесь удаётся перейти от квадратичного потребления памяти (n^2) к более экономному соотношению nw (предполагается, что $w<<n$) [49].

В случае seq2seq [50] моделей используется комбинация локального и глобального двунаправленного внимания. Потребление памяти снижается с

зависимости $n_s^2 + n_s n_t + n_t^2$ до зависимости $wn_s + n_s n_t + n_t^2$, где n_s - длина последовательности на входе энкодера и n_t – длина последовательности на выходе у декодера. Таким образом, для того чтобы такая модель была эффективна, предполагается, что значение n_s должно быть значительно больше n_t [49].

1.2.4 Существующие датасеты для обучения суммаризации научных статей

Датасетов для обучения задаче суммаризации существует не так много. Рассмотрим наиболее популярные датасеты.

Датасет TAC2014 BiomedSumm, который посвещён биологической и медицинской тематике, [51] включает в себя 20 тематик, для которых предоставлено по четыре резюме, составленные четырьмя экспертами в предметной области. Каждое резюме содержит не более 250 слов.

CL-SciSumm [52] – датасет со статьями по компьютерной лингвистике и обработке естественного языка, включает 30 тематик и имеет 3 части: тренировочную, валидационную и тестоввую, с одним резюме по каждой теме и набором цитирующих статей. Все статьи имеют формат XML. Все резюме созданы одним лишь экспертом.

ACL Anthology (ANN) [53] представляет из себя большую поддерживаемую базу данных цитат и резюме в области компьютерной лингвистики. На момент написания работы данная база содержит более 23 тыс. статей и более 124 тыс. цитирований.

Существует датасет для разработки систем суммаризации от Microsoft Academic Search [54] Он содержит аннотации статей, цитируемых предложениях, авторах, месте публикации и приложенной статье цитатных предложений.

Коллекция cmp-lg corpus [55] состоящая из 183 документов, размеченных в формате XML, он используется в качестве ресурса для обобщения, извлечения и поиска информации. Документы представляют собой научные статьи из ACL. Они охватывают ключевую информацию для каждой статьи (такую как название,

автор и дата), помимо основных структурных элементов, таких как аннотация, основная часть, разделы и списки.

PLOS [56] – это набор из 50 научных статей на тему медицины, по каждой статье имеется «золотой» стандарт резюме.

ScisummNet [57] - первый крупномасштабный набор данных Scisumm, аннотированный человеком, ScisummNet. Он предоставляет более 1000 статей из сети ACL с их сетями цитирования (как-то цитирующие предложения и информация о количестве цитирований) и их подробными резюме, созданными вручную.

Датасет для обучения автоматической суммаризации научных «scientific_papers» [58], размещённый в Hugging Face содержит выборки статей из базы Pubmed (более 11 Гб) и arxiv (более 6 Гб), каждая выборка содержит три подвыборки: обучающую, валидационную и тестовую. Среди доступных полей: аннотация, текст статьи и заголовки разделов.

1.2.5 Основные проблемы автоматической суммаризации

Основной проблемой для задачи автоматической суммаризации научных статей является небольшое количество данных для обучения или же их отсутствие для построения алгоритмов суммаризации для специфики некоторых предметных областей.

Также на данный момент стоит проблема метрик для оценки качества суммаризации и общих систем бейслайна для сравнения алгоритмов между собой [8].

Большая часть рассмотренных алгоритмов нацелены на построении резюме на основе цитирующих предложений для одной статьи. Для получения качественно новых результатов в области многодокументной суммаризации с хорошей читаемостью и согласованностью сгенерированных текстов потребуется большее количество исследований [8].

2 ИЗВЛЕЧЕНИЕ ДАННЫХ ИЗ НАУЧНЫХ СТАТЕЙ

В ходе извлечения данных из научных статей было необходимо решить две задачи:

1. извлечение текстовой информации научных статей;
2. извлечение нетекстовой информации.

Под «текстовой информацией» подразумеваются заголовки, абзацы, аффилиация, ссылки и контакты.

Под «нетекстовой информацией» подразумеваются рисунки, таблицы, математические и химические формулы. Также при выделении рисунков и таблиц, необходимо было получить их описания, для таблиц содержание их ячеек в структурированном виде (в формате json), для формул – номер формулы (при наличии).

2.1 Исходные данные

Исходный набор данных представляет из себя коллекцию статей в формате PDF с наличием текстового слоя.

Структура данных в конечной системе будет в виде MAG/OAG (графовидная структура) с возможностью добавления новых данных. Финальный объём базы знаний оценивается в 100 Тб.

Было принято решение получать текстовую информацию, используя инструменты библиотеки машинного обучения GROBID (или Grobid, расшифровывается как GeneRation Of Bibliographic Data [59], что переводится как генерация библиографических данных), для извлечения данных (извлечение тестовой информации ссылок, колонтитулов, координат объектов документов и пр.) из документов PDF и дальнейшей её структуризации.

Всё же, для решения задачи получения «нетекстовой информации» система GROBID показала неудовлетворительные результаты для всех объектов кроме формул на отсканированных статьях. Инструменты GROBID систематически неверно идентифицировали объекты и описания к ним, а по

полученной содержательной части таблицы сложно было восстановить её исходную структуру.

Исходя из сказанного выше, для повышения качества результата было принято решение написать свой обработчик таблиц и рисунков.

2.2 Извлечение рисунков

Алгоритм извлечения рисунков из статей имеет следующий общий вид:

Шаг 1. Детектирование объектов на изображении страницы;

Шаг 2. Сортировка объектов по расположению на странице;

Шаг 3. Для каждого объекта изображения выполняются шаги 4-6;

Шаг 4. В зависимости от класса и относительного расположения объектов для рисунка происходит поиск блока с описанием и объединяется с блоком рисунка (Рисунок 2);

Шаг 5. Производится распознавание текста описания по совместному изображению, полученному на шаге 5;

Шаг 6. Изображение и его описание сохраняются в базе данных по его идентификатору.

Для детектирования изображений и таблиц, а также распознавания содержимого последних используется модель mask_rcnn_X_101_32x8d_FPN_3x [60], обученная на датасете PubLayNet [61], входящая в состав библиотеки Layout Parser [62]. Данная модель предназначена для детектирования таких классов, как текст, заголовок, список, таблица и рисунок.

Для распознавания текста на шаге 5 также используются инструменты агент TesseractOCR, встроенный в Layout Parser. В качестве идентификатора изображения использовался его номер. Например, Figure n преобразовалось бы в figure_n.

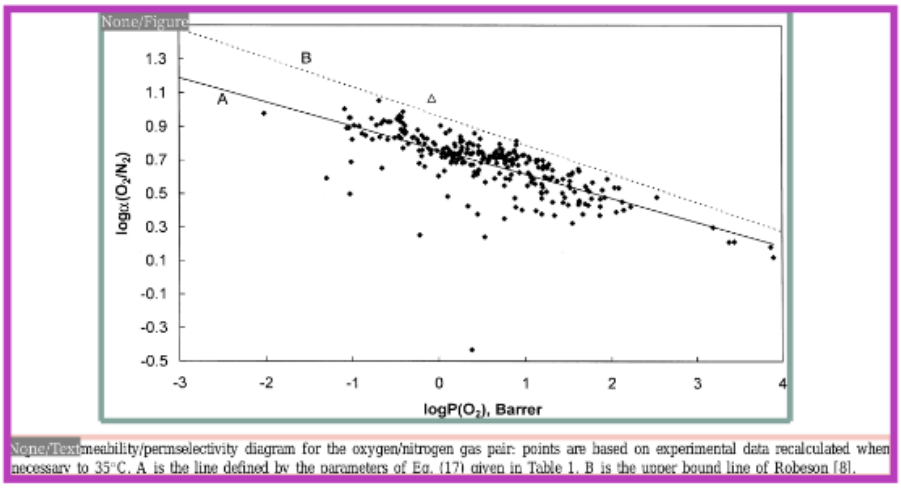


Рисунок 2 - Объединение блоков рисунка и описания для распознавания:

зелёный прямоугольник – рисунок; розовый – описание к рисунку; фиолетовый – объединение блоков

На рисунке 3 представлена общая блок-схема алгоритма.

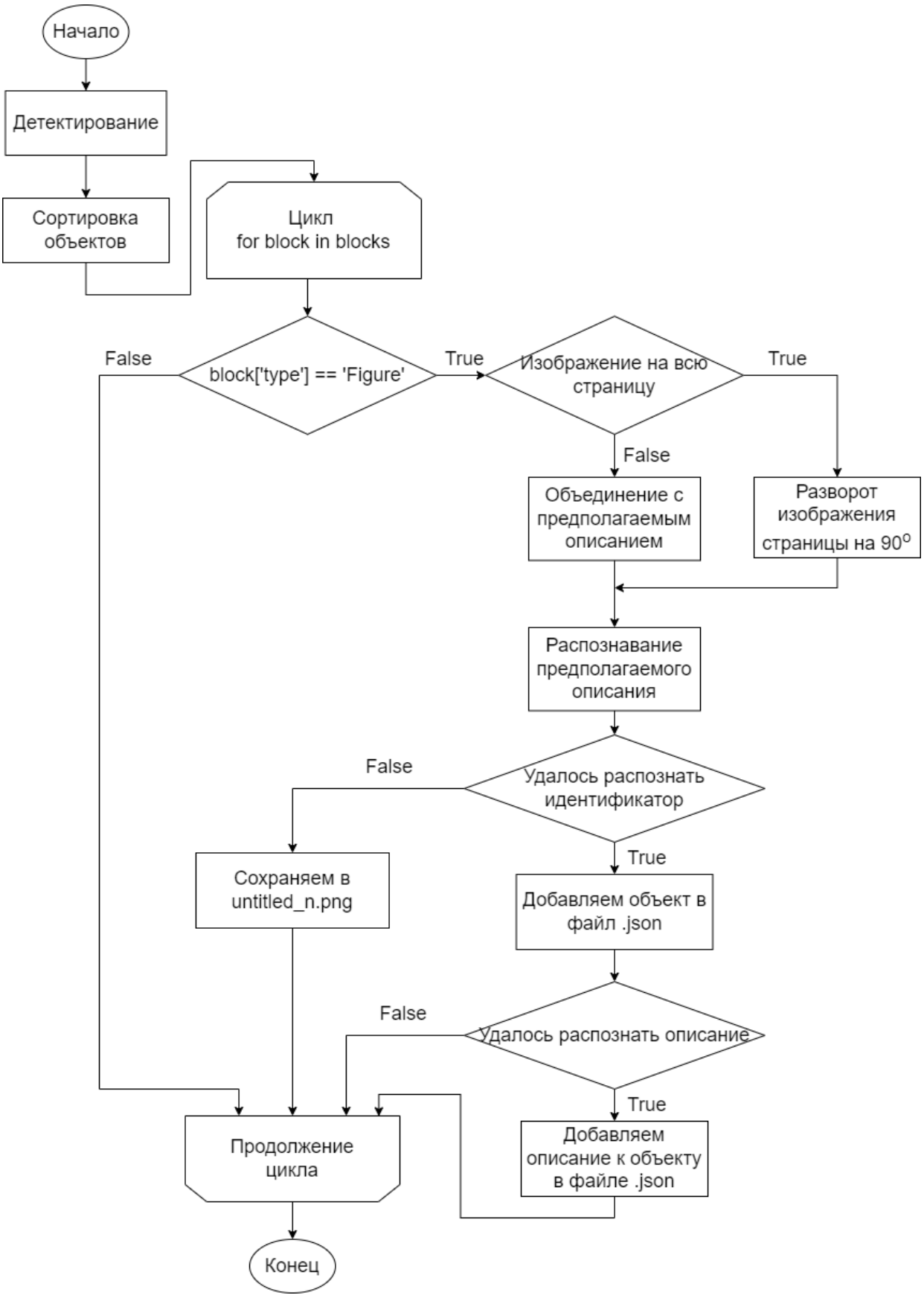


Рисунок 3 – Общая блок-схема алгоритма извлечения изображений научных статей

2.3 Извлечение таблиц

Извлечение таблиц и их описаний происходит по такому же алгоритму, главное отличие заключается в том, что у таблиц описание находится строго сверху.

После получения изображения таблицы происходит извлечение информации из самой таблицы.

Алгоритм для получения данных таблиц имеет следующий вид:

Шаг 1. таблица делится на две части – по горизонтальным линиям с нахождением границы раздела при помощи метода Hough Lines [63];

Шаг 2. распознавание оглавления таблицы при помощи агента TesseracOCR;

Шаг 3. агрегация распознанных блоков и организация мультииндексов;

Шаг 4. распознавание тела таблицы;

Шаг 6. организация элементов тела таблицы в соответствии с расстоянием и координатами столбцов;

Шаг 7. объединение таблицы.

На рисунке 4 представлена блок-схема алгоритма по извлечению формул.

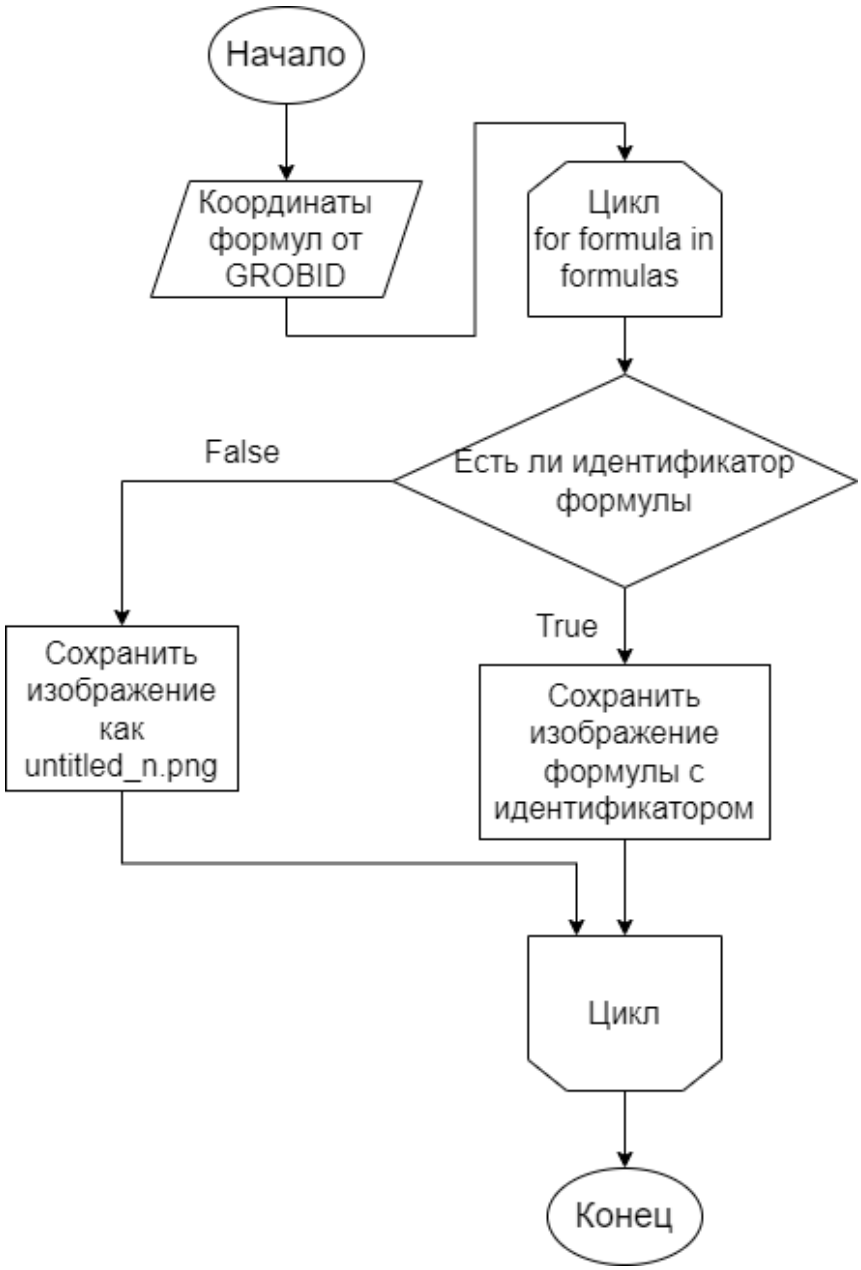


Рисунок 4 – Алгоритм по извлечению формул в общем виде

Алгоритм был протестирован на выборке из 50 статей. В случае извлечения изображений рисунков, таблиц и формул точность всего алгоритма была обусловлена точностью детектирования объектов (построения рамки и классификации объектов), которая известна (88.98% mAP).

Результаты тестов по извлечению данных из таблиц записывались в файл формата CSV. На этапе извлечения изображений таблиц удалось получить 170 предполагаемых изображений таблиц.

На рисунке 5 представлена таблица с соотношением типов объектов полученных при извлечении таблиц

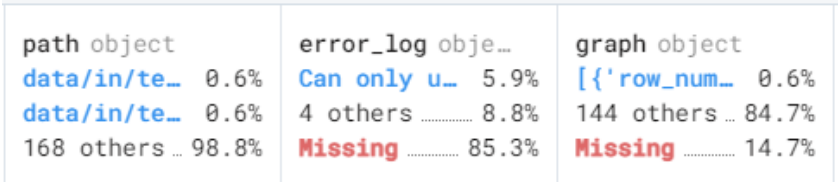


Рисунок 5 – Столбцы таблицы с результатами тестов

Определим процент предполагаемых таблиц, для которых удалось построить таблицы. Во время тестов, в случае успеха обработки таблицы, в ячейку graph для таблицы по данному пути (path) записывались извлечённые данные. В случае возникновения каких-либо ошибок в ячейку error_log записывалось сообщение об ошибке, а в ячейку graph записывалось пустое значение. Таким образом, примерно 85,3% изображений предполагаемых таблиц были обработаны, из этого множества обработанных объектов только 89,63% были маркированы как таблиц (т.е. это те таблицы, для которых удалось получить идентификатор и описание).

Алгоритм был проверен на другой выборке, состоящей из 90 статей. Содержания извлечённой из таблиц информации проверялось вручную, как и результаты извлечения других объектов. В результате ручной проверки, точность алгоритмов признана удовлетворительной для дальнейшего использования в проекте на данном этапе его развития.

2.4 Структура извлечённых данных

С использованием разработанных алгоритмов, работающих с функционалом GROBID и Layout Parser, был произведён сбор данных научных статей.

Было обработано полностью около 17 тыс. статей на тематику, связанную с химией и химической технологией.

На рисунке 6 условно представлен вид графа знаний (точнее его ветвь).

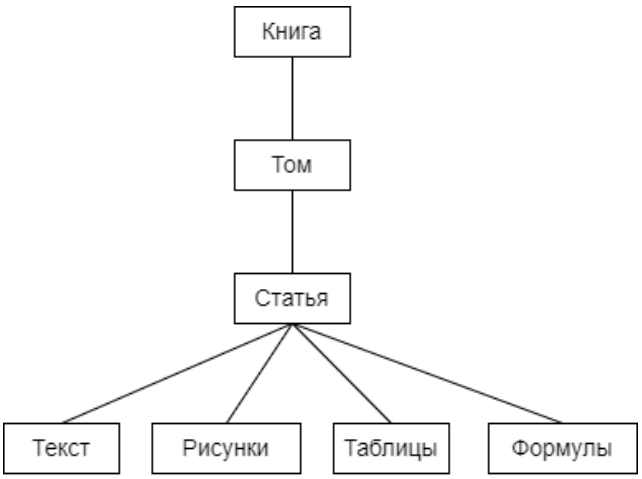


Рисунок 6 - Ветвь разработанного графа знаний

2.5 Вывод по разделу

В результате проведённой работы была подготовлена база знаний для создания прототипа системы интеллектуального поиска и генерации обзорных статей по ключевым словам. К созданию прототипа было переработано около 17 тыс. научных статей и получено более 140 Гб данных.

3 РАЗРАБОТКА СИСТЕМЫ ПОИСКА

3.1 Общий алгоритм

После совершения запроса в виде строки алгоритм научного поиска выполняет следующие общие шаги:

1. классификация запроса, отнесение его к одному из кластеров;
2. поиск наиболее схожих понятий внутри кластера;
3. суммаризация наиболее релевантных статей (получение рефератов статей);
4. суммаризация полученных рефератов статей.

На первом этапе строка запроса получает векторное представление, далее определяется, в каком из заранее вычисленных кластеров будет происходить поиск. На втором этапе происходит поиск наиболее схожих с запросом документов по их векторному представлению. Затем статьи с наиболее схожим векторным представлениям получают краткий реферат при помощи модули суммаризации. В конце собирается обзорный текст на основе сгенерированных рефератов.

Далее алгоритмы в основе этапов поиска будут рассмотрены отдельно более детально.

3.2 Данные

Перед тем как начать применять данные для обучения моделей и построения алгоритмов необходимо получить хотя бы общее представления о свойствах данных которые имеются в данной коллекции.

Для построения, обучения и тестирования алгоритмов были взяты публикации из научного журнала «Journal of Membrane Science» [62] по 2019 год. В журнале публикуются работы, связанные с мембранным транспортом, со структурой и образованием мембран, с мембранной очисткой, разработкой технологических процессов и их применением.

Изначальное число документов было около 17 тыс., но для некоторых статей не удалось извлечь текст тела статьи или аннотацию. После того, как неполные объекты были исключены из выборки, осталось 14529 документов.

Статьи имеют следующее распределение по годам публикации (Рисунок 7).

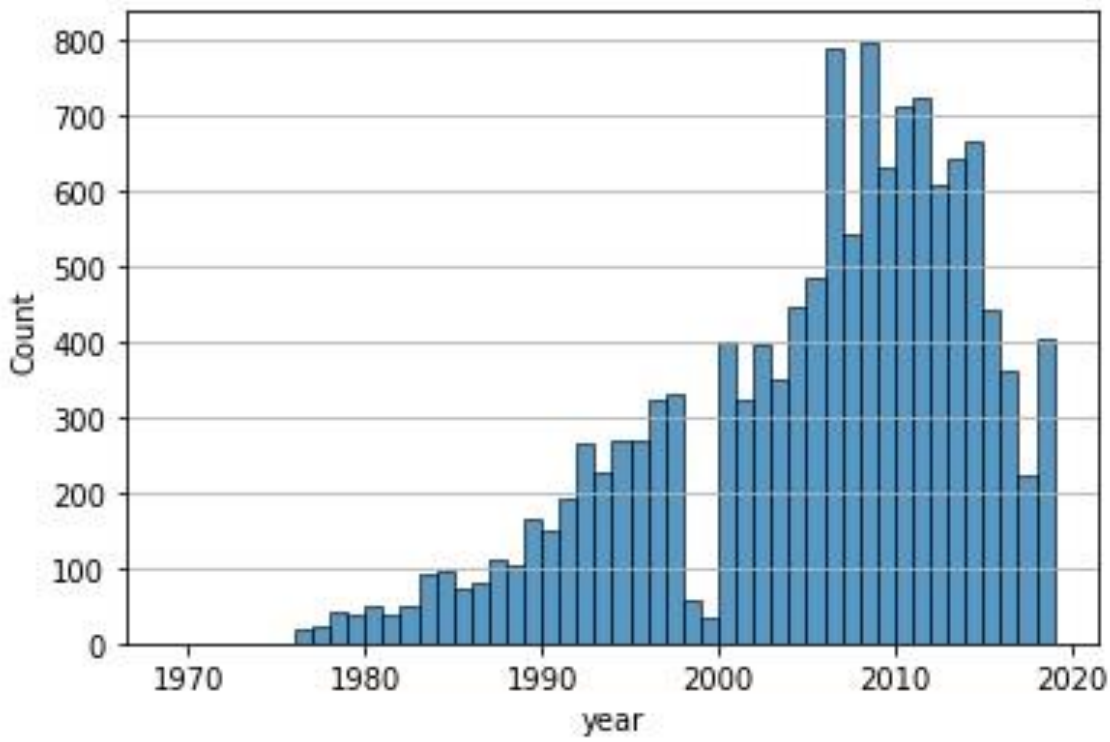


Рисунок 7 – Распределение публикаций по годам

На рисунке 8 представлено распределение ключевых слов в статьях. Большинство научных работ в коллекции документов имеют около пяти ключевых слов, но при этом есть значительное в сравнении с размером датасета количество статей, для которых не удалось извлечь ключевые слова.

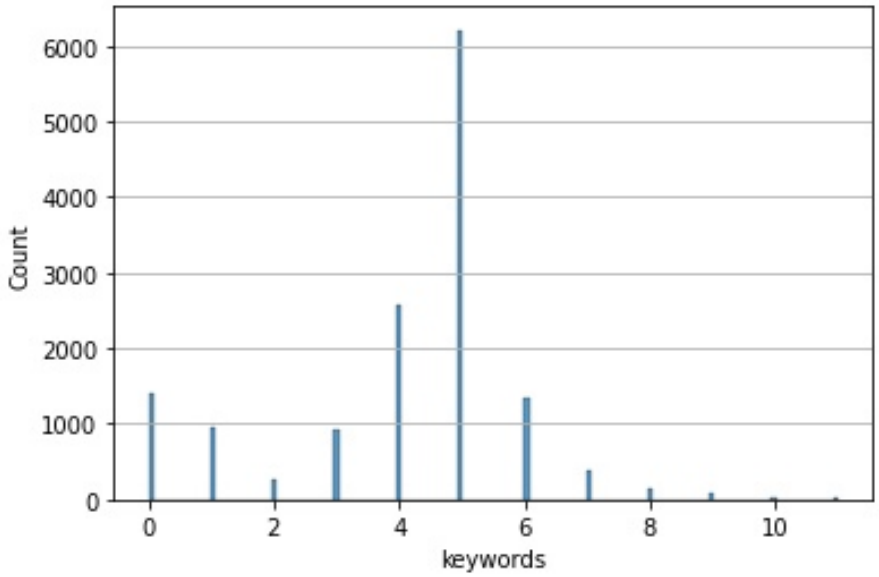


Рисунок 8 – Распределения количества ключевых слов статей

С целью представления наиболее широкого числа понятий, используемых в статьях, были получены именованные сущности на основе поиска по тексту документов понятий, встречающихся в общем списке ключевых слов. Распределение именованных сущностей (Рисунок 9) показывает, что их количество в десятки раз превосходит количество ключевых слов, при этом для каждого документа в коллекции число именованных сущностей не равно нулю. Среднее количество именованных сущностей на документ составляет 167±65, что в значительной мере больше, чем ключевых слов.

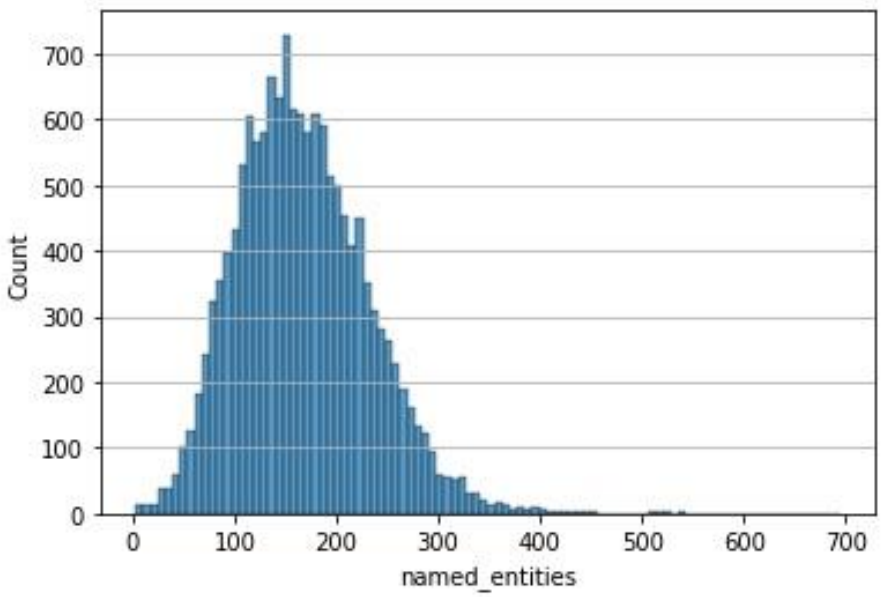
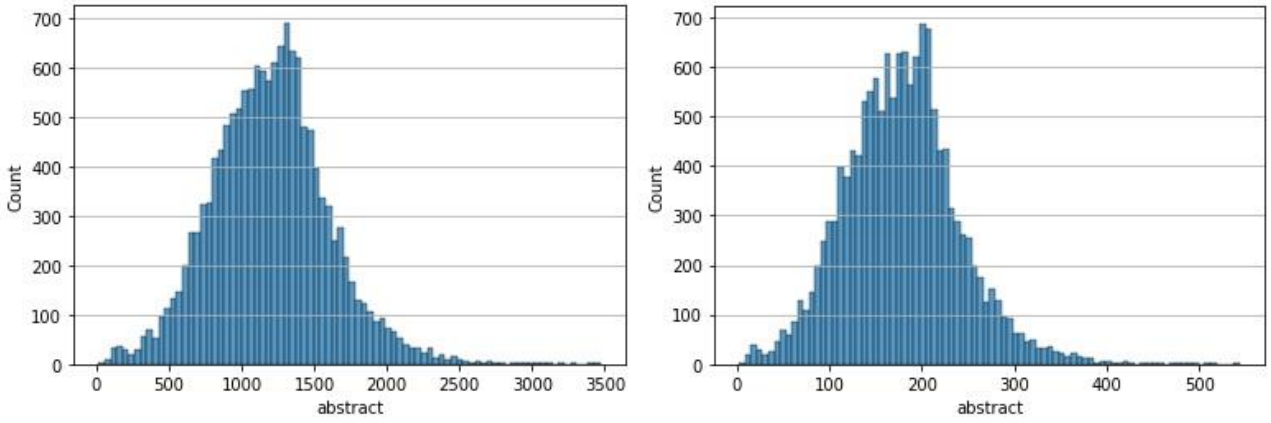


Рисунок 9 - Распределение именованных сущностей

Текст тела статьи и её аннотация используются при обучении модели суммаризации, поэтому в высокой степени важно изучить их свойства.

Средняя длина аннотации в словах (токенах) 179±62 составила и 1199±401 в символах. На рисунке 10 представлены распределения длины аннотации в символах и в словах (токенах).



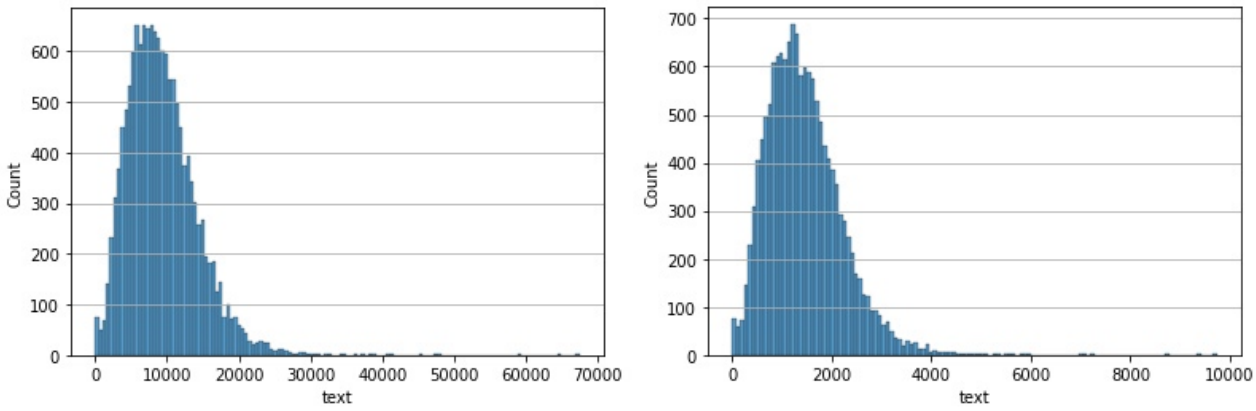
а б

Рисунок 10 – Распределение длин аннотаций:

а – распределение по количеству символов; б – распределение по количеству токенов

Средняя длина текста в токенах составила 1448±754 и 9328±4867 в символах. На рисунке 11 представлены распределения длины текста в символах и в токенах.

Средняя длина текста тела статьи



а б

Рисунок 11 – Распределение длин текстов тела статьи:

а – распределение по количеству символов; б – распределение по количеству токенов

3.3 Векторное представление для статей

В качестве энкодера для составления векторного представления слов была использована модель из библиотеки genism, которая была обученная на корпусе из собственных данных.

Перед обучением энкодера текст получил специальную предобработку, которая состояла из приведения всего текста к нижнему регистру, удаления стоп-слов, лемматизации и удалению неалфавитных символов. Для включения в память модели биграмм использовался модуль библиотеки Phraser.

3.3.1 Исследование модели

После обучения модели целесообразно провести проверку её адекватности. Десятью наиболее популярными токенами (после очистки и лемматизации) является следующий набор понятий:

```
['membrane', 'water', 'c', 'solution', 'high', 'increase', 'h', 'm',
'concentration', 'fig']
```

Можно проверить, какие наиболее близкие понятия в векторном пространстве к указанным выше словам вернёт модель. Сходство понятий оценивается при помощи косинусного расстояния.

На центральную в коллекции статей на понятие «membrane» (мембрана) модель вернёт:

```
word2vector_model.most_similar('membrane')
```

```
[('composite', 0.5857453346252441),
 ('membranes', 0.5775184631347656),
 ('support', 0.5489179491996765),
 ('layer', 0.5259989500045776),
 ('hollow_fiber', 0.5142043232917786),
 ('surface', 0.5007392168045044),
 ('performance', 0.4742079973220825),
 ('substrate', 0.461181640625),
 ('asymmetric', 0.44165390729904175),
 ('film', 0.44153258204460144)]
```

Таким образом все понятия схожие с понятием «membrane» относятся либо непосредственно к мембране, либо к поверхностным явлениям (которые во

многом и определяют свойства биологических и искусственных мембран) либо к композитным материалам (к которым относятся мембранны).

Самое неочевидное понятие в списке «h» имеет следующие схожие понятие:

```
word2vector_model.most_similar('h')
```

```
[('c', 0.731769323348999),
 ('min', 0.6994180679321289),
 ('o', 0.624107837677002),
 ('day', 0.5981701016426086),
 ('room_temperature', 0.5866249799728394),
 ('hour', 0.5820997357368469),
 ('k', 0.518334150314331),
 ('n', 0.5175007581710815),
 ('respectively', 0.5152209997177124),
 ('overnight', 0.5024268627166748)]
```

Из схожих понятий можно заключить, что «h» – это обозначение химического элемента «водород» (H - hydrogen) и единица измерения времени час (от английского hour).

3.4 Кластеризация на основе векторного представления

С целью повышения точности и скорости ранжирования предполагалось использовать кластеризацию статей.

Для кластеризации данных были опробованы самоорганизующиеся карты Кохонена в виду имеющейся возможности задать количество кластеров посредствам конфигурирования количества нейронов в матрице самоорганизующейся карты.

Самоорганизующиеся карты были взяты из библиотеки sklearn-som [63].

Были проведены эксперименты по выделению кластеров документов на основе полученных именованных сущностей и также на основе списков, полученных объединением именованных сущностей с ключевыми словами. Так как ключевые слова удалось получить не во всех случаях, то отдельных экспериментов с выделением кластеров на основе ключевых слов не проводилось.

Векторное представление документов было получено путём сложения всех векторных представлений списка (списка именованных сущностей или объединённого списка). В случае, если сущность не имела векторного представления, то она пропускалась. Так как данные не имеют разметки (ведь кластеры неизвестны, а их разметка излишне трудоёмка), то для оценки качества кластеризации использовался индекс Дэвиса-Болдуина [64]. В таблице 1 представлен результат оценки.

Таблица 1 – Оценка качества кластеризации

Количество нейронов в картах Кохонена	3	4	5	6	7	8
Индекс Дэвиса-Болдуина	2,51	2,62	3,03	3,05	3,04	2,95

Наиболее низкое значение индекса Дэвиса-Болдуина соответствует лучшему разделение на кластеры, но при тестировании, возвращаемых значений в дальнейшем наиболее адекватные ответы на поисковые запросы удалось получить при конфигурации в 6 нейронов на самоорганизующихся картах.

На рисунках 13 и 14 представлены визуализации двумерных проекции векторов документов для именованных сущностей и комбинированного списка соответственно. Проекция была получена при помощи метода выделения главных компонент [65].

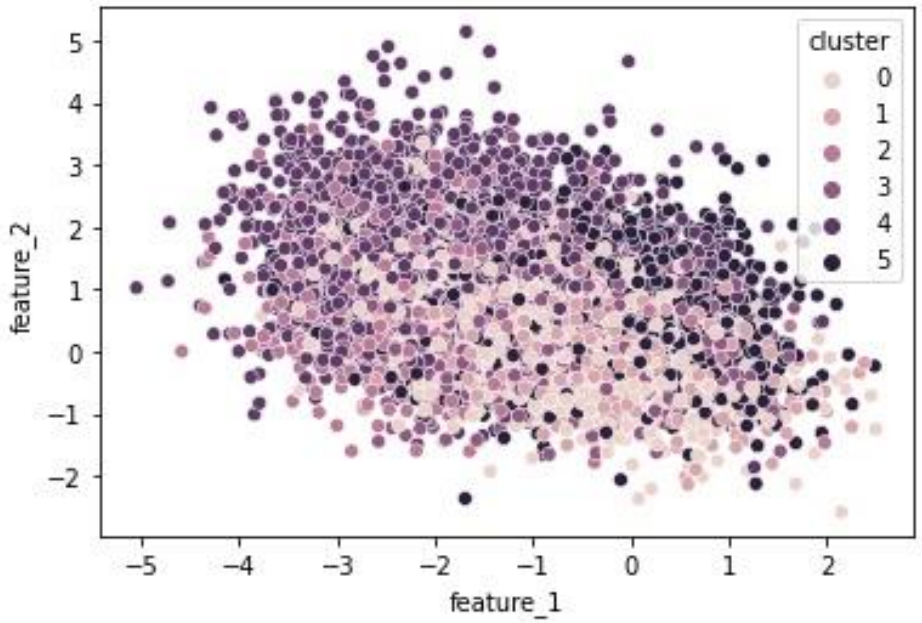


Рисунок 13 – Кластеры именованных сущностей в двумерной проекции

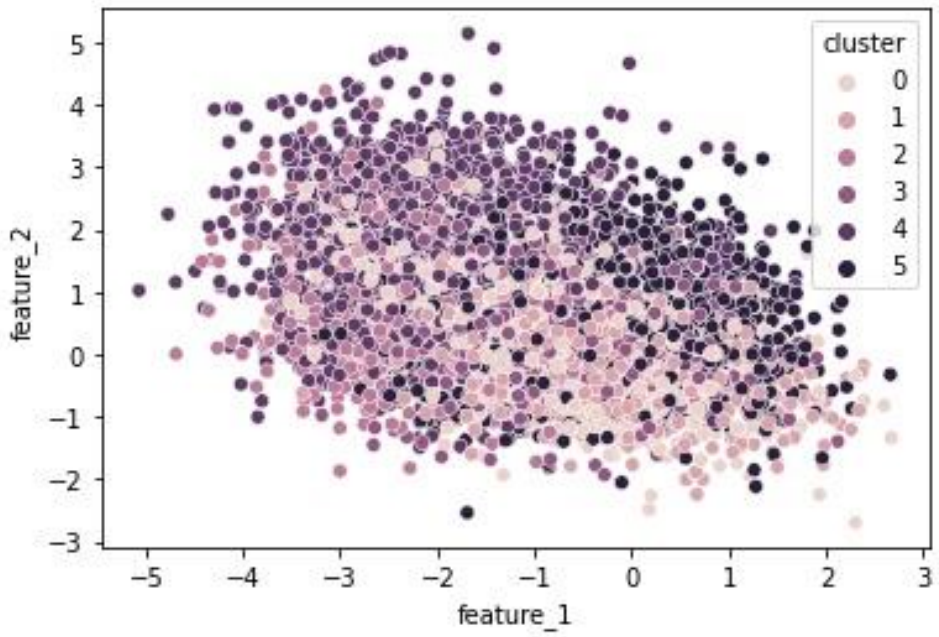


Рисунок 14 – Кластеры комбинации именованных сущностей и ключевых слов в двумерной проекции

На рисунках 15 и 16 представлены гистограммы распределения документов по кластерам.

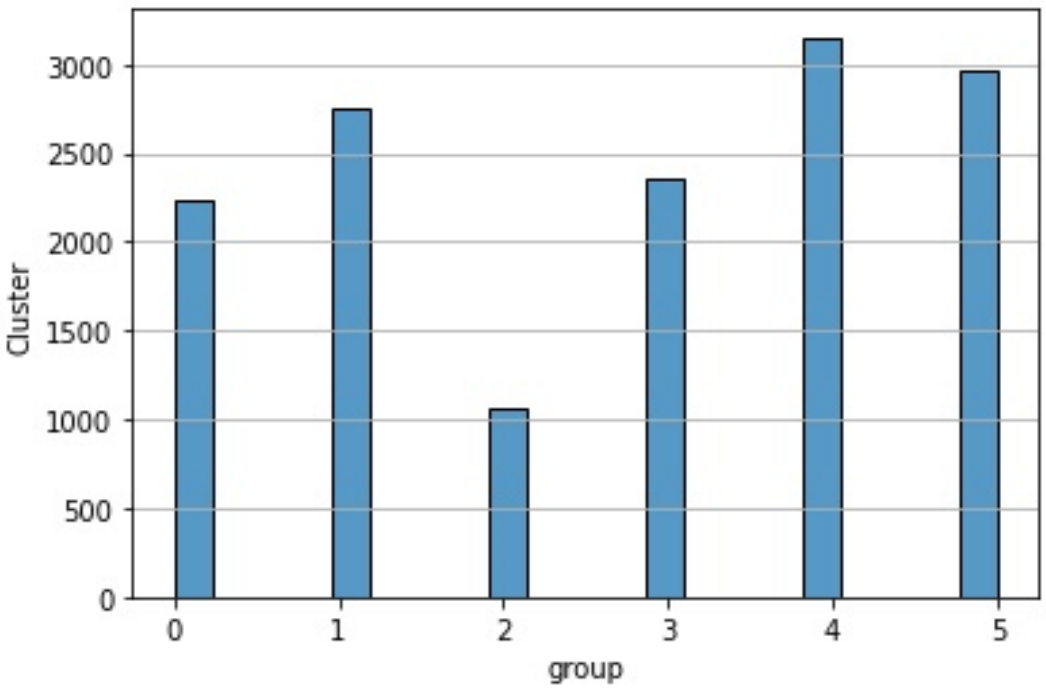


Рисунок 15 – Распределение объектов по кластерам для именованных сущностей

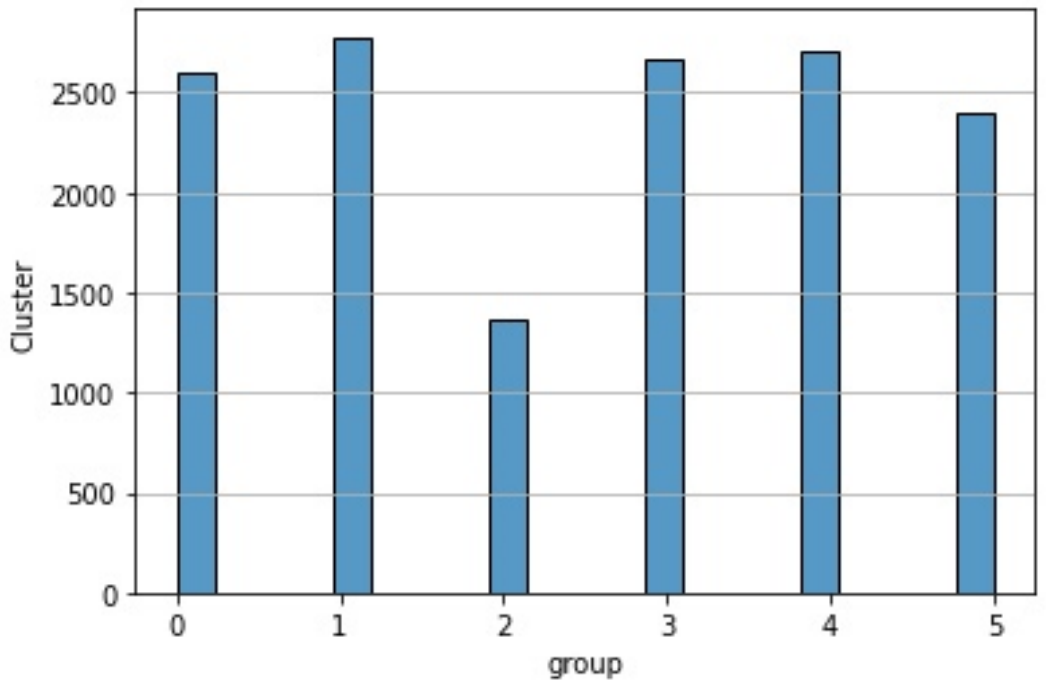


Рисунок 16 – Распределение объектов по кластерам для комбинации именованных сущностей и ключевых слов

В качестве примера работы системы поиска документов через векторное представление рассмотрим то, какие документы удаётся найти на запрос *«reverse*

osmosis for seawater desalination», что переводится как «опреснение морской воды при помощи обратного осмоса». В таблице 2 представлен результат данного запроса (точнее, его часть с косинусным расстоянием между представлением запроса и id документа для иллюстрации уникальности найденного документа) для ранжирования с применением кластеризации на основе именованных сущностей и комбинированного (кластеризация по объединённому списку ключевых слов и именованных сущностей).

Таблица 2 – Сравнение ранжирования с применением кластеризации по именованным сущностям и их комбинации с ключевыми словами

Ранжирование с применением комбинированного метода кластеризации		Ранжирование с применением кластеризации по именованным сущностям	
id	Косинусное расстояния	id	Косинусное расстояния
13947	0,616	8555	0,573
3613	0,621	1685	0,579
4474	0,629	8273	0,59
2985	0,637	241	0,612
7145	0,638	3613	0,621
3774	0,656	7145	0,638
12302	0,661	3774	0,656
13672	0,664	12302	0,661
3844	0,686	3844	0,686
5138	0,698	5138	0,698

В данном конкретном случае ранжирование с использованием кластеризации по вектору для комбинированного списка даёт более релевантные результаты, если судить по величине косинусного расстояния.

В таблице 3 представлено сравнение ранжирования с применением кластеризации по комбинированному списку и ранжирования без применения кластеризации, но поиск при этом также производится по комбинированному списку.

Таблица 3 – Сравнение ранжирования с применением кластеризации и без

Ранжирование без применением кластеризации (комбинированный список)		Ранжирование с применением кластеризации (комбинированный список)	
id	Косинусное расстояния	id	Косинусное расстояния
7883	0,632	13947	0,616
2985	0,637	3613	0,621
7145	0,638	4474	0,629
9858	0,644	2985	0,637
3431	0,654	7145	0,638
3774	0,656	3774	0,656
12302	0,661	12302	0,661
13672	0,664	13672	0,664
3844	0,686	3844	0,686
5138	0,698	5138	0,698

Для данного запроса лучшие результаты в процессе ранжирования были произведены при помощи поиска в векторном представлении комбинированного списка без использования кластеризации.

С целью повышения качества ранжирования при использовании алгоритмов кластеризации было опробовано использование самоорганизующихся карт с меньшим количеством нейронов (соответственно и кластеров тоже), но такой подход привёл к драматическому снижению релевантности найденных документов.

На рисунке 17 представлена диаграмма, содержащая пример ответа на запрос *«reverse osmosis for seawater desalination».*

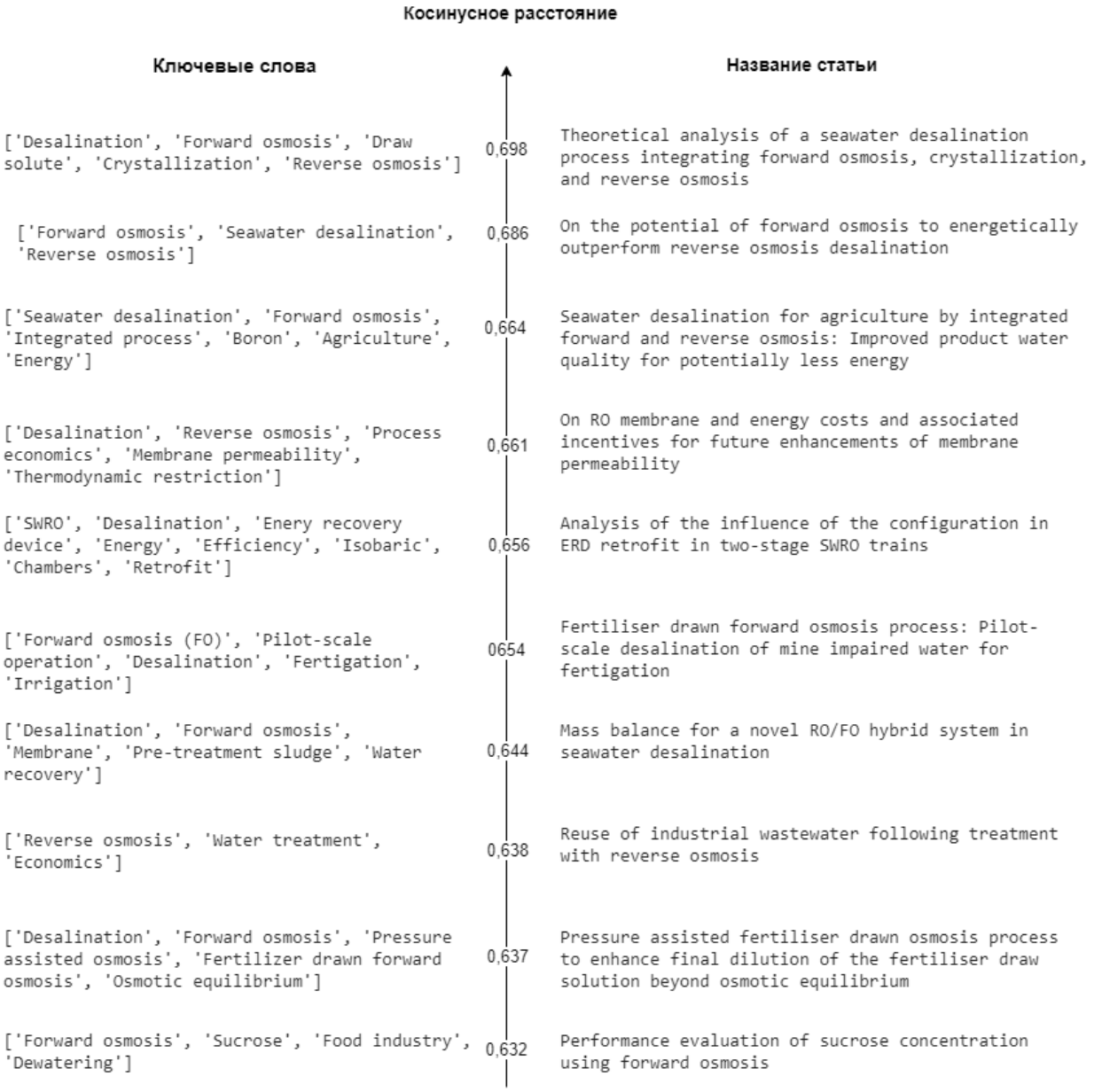


Рисунок 17 – Полученные результаты на запрос *«reverse osmosis for seawater desalination»*

Таким образом, при помощи эксперта было сформулировано 10 разносторонних вопросов (насколько это позволяет тематика собранного датасета) запросов, которые позволили также количественно оценить релевантность документов, возвращаемых поисковой системой по запросу. Результаты представлены в таблице 4.

Таблица 4 – Среднее косинуснинусное расстояние десяти наиболее релевантных документов

Ранжирование с применением кластеризации по именованным сущностям	Ранжирование с применением комбинированного метода кластеризации	Ранжирование без применением кластеризации
0.615±0,087	0.624±0,092	0.642±0,078
0.605±0,089	0.629±0,09	0.644±0,078
0.607±0,088	0.634±0,089	0.649±0,076
0.611±0,087	0.636±0,089	0.651±0,075
0.616±0,086	0.638±0,09	0.654±0,074
0.625±0,087	0.645±0,087	0.656±0,074
0.628±0,087	0.651±0,085	0.662±0,074
0.637±0,083	0.661±0,085	0.667±0,073
0.642±0,086	0.669±0,088	0.675±0,071
0.667±0,095	0.672±0,099	0.684±0,070

Таким образом, в случае с имеющейся базой знаний при применении самоорганизующиеся алгоритма кластеризации подбираются в некоторой степени мене релевантные документы, что обусловлено ограничением поиска только внутри заданного кластера. Объяснить это явление можно тем, что внутри журнала тематики исследований в высокой степени схожи и могу пересекаться, а выигрыш от повышения скорости обработки запроса не столь значителен при размере коллекции всего в 14,5 тыс. документов, поэтому деление на кластеры в рамках данного журнала не целесообразно. Вероятно, к кластеризации можно прибегнуть при расширении коллекции и добавлении других журналов.

3.5 Суммаризация

В качестве модели для суммаризации была взята конфигурация Longformer allenai/led-large-16384-arxiv [66], обученная на датасете pubmed [67] и дообучена на собственном датасете.

Выбор данной модели был обусловлен её сниженным потреблением памяти относительно длины обрабатываемой последовательности [68] при том, что модель выдаёт результаты на уровне state-of-the-art [42]. Модель предобученная на корпусе pubmed была выбрана из-за некоторой схожести тематики решаемой задачи, ведь модель была обучена на корпусе из научных статей по темам, связанным с медициной.

Для обучения было выделено около 12,5 тыс. статей, для валидации около одной тысячи. Выборки не имели совпадающих текстов и аннотаций.

На рисунке 18 представлен график обучения модели. На графике представлена динамика метрик ROUGE-2 precision, ROUGE-2 recal и ROUGE-2 F1-score на валидационной выборке.

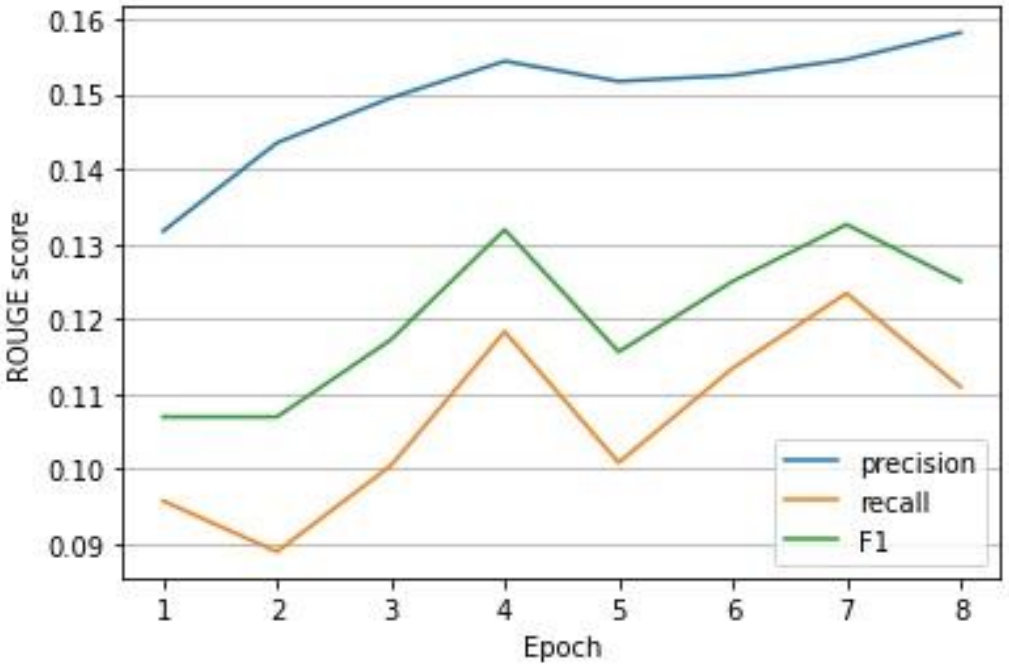


Рисунок 18 – Динамика обучения модели

Модель обучалась в среде Google Colaboratory в среде с использованием графического процессора. Тестирование производилось на текстах, которые не были включены ни в обучающую, ни в валидационную выборку. Всего для тестирования было отобрано 256 статей.

В таблице 5 представлены результаты тестов дообученной и исходной моделей.

Таблица 5 – Тест моделей по суммаризации статьи

Модель	ROUGE-2			FKGL	GFI	CLI
	Precision	Recall	F1			
Дообученная модель	**16,28**	**12,80**	**13,62**	**43,11±11,43**	**13,15±2,35**	**13,64±2,44**
Исходная модель	15,22	12,72	13,07	35,92±14,93	16,52±3,44	14,39±3,18
Характеристики аннотаций статей	-	-	-	30,16±21,68	16,42±7,48	15,51±2,93

Таким образом, за счёт дообучения модели удалось в некоторой мере повысить показатели качества суммаризации. У обученной модели несколько выше показатели ROUGE-2 precision, recall и F1

Чем выше значение индекса удобочитаемости Флеша-Кинкейда, тем выше читаемость текста (от 0 до 100). Что касается индекса туманности Ганнинга и индекса Колман-Лиау, то тут наоборот - чем выше значения этих индексов, тем сложнее даётся прочтение данного текста.

Из таблицы также видно, что аннотация, которая была использована в качестве примера для обучения суммаризации, что текст сам по себе был трудночитаемым.

3.6 Структура программы

На рисунке 19 представлена диаграмма классов.

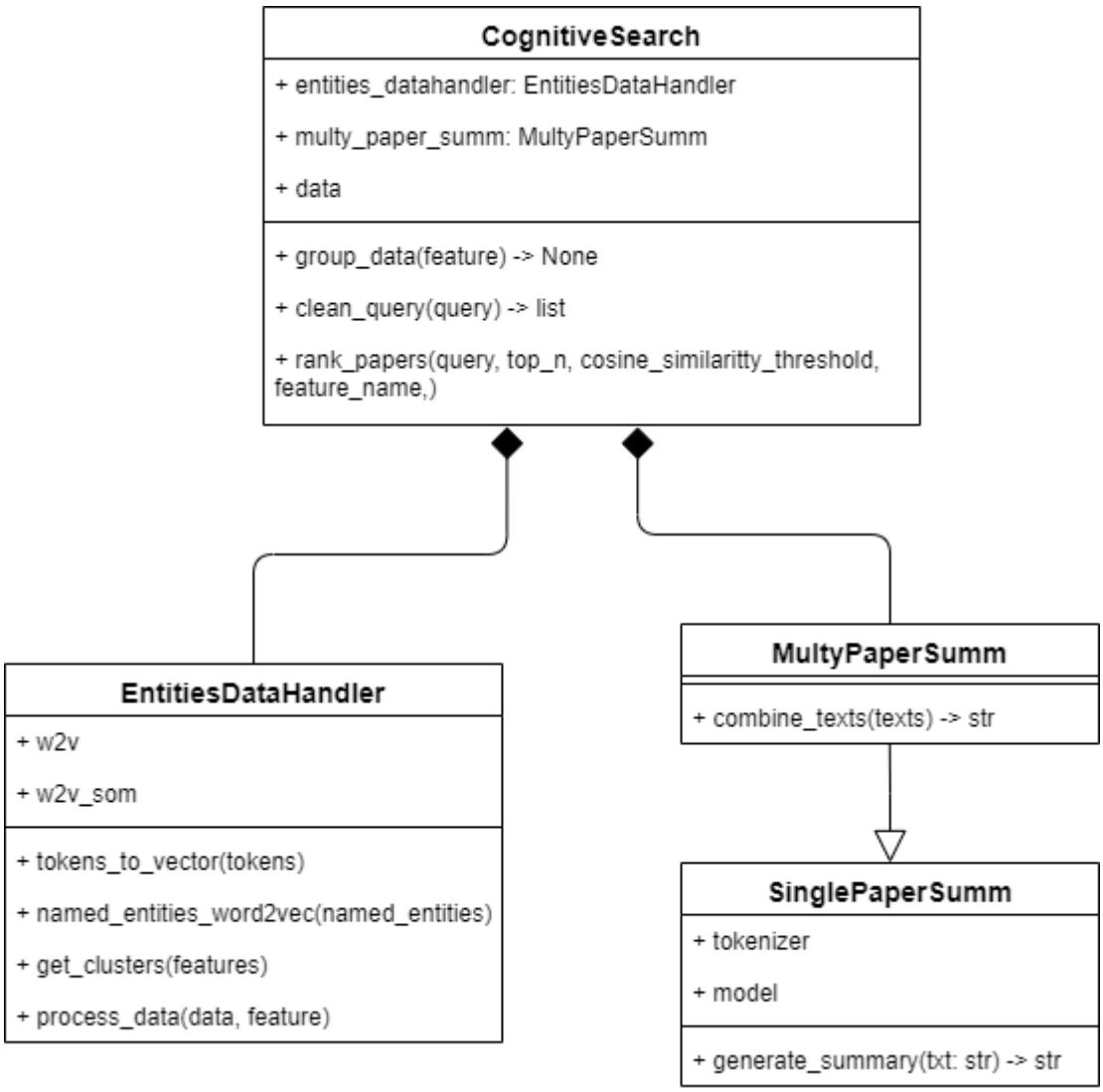


Рисунок 19 – Диаграмма классов разрабатываемой системы поиска

Основная логика прописана в классе CognitiveSearch. Обработка данных (кластеризация и получение векторных представлений).

На данный момент суммаризация статей как общая, так и единичная, производится при помощи одной ранее обученной модели.

3.7 Выводы по разделу

В результате разработки алгоритма поиска по контексту был обучен на собственном датасете эмбединг для создания векторного представления текста, модель в последствии была изучена на предмет понимания контекста предметной области, в результате чего было установлено удовлетворительное понимание понятий области применения.

Далее был обучен алгоритмы кластеризации на основе самоорганизующихся карт, которые были обучены на различных векторных представлениях документов.

В качестве векторных представлений документов использовались сумма векторов ключевых слов, именованных сущностей, а также их комбинация. Тестирование показало наиболее удачным способом – векторное представление документа, на основе объединённого (комбинированного) списка.

В качестве алгоритма кластеризации для разбиения пространства поиска на кластеры и отнесения векторного представления запроса к одному из кластеров были выбраны карты Кохонена. Также лучшие результаты (по тому, как изменилось косинусное расстояние от векторного представления запроса до представления документа) показала конфигурация карт с шестью нейронами, хотя согласно индексу качества Дэвиса-Болдуина, который составил 3,05, данная конфигурация наихудшая из опробованных. Всё же результаты ещё несколько лучше удалось объяснить без применения кластеризации данных. Последнее может быть вызвано тем, что у журнала, на основе которого происходило обучение и тестирование, узкая специализация.

Для решения задачи суммаризации отдельное научной статьи трансформер Longformer encoder-decoder был дообучен собственном датасете, что позволило улучшить показатель ROUGE-2 F1-метрики с 13,07 до 13,62. На данном этапе развития проекта для построения общего реферата по теме была выбрана та же модель.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА
«ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ»

Студенту:

Группа	ФИО
8ВМ03	Хайров Марк Альбертович

Школа	ИШИТР	Отделение школы (НОЦ)	ОИТ
Уровень образования	магистр	Направление/специальность	Информатика и вычислительная техника

Исходные данные к разделу «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение»:

1. *Стоимость ресурсов научного исследования (НИ): материально-технических, энергетических, финансовых, информационных и человеческих*	*1.Рыночная стоимость материальных ресурсов и специального оборудования;* *2.Тарифные ставки исполнителей определены штатным расписанием НИ ТПУ.*
2. *Нормы и нормативы расходования ресурсов*	*Норма амортизации 10%.* *30% премии; 20% надбавки; 13,5% дополнительная заработная плата; 16% накладные расходы; 1,3 районный коэффициент.*
3. *Используемая система налогообложения, ставки налогов, отчислений, дисконтирования и кредитования*	*Отчисления во внебюджетные фонды (30%).*
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. *Оценка коммерческого и инновационного потенциала НТИ*	*1.Потенциальные потребители результатов исследования;* *2.Анализ конкурентных технических решений;* *3.SWOT – анализ.*
2. *Разработка устава научно-технического проекта*	*1.Цели и результат проекта.* *2.Организационная структура проекта.* *3.Ограничения и допущения проекта.*
3. *Планирование процесса управления НТИ: структура и график проведения, бюджет, риски и организация закупок*	*1.Структура работ в рамках научного исследования;* *2.Определение трудоемкости выполнения работ и разработка графика проведения научного исследования;* *3.Расчет бюджета научно - технического исследования (НТИ).*
4. *Определение ресурсной, финансовой, экономической эффективности*	*1.Определение интегрального финансового показателя разработки;* *2.Определение интегрального показателя ресурсоэффективности разработки;* *3.Определение интегрального показателя эффективности.*
Перечень графического материала *(с точным указанием обязательных чертежей):*	
1. *Сегментирование рынка* 2. *Оценка конкурентоспособности технических решений* 3. *Диаграмма FAST* 4. *Матрица SWOT* 5. *График проведения и бюджет НТИ* 6. *Потенциальные риски*	

7. *Оценка ресурсной, финансовой и экономической эффективности НТИ*

Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
доцент	Былкова Т.В.	к.э.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ03	Хайров Марк Альбертович		

4 ФИНАНСОВЫЙ МЕНЕДЖМЕНТ, РЕСУРСОЭФФЕКТИВНОСТЬ И РЕСУРСОСБЕРЕЖЕНИЕ

В качестве объекта исследования выступает система научного поиска с функцией генерации обзорной научной статьи

Цель дипломной работы: разработка системы для автоматизации анализа научной литературы.

Целью раздела «Финансовый менеджмент, ресурсоэффективность и ресурсосбережение» является определение эффективности научно-исследовательского проекта.

4.1 Предпроектный анализ

В процессе написания магистерской диссертации были определены потенциальные потребители результатов исследования. К ним можно отнести научно-исследовательские институты, различные проектные организации и университеты. В качестве основных критериев сегментирования можно выделить размер организации и её тип (НИИ, университет или коммерческая проектная организация). В таблице 6 представлена карта сегментации рынка химических баз знаний.

Таблица 6 – Карта сегментирования рынка химических баз знаний

Критерий		Вид компании		
		НИИ	Университеты	Проектные организации
Размер компании	Крупные			
	Средние			
	Малые			

	Reaxys		SciFinder (CAS)		PubChem

В результате анализа сегментов рынка химических баз знаний в качестве основных клиентов были выявлены потенциальные потребители разрабатываемого продукта, ими оказались средние и малые проектные организации. По мере развития продукта возможен выход и на другие сегменты рынка с использованием, например, подписок разного уровня.

На данный момент на рынке существуют следующие аналоги разрабатываемой системы:

- Reaxys;
- SciFinder (CAS);
- PubChem.

Таблица 7 – Оценочная карта для сравнения конкурентных технических решений

Критерии оценки	Вес критерия	Баллы				Конкурентноспособность			
		$Б_ф$	$Б_{К1}$	$Б_{К2}$	$Б_{К3}$	$К_ф$	$К_{К1}$	$К_{К2}$	$К_{К3}$
1	2	3	4	5	6	7	8	9	10
Технические критерии оценки ресурсоэффективности									
Функциональность	0,25	4	4	4	4	1	1	1	1
Устойчивость	0,1	3	4	4	4	0,3	0,4	0,4	0,4
Безопасность	0,1	4	4	4	4	0,4	0,4	0,4	0,4
Простота интерфейса	0,15	4	3	3	4	0,6	0,45	0,45	0,45
Экономические критерии оценки ресурсоэффективности									
Конкурентоспособность продукта	0,1	2	3	4	4	0,2	0,3	0,4	0,4
Область применения	0,15	3	2	4	3	0,45	0,3	0,6	0,45
Уровень проникновения на рынок	0,05	0	3	4	4	0	0,15	0,2	0,2
Поддержка продукта	0,1	2	5	3	4	0,2	0,5	0,3	0,4
Итого	1					3,15	3,5	3,75	3,85

Экспертная оценка основных технических и экономических характеристик конкурентных решений позволяет сказать, что разрабатываемая система является конкурентной по сравнению с аналогами.

Основными достоинствами разрабатываемой системы являются функциональность, в том числе особые функции системы как генерация обзорной научной статьи, весьма широкая область применения и простота пользовательского интерфейса.

Основными недостатками разрабатываемого продукта являются отсутствие базового проникновения на рынок, слабо налаженная техническая поддержка продукта.

Таким образом появляется необходимость в разработке стратегии продвижения продукта, а также организации сервиса поддержки.

В качестве объекта FAST-анализа выступает объект исследования. В глобальном смысле – это разрабатываемая система научного поиска с функцией генерации обзорных научных статей.

Главной функцией объекта является автоматизированный научный поиск.

В системе имеются три основные подсистемы:

1. Система сбора данных;
2. База знаний;
3. Система генерации обзорной научной статьи.

В качестве вспомогательной подсистемы можно выделить систему навигации пользователя, которая предназначена для предоставления пользователям доступа к функционалу.

В таблице 8 представлено описание функций, выполняемых объектом.

Таблица 8 – Классификация функций, выполняемых объектом исследования

Функция №	Наименование элемента	Выполняемая функция	Ранг функции		
			Главная	Основная	Вспомогательная
0	Система научного поиска	Автоматизированный научный поиск	X		
1	Система сбора данных	Формирование графа знаний из файлов с научными статьями		X	
2	База знаний	Хранение знаний и доступ к ним		X	
3	Система генерации обзорной статьи	Генерация обзорной научной статьи по заданной теме		X	
4	Система навигации пользователя	Предоставления предоставление пользователям доступа к функционалу			X

В дальнейшем анализе будут рассмотрены только основные и вспомогательная функция (Функция 1-4), т.к. объект и выполнение главной функции обеспечивается работой его подсистем.

Для оценки значимости функций используется метод расстановки приоритетов. В основу данного метода положено расчётно-экспертное определение значимости функции.

На первом этапе строится матрица смежности функций (Таблица 9).

Таблица 9 – Матрица смежности

	Функция 1	Функция 2	Функция 3	Функция 4
Функция 1	=	=	=	>
Функция 2	=	=	=	>
Функция 3	=	=	=	>
Функция 4	<	<	<	=

Второй этап связан с преобразованием матрицы смежности в матрицу количественных соотношений функции (Таблица 10).

Таблица 10 – Матрица количественных соотношений функции

	Функция 1	Функция 2	Функция 3	Функция 4	Итого
Функция 1	1	1	1	0,5	4
Функция 2	1	1	1	0,5	4
Функция 3	1	1	1	0,5	4
Функция 4	0,5	0,5	0,5	1	3
Сумма					15

В таблице 11 представлены коэффициенты значимости функций и примерное определение их стоимости разработки исходя из оценки времени на реализацию функционала.

Таблица 11 – Определение стоимости функций, выполняемых объектом исследования

Функция №	Наименование элемента	Выполняемая функция	Коэффициент значимости	Трудоёмкость элемента (дни)	Себестоимость
1	Система сбора данных	Формирование графа знаний из файлов с научными статьями	0.26	15	28578,66
2	База знаний	Обеспечивает хранение знаний и доступ к ним	0.26	12	14822,52
3	Система генерации обзорной статьи	Генерация обзорной научной статьи по заданной теме	0.28	26	32115,46
4	Система навигации пользователя	Предоставления предоставление пользователям доступа к функционалу	0.20	10	19052,66

На рисунке 20 представлена функционально стоимостная диаграмма.

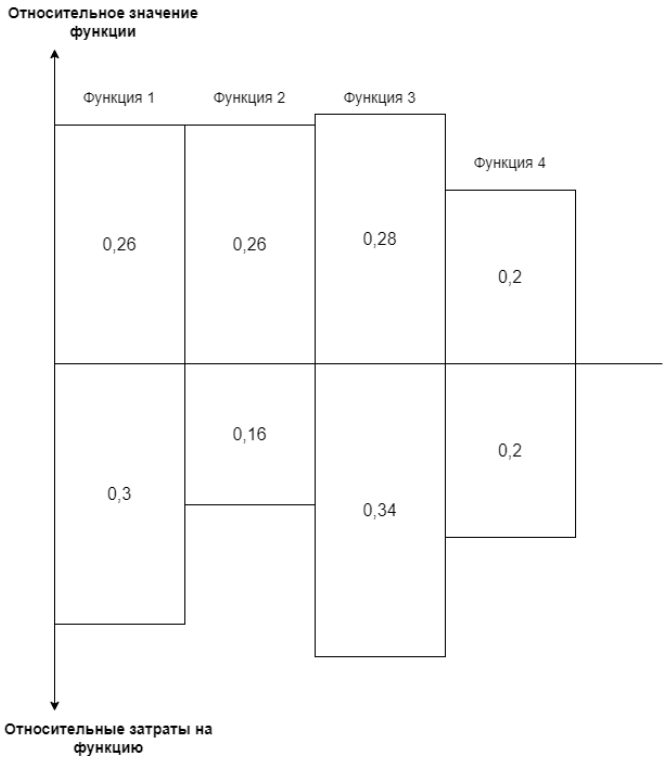


Рисунок 20 – Функционально-стоимостная диаграмма

Исходя из диаграммы (Рисунок 20) видно, что относительные значение функций, выполняемых системой сбора данных, базой знаний и системой генерации обзорной научной статьи (основных функций) имеют приблизительно одинаковые значения. Связано это с тем, что между основными функциями и главной имеется тесная взаимосвязь. Таким образом, невозможно решить поставленную задачу без качественной реализации каждой из основных функций.

Что касается системы навигации пользователя, то в первом приближении (для прототипа) можно пренебречь данной функцией и выделить её в качестве побочной.

Приведем матрицу SWOT-анализа для разрабатываемого продукта (Таблица 12).

Таблица 12 - Матрица SWOT

	Сильные стороны С1. Широкая область применения; С2. Возможность генерации обзорной научной статьи; С3. Простота пользовательского интерфейса; С4. Актуальность разработки; С5. Низкая стоимость производства по сравнению с другими технологиями.	**Слабые стороны** Сл1. Небольшой опыт команды разработчиков; Сл2. Ограниченный бюджет разработки; Сл3. Сложность обновления базы знаний; Сл4. Отсутствие узнаваемости продукта; Сл5. Неопределённость в коммерческом успехе продукта.
Возможности В1. Захват смежных сегментов рынка; В2. Спрос на разрабатываемый продукт и используемые технологии; В3. Дотации и льготы от государства и/или гранты.	В1С1 – продвижение продукта на смежном рынке;	В2Сл1 – наем квалифицированного персонала; В3Сл2 – привлечение дополнительного финансирования за счёт грантов и программ от государства;
Угрозы У1. Отсутствие дополнительного спроса на подобного рода системы; У2. Появление и развитие аналогичных систем; У3. Дефицит или высокая стоимость вычислительных ресурсов.	У1С1С4 – продвижение продукта на смежных рынках; У2С1С2С3 – опора на базовые преимущества проекта; У2С5 – изменение стоимости подписки, создание специальных предложений.	У3Сл3Сл5 – слияние с конкурентным проектом; У2Сл4 – организация рекламной компании.

Таким образом, можно сделать вывод, что для успешной реализации рассматриваемого программного продукта прежде всего стоит обеспечить рекламную компанию, также предусмотреть специальные предложения для привлечения клиентов в том числе со смежных рынков.

Представим результат оценки степень проекта готовности к коммерциализации. В таблице 13 представлен перечень вопросов, позволяющих выяснить проработанность проекта с точки зрения коммерциализации и компетенций разработчика.

Таблица 13 - Бланк оценки степени готовности научного проекта к коммерциализации

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
1.	Определен имеющийся научно-технический задел	3	4
2.	Определены перспективные направления коммерциализации научно-технического задела	3	3
3.	Определены отрасли и технологии (товары, услуги) для предложения на рынке	3	2
4.	Определена товарная форма научно-технического задела для представления на рынок	2	2
5.	Определены авторы и осуществлена охрана их прав	3	3
6.	Проведена оценка стоимости интеллектуальной собственности	2	2
7.	Проведены маркетинговые исследования рынков сбыта	3	2

Продолжение таблицы 13

№ п/п	Наименование	Степень проработанности научного проекта	Уровень имеющихся знаний у разработчика
8.	Разработан бизнес-план коммерциализации научной разработки	2	2
9.	Определены пути продвижения научной разработки на рынок	3	3
10.	Разработана стратегия (форма) реализации научной разработки	3	2
11.	Проработаны вопросы международного сотрудничества и выхода на зарубежный рынок	2	2
12.	Проработаны вопросы использования услуг инфраструктуры поддержки, получения льгот	2	2
13.	Проработаны вопросы финансирования коммерциализации научной разработки	3	3
14.	Имеется команда для коммерциализации научной разработки	4	3
15.	Проработан механизм реализации научного проекта	2	2
	ИТОГО БАЛЛОВ	40	37

Согласно таблице 13 готовность проекта к коммерциализации находится на среднем уровне. Этот уровень можно повысить более детально, проработав бизнес-план и вопросы международного сотрудничества, доработав маркетинговые исследования, а также за счёт повышением квалификации

разработчика в области интеллектуального права, продвижения продукта и других вопросов.

Для разрабатываемого проекта лучше всего подойдёт такой метод коммерциализации как торговля патентными лицензиями. Передача третьим лицам права использования системы на основе лицензий с различным уровнем прав и доступным функционалом.

Многие конкурентные решения также работают согласно модели торговли патентными лицензиями.

4.2 Инициация проекта

Устав научного проекта магистерской работы:

1. Цели и результат проекта.

Приведем информацию о заинтересованных сторонах проекта, иерархии целей проекта и критериях достижения целей.

В таблице 14 представлены заинтересованные стороны и их ожидания от проекта.

Таблица 14 - Заинтересованные стороны проекта

Заинтересованные стороны проекта	Ожидания заинтересованных сторон
Научные и проектные организации	Появление системы, позволяющей автоматизировать процесс научного поиска
Инвесторы	Получение прибыли от ранее предоставленного капитала

Представим информацию об иерархии целей проекта и критерия достижения целей в таблице 15.

Таблица 15 - Цели и результаты проекта

Цели проекта:	Разработка системы научного поиска с возможностью генерации обзорной научной статьи

Продолжение таблицы 15

Ожидаемые результаты проекта:	База знаний научных статей с системой генерации обзорной научной статьи
Критерии приемки результата проекта:	Объём базы знаний, метрики качества суммаризации
Требования к результату проекта:	Требования:
	Время работы алгоритма не превышает заданное техническим заданием
	Качество суммаризации удовлетворяет требованиям, поставленным в техническом задании
	Алгоритмы должны иметь определённый уровень толерантности к некорректным вводным

2. Организационная структура проекта. Определим участников рабочей группы данного проекта, роль каждого участника в данном проекте, а также функции, выполняемые каждым из участников и их трудозатраты в проекте. Представим эту информацию в таблице 16.

Таблица 16 - Рабочая группа проекта

№ п/п	ФИО, основное место работы, должность	Роль в проекте	Функции	Трудозатраты, час.
1	Хайров М.А., ТПУ, магистрант	Инженер-программист	Основной разработчик проекта	540
2	Иванова Ю.А., ТПУ, доцент	Консультации по основным вопросам темы	Руководитель проекта	30
ИТОГО:				570

3. Ограничения и допущения проекта. Ограничения проекта – все факторы, которые могут послужить ограничением степени свободы участников команды проекта, а также «границы проекта» – параметры проекта или его продукта, которые не будут реализованных в рамках данного проекта. Представим эту информацию в таблице 17.

Таблица 17 - Ограничения проекта

Фактор	Ограничения/ допущения
Сроки проекта:	4 месяца
Дата утверждения плана управления проектом	31.01.2022
Дата завершения проекта	31.05.2022

4.3. Планирование управления научно-техническим проектом

Итогом планирования станет график проведения работ с указанием структуры и продолжительности, а также участников работ.

4.3.1 План проекта

В таблице 18 приведены этапы, работы, руководители и исполнители.

Таблица 18 – Структура запланированных работ

Основные этапы	№ раб	Содержание работ	Должность исполнителя
Разработка технического задания	1	Составление и утверждение технического задания	Руководитель
Выбор направления исследований	2	Подбор и изучение материалов по теме	Руководитель, инженер
Выбор направления исследований	3	Выбор направления исследований	Руководитель, инженер
	4	Календарное планирование работ	Руководитель
Формирование базы знаний	5	Разработка алгоритма сбора данных для базы знаний	Инженер
	6	Утверждение алгоритма сбора данных для базы знаний	Руководитель
	7	Сбор данных и формирование баз знаний	Инженер
Исследования в рамках проекта	8	Изучение алгоритмов суммаризации текста, Адаптация выбранного алгоритма суммаризации под свою задачу	Инженер
	9	Утверждение и внесение изменений в выбранную технологию	Руководитель
	10	Разработка алгоритмов суммаризации	Инженер
	11	Анализ результатов, исследование и подтверждение адекватности принятых решений	Руководитель, инженер
	12	Изменение утвержденной технологии согласно проведенным исследованиям	Руководитель, инженер
	13	Обучение алгоритмов	Инженер
Обобщение и оценка результатов	14	Оценка эффективности полученных результатов	Руководитель, инженер
Разработка технической документации	15	Составление документации	Инженер
	16	Составление пояснительной записки	Инженер

Таким образом, основная работа выполняется инженером, при этом основные решения принимаются исходя из профессионального опыта руководителя проекта.

В рамках планирования научного проекта необходимо построить календарный график проекта. Линейный график представлен в таблице 19.

Таблица 19 - Календарный план проекта

Код работы	Название	Длительность, дни	Дата начала работ	Дата окончания работ	Состав участников
1	Составление и утверждение технического задания	1	31.01.2022	31.01.2022	Иванова Ю.А., Хайров М.А.
2	Подбор и изучение материалов по теме	4	01.02.2022	04.02.2022	Хайров М.А.
3	Выбор направления исследований	2	07.02.2022	08.02.2022	Хайров М.А.
4	Календарное планирование работ	2	09.02.2022	10.02.2022	Хайров М.А.
5	Разработка алгоритма сбора данных для базы знаний	12	11.02.2022	28.02.2022	Хайров М.А.
6	Утверждение алгоритма сбора данных для базы знаний	3	01.03.2022	03.03.2022	Иванова Ю.А., Хайров М.А.
7	Сбор данных и формирование баз знаний	12	04.03.2022	18.03.2022	Хайров М.А.
8	Изучение алгоритмов суммаризации текста, Адаптация выбранного алгоритма суммаризации под свою задачу	9	21.03.2022	31.03.2022	Хайров М.А.
9	Утверждение и внесение изменений в выбранную технологию	2	01.04.2022	04.04.2022	Иванова Ю.А., Хайров М.А.
10	Разработка алгоритмов суммаризации	15	05.04.2022	25.04.2022	Хайров М.А.

Продолжение таблицы 19

Код работы	Название	Длительность, дни	Дата начала работ	Дата окончания работ	Состав участников
11	Анализ результатов, исследование и подтверждение адекватности принятых решений	4	26.05.2022	29.05.2022	Иванова Ю.А., Хайров М.А.
12	Изменение утвержденной технологии согласно проведенным исследованиям	7	04.05.2022	16.05.2022	Иванова Ю.А., Хайров М.А.
13	Обучение алгоритма	3	17.05.2022	19.05.2022	Хайров М.А.
14	Оценка эффективности полученных результатов	1	20.05.2022	20.05.2022	Иванова Ю.А., Хайров М.А.
15	Составление документации	1	23.05.2022	23.05.2022	Хайров М.А.
16	Составление пояснительной записки	2	24.05.2022	25.05.2022	Хайров М.А.
ИТОГО:		80			

По этим данным строим диаграмму Ганта для максимальной длительности работ (Таблица 20).

Таблица 20 – Календарный план-график выполнения работ по проекту

№	Вид работ	Испол-нители	$T_{кi}$, кал. дн.	Продолжительность выполнения работ											
				февр.			март			апрель			май		
				1	2	3	1	2	3	1	2	3	1	2	3
1	Составление и утверждение технического задания	И	1												
2	Подбор и изучение материалов по теме	И, Р	4												
3	Выбор направления исследований	И	2												

Продолжение таблицы 20

№	Вид работ	Испол-нители	$T_{кi}$, кал. дн.	Продолжительность выполнения работ											
				февр.			март			апрель			май		
				1	2	3	1	2	3	1	2	3	1	2	3
4	Календарное планирование работ	И	2												
5	Разработка алгоритма сбора данных для базы знаний	И	12												
6	Утверждение алгоритма сбора данных для базы знаний	Р, И	3												
7	Сбор данных и формирование баз знаний	И	12												
8	Изучение алгоритмов суммаризации текста, Адаптация выбранного алгоритма суммаризации под свою задачу	И	9												
9	Утверждение и внесение изменений в выбранную технологию	И, Р	2												
10	Разработка алгоритмов суммаризации	И	15												
11	Анализ результатов, исследование и подтверждение адекватности принятых решений	И, Р	4												
12	Изменение утвержденной технологии согласно проведенным исследованиям	И, Р	7												
13	Обучение алгоритма	И	3												
14	Оценка эффективности полученных результатов	И, Р	1												
15	Составление документации	И	1												
16	Составление пояснительной записки	И	2												

■ - работы, выполняемые руководителем (Р);

■ - работы, выполняемые инженером (И).

4.3.2 Бюджет научного исследования

В процессе формирования бюджета, планируемые затраты группируются по статьям, представленным в таблице 21.

Таблица 21 – Группировка затрат по статьям

	Статьи									
Вид работ	Сырье, материалы (за вычетом возвратных отходов), покупные изделия и полуфабрикаты	Специальное оборудование для научных (экспериментальных) работ	Основная заработная плата	Дополнительная заработная плата	Отчисления на социальные нужды	Научные и производственные командировки	Оплата работ, выполняемых сторонними организациями и предприятиями	Прочие прямые расходы	Накладные расходы	Итого плановая себестоимость
1	932	0	145647,98	29129,60	52433,27	0	0	0	139822,07	367964,92

Расчет стоимости материальных затрат производится по действующим прейскурантам или договорным ценам.

Стоимость материалов представлена в таблице 22.

Таблица 22 - Материалы

Наименование	Единица измерения	Кол-во	Цена за ед., руб.	Затраты на материалы, руб.
Бумага, формат А4	Шт	500	1,27	633
Шариковые ручки	Шт	8	37,37	299
Итого:				932

Основная заработная плата ($З_{осн}$) руководителя (инженера) от предприятия рассчитывается по следующей формуле:

$$З_{осн} = З_{дн} \cdot T_p, \quad (7)$$

где

T_p – продолжительность работ, выполняемых научно-техническим работником, раб. дн.;

$З_{дн}$ – среднедневная заработная плата работника, руб.

Среднедневная заработная плата рассчитывается по формуле:

$$З_{дн} = \frac{З_м \cdot М}{F_д}, \quad (8)$$

где

$З_м$ – месячный должностной оклад работника, руб.;

М – количество месяцев работы без отпуска в течение года: 10,4;

$F_д$ – действительный годовой фонд рабочего времени научно-технического персонала, раб. дн.

Месячный должностной оклад работника:

$$З_м = З_{тс} \cdot (1 + k_{пр}) \cdot k_p \quad (9)$$

где $З_б$ – заработная плата по тарифной ставке, руб.: для руководителя 37700 руб., для инженера – 13900 руб.;

$k_{пр}$ – премиальный коэффициент, равный 0,3;

$k_р$ – районный коэффициент, равный 1,3 (г. Томск).

В таблице 23 представлен баланс рабочего времени по проекту.

Таблица 23 - Баланс рабочего времени

Показатели рабочего времени	Руководитель	Инженер
Календарное число дней	365	365
Количество нерабочих дней – выходные дни – праздничные дни	104/14	104/14
Потери рабочего времени: – отпуск – невыходы по болезни	24/10	24/10
Действительный годовой фонд рабочего времени	213	213

Месячный должностной оклад руководителя и инженера соответственно:

$$З_{м\ рук.} = 37700 \cdot (1 + 0{,}3) \cdot 1{,}3 = 63713 \text{ руб.}$$

$$З_{м\ инж.} = 13900 \cdot (1 + 0{,}3) \cdot 1{,}3 = 23491 \text{ руб.}$$

Среднедневная заработная плата руководителя и инженера соответственно:

$$З_{дн.\ рук.} = \frac{63713 \cdot 11{,}2}{213} = 3350{,}17 \text{ руб.}$$

$$З_{дн.\ инж..} = \frac{23491 \cdot 11{,}2}{213} = 1235{,}21 \text{ руб.}$$

Основная заработная плата руководителя и инженера соответственно:

$$З_{осн\ рук.} = 3350{,}17 \cdot 14 = 46902{,}38 \text{ руб.}$$

$$З_{осн\ инж.} = 1234{,}32 \cdot 80 = 98745{,}60 \text{ руб.}$$

Расчет основной заработной платы приведем в таблице 24.

Таблица 24 – Основная заработная плата исполнителей проекта

Исполнители НИ	$З_т$, руб.	$k_{пр}$	$k_д$	$k_р$	$З_м$, руб.	$З_{дн}$, руб.	$Т_р$, раб. дн.	$З_{осн}$, руб.
Руководитель	37700	0,3	0,2	1,3	63713	3350,17	14	46902,38
Инженер	13900	0,3	0,2	1,3	23491	1235,21	99	98745,60
Итого:								145647,98

Дополнительная заработная плата рассчитывается следующим образом:

$$З_{доп} = k_д \cdot З_{осн} \quad (10)$$

$k_д$ – коэффициент доплат и надбавок, 20 % от $З_{тс}$ (за расширение сфер обслуживания).

В таблице 25 приведена дополнительная заработная плата для исполнителей.

Таблица 25 – Дополнительная заработная плата исполнителей

Заработная плата	Руководитель	Инженер
Основная зарплата	46902,38	98745,60
Дополнительная зарплата	9380,48	19749,12
Зарплата исполнителя	56282,86	118494,72
Итого по статье $С_{зп}$	174777,58	

Величина отчислений во внебюджетные фонды определяется исходя из следующей формулы:

$$С_{внеб} = k_{внеб} \cdot (З_{осн} + З_{доп}), \quad (11)$$

$k_{внеб}$ – коэффициент отчислений на уплату во внебюджетные фонды.

Вычисляется согласно нормам государственного социального страхования (ФСС), пенсионного фонда (ПФ) и медицинского страхования (ФФОМС). Общая ставка взносов составляет 30 %:

- 22 % – на пенсионное страхование;
- 5,1 % – на медицинское страхование;
- 2,9 % – на социальное страхование.

Расчет отчислений во внебюджетные фонды приведем в таблице 26.

Таблица 26 - Отчисления во внебюджетные фонды

Исполнитель	Заработная плата, руб.	Отчисления на социальные нужды, руб.
Руководитель	56282,86	16884,86
Инженер	118494,72	35548,416
Итого:	174777,58	52433,27

Накладные расходы составляют 80% от суммы основной и дополнительной заработной платы работников, непосредственно участвующих в выполнение темы.

$$C_{накл} = k_{накл} \cdot (З_{осн} + З_{доп}) \quad (12)$$

где $k_{накл}$ – коэффициент накладных расходов.

Таким образом статья накладных расходов составит:

$$C_{накл} = 0{,}8 \cdot 174777{,}58 = 139822{,}07 \; руб.$$

3.2.5 Общий бюджет затрат НТИ

Общий бюджет затрат НТИ состоит из затрат на покупные изделия, заработную плату, отчисления на социальные нужды и накладных расходов:

$$C_{\text{проект}} = \text{С}_{\text{пок.изд.}} + \text{С}_{\text{зп}} + \text{С}_{\text{внеб}} + C_{\text{накл}}$$
$$= 932 + 174777{,}58 + 52433{,}27 + 139822{,}07$$
$$= 367964{,}92 \text{ руб.}$$

Таким образом, бюджет затрат НТИ составил 410447,99 руб.

Обратимся к расчету бюджета вариантов исполнения проекта. В качестве аналога разработки рассматривались два технических решения для суммаризации научных статей.

В качестве аналога 1 рассматривалась система с последовательным использованием нескольких моделей для суммаризации: класстеризации, классификации и алгоритм ранжирования предложений.

В качестве аналога 2 рассматривалась модель на основе генератора указателей.

Для реализации было принято построить алгоритм суммаризации на основе трансформера для работы с длинными текстовыми последовательностями.

В случае, если бы в разработку был принят один из аналогов, то пришлось бы пересмотреть сроки и, как следствие, бюджет проекта. Так для аналога 2 пришлось бы увеличить время обучения на 5 дней, а в случае с аналогом 1 необходимо было бы повысить количество дней для разработчика на 12, а для руководителя на 3 (для контроля и оценки качества возросшего числа алгоритмов).

В таблице 27 представлены бюджеты текущего проекта и его аналогов.

Таблица 27 – Группировка затрат по статьям для вариантов исполнения проекта

	Статьи					
Вид работ	Сырье, материалы (за вычетом возвратных отходов), покупные изделия и полуфабрикаты	Основная заработная плата	Дополнительная заработная плата	Отчисления на социальные нужды	Накладные расходы	Итого плановая себестоимость
Текущее исполнение	932	145647,98	29129,60	52433,27	139822,07	367964,92
Аналог 1	932	170510,33	34102,66	61383,72	163689,92	430618,63
Аналог 2	932	151819,58	30363,92	54655,05	145746,80	383517,35

4.3.3 Организационная структура проекта

Выберем структуру, согласно предложенной таблице 28. Так как проект разрабатывается при высокой степени неопределённости, технология проекта новая, сложность проекта высокая, проект находится в жёстких рамках по времени, но при этом, взаимная зависимость проекта от организаций более высокого уровня невысокая, то рассматриваемый научный проект идеально вписывается в проектный тип организации (в таблице представлены критерии выбора).

Таблица 28 – Выбор организационной структуры научного проекта

Критерии выбора	Функциональная	Матричная	Проектная
Степень неопределенности условий реализации проекта	Низкая	Высокая	Высокая
Технология проекта	Стандартная	Сложная	Новая
Сложность проекта	Низкая	Средняя	Высокая
Взаимозависимость между отдельными частями проекта	Низкая	Средняя	Высокая

Продолжение таблицы 28

Критерии выбора	Функциональная	Матричная	Проектная
Критичность фактора времени (обязательства по срокам завершения работ)	Низкая	Средняя	Высокая
Взаимосвязь и взаимозависимость проекта от организаций более высокого уровня	Высокая	Средняя	Низкая

На рисунке 21 представлена примерная схема организационной структуры проекта.

Рисунок 21 – Проектная организационная структура

4.3.4 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. Реестр рисков проекта приведен в таблице 29.

Таблица 29 – Реестр рисков

№	Риск	Потенциальное воздействие	Вероятность наступления (1-5)	Влияние риска (1-5)	Уровень риска	Способы смягчения риска	Условия наступления
1	Низкое качество итоговой системы генерации обзорных научных статей	Отказ от проекта	4	5	Высокий	Увеличение обучающей выборки, смена архитектуры алгоритма суммаризации	Некорректная оценка сложности проекта
2	Отставание от сроков	Выход за рамки бюджета проекта	3	3	Средний	Отказ от некритичного функционала	Ошибки планирования и управления
	Создание более конкурентного продукта другими организациями	Вытеснение разработки аналогом	2	4	Средний	Доработка проекта и расширение функционала	Создание аналогичного проекта другими организациями

4.4 Оценка сравнительной эффективности исследования

Эффективность проекта определяется путём сравнения выбранного варианта исполнения с другими гипотетическими аналогами.

Определение эффективности происходит на основе расчета интегрального показателя эффективности научного исследования. Его нахождение связано с определением двух средневзвешенных величин:

Интегральный финансовый показатель разработки:

$$I_{\text{финр}}^{исп.i} = \frac{\Phi_{\text{p}i}}{\Phi_{\max}}, \quad (13)$$

Φ_{pi} – стоимость *i*-го варианта исполнения;

$\Phi_{\max}$ – максимальная стоимость исполнения научно-исследовательского проекта.

Интегральный показатель ресурсоэффективности:

$$\text{I}_{\text{pi}} = \sum a_i \cdot b_i, \quad (14)$$

где

a_i – весовой коэффициент *i*-го варианта исполнения разработки;

b_i – бальная оценка *i*-го варианта исполнения разработки.

Расчет интегрального показателя ресурсоэффективности приведен в таблице 30.

Таблица 30 - Сравнительная оценка характеристик вариантов исполнения проекта

ПО / Критерии	Весовой коэффициент параметра	Текущий проект	Аналог 1	Аналог 2
1. Способствует росту производительности труда пользователя	0,25	5	5	5
2. Удобство в эксплуатации	0,15	4	4	4
3. Потребность в ресурсах	0,2	4	2	3
4. Устойчивость	0,2	4	3	5
5. Безопасность	0,2	4	3	2
Итого:	1	21	17	19

$$I_{\text{тп}} = 5 \cdot 0{,}25 + 4 \cdot 0{,}15 + 4 \cdot 0{,}2 + 4 \cdot 0{,}2 + 4 \cdot 0{,}20 = 4{,}25$$

$$I_{\text{a1}} = 5 \cdot 0{,}25 \ + 4 \cdot 0{,}15 + 2 \cdot 0{,}2 + 3 \cdot 0{,}2 + 3 \cdot 0{,}20 = 3{,}45$$

$$I_{\text{a2}} = 5 \cdot 0{,}25 \ + 4 \cdot 0{,}15 + 3 \cdot 0{,}2 + 5 \cdot 0{,}2 + 2 \cdot 0{,}20 = 3{,}85$$

Интегральный показатель эффективности вариантов исполнения разработки $I_{\text{исп}.i}$ определяется на основании интегрального показателя ресурсоэффективности и интегрального финансового показателя по формуле:

$$I_{\text{исп}.i} = \frac{I_{p-\text{исп}.i}}{I_{\text{финр}}^{\text{исп}.i}} \tag{15}$$

Сравнительная эффективность проекта:

$$\text{Э}_{\text{ср}} = \frac{I_{\text{исп.1}}}{I_{\text{исп.2}}} \tag{16}$$

В таблице 31 представлен анализ сравнительной эффективности разных вариантов исполнения разработки.

Таблица 31 - Сравнительная эффективность разработки

№ п/п	Показатели	Разработка	Аналог 1	Аналог 2
	Интегральный финансовый показатель разработки	1	1,17	1,04
	Интегральный показатель ресурсоэффективности разработки	4,25	3,45	3,85
	Интегральный показатель эффективности	4,25	2,95	3,70
	Сравнительная эффективность вариантов исполнения	1,44	1	1,25

Таким образом, сравнение значений интегральных показателей эффективности позволил выбрать более эффективный вариант решения поставленной в магистерской диссертации технической задачи с позиции финансовой и ресурсной эффективности. Исходя из интегральных показателей эффективности исполнение разработки является эффективней предложенных аналогов. Эффект достигается за счёт использования технически

сбалансированного решения, которое, хотя и весьма требовательно к ресурсам, является высокоэффективным с точки зрения качества итогового продукта, надёжным и при этом наиболее простым в реализации, значит и наиболее эффективным с точки зрения финансовых затрат на реализацию.

Итак, в качестве потенциальных клиентов разрабатываемого продукта с наибольшей вероятностью выступят малые и средние проектные организации.

Разрабатываемый продукт является относительно конкурентным, но при это имеет ряд слабых сторон, которые следует доработать для повышения конкурентоспособности.

Продукт имеет средний уровень готовности к коммерциализации, это касается как степени проработанности, так и компетенции команды разработчиков.

В результате работы над разделом был установлен календарный план проекта и рассчитан его бюджет, который составил 367964,92 руб..

Был составлен реестр рисков и меры по смягчению негативного воздействия.

Исходя из интегральных показателей эффективности, выбранный вариант исполнения проекта является предпочтительным из-за в виду выбора алгоритма суммаризации высокого качества, надёжности и финансовой эффективности, что достигается сравнительной простотой реализации технического решения.

ЗАДАНИЕ ДЛЯ РАЗДЕЛА «СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ»

Студенту:

Группа		ФИО	
8ВМ03		Хайров Марк Альбертович	
Школа	**Инженерная школа информационных технологий и робототехники**	**Отделение (НОЦ)**	**Отделение информационных технологий**
Уровень образования	магистратура	**Направление/ специальность**	09.04.01 Информатика и вычислительная техника

Тема ВКР:

Разработка системы умного поиска с генерацией обзорных научных статей	
Исходные данные к разделу «Социальная ответственность»:	
Введение – Характеристика объекта исследования (вещество, материал, прибор, алгоритм, методика) и области его применения. – Описание рабочей зоны (рабочего места) при разработке проектного решения/при эксплуатации	*Объект исследования:* математическая модель по суммаризации текста научных статей *Область применения:* информационные технологии, машинное обучение *Рабочая зона:* офис *Размеры помещения:* 5х4 м Система отопления Система вентиляции: естественная Система освещения: совмещённая *Количество и наименование оборудования рабочей зоны:* персональный компьютер *Рабочие процессы, связанные с объектом исследования, осуществляющиеся в рабочей зоне:* разработка математической модели и программного продукта при помощи персонального компьютера
Перечень вопросов, подлежащих исследованию, проектированию и разработке:	
1. Правовые и организационные вопросы обеспечения безопасности <u>при разработке проектного решения:</u> – специальные (характерные при эксплуатации объекта исследования, проектируемой рабочей зоны) правовые нормы трудового законодательства; – организационные мероприятия при компоновке рабочей зоны.	- Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 27.12.2018) - ГОСТ 12.0.003-2015 Опасные и вредные производственные факторы. Классификация. Перечень опасных и вредных факторов - ГОСТ 22269-76 «Рабочее место оператора. Взаимное расположение элементов рабочего места»; - ГОСТ 12.2.032-78 «ССБТ. Рабочее место при выполнении работ сидя. Общие эргономические требования»; - ГОСТ Р 50923-96. Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения; - ГОСТ 21889-76 «Система "Человек-машина". Кресло человека-оператора. Общие эргономические требования» - СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» - СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95*

	- ГОСТ 12.1.003-2014 ССБТ. «Шум. Общие требования безопасности» - ГОСТ 12.1.029-80 ССБТ. Средства и методы защиты от шума. Классификация - ГОСТ 12.1.030-81 Система стандартов безопасности труда (ССБТ). Электробезопасность. Защитное заземление. Зануление - ГОСТ 12.1.038-82 ССБТ. «Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов» - ГОСТ 12.1.005-88 ССБТ. «Общие санитарно-гигиенические требования к воздуху рабочей зоны» - ГОСТ 12.1.007-76 ССБТ. «Вредные вещества. Классификация и общие требования безопасности»
2. Производственная безопасность при разработке проектного решения: – Анализ выявленных вредных и опасных производственных факторов – Расчет уровня опасного или вредного производственного фактора	**Опасные факторы:** – опасность поражения электрическим током; **Вредные факторы:** – повышенный уровень шума; – недостаток необходимого искусственного освещения; – отклонение показателей микроклимата; – повышенный уровень излучения электромагнитных полей; **Средства защиты:** – звукопоглощающие конструкции (подвесной потолок, перегородки); – защитное заземление; – знаки безопасности. **Расчёт:** системы искусственного освещения
3. Экологическая безопасность при разработке проектного решения	Воздействие на селитебную зону, атмосферу и гидросферу не выявлено Воздействие на литосферу при утилизации компьютера и периферийных устройств (аккумуляторы, батарейки, кабели); люминесцентных ламп;
4. Безопасность в чрезвычайных ситуациях при разработке проектного решения	Возможные ЧС: Пандемия; Аварии на коммунальных системах жизнеобеспечения населения; Пожар. Наиболее типичная ЧС: Пожар
Дата выдачи задания для раздела по линейному графику	

Задание выдал консультант:

Должность	ФИО	Ученая степень, звание	Подпись	Дата
Профессор ООД ШБИП	Федоренко О.Ю.	д.м.н.		

Задание принял к исполнению студент:

Группа	ФИО	Подпись	Дата
8ВМ03	Хайров Марк Альбертович		

5 СОЦИАЛЬНАЯ ОТВЕТСТВЕННОСТЬ

В рамках магистерской диссертации разрабатывался проект по созданию системы умного поиска с возможностью генерации обзорных научных статей.

Областью применения данного проекта являются информационные технологии и машинное обучение. Потенциальными пользователями разработки могут стать научно-исследовательские институты, университеты и различные проектные организации.

Обеспечение производственной и экологической безопасности является необходимым условием реализации любых проектов, в том числе конструкторских и исследовательских. В общем, обеспечение безопасности предполагает создание безопасных и благоприятных рабочих условий для лиц, задействованных в работе над проектом, а также условий, обеспечивающих экологическую безопасность окружающей среды.

Данный раздел направлен на анализ вредных и опасных факторов, негативно влияющих на человека и окружающую среду, анализ возможных чрезвычайных ситуаций и выявление мер по защите окружающей среды и персонала. Также в разделе рассматриваются правила поведения при возникновении чрезвычайной ситуации.

Разработка программного продукта производится в офисном помещении с использованием современного персонального компьютера (ПК) по адресу г. Томск, пр. Ленина 45. Размеры помещения - 5х4 м. Помещение оборудовано системой естественной вентиляции и совмещённой системой освещения.

5.1 Правовые и организационные вопросы обеспечения безопасности

Регулирование отношений между работником и работодателем, касающихся оплаты труда, трудового распорядка, особенности регулирования труда женщин, детей, людей с ограниченными способностями и проч., осуществляется трудовым кодексом РФ (ТК РФ) [71].

Согласно ТК РФ, продолжительность рабочей не дели не должна быть меньше установленной в трудовом договоре, но не больше 40 часов в неделю. В течении дня работнику должен быть предоставлен перерыв продолжительностью не более двух часов и не менее 30 минут.

Разработка системы интеллектуального поиска ведётся за рабочим местом с использованием персонального компьютера. Основными документами регламентирующее условия и организацию работы с ПК являются следующие стандарты:

- ГОСТ 22269-76 «Рабочее место оператора. Взаимное расположение элементов рабочего места» [72];
- ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования» [73];
- ГОСТ Р 50923-96 «Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения» [74];
- ГОСТ 21889-76 «Система "Человек-машина". Кресло человека-оператора. Общие эргономические требования» [75].

Соблюдение требований представленных стандартов позволяет снизить влияние вредных факторов при осуществлении работ в положении сидя с использованием ПК.

5.1.1 Организационные мероприятия при компоновке рабочей зоны

Взаимное расположение элементов рабочего места должно способствовать оптимальному режиму труда и отдыха, снижению утомления оператора, предупреждению появления ошибочных действий [72].

Для профилактики статических физических перегрузок сотрудников необходима грамотная компоновка рабочей зоны. Основные элементы рабочей зоны, организация которых влияет на комфорт разработчика: стол, кресло и дисплей.

Рабочие столы по конструктивному исполнению подразделяют на регулируемые и нерегулируемые по изменению высоты рабочей поверхности.

Регулируемая высота рабочей поверхности стола должна изменяться в пределах от 680 до 800 мм. Механизмы для регулирования высоты рабочей поверхности стола должны быть легко досягаемыми в положении сидя, иметь легкость управления и надежную фиксацию.

Высота рабочей поверхности стола при нерегулируемой высоте должна составлять 725 мм [74].

Рабочий стол должен иметь пространство для ног высотой не менее 600 мм, шириной - не менее 500 мм, глубиной на уровне колен - не менее 450 мм и на уровне вытянутых ног - не менее 650 мм [74, 75].

Экран видеомонитора должен находиться от глаз пользователя на расстоянии 600-700 мм, но не ближе 500 мм с учетом размеров алфавитно-цифровых знаков и символов [74].

Конструкция рабочего стула (кресла) должна обеспечивать поддержание рациональной рабочей позы при работе с ПК, позволять изменять позу с целью снижения статического напряжения мышц шейно-плечевой области и спины для предупреждения развития утомления. Тип рабочего стула (кресла) следует выбирать с учетом роста пользователя, характера и продолжительности работы с ПК [75].

5.1.2 Влияние разрабатываемого программного продукта на рабочий процесс

Разрабатываемый программный продукт призван оптимизировать работы по анализу научной литературы, таким образом сократив усилия и время исследователя, затрачиваемые на поиск необходимой информации. Высвобожденное время на поиск и анализ научной литературы позволяет снизить умственные и зрительные нагрузки, а также повысить частоту смены деятельности исследователя.

5.2 Производственная безопасность

В процессе разработки программного продукта, его эксплуатации и поддержки разработчик может подвергнуться влиянию вредных и опасных факторов.

Предполагаемые опасные и вредные факторы, возникающие в разное время жизненного цикла продукта представлены в таблице 32.

Таблица 32 – Перечень опасных и вредных факторов производственной среды

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Разработка	Изготовление	Эксплуатация	
1. Отклонение показателей микроклимата в закрытом помещении	+	+	+	СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [75]
2. Недостаточная освещенность рабочей зоны	+	+	+	СП 52.13330.2016 Естественное и искусственное освещение. Актуализированная редакция СНиП 23-05-95* [76]

Продолжение таблицы 32

Факторы (ГОСТ 12.0.003-2015)	Этапы работ			Нормативные документы
	Разработка	Изготовление	Эксплуатация	
3. Повышенный уровень шума на рабочем месте	+	+	+	ГОСТ 12.1.003-2014 ССБТ. «Шум. Общие требования безопасности» [77] ГОСТ 12.1.029-80 ССБТ. Средства и методы защиты от шума. Классификация [78]
4. Повышенный уровень электромагнитных излучений	+	+	+	СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания» [75] ГОСТ 12.1.006-84 ССБТ. «Электромагнитные поля радиочастот. Допустимые уровни на рабочих местах и требования к проведению контроля» [79]
5. Повышенное значение напряжения в электрической цепи, замыкание которой может произойти через тело человека	+	+	+	ГОСТ 12.1.038-82 ССБТ. «Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов» [80] ГОСТ 12.1.030-81 Система стандартов безопасности труда (ССБТ). Электробезопасность. Защитное заземление. Зануление [81]

5.2.1 Вредные производственные факторы

Далее рассмотрим каждый вредный фактор отдельно, а также проанализируем источники их возникновения, воздействия на организм человека, регламентированные нормы и решения для снижения влияния данного фактора.

5.2.1.1 Отклонения показателей микроклимата

Микроклимат определяется действующими на организм человека показателями температуры, влажности и скорости движения воздуха. Длительное воздействие на человека неблагоприятных показателей микроклимата приводит к снижению производительности труда, а также и к заболеваниям, таким образом, в офисном помещении, где происходят работы над проектом, следует поддерживать параметры микроклимата, регламентируемые СанПиН 1.2.3685-21 [76].

Согласно таблице 5.28 СанПиН 1.2.3685-21, помещение, в котором производятся работы по проекту относится ко второй категории - помещения, в которых люди заняты умственным трудом, учебой. Нормы показателей микроклимата для помещений этого типа представлены в таблице 33.

Таблица 33 – Оптимальные и допустимые параметры микроклимата помещения, в которых люди заняты умственным трудом, учёбой [76]

Период года	Температура воздуха, °С		Результирующая температура, °С		Относительная влажность, %		Скорость движения воздуха, м/с	
	оптимальная	Допустимая	оптимальная	Допустимая	оптимальная	Допустимая	оптимальная	Допустимая
Холодный	19-21	18-23	18-20	17-22	45-30	60-30	0,2	0,3
Тёплый	23-25	18-28	22-24	19-27	60-30	65-30	0,15	0,25

Для поддержания требуемых параметров микроклимата, офисное помещение, где происходит разработка, должно быть оборудовано системой центрального отопления, кондиционером и вентиляцией.

5.2.1.2 Освещение

В комфортном рабочем пространстве важна правильная организация освещения. Недостаточная освещённость рабочей зоны может привести к ухудшению снижению зрения программиста.

Способы освещения помещения могут быть искусственными, естественными и совместным.

В рабочей зоне используется совместный способ освещения.

Работа программиста относится к разряду зрительных работ высокой точности. В таблице 34 представлены требования к искусственному освещению для данного вида работ.

Таблица 34 – Требования к искусственному освещению работ высокой точности [75]

Наименьший или эквивалентный размер объекта различения, мм	Разряд зрительной работы	Подразряд зрительной работы	Контраст объекта с фоном	Средняя освещённость на рабочей поверхности от системы общего освещения, лк, не менее	Коэффициент пульсации $K_{\text{п}}$, %
от 0,3 до 0,5	III	Г	Средний Большой	400	15

Произведём расчёт искусственного освещения для помещения, в котором производилась разработка проекта.

Рабочее помещение имеет следующие параметры: длина А = 5 м, ширина В = 4 м, высота Н = 3 м.

В качестве светильников в помещении использовались открытые двухламповые светильники типа ШОД – 2-80. Мощность лампы 80 Вт. Длина светильника ($l_{\text{св}}$) 1530 мм. Ширина светильника 284 мм. Для создания благоприятных зрительных условий, наименьшая допустимая высота подвеса светильника ШОД над полом 2,5 м.

Принимаем высоту свеса светильника $h_{\text{с}} = 0{,}5$ м, высота нерегулируемой рабочей поверхности была ранее определена как $h_{\text{рп}} = 0{,}725$ м. Таким образом, высота светильника над полом:

$$h_{\text{п}} = H - h_{\text{с}} = 3{,}0 - 0{,}5 = 2{,}5 \text{ м} \quad (1)$$

Тогда расчётная высота светильника над рабочей поверхностью составит:

$$h = h_{\text{п}} - h_{\text{рп}} = 2{,}5 - 0{,}725 = 1{,}875 \text{ м} \quad (2)$$

Расстояние между светильниками определяется как:

$$L = \lambda \cdot h \quad (3)$$

где λ – коэффициент расположения светильников (для светильников ШОД значение коэффициента варьируется в пределах 1.1-1.3).

Таким образом, расстояние между светильниками составит:

$$L = 1{,}1 \cdot 1{,}875 \approx 2.06 \text{ м} \quad (4)$$

Определим количество рядов светильников с люминесцентными лампами:

$$n_{\text{ряд}} = \frac{(B - \frac{2}{3}L)}{L} + 1 = \frac{(4 - \frac{2}{3} \cdot 2{,}06)}{2{,}2} + 1 = 2{,}27 \quad (5)$$

Определим количество светильников с люминесцентными лампами для одного ряда:

$$n_{\text{св}} = \frac{(A - \frac{2}{3}L)}{l_{\text{св}} + 0{,}5} = \frac{(5 - \frac{2}{3} \cdot 2{,}06)}{1{,}53 + 0{,}5} = 1{,}77 \quad (6)$$

Оптимальное расстояние l от крайнего ряда светильников до стены рекомендуется принимать равным L/3, в данном случае это приблизительно 0,686 м.

Так как размещение светильников происходит в два ряда, то на расстояние от торца светильника до стены остаётся $l = 0.686$ м.

На рисунке 22 представлен план размещения светильников

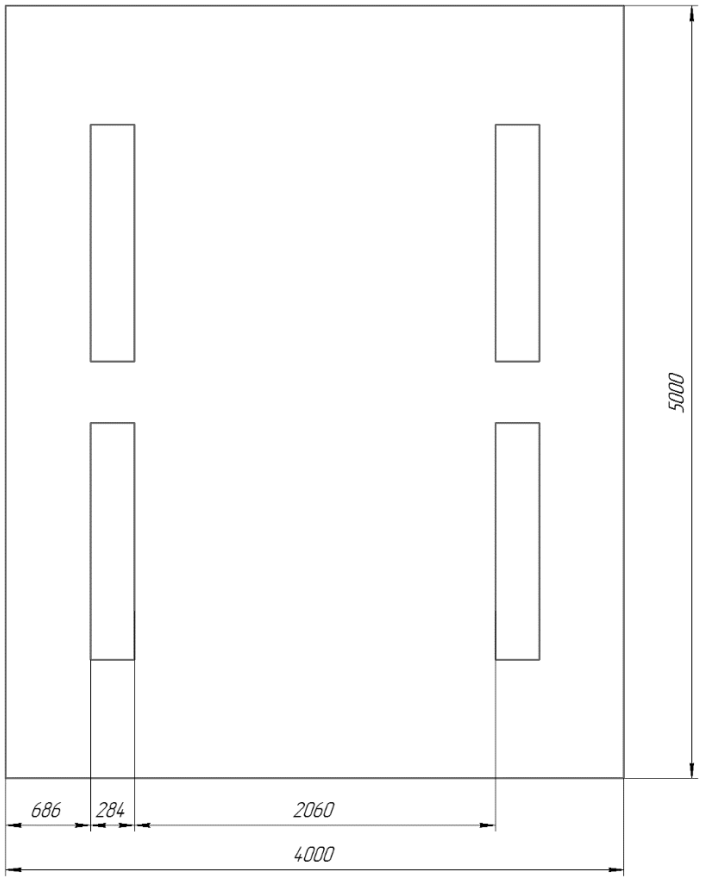


Рисунок 22 – План размещения светильников

Коэффициент использования светового потока показывает, какая часть светового потока ламп попадает на рабочую поверхность. Он зависит от индекса помещения i, типа светильника, высоты светильников над рабочей поверхностью h и коэффициентов отражения стен и потолка.

Индекс помещения определяется по формуле:

$$\mathrm{i} = \frac{\mathrm{S}}{\mathrm{h(A+B)}} = \frac{5 \cdot 4}{2{,}06 \cdot (5+4)} \approx 1.1 \qquad (7)$$

Значение коэффициента отражения поверхности потолка $\rho_{\text{п}}$ для расчёта принимаем равным 50% как для побеленного потолка. Стены в рабочем помещении оклеены светлыми обоями, таким образом коэффициент отражения

стен $\rho_{\text{п}}$ составит 30%. Коэффициент использования светового потока η составит 35%.

Световой поток лампы Ф определяется по следующей формуле:

$$\Phi = \frac{E_n \cdot S \cdot K_{\text{з}} \cdot Z}{N_{\text{л}} \cdot \eta} \quad (8)$$

где E_n – нормативная освещённость по СП 52.13330.2016, лк; S – площадь освещаемого помещения, м^2 ; $K_{\text{з}}$ – коэффициент запаса, учитывающий загрязнение светильника (источника света, светотехнической арматуры, стен и пр., т. е. отражающих поверхностей), наличие в атмосфере цеха дыма, пыли; Z – коэффициент неравномерности освещения (для люминисцентных ламп 1.1), $N_{\text{л}}$ – число ламп в помещении (необходимо учесть число ламп в светильнике).

Принимаем коэффициент запаса $K_{\text{з}}$ равным 1,5 как для помещения с малым выделением пыли.

Таким образом световой поток лампы

$$\Phi = \frac{400 \cdot 4 \cdot 5 \cdot 1{,}5 \cdot 1{,}1}{8 \cdot 0{,}35} = 4714 \text{ лм} \quad (9)$$

Полученное расчётное значение должно удовлетворять следующему условию:

$$-10\% \leq \frac{\Phi_{\text{ст}} - \Phi_{\text{расч}}}{\Phi_{\text{ст}}} \cdot 100\% \leq +20\% \quad (10)$$

Стандартный световой поток люминесцентной лампы белой цветности (ЛБ) мощностью 80 Вт при напряжении сети 220 В составляет $\Phi_{\text{ст}} = 5200$ лм. Таким образом отклонение рассчитанного светового потока составит:

$$\frac{5200 - 4715}{5200} \cdot 100\% = 9{,}35\% \quad (11)$$

Следовательно, необходимые параметры освещённости обеспечиваются с допустимым отклонением.

Номинальная электрическая мощность всей осветительной системы составит:

$$P = N_{л} \cdot p_{л} = 4 \cdot 80 = 320 \text{ Вт} \quad (12)$$

2.1.3 Шум на рабочем месте

Согласно ГОСТ 12.1.003-2014 [77] Шум - это звуковые колебания в диапазоне слышимых частот, способные оказать вредное воздействие на безопасность и здоровье работника.

В случае разработки с использованием ПК источниками шума выступают вентиляторы в системном блоке и накопители типа HDD.

Шум на рабочем месте оказывает раздражающее влияние на работника, повышает его утомляемость, а при выполнении задач, требующих внимания и сосредоточенности, способен привести к росту ошибок и увеличению продолжительности выполнения задания. Длительное воздействие шума влечет тугоухость работника вплоть до его полной глухоты [78].

В таблице 35 представлены допустимые уровни звука на рабочем месте как для офисов, проектных и научно-исследовательских организаций.

Таблица 35 – Допустимые уровни звука на рабочем месте [78]

Уровни звукового давления, дБ, в октавных полосах частот со среднегеометрическими частотами, Гц									Уровень звука, дБ
31,5	631	125	250	500	1000	2000	4000	8000	
86	71	61	54	49	45	42	40	38	50

Для снижения вредного воздействия шума на рабочем месте следует проводить своевременное обслуживание элементов системного блока: чистить их от пыли, смазывать и менять термопасту. Также помещение оборудовано такими средствами звукопоглощения как подвесной потолок с облицовкой из звукопоглощающих материалов и звукопоглощающими перегородками [78].

5.2.1.4 Электромагнитное излучение

Все электрические приборы излучают такие волны, однако наибольший вклад вносит экран монитора.

Воздействие электромагнитного излучения на человека зависит от напряженностей электрического и магнитного полей, потока энергии, частоты колебаний, размера облучаемого тела. При определённых уровнях такие поля оказывают вредное влияние на человека: нарушение функционального состояния нервной и сердечно-сосудистой систем, это проявляется в повышенной утомляемости, понижении качества выполнения рабочих операций, изменении кровяного давления и пульса

В таблице 36 представлены допустимые уровни электрических и магнитных полей промышленной частоты 50 Гц.

Таблица 36 – Предельно допустимые уровни электрических и магнитных полей промышленной частоты 50 Гц [76]

Тип воздействия	Напряженность электрического поля, кВ/м	Индукция (напряженность магнитного поля), мкТл (А/м)
В общественных зданиях	0,5	10,0 (8,0)

Для снижения вредного воздействия следует увеличить расстояние от источника электромагнитного излучения (не менее 50 см от пользователя).

Также следует регулярно проводить контроль напряжённости электромагнитных полей. В ГОСТ 12.1.006-84 ССБТ [80] рекомендуется проводить измерения напряжённости не реже одного раза в год при наибольшей используемой мощности источника на расстоянии соответствующим нахождению тел рабочих на нескольких уровнях от земли.

5.2.2 Опасные производственные факторы

5.2.2.1 Опасность поражения электрическим током

Действие электрического тока на живую ткань носит разносторонний и своеобразный характер. Проходя через организм человека, электрический ток производит термическое, электролитическое, механическое и биологическое воздействия [83].

Исход поражения человека электрическим током зависит от силы тока и времени его прохождения через организм, характеристики тока (переменный или постоянный), пути тока в теле человека, при переменном токе – от частоты колебаний [83].

Поражение электрическим током при работе с ПК может произойти при прикосновении к оголённым токоведущим элементом, что может произойти при нарушении электрической изоляции этих элементов или их пробоя.

Напряжения прикосновения и токи, протекающие через тело человека при нормальном (неаварийном) режиме электроустановки, не должны превышать значений, указанных в таблице 37.

Таблица 37 – Предельно допустимые напряжение и токи прикосновения [81]

Род тока	не более	
	U, В	I, мА
Переменный, 50 Гц	2,0	0,3
Постоянный	8,0	1,0

Для того, чтобы минимизировать риск поражения электрическим током, следует проверять исправностью изоляций проводов перед началом работы, не допускать перегрева приборов, а перед началом разработки провести инструктаж по технике безопасности с исполнителями.

Также в электроустановках переменного тока в сетях с изолированной нейтралью или изолированными выводами однофазного источника питания

электроэнергией защитное заземление должно быть выполнено в сочетании с контролем сопротивления изоляции [82].

В качестве дополнительного средства коллективной защиты можно использовать предостерегающие знаки (например, о том, что объект находится под напряжением).

Исполнителям проекта может быть присвоена I категория персонала по электробезопасности согласно правилам по охране труда при эксплуатации электроустановок [84], которая присваивается неэлектротехническому персоналу после проведения инструктажа специалистом по охране труда, имеющим группу IV и выше.

5.3 Экологическая безопасность

Разработка, эксплуатация и поддержка программного продукта происходит главным образом с использованием ПК, в офисном помещении, отсюда влияние на окружающую среду обусловлено жизненными циклами офисной и компьютерной техники и её эксплуатации.

Вредные вещества, содержащиеся в отработанных люминесцентных лампах и аккумуляторах, могу привести к загрязнению почв или грунтовых вод, если избавляться от них таким же образом, как и от бытовых отходов, к тому же рассматриваемый тип изделий содержит ценные материалы, добыча которых также наносит вред окружающей среде.

Люминесцентные лампы хоть и позволяют сберегать электроэнергию, но содержат пары ртути, которые по степени воздействия на организм человека относится к 1-му классу опасности в соответствии с требованиями ГОСТ 12.1.005-88 [85] и ГОСТ 12.1.007-76 [86].

Вышедшие из строя люминесцентные лампы, должны начинаться с помещения их на хранение в специальные контейнеры в оборудованных под эти нужды помещениях. При накоплении определенного количества ртутьсодержащих и прочих опасных видов ламп их сортируют, помещают в

отдельные ячейки и отправляют в профильную компанию для последующей нейтрализации и переработки [87].

Аккумуляторы от ноутбуков и телефонов содержат тяжёлые металлы, которые, поэтому их нужно передавать в организации, осуществляющие утилизацию аккумуляторов.

Любые отработанные и ненужные кабели нужно сдавать в организации, обладающие лицензией на хранение и переработку лома чёрных и цветных металлов [88]. Кабеля как правило содержат металлы высокого качества, так что организации, занимающиеся утилизацией такого рода отходов даже готовы платить за сдачу лома.

Согласно постановлению правительства РФ от 31 декабря 2020 года, N 2398 объект относится к IV типу объектов, оказывающих негативное воздействие на окружающую среду, как объект с отсутствием загрязняющих выбросов, отсутствием сбросов загрязняющих веществ в сточные воды, с использованием и пр. [89].

Так как рассматриваемое производство не создаёт повышенного уровня шума, электромагнитных излучений за пределами производственного помещения или выбросов вредных веществ, то установление санитарно-защитной зоны не требуется.

5.4 Безопасность в чрезвычайных ситуациях

В процессе разработки программного продукта могут возникнуть такие виды чрезвычайных ситуаций (ЧС) как внезапное обрушение здания, аварии на коммунальных системах жизнеобеспечения населения, пожар, угроза пандемии.

Наиболее вероятным из перечисленных ЧС может стать пожар, который может возникнуть в следствии возгорания неисправности сети электропитания или используемой техники.

5.4.1 Помещение, в котором может возникнуть пожар

Согласно СП 12.13130.2009 [90] помещение, в котором происходит разработка по взрывопожарной и пожарной опасности относится к категории (Д) с пониженной пожароопасностью, так как материалы, которые находятся в обращении в помещении, относятся к негорючим веществам, а все материалы находятся в холодном состоянии при нормальных условиях эксплуатации.

Пожар, который может возникнуть в рассматриваемом помещении – это пожар твёрдых горючих веществ и материалов (класс А) [90].

Первичные средства пожаротушения - средства пожаротушения, используемые для борьбы с пожаром в начальной стадии его развития.

В помещении, в котором ведётся разработка нет таких средств, но они есть в зданиях и расположены на лестничных клетках, где имеются пожарные краны, пожарный инвентарь и переносные огнетушители.

При возникновении задымления в здании срабатывают датчики, которые активируют систему оповещения о возгорании, после чего начинается эвакуация всех, кто находился в помещении, в соответствии с планом эвакуации при пожарах и других ЧС (Рисунок 23).

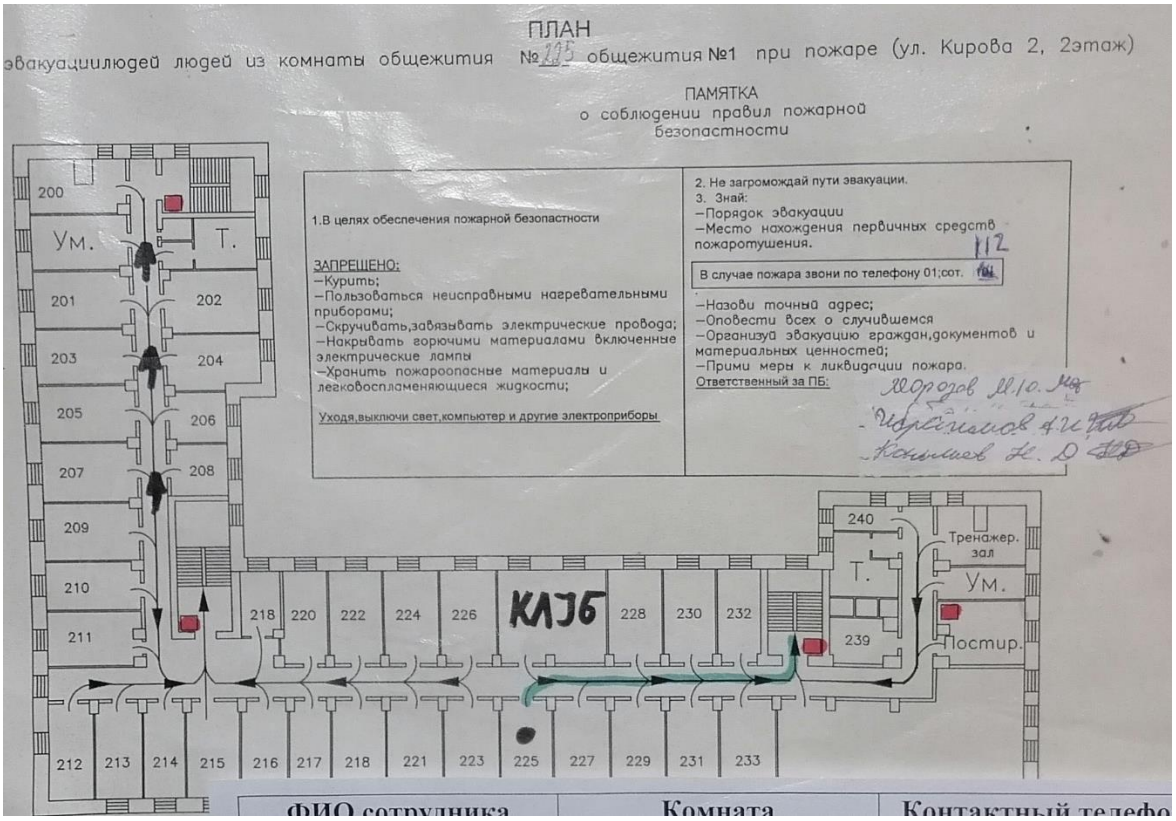


Рисунок 23 – План эвакуации при пожарах и других ЧС (красными прямоугольниками показаны огнетушители, стрелками показаны направления эвакуации)

5.4.2 Превентивные меры против возникновения пожара

Согласно ГОСТ 12.1.004-91 [91] предотвращение пожара должно достигаться и (или) предотвращением образования в горючей среде источников зажигания. На основе указанного стандарта администрация здания предъявляет требования к пожарной безопасности. Данные требования жёстко регламентируются правилами использования помещений в здании. Согласно этим правилам в помещении строго запрещается хранить пожаровзрывоопасные вещества, а к технике, сетевым фильтрам, электрическим обогревателям и источникам освещения предъявляются жёсткие требования, в случае невыполнения которых накладываются санкции (замечание, штраф или запрет на использование помещений). Администрация проводит регулярные проверки на соответствие помещений требованиям.

С целью предотвращения пожара необходимо со стороны персонала также следует обеспечить следующие превентивные меры:

- периодическая проверка проводки и исправности электроприборов;
- отключение электроприборов при уходе из помещения;
- инструктаж персонала по пожарной безопасности.

В особенности важно соответствующим образом обучить персонал соблюдать технику пожарной безопасности и правилу действий во время пожара.

При возгорании в помещении, где происходит разработка, порядок действий следующий:

1. закрыть окно в случае, если оно открыто (для уменьшения потока воздуха);
2. оценить возможность тушения своими силами (если нет возможности потушить, то п.4);
3. плотно закрыть дверь (не на ключ), отключить от сети электроприбор, возгорание которого произошло, накрыть плотной тканью всю площадь возгорания;
4. если ликвидировать очаг возгорания не удаётся, то необходимо покинуть помещение;
5. при наличии телефона позвонить по номеру 112 с сотового телефона или 01 со стационарного, сообщить:
    a. фамилию имя;
    b. что произошло (например, произошло возгорание чайника);
    c. адрес;
    d. этаж;
    e. номер кабинета;
    f. есть ли пострадавшие;
6. плотно закрыть за собой дверь, уплотнив щели;
7. сообщить администрации здания о возгорании;

8. покинуть опасную зону (эвакуироваться на 50 метров от здания).

5.5 Выводы по разделу

Входе написания раздела «социальная ответственность» были исследованы следующие вредные факторы как отклонение показателей микроклимата, недостаточная освещённость рабочей зоны, повышенный уровень шума, повышенный уровень электромагнитных полей, а также опасность поражения электрическим током.

Были предложены меры защиты от вышеперечисленных вредных и опасных факторов.

Персонал имеет категории I по электробезопасности [81].

Был произведён анализ проекта с точки зрения экологической безопасности. В основном, негативное воздействие разработки на окружающую среду происходит в следствии возникновения отходов, наиболее вредные из которых отработанные люминесцентные лампы и аккумуляторы.

Рассматриваемое в разделе помещение, в котором ведётся разработка, относится к IV категории объектов, оказывающих негативное воздействие на окружающую среду.

Наиболее вероятное ЧС – пожар. Помещение, в котором происходит разработка относится к классу помещений с пониженной пажароопасностью.

Помещение, в котором происходит разработка по взрывопожарной и пожарной опасности относится к категории (Д) с пониженной пожароопасностью Согласно СП 12.13130.2009.

Были разработаны рекомендации для предотвращения возникновения пожара и инструкция для персонала в случае возникновения пожара.

ЗАКЛЮЧЕНИЕ

В результате проведённой работы удалось разработать систему научного поиска. Исследование моделей, которые используются в системе, и ответов на запросы позволяют подтвердить адекватность моделей и применимость системы для использования в практических целях.

В качестве алгоритма кластеризации для разбиения пространства поиска на кластеры и отнесения векторного представления запроса к одному из кластеров были выбраны карты Кохонена. Лучшие результаты по косиносному расстоянию показала конфигурация карт с шестью нейронами, хотя согласно индексу качества Дэвиса-Болдуина, который составил 3,05, данная конфигурация наихудшая из опробованных.

Выяснилось, что для небольшой коллекции документов целесообразно опустить этап с кластеризацией, так как некоторые релевантные документы могут не быть найдены, если поиск осуществляется только по одному кластеру.

Для решения задачи суммаризации отдельное научной статьи трансформер Longformer encoder-decoder был дообучен собственном датасете, что позволило улучшить показатель ROUGE-2 F1-метрики с 13,07 до 13,62, также ROUGE-2 precision с 15,22 до 16,28 и ROUGE-2 recall с 12,72 до 12,80. На данном этапе развития проекта для построения общего реферата по теме была выбрана та же модель.

В ходе работы также удалось собрать датасет из научных статей по различным исследования, связанным с мембранами.

Нужно отметить, что разработка имеет перспективу не только в рамках автоматизации научного поиска, но также может быть применена на другие сферы как анализ новостей и технических или других документов, при должной конфигурации.

В дальнейшем планируется вовлечение сети цитирований в генерацию рефератов и добавление различной нетекстовой информации.

СПИСОК ИСТОЧНИКОВ

1. Automatic Text Summarization with Machine Learning — An overview [Электронный ресурс]: https://medium.com [сайт]. Режим доступа: https://medium.com/luisfredgs/automatic-text-summarization-with-machine-learning-an-overview-68ded5717a25., свободный (дата обращения 28.05.2022)
2. Yue Guo, Weijian Qiu, Yizhong Wang, and Trevor A. Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. ArXiv abs/2012.12573 (2021).
3. C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004.
4. A. Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In HLTNAACL, pages 145–152, 2004.
5. Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.
6. Gunning, R.; et al. 1952. Technique of clear writing. McGraw-Hill.
7. Coleman, M.; and Liau, T. L. 1975. A computer readability formula designed for machine scoring. Journal of Applied Psychology 60(2): 283.
8. Altmami, N. I.; and Menai, M. E. B. 2020. Automatic summarization of scientific articles: A survey. Journal of King Saud University - Computer and Information Sciences
9. Lloret, Elena, Romá-Ferri, María Teresa, Palomar, Manuel, 2011. COMPENDIUM: a text summarization system for generating abstracts of research papers. In: International Conference on Application of Natural Language to Information Systems. Springer, pp. 3–14.
10. Ferrández, Oscar, Micol, Daniel, Munoz, Rafael, Palomar, Manuel, 2007. ‘A Perspective-Based Approach for Solving Textual Entailment Recognition’. In:

Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing, Association for Computational Linguistics, 66–71.

11. Blake, Barry J., 1992. T. Givón, Syntax: A Functional-Typological Introduction, Volume II. Amsterdam: John Benjamins, 1990. pp. Xxv+ 552.' J. Linguist. 28(2), 495–500.
12. Luhn, Hans Peter, 1958. The automatic creation of literature abstracts. IBM J. Res. Dev. 2 (2), 159–165.
13. Yang, Shansong et al., 2016. Amplifying scientific paper's abstract by leveraging data-weighted reconstruction. Inf. Process. Manage. 52 (4), 698–719.
14. Slamet, Cepi et al., 2018. Automated Text Summarization for Indonesian Article Using Vector Space Model. In: IOP Conference Series: Materials Science and Engineering. IOP Publishing, p. 012037.
15. Lovins, Julie Beth. Development of a Stemming Algorithm // Mechanical Translation and Computational Linguistics. — 1968. — T. 11.
16. Hetami, A., 2015. Perancangan Information Retrieval (IR) Untuk Pencarian Ide Pokok Teks Artikel Berbahasa Inggris dengn Pembobotan Vector Space Model. Jurnal Ilmiah Teknologi Informasi Asia 9 (1), 53–59.
17. A. Nenkova and R. J. Passonneau. Evaluating content selection in summarization: The pyramid method. In HLTNAACL, pages 145–152, 2004.
18. Elkiss, Aaron et al., 2008. Blind men and elephants: what do citation summaries tell us about a research article? J. Am. Soc. Inf. Sci. Technol. 59 (1), 51–62.
19. Qazvinian, Vahed, Radev, Dragomir R., 2008. Scientific Paper Summarization Using Citation Summary Networks. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pp. 689–696.
20. Abu-Jbara, Amjad, Radev, Dragomir, 2011. In: Coherent citation-based summarization of scientific papers. Association for Computational Linguistics, pp. 500–509.

21. Cohan, Arman, Goharian, Nazli, 2015. Scientific article summarization using citation-context and article's discourse structure. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal: Association for Computational Linguistics, pp. 390–400.
22. Cohan, Arman, Goharian, Nazli, 2018. Scientific document summarization via citation contextualization and scientific discourse. Int. J. Digital Lib. 19 (2–3), 287–303.
23. Mei, Qiaozhu, Zhai, ChengXiang, 2008. Generating Impact-Based Summaries for Scientific Literature. Proceedings of ACL-08: HLT: 816–824.
24. Qazvinian, Vahed, Radev, Dragomir R., Ozgur, Arzucan, 2010. Citation Summarization through Keyphrase Extraction. In: Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), 895–903.
25. Galgani, Filippo, Compton, Paul, Hoffmann, Achim, 2015. Summarization based on bi-directional citation analysis. Inf. Process. Manage. 51 (1), 1–24.
26. De Waard, Anita, Maat, Henk Pander, 2012. 'Epistemic Modality and Knowledge Attribution in Scientific Discourse: A Taxonomy of Types and Overview of Features'. In: Proceedings of the Workshop on Detecting Structure in Scholarly Discourse, Association for Computational Linguistics, 47–55.
27. Ronzano, Francesco, Saggion, Horacio, 2016. An Empirical Assessment of Citation Information in Scientific Summarization. In: International Conference on Applications of Natural Language to Information Systems, Springer, 318–325.
28. Wang, Jie, Ma, Shutian, Zhang, Chengzhi, 2017. Citationas: a summary generation tool based on clustering of retrieved citation content. Framework 7 (8).
29. Zamir, Oren, Etzioni, Oren, 1999. Grouper: a dynamic clustering interface to web search results. Comput. Networks 31 (11–16), 1361–1374

30. Osin´ ski, Stanis'law, Stefanowski, Jerzy, Weiss, Dawid, 2004. Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition. In: Intelligent Information Processing and Web Mining, Springer, 359–368.
31. Steinbach, Michael, Karypis, George, Kumar, Vipin, 2000. A Comparison of Document Clustering Techniques. In: KDD Workshop on Text Mining, Boston, 525–526.
32. Yang, Yiming, Pedersen, Jan O., 1997. A comparative study on feature selection in text categorization. In: Icml, 35.
33. Luhn, Hans Peter, 1958. The automatic creation of literature abstracts. IBM J. Res. Dev. 2 (2), 159–165.
34. Mikolov, Tomas, Le, Quoc V., Sutskever, Ilya, 2013. Exploiting Similarities among Languages for Machine Translation. arXiv preprint arXiv:1309.4168.
35. Miller, George A., 1998. WordNet: An Electronic Lexical Database. MIT Press.
36. Al Saied, Hazem, Dugué, Nicolas, Lamirel, Jean-Charles, 2018. Automatic summarization of scientific publications using a feature selection approach. Int. J. Digital Lib., 1–13
37. Agrawal, Kritika, Mittal, Aakash, Pudi, Vikram, 2019. Scalable, semi-supervised extraction of structured information from scientific literature. In: Proceedings of the Workshop on Extracting Structured Knowledge from Scientific Publications, pp. 11–20.
38. Lamirel, Jean-Charles, Cuxac, Pascal, Chivukula, Aneesh Sreevallabh, Hajlaoui, Kafil, 2013. A new feature selection and feature contrasting approach based on quality metric: application to efficient classification of complex textual data. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining. Springer, pp. 367–378.
39. Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, Toutanova, Kristina, 2018. 'Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding'. arXiv preprint arXiv:1810.04805.

40. Gupta, Sonal, Manning, Christopher, 2014. ‘Improved Pattern Learning for Bootstrapped Entity Extraction’. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, 98–108.
41. Ma, C., Zhang, W. E., Guo, M., Wang, H., and Sheng, Q. Z. (2020). Multi-document summarization via deep learning techniques: A survey. arXiv preprint arXiv:2011.04843
42. Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. arXiv:2004.05150.
43. Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, C. Alberti, S. Ontan˜on, ´ Philip Pham, Anirudh Ravula, Qifan Wang, L. Yang, and A. Ahmed. 2020. Big bird: Transformers for longer sequences. ArXiv, abs/2007.14062.
44. Athar Sefid and C. Lee Giles. 2022. SciBERTSUM: Extractive Summarization for Scientific Documents. CoRR abs/2201.08495 (2022).
45. Qizhe Xie, Zihang Dai, Eduard H. Hovy, Minh-Thang Luong, and Quoc V. Le. 2019. Unsupervised data augmentation for consistency training. arXiv preprint, abs/1904.12848.
46. Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval2019 task 4: Hyperpartisan news detection. In Proceedings of the 13th International Workshop on Semantic Evaluation, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.
47. Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. In ACL.
48. Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 268–284, Online. Association for Computational Linguistics

49. Hugging Face Reads, Feb. 2021 - Long-range Transformers [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/blog/long-range-transformers., свободный (дата обращения 28.05.2022)
50. Sequence to Sequence (seq2seq) and Attention [Электронный ресурс]: https://lena-voita.github.io [сайт]. Режим доступа: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html., свободный (дата обращения 28.05.2022)
51. TAC 2014 Biomedical Summarization Track [Электронный ресурс]: https://tac.nist.gov [сайт]. Режим доступа: https://tac.nist.gov/2014/BiomedSumm/., свободный (дата обращения 28.05.2022)
52. WING-NUS/scisumm-corpus [Электронный ресурс]: https://github.com [сайт]. Режим доступа: https://github.com/WING-NUS/scisumm-corpus., свободный (дата обращения 28.05.2022)
53. ACL Anthology Network (All About NLP) [Электронный ресурс]: https://clair.eecs.umich.edu [сайт]. Режим доступа: https://clair.eecs.umich.edu/aan/index.php., свободный (дата обращения 28.05.2022)
54. Microsoft Academic Search (All About NLP) [Электронный ресурс]: http://academic.research.microsoft.com/ [сайт]. Режим доступа: http://academic.research.microsoft.com/., свободный (дата обращения 28.05.2022)
55. TIPSTER Text Summarization Evaluation Conference (SUMMAC) [Электронный ресурс]: http://www-nlpir.nist.gov [сайт]. Режим доступа: http://www-nlpir.nist.gov/related_projects/tipster_summac/cmp_lg.html., свободный (дата обращения 28.05.2022)

56. PLOS [Электронный ресурс]: http://www.ncbi.nlm.nih.gov/pubmed [сайт]. Режим доступа: http://www.ncbi.nlm.nih.gov/pubmed., свободный (дата обращения 28.05.2022)
57. ScisummNet [Электронный ресурс]: https://cs.stanford.edu [сайт]. Режим доступа: https://cs.stanford.edu/~myasu/projects/scisumm_net/., свободный (дата обращения 28.05.2022)
58. scientific_papers [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/datasets/scientific_papers., свободный (дата обращения 28.05.2022)
59. GROBID Documentation [Электронный ресурс]: https://grobid.readthedocs.io [сайт]. Режим доступа: https://grobid.readthedocs.io/en/latest/., свободный (дата обращения 21.12.2021)
60. K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask ´ R-CNN. arXiv:1703.06870, 2017.
61. ibm-aur-nlp/PubLayNet [Электронный ресурс]: https://github.com [сайт]. Режим доступа: https://github.com/ibm-aur-nlp/PubLayNet., свободный (дата обращения 21.12.2021)
62. What is Layout Parser? [Электронный ресурс]: https://layout-parser.github.io/ [сайт]. Режим доступа: свободный (дата обращения 21.12.2021)
63. Hough Line Transform [Электронный ресурс]: https://docs.opencv.org [сайт]. Режим доступа: https://docs.opencv.org/3.4/d9/db0/tutorial_hough_lines.html., свободный (дата обращения 21.12.2021)
64. Davies-Bouldin Index [Электронный ресурс]: https://docs.opencv.org [сайт]. Режим доступа: https://scikit-learn.org/stable/modules/clustering.html#davies-bouldin-index., свободный (дата обращения 21.12.2021)
65. Journal of Membrane Science — An overview [Электронный ресурс]: https://www.sciencedirect.com [сайт]. Режим доступа:

https://www.sciencedirect.com/journal/journal-of-membrane-science., свободный (дата обращения 28.05.2022)

66. textstat 0.7.3 — An overview [Электронный ресурс]: https://pypi.org [сайт]. Режим доступа: https://pypi.org/project/sklearn-som/., свободный (дата обращения 28.05.2022)
67. Decomposing signals in components (matrix factorization problems) — An overview [Электронный ресурс]: https://scikit-learn.org [сайт]. Режим доступа: https://scikit-learn.org/stable/modules/decomposition.html., свободный (дата обращения 28.05.2022)
68. allenai/led-large-16384-arxiv — An overview [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/allenai/led-large-16384-arxiv., свободный (дата обращения 28.05.2022)
69. Dataset: scientific_papers / pubmed — An overview [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/datasets/viewer/?dataset=scientific_papers., свободный (дата обращения 28.05.2022)
70. Hugging Face Reads, Feb. 2021 - Long-range Transformers [Электронный ресурс]: https://huggingface.co [сайт]. Режим доступа: https://huggingface.co/blog/long-range-transformers., свободный (дата обращения 28.05.2022)
71. Трудовой кодекс Российской Федерации от 30.12.2001 N 197-ФЗ (ред. от 27.12.2018)
72. ГОСТ 22269-76 «Рабочее место оператора. Взаимное расположение элементов рабочего места»
73. ГОСТ 12.2.032-78 ССБТ «Рабочее место при выполнении работ сидя. Общие эргономические требования»

74. ГОСТ Р 50923-96 «Дисплеи. Рабочее место оператора. Общие эргономические требования и требования к производственной среде. Методы измерения»
75. ГОСТ 21889-76 «Система "Человек-машина". Кресло человека-оператора. Общие эргономические требования»
76. СанПиН 1.2.3685-21 «Гигиенические нормативы и требования к обеспечению безопасности и (или) безвредности для человека факторов среды обитания»
77. СП 52.13330.2016 «Естественное и искусственное освещение». Актуализированная редакция СНиП 23-05-95*
78. ГОСТ 12.1.003-2014 ССБТ. «Шум. Общие требования безопасности»
79. ГОСТ 12.1.029-80 ССБТ. «Средства и методы защиты от шума. Классификация»
80. ГОСТ 12.1.006-84 ССБТ. «Электромагнитные поля радиочастот. Допустимые уровни на рабочих местах и требования к проведению контроля»
81. ГОСТ 12.1.038-82 ССБТ. «Электробезопасность. Предельно допустимые уровни напряжений прикосновения и токов»
82. ГОСТ 12.1.030-81 Система стандартов безопасности труда (ССБТ). «Электробезопасность. Защитное заземление. Зануление»
83. Белов, Сергей Викторович. Безопасность жизнедеятельности и защита окружающей среды (техносферная безопасность): учебник для академического бакалавриата / С. В. Белов. - 5-е изд., перераб. и доп.. - Москва: Юрайт ИД Юрайт, 2015. - 703 с.
84. ПРИКАЗ от 15 декабря 2020 года N 903н Об утверждении Правил по охране труда при эксплуатации электроустановок.
85. ГОСТ 12.1.005-88 ССБТ. «Общие санитарно-гигиенические требования к воздуху рабочей зоны»

86. ГОСТ 12.1.007-76 ССБТ. «Вредные вещества. Классификация и общие требования безопасности»
87. Порядок утилизации ламп: требования и правила утилизации ртутьсодержащих (ртутных), люминесцентных и других видов ламп [Электронный ресурс]: Режим доступа http://www.ecobasis.ru [сайт]. Режим доступа http://www.ecobasis.ru/2016/06/15/poryadok-utilizacii-lamp-trebovaniya-pravila/., свободный (дата обращения 07.05.2022)
88. Постановление Правительства РФ от 12 декабря 2012 г. N 1287 «О лицензировании деятельности по заготовке, хранению, переработке и реализации лома черных и цветных металлов»
89. Постановление правительства РФ от 31 декабря 2020 года N 2398 «Об утверждении критериев отнесения объектов, оказывающих негативное воздействие на окружающую среду, к объектам I, II, III и IV категорий» (с изменениями на 7 октября 2021 года)
90. СП 12.13130.2009 ОПРЕДЕЛЕНИЕ КАТЕГОРИЙ ПОМЕЩЕНИЙ, ЗДАНИЙ И НАРУЖНЫХ УСТАНОВОК ПО ВЗРЫВОПОЖАРНОЙ И ПОЖАРНОЙ ОПАСНОСТИ
91. ФЗ от 22.07.2008 N 123-ФЗ (ред. от 30.04.2021) «Технический регламент о требованиях пожарной безопасности»

Приложение А

(справочное)

THE LITERATURE OVERVIEW DEVOTED TO THE PROBLEM OF SCIENTIFIC ARTICLES SUMMARIZATION

Студент

Группа	ФИО	Подпись	Дата
8ВМ03	Хайров Марк Альбертович		

Руководитель ВКР

Должность	ФИО	Учёная степень	Подпись	Дата
Доцент ОИТ	Иванова Юлия Александровна	к.т.н.		

Консультант-лингвист ОИЯ ШБИП

Должность	ФИО	Учёная степень	Подпись	Дата
Старший преподаватель	Ануфриева Татьяна Николаевна			

1 Automatic Text Summarization

Summarization is the task of condensing a piece of text to a shorter version, reducing the size of the initial text while at the same time preserving key informational elements and the meaning of content [1].

1.1The Types of Summarization

From the point of view of the automatic scientific paper summarization problem, it is important to single out two features by which summation types are classified, these two features are: the method of summation and the number of documents used for summation.

According to the method summarization is divided into:

- extractive;
- abstractive;
- hybrid.

Extractive summarization picks up sentences directly from the document based on a scoring function to form a coherent summary. This method works by identifying important sections of the text cropping out and portions together portions of the content to produce a condensed version [1].

The advantage of the extractive approach is the ability to extract important information with correct facts.

The disadvantage is that the resulting abstract can often be inconsistent, the sentences can be incoherent, resulting in a significant deterioration in the readability of the generated text.

Abstractive summarization methods aim at producing summary by interpreting the text using advanced natural language techniques in order to generate a new shorter text — parts of which may not appear as part of the original document, that conveys the most critical information from the original text, requiring rephrasing sentences and incorporating information from full text to generate summaries such as a human-written abstract usually does [1].

The advantage of the abstractive approach lies in the possibility of obtaining a consistent, coherent and readable text.

Among the shortcomings, it can be noted that as a result of abstract summarization, some of the facts may be lost, and some may be distorted.

Hybrid method combines both extractive and abstractive methods. As a rule, two models are used in such methods: extractor and abstractor [1].

This approach allows you to take advantage of the two summarization methods. Among the disadvantages is the increasing complexity of the solution.

By the number of documents, the summarization is:

• single-document;

• multi-document.

Single-document summarization is carried out according to one source, and in multi-document summation there are several of them at once.

1.2 Summarization Evaluation Metrics

There are three groups of methods for evaluating the quality of automatic summation: manual, automatic and semi-automatic methods.

1.2.1 Manual Methods

A manual evaluation of an automatically generated summary requires a group of human experts to rate the quality of a summary according to some specified criteria. In particular, an automatically generated text can be compared with the text written by an expert for the same topic. It is possible that an automatically generated summary may be better in some way than the text of an expert [2], and manual evaluation allows to identify such cases.

1.2.2 ROUGE (automatic methods)

The most popular metrics for automatic summarization evaluation are ROUGE metrics [3] and one of the most popular varieties is ROUGE-N.

ROUGE-N is the proportion of n-grams from the reference example resume that ended up in the auto-generated resume:

$$ROUGE_N = \frac{\sum_{s\in\{RefSummaries\}} \sum_{gram_n\in S} count_{match}(gram_n)}{\sum_{s\in\{RefSummaries\}} \sum_{gram_n\in S} count(gram_n)}, \quad (1)$$

where n stands for the length of the n-gram, gramn, and $count_{match}$ is the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries

There are also metrics such as: ROUGE-L - longest common sequence, ROUGE-W - weighted longest common sequence and ROUGE-S - skipgram match statistics.

1.2.3 PYRAMID (semi-automatic methods)

The PYRAMID summarization evaluation method [4] is based on the hypothesis that a reference standard for text abstracts exists and it is possible to obtain it.

Although people tend to summarize text in different ways, extracting different information from the source text, it is assumed that it is possible to assemble a reference abstract from many abstracts compiled by people using information about the frequency of occurrence of this or that information. By frequency, one can judge the significance of information and evaluate the quality of summation.

1.2.4 Readability Automatic Evaluation

The above-mentioned automatic and semi-automatic methods make it possible to understand, among other things, the amount important information that was extracted, but there are methods that also allow you to evaluate the readability of the text. There are the following methods for automatic evaluation of text readability:

- Flesch-Kincaid grade level [5]:

$$FKGL = 0{,}39 \cdot \left(\frac{total\ words}{total\ sentences}\right) + 11{,}8 \cdot \left(\frac{total\ sylables}{total\ words}\right) - 15{,}59, \quad (2)$$

- Gunning fog index [6]:

$$GFI = 0{,}4 \cdot \left[\left(\frac{words}{sentences}\right) + 100 \cdot \left(\frac{complex\ words}{words}\right)\right], \quad (3)$$

- Coleman-Liau index [7]:

$$CLI = 0.0588L - 0.296S - 15.8, \quad (4)$$

where L is the average number of letters per 100 words and S is the average number of sentences per 100 words.

2 Scientific Articles Automatic Summarization

The automatic summarization of many scientific articles is somewhat different from the general summation of a set of texts.

Scientific articles have the following features [8]:

- A clear overall structure;
- A significant amount of text;
- the purpose of the summary is not unique;
- the presence of figures, tables, formulas, diagrams and algorithms;
- domain-specific vocabulary.

Firstly, the general structure of any scientific article:

- annotation;
- motivation;
- background;
- methodology;
- experiment;
- results and discussion;
- references.

Secondly, scientific articles tend to have a large amount of text for automatic summarization, even if in case of a single-document summarization.

Thirdly, the information is required to get from a scientific study may vary, it may be methods, results, limitations, or any other aspects of the work.

Fourthly, some of the information can be presented in the form of tables, figures, formulas or pseudocode, which is already a separate task [8].

Fifthly, the summarization of a scientific article requires the ability to highlight the common, for which the summarizing model must have some comprehension of the

texts and, presumably, some domain knowledge. For example, in the case of a study of the same topic, different articles may provide different arguments and even different conclusions. The task of summarizing news involves covering the same event from different points of view.

At the moment, there are two main areas related to the summarization of scientific articles: automatic abstract generation and citation-based summarization. The further review includes the achievements of both approaches.

2.1 Automatic Abstract Generation

Lloret et al. (2011) [9] have suggested two approaches for generating research article abstracts. The first is a purely extractive summarizer (COMPENDIUM$_E$), and the second is based on the extractive and abstractive technique (COMPENDIUM$_{E\text{-}A}$). The extractive summarizer, COMPENDIUM$_E$, relies on four main stages: 1) preprocessing (i.e., tokenization, sentence segmentation, stop words elimination, and part-of-speech tagging); 2) redundancy removal using a textual entailment (TE) tool [10]; 3) sentence relevance identification, which assigns to each sentence a score that reflects its importance based on two features – code quantity principle (CQP) [11] and term frequency (TF) [12] – and then ranks them according to their scores; and 4) summary generation, which selects the highest-ranked sentences to generate a final summary in the same order as the sentences appear within the original document. Thus, the generated summary is an extractive one. By contrast, COMPENDIUM$_{E\text{-}A}$ is based on the extractive and abstractive technique. This method takes COMPENDIUM$_E$ as a base and integrates an information compression and fusion stage between its third and fourth steps to generate an abstractive summary. New sentences are created by either combining information from two sentences or by shortening a long sentence into smaller ones.

Yang et al. (2016) [13] proposed a system for an expanded abstract that describes the most important aspects of a scientific article using a data-weighted reconstruction approach. This consists of two phases: weight learning and salient sentence selection. During the first phase, semantic information from the citation sentences and social

structure are considered. The authors used the target article abstract, which contains the main aspects, and the set of sentences that cite the target article to provide complementary aspects as an input to their system. First, they built a heterogeneous bibliographic network. They then identified social relations such as paper-coauthor-paper and paper-cite-paper, as well as similar semantic relations between sentences. Furthermore, they proposed a data-weighted objective function based on the learned sentence's weight and weighted reconstruction error. Thus, from the viewpoint of data reconstruction, they can detect salient sentences.

Slamet et al. (2018) [14] proposed a simple system that automatically generates an article abstract for the Indonesian language. Four main steps are used in their system. First, a preprocessing step (consisting of sentence extraction, case folding, tokenization, filtering, and stemming [15]) is used to prepare the input text for the next step. Next, computing Term Frequency-Inverse Document Frequency (TF-IDF) [16] for each term in the preprocessed text. Using cosine similarity [17] and vector space modeling (VSM)[16], the similarity between the text and the 20 keywords of the TF-IDF output are computed, and the sentences are ranked based on their similarity scores. Finally, the final abstract is compiled from the top ten sentences.

2.2 Citation-based summarization

The citing text of scientific articles often contains the most important information about the cited work. With the help of citations, one can evaluate the contribution of a particular work and get an idea of its advantages and disadvantages from the points of view of different authors.

In early work on automatic citation-based summarization [18-20], articles were summarized by means of extracting a set of sentences from a set of citing text. More recent work has pointed out [21-24] some problems with the use of citation suggestions. When quoting sentences, the discussion of the target article usually overlaps with discussion of other cited articles or with the content of irrelevant information in the citing article. As a solution to this problem, the use of various parts of the text from the cited article (text spans) was proposed.

Next, an overview of methods for automatically summarizing scientific articles based on citations only and span-based summarization algorithms will be presented.

2.2.1 Automatic Summarization Based Only on Citations

Qazvinian and Radev (2008) proposed a citation-based summarization system called C-LexRank [19]. In this model, citing sentences are represented as graph vertices connected by edges with weights, which represent the measures of similarity between sentences. Sentences are divided into clusters according to their similarity. The summary is made by selecting a sentence from each cluster.

In 2011, Abu-Jbara and Radev [20] noted some of the problems with the model mentioned above. They stated that citing sentences may contain irrelevant information, resulting in decreased readability, consistency, increased abstract size, and loss of important information during sentence ranking. The authors have improved this algorithm by labeling links, introducing a classifier for the links (delete, save or replace with a pronoun) and introducing a system for grouping citing sentences in relation to the section (introduction, problem statement, method, results and restrictions).

2.2.2 Automatic Cited Text Span-based Summarization

Galgani et al. (2015) have suggested that such strategies face limitations when only the set of citing sentences is used to generate a summary [25]. They introduced a new trend of automatic summarization methods that combine incoming and outgoing citations along with different elements of both the citing and cited articles, in addition to the full target text. They proposed two methods to generate a summary, both of which use the same input: a target article to be summarized and a collection of its citances and citphrases extracted from connected documents. They generated the final summary by either (1) ranking the extracted text from all citances and citphrases to find common concepts over several citations and generate the final summary or (2) measuring the similarity between each sentence in the target article and the citations (either citphrases or citances) and then ranking the sentences in decreasing order. Thus,

the summary sentences are those that are highly similar to the citation sentences. The underlying idea is that citations represent the main issues of the target article and can therefore be used to select the segments that represent these issues.

The citing sentences often lack context of the research [26]. To solve this problem, Cohan and Goharian (2015) [21] proposed a solution that provides context for citations from the cited text. Authors used a vector of their n-grams as a representation of citing sentences to find the most relevant text spans, which are then grouped by topic; sentences are further ranked by information content. The abstract is compiled either by iteratively searching for the highest rated proposals or by using a greedy strategy.

Ronzano and Saggion (2016) [27] investigated the influence of citation context and which part of the article it belongs to on the quality of summarization. It was found that the use of citation context improves the quality of summarization. The best results (mean ROUGE-2) were obtained when the abstract was used as a reference summary and sentences for the resulting summary were taken from the article body and citation context.

In 2018, a framework was presented by Cochan and Goharian that solves the problem of inaccurate citation text [22]. The solution was also based on embedding and a classifier that allows finding the appropriate context for each citing sentence.

In 2017, CitationAS, an automatic tool for summary generation, was built by Wang and Zhang [28]. It uses a set of rules to identify citation sentences and consists of three core stages. The first is clustering, in which three algorithms are used: suffix tree clustering (STC) [29], Lingo [30], and bisecting K-means [31]. During this stage, citation sentences are first represented using VSM [32], and TF-IDF [33] is used to calculate feature weights; similar sentences are grouped into one cluster. Next, Word2Vec [34], WordNet [35], and a combination of the two are used to generate cluster labels, following which the clusters with similar labels are merged. Finally, the clusters are sorted by size, and sentences are extracted to form the final summary. One

advantage of CitationAS is that the summary generated is comprehensive and representative of the topic, although it contains some redundant content.

In 2018, a solution was proposed for the summation of scientific articles [44] [36] using the feature maximization method [37]. The proposed summation systems are statistical, have no parameters, are language independent, and do not need additional corpora. The general structure of the proposed system consists of five main stages. First, the input text is pre-processed (i.e., stop-word removal and stemming) and the word weight is calculated for a set of keywords in the article's title, subtitles, and abstract. The sentence weights are then calculated using the average of its word weights. The size of the final abstract is determined in the third stage based on the distribution of weights. At the end, text processing can be applied to remove redundant content.

One of the recent works [38] has proposed a system for summarizing article. It is based on the hypothesis that the contribution of any article is best described by its goal, the method used and the results section. Based on this hypothesis, the authors used the k-nearest neighbor classifier [39] and bootstrap [40] to extract the three concepts identified above from the title of the target article, the abstract, and the citation context (i.e., the cited text). Based on the information received, a knowledge graph is built for a graphical representation of the connection between the extracted concepts and citations.

2.2.3 Scientific Article Automatic Summarization with Transformers

Transformers have been successfully used for the summarization of relatively short texts in comparison with scientific articles [41]. There have been improvements in summarization of long sequences of texts recently [42, 43]. Attempts to adapt both sort- and long-sequenc for the tasks related to the summarization of scientific articles are being made.

SciBERTSUM [44] is a model based on BERT [45] adapted to solve the problem of scientific articles summarization. In this model, both the local memory mechanism

and the global one are used. During model training, the authors used presentation slides as a summarization reference.

There are also cases of training transformers on corpora from scientific articles for summarization [2].

3 The Main Issues of Automatic Summarization of Scientific Articles

The main issue of automatic summation of scientific articles is that there is a small amount of data for training summarization algorithms for the specifics of some subject areas and it is often problematic to obtain that data.

Also at the moment there is a problem of metrics for assessing the quality of summarization and general baseline systems for comparing algorithms to each other [8].

Most of the considered algorithms are aimed at building a summary based on citing sentences for one article. To obtain qualitatively new results in the field of multi-document summarization with good readability and consistency of the generated texts, more research will be required [8].