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

Направление	09.04.04 – Программная инженерия
Профиль	Разработка распределенных программных систем
Факультет	КТИ
Кафедра	МО ЭВМ

К защите допустить

Зав. кафедрой Кринкин К.В.

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА МАГИСТРА

ТЕМА: РАЗРАБОТКА СИСТЕМЫ АВТОМАТИЗИРОВАННОЙ ПРОВЕРКИ НАИБОЛЕЕ ЧАСТЫХ ОШИБОК В НАУЧНЫХ ТЕКСТАХ

Студент		_____ *подпись*	Блеес Э.И.
Руководитель	*(Уч. степень, уч. звание)*	_____ *подпись*	Заславский М.М.
Консультанты	*(Уч. степень, уч. звание)*	_____ *подпись*	Иванов А.Н.
	(Уч. степень, уч. звание)	_____ *подпись*	Чередниченко А.И.

Санкт-Петербург

2019

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Утверждаю
Зав. кафедрой МО ЭВМ
_____ Кринкин К.В.
«___»_____20___ г.

Студент Блеес Э.И. Группа 3304

Тема работы: Разработка системы автоматизированной проверки наиболее частых ошибок в научных текстах

Место выполнения ВКР: СПбГЭТУ «ЛЭТИ», кафедра МО ЭВМ

Исходные данные (технические требования):

TO DO

Содержание ВКР:

TO DO

Перечень отчетных материалов: пояснительная записка, иллюстративный материал

Дополнительные разделы: Безопасность жизнедеятельности

Дата выдачи задания
«___»_____20___ г.

Дата представления ВКР к защите
«___»_____20___ г.

Студент _____ Блеес Э.И.

Руководитель _____ Заславский М.М.
(Уч. степень, уч. звание)

КАЛЕНДАРНЫЙ ПЛАН ВЫПОЛНЕНИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Утверждаю
Зав. кафедрой МО ЭВМ
_____ Кринкин К.В.
«___»_____20___ г.

Студент Блеес Э.И. Группа 3304

Тема работы: Разработка модуля автоматизации импорта и геоконтекстной разметки открытых данных

№ п/п	Наименование работ	Срок выполнения
1	Обзор литературы по теме работы	24.04 – 30.04
2	Наименование раздела	01.05 – 04.05
3	Наименование раздела	05.05 – 19.05
4	Наименование раздела	20.05 – 24.05
5	Предзащита	30.05
6	Оформление пояснительной записки	25.05 – 01.06
7	Оформление иллюстративного материала	27.05 – 15.06

Студент _____ Блеес Э.И.

Руководитель _____ Заславский М.М.
(Уч. степень, уч. звание)

РЕФЕРАТ

ABSTRACT

Содержание

ВВЕДЕНИЕ.....5
1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ 8
2. ПОСТАНОВКА ЗАДАЧИ И ВЫБОР МЕТОДА РЕШЕНИЯ 16
3. ОПИСАНИЕ МОДЕЛИ ПРОВЕРКИ СТАТЬИ.....17
4. ОПИСАНИЕ РЕШЕНИЯ.....28
ЗАКЛЮЧЕНИЕ 40

ОПРЕДЕЛЕНИЯ, ОБОЗНАЧЕНИЯ И СОКРАЩЕНИЯ

В настоящей пояснительной записке применяют следующие термины с соответствующими определениями:

ВВЕДЕНИЕ

Актуальность.

Соответствие статьи научному стилю является одним из основных критериев принятия статьи к публикации. В текущем виде, процесс проверки представляет собой отправку статьи на рецензирование, ожидание ответа,

исправление недочетов и отправка на повторную проверку – данные этапы могут занимать достаточно много времени. В связи с этим, автоматизация данного процесса является актуальной задачей, позволяющей значительно ускорить процесс выявления ошибок для исправления, и в следствие этого ускорить сам процесс публикации статьи, а также ускорить обучение начинающих авторов.

ПОДРОБНЕЕ

Цель работы.

Разработать программу для проверки статьи на соответствие научному стилю и поиску наиболее частых ошибок в ней.

Постановка задачи.

Для достижения поставленной цели необходимо решить следующие задачи:

- Исследовать возможности автоматизации проверки научных статей на соответствие научному стилю;
- Провести экспериментальное исследование на статьях для определения допустимых значений критериев;
- Разработать решение.

Объект исследования.

Научные статьи.

Предмет исследования.

Автоматизация проверки научных статей на соответствие научному стилю

Практическая значимость.

Решение позволяет ускорить процесс рецензирования статьи за счет своевременных исправлений наиболее частых ошибок до отправки статьи рецензенту. Решение будет применяться для проверки статей студентов СПбГЭТУ кафедры МОЭВМ в рамках курса обучения написанию научных статей студентов.

Опубликованные работы по теме.

1. Блеес Э.И., Заславский М.М., Андросов В.Ю. Автоматизация процесса проверки текста на соответствие научному стилю // Современные технологии в теории и практике программирования: материалы научно-практической конференции студентов, аспирантов и молодых ученых -2018. - С. 118-121;
2. Блеес Э.И., Заславский М.М. Исследование критериев соответствия текста научному стилю // Научно-технический вестник информационных технологий, механики и оптики. 2019. Т. 19. № 2. С. 299–305. doi: 10.17586/2226-1494-2019-19-2-299-305

1. ОБЗОР ПРЕДМЕТНОЙ ОБЛАСТИ

1.1. Основные понятия

Научный стиль — наиболее строгий стиль речи, используемый для написания научных статей. Стиль научных работ определяется содержанием и целями научного сообщения: точно и полно объяснить факты, показать причинно-следственные связи между явлениями, выявить закономерности, доказать утверждения.

В научных журналах существуют требования к структуре статьи, но отсутствуют структурированные требования к её стилю. В связи с этим, характеристика научного стиля получена из пособий, посвященных определению стилей русского языка и речи.

1.1.1. Стилистические особенности научного стиля

Научный стиль характеризуется логической последовательностью изложения, упорядоченной системой связи между частями высказывания, стремлением авторов к точности, сжатости, однозначности при сохранении насыщенности содержания [Трофимова Г. К. Русский язык и культура речи. – 2012.]. Выделяются следующие стилистические особенности научного стиля:

- Логичность — наличие смысловых связей между последовательными единицами (блоками) текста. Логичность, тесно связанная с последовательностью, доказательностью и аргументированностью изложения, выражается на синтаксическом уровне и на уровне текста. Для создания логичности используют полнооформленность высказывания — полнота грамматического оформления предикативных единиц, что выражается в преобладании союзных предложений над бессоюзными, так как союзы четче передают смысловые и логические связи частей предложения. Также для выражения логичности в научной речи используются рассуждение и доказательство [ссылка];

- Последовательность — характеристика текста, в котором выводы вытекают из содержания и непротиворечивы, текст разбит на отдельные смысловые отрезки, отражающие движение мысли от частного к общему или от общего к частному. В простом и сложном предложениях используются вводные слова и словосочетания, подчеркивающие логику мысли и последовательность изложения (во-первых, во-вторых, следовательно, итак, таким образом, с одной стороны, с другой стороны и т.п.) [ссылка];
- Точность (а также ясность) научного стиля — употребление большого числа терминов, как правило, слов однозначных, строго определенных в пределах конкретной науки. Нежелательна и даже недопустима замена терминов синонимами, для научной речи характерно ограничение синонимических замен; важно давать четкие определения вновь вводимым понятиям; слова — однозначны, высказывания — недвусмысленны (явление многозначности слов несвойственно научной речи). Используются вводные слова и обороты, вводные и вставные конструкции в функции уточнения; употребляются обособленные согласованные определения, в том числе причастные обороты (в синтаксической функции уточнения); необходима четкость оформления синтаксических связей; кроме того, — точные библиографические ссылки и сноски [ссылка];
- Некатегоричность изложения — взвешенность оценок в отношении степени изученности темы, действенности теории и путей решения исследуемых проблем, степени завершенности результатов исследования, так и упоминаемых в работе и цитируемых мнений других авторов-ученых и личных [ссылка];
- Диалогичность — коммуникативная направленность научной речи, необходимость учета адресата. Хотя научный текст квалифицируется как монологический, ему свойственна диалогичность, т.е. направленность речи на адресата [ссылка];

- Аргументированность научной речи — обоснованность; отсутствие или слабость аргументов в научной речи — логическая и стратегическая ошибкой [ссылка].

1.1.2. Подстили научного стиля

Научный стиль речи подразделяется на подстили: собственно-научный, научно-информативный, научно-технический, учебно-научный, научно-популярный.

Отличительная черта собственно-научного стиля — академическое изложение, адресованное специалистам. Признаки этого подстиля — точность передаваемой информации, убедительность аргументации, логическая последовательность изложения, лаконичность. Цель стиля — выявление и описание новых фактов, закономерностей, открытий. К собственно-научному подстилю относятся такие жанры, как статья, доклад, монография.

Назначение научно-информативного подстиля — сообщение научной информации с точным объектным описанием фактов. К стереотипность композиции, к особенностям относятся стандартизация языковых средств, унификация синтаксических конструкций. Этот подстиль реализуется в рефератах, аннотациях, каталогах, специальных словарях, патентных и технологических описаниях.

Научно-технический стиль направлен на применение достижений фундаментальной науки на практике. Адресат — профессионалы технико-технического профиля. Используется в руководствах, справочниках.

В учебно-научном подстиле излагаются основы наук в учебной литературе. Отличительные признаки подстиля определяются задачами, вытекающими из направленности адресату — будущему специалисту: тематическое ограничение в освещении основ научных дисциплин; обучающий характер; обилие определений, примеров, иллюстраций, пояснений, толкований. Подстиль объединяет жанры учебников (учебных монографий), учебных и учебно-методических пособий, учебных словарей,

лекций, конспектов и другого и предполагает последовательное, системное раскрытие основных вопросов предмета или учебной темы с подробным изложением устоявшейся в науке точки зрения.

Произведения научно-популярного подстиля адресованы широкому кругу читателей, поэтому научные данные излагаются в доступной и занимательной форме. Научно-популярное сообщение по характеру близко к художественной прозе — допускается эмоциональная окрашенность, образность языковых средств, замена узкоспециальной лексики общедоступной, обилие конкретных примеров и сравнений, употребление элементов устной (разговорной) речи. К подстилю относятся такие жанры, как очерк, эссе, книга, лекция научно-популярного характера, статья в периодическом издании. Цель стиля — ознакомление с описываемыми явлениями и фактами. Употребление цифр и специальных терминов минимально (каждый из них подробно поясняется). Особенности стиля: относительная лёгкость чтения, использование сравнения с привычными явлениями и предметами, упрощения, рассматривание частных явлений без обзора и классификации.

В рамках данной работы, статьи будут проверяться на соответствие собственно-научному подстилю.

1.1.3. Морфологические особенности научного стиля

Из-за наличия стилистических особенностей, описанных выше, научному стилю характерны морфологические особенности написания текста. Часть этих особенностей выражается в ограничениях:

- Использование личных местоимений. Личные и притяжательные местоимения (я, ты, мною, вы, наш) имеют отвлеченно-обобщенный характер и их употребление необходимо избегать, но некоторые формы употреблять для связи допускается (их, своих);

- Использование неопределенных местоимений (кое-что, что-нибудь). Эти местоимения, в силу их неопределенности, не употребляются;

Соблюдение перечисленных ограничений является частью проверки статьи на соответствие научному стилю.

1.2. Проверка качества текста

Проверяя текст на соответствие научному стилю, следует в первую очередь реализовать и базовую проверку на качество [2-3] текста. К такого рода анализу относится SEO-анализ. SEO (search engine optimization) анализ [2-3] популярен и актуален в связи с необходимостью продвижения ресурсов, товаров и услуг в сети Интернет. SEO анализ текста дает возможность понять, насколько часто употребляются ключевые слова в тексте, как много в тексте слов, не имеющих смысловой нагрузки и другое.

1.2.1. Числовые критерии проверки

SEO-анализе вводит следующие термины для двух критериев, которые проверяются в данной работе: Тошнота – это показатель повторений в текстовом документе ключевых слов и фраз. Синонимом тошноты является термин плотность [3]. Стоп-слова – это слова в тексте, которые не несут смысловой нагрузки [3]. Вода - процентное соотношение стоп-слов и общего количества слов в тексте [3]. Так как эти критерии вычисляемы, то можно автоматизировать их получение. Так же существует эмпирическая закономерность распределения частоты слов естественного языка - Закон Ципфа: если все слова языка или достаточно длинного текста упорядочить по убыванию частоты их использования, то частота n-го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n [4-5]. Соответствие распределения слов в тексте закону Ципфа говорит об уровне его естественности. Расчет этого критерия так же можно автоматизировать.

1.2.2. Морфологические ограничения

Одна из главных задач научного текста - донесение информации. В связи с чем, каждый научный текст является информационным. Информационный стиль в виду его главной цели – лаконичного донесения информации, также обладает морфологическими ограничениями:

- Использование слов усилителей (безусловно, очень, абсолютно и др.);
- Использование обобщений (со всего мира, весь, в общем);
- Необъективная оценка (уникальный, новейший);
- Использование риторических вопросов.

1.2.3. Качество содержания текста

Помимо описанных критериев важными показателями качества научной статьи являются её экспертность и полезность. На данный момент верификация этих критериев возможна только силами человека, однако ведутся разработки инструментов, способных выполнить данную задачу с помощью методов машинного обучения [6]. Недостатком подобных систем является сложность настройки, необходимость больших обучающих выборок и узкая ориентация в смысле предметной области.

1.3. Обзор аналогов

Важно знать о наличии программ или сервисов, предоставляющих услуги проверки текста по вышеописанным параметрам. Существуют веб сервисы, позволяющие провести SEO анализ текста, например анализатор качества контента 1y.ru [5], сервис проверки текстов text.ru [6], сервис, осуществляющий поиск стоп-слов и подсчет их процентного соотношения к общей длине текста contentmonster.ru [7].

Также существует веб ресурс glvd.ru [ССЫЛКА] – сервис «помогающий очистить тест от словесного мусора и проверяющий его на соответствие информационному стилю». Информационный и научный стили имеют общую цель – донесение информации. Научный стиль является подмножеством информационного стиля [ССЫЛКУ НАЙТИ НА ЧТО НИБУДЬ].

Сравнение аналогов будет проводиться по следующим критериям:

• Многокритериальная проверка - как много критериев проверки использует сервис;

• Ограничение длины текста - отсутствие ограничения длины текста, поступающего на проверку;

• Проверка стиля - проверка текста на соответствие научному или информационному стилю;

• Возможность загрузки файлов для проверки.

В табл.1 представлено сравнение аналогов.

Аналог	Многокритериальная проверка	Нет ограничения на длину текста	Проверка стиля	Возможность загрузки файлов для проверки
1y.ru	-	+	-	-
text.ru	+	-	-	-
contentmonster.ru	+	+	-	-
glvrd.ru	+	+	+	-

Как показывает сравнение аналогов, ни один из аналогов не имеет возможности проанализировать текст из файла. Эту возможность необходимо будет реализовать.

1.4. Используемые правила проверки научных статей в существующем курсе

В связи с отсутствием формализованных правил проверки научных статей в научных журналах, для обучения студентов СпбГЭТУ на кафедре МОЭВМ был создан онлайн курс на платформе Stepik [ссылка], в котором

используются правила проверки, полученные обобщением требований к статьям в журналах:

1. Термины из названия упоминаются равномерно по тексту статьи.
2. Каждое ключевое слово упоминается в основном тексте хотя бы один раз.
3. Аннотация написана в совершенном времени.
4. Во Введении выполнена постановка цели, кратко описана решаемая проблема, обозначены задачи.
5. В основной части вашей работы присутствует развернутая постановка цели исследования, описание методов решения и результатов их применения.
6. В Выводах описан краткий результат решения каждой из поставленных задач.
7. В Выводах обозначены направления для дальнейших исследований.
8. Более половины элементов списка литературы - актуальные и значимые научные работы.
9. Все элементы списка литературы имеют минимум одно упоминание в тексте.
10. Все рисунки и таблицы имеют подрисуночные подписи и ссылки в тексте.
11. Все формулы имеют ссылки в тексте и описание используемых обозначений.
12. Иллюстративный материал занимает не более 30-40% от общего объема работы.

Требования 2, 8, 9, 10 автоматизируемы, так как являются структурными.

1.5. Выводы

В разделе было дано описание научного стиля и описание критериев проверки качества текста, в результате которого было принято решение о том, что программа должна осуществлять проверку текста на соответствие собственно-научному подстилю, в том числе проверяя соблюдение ограничений, накладываемых морфологическими особенностями научного стиля. Также был проведен анализ аналогов, проверяющих качества текста, в результате которого было получено еще одно требование к решению – анализ текста статьи, полученного из файла.

2. ПОСТАНОВКА ЗАДАЧИ И ВЫБОР МЕТОДА РЕШЕНИЯ

2.1. Задача

Реализовать автоматизированное решение — информационную систему проверки статей на соответствие научному стилю, в том числе, как давая числовую оценку работе, так и показывая ошибки, допущенные автором. Требования к системе сформулированы на основе обзора предметной области и последующего использования приложения — для проверки работ студентов в рамках курса по написанию научных статей.

2.2. Требования к решению

Были сформированы следующие требования к решению:

- Выполнение проверки на соответствие научному стилю и поиск наиболее частых ошибок;
- Простота использования решения – интерактивный, понятный пользовательский интерфейс;
- Наглядное представление результатов;
- Возможность контейнеризации решения для быстрого развертывания в любой среде.

Также необходимо реализовать возможность получать текст статьи из файла, так как это удобно пользователю. Принято решение реализовать получение текста из файлов формата PDF, в связи с тем, что любой другой формат легко форматируется в него.

2.3. Выбор метода решения

Принято решение реализовать веб-сервис, так как такой вид решения обладает следующими преимуществами:

- Пользователю не нужно ничего устанавливать;
- Возможность отображать много информации на экране в удобном для восприятия виде, используя контекстные меню, всплывающие подсказки и продуманный пользовательский интерфейс;
- Интерактивность – приложение может модифицировать экран, реагируя на действия пользователя.

Выбранный метод решения позволит соблюсти требования, относящиеся к пользовательскому интерфейсу.

3. ОПИСАНИЕ МОДЕЛИ ПРОВЕРКИ СТАТЬИ

3.1. Числовые критерии проверки

Выше были описаны три числовых критерия проверки статьи, которые можно автоматизировать. Для удобства обозначим их:

• Тошнота или уровень ключевых слов в тексте – α ,

• Уровень воды в тексте или процентное соотношение стоп-слов и общего количества слов в тексте – β ,

• Значение отклонения текста статьи от идеальной кривой по Ципфу [4-5] – λ .

Однако, для использования числовых критериев для оценки качества статьи, необходимо установить, как качество статьи связано со значениями этих числовых критериев.

3.1.1. Исследование взаимосвязи значений числовых критериев с качеством научной статьи

Поскольку требования научного стиля плохо формализуемы, то будем рассматривать экспериментальные свидетельства качества научных текстов — факты публикации определенных текстов в научных изданиях, индексируемых в ВАК [8] и РИНЦ [9], так как к изданиям, индексируемых ими, предъявляются жесткие требования как к оформлению, так и к содержанию и структуре статей. Для простоты анализа установили, что качество научной статьи можно выразить булевой переменной (1 - текст соответствует нормам научного стиля, 0 - текст не соответствует нормам научного стиля). Рассмотрим, статистические свойства распределений значений критериев α , β и λ для научных статей, опубликованных в изданиях ВАК и/или РИНЦ. Была исследована выборка из 2500 статей в формате PDF, полученная с помощью исполняемого сценария [10], который выполняет веб-скрэпинг [11] научной интернет-библиотеки "Киберленика" [12]. Были загружены и проанализированы статьи технической направленности, специальностей «Информатика» и «Вычислительная техника», опубликованные в изданиях ВАК и/или РИНЦ.

Для исследования требовалось быстрое в разработке и внесению изменений, легкое в использовании решение, получающее текст из PDF файла, и рассчитывающее значения числовых критериев по полученному тексту.

Был реализован исполняемый сценарий [ссылка] на языке Python. Выбор обоснован Python легкостью разработки исполняемых сценариев на языке, а также наличием большого количества модулей для разнообразных задач.

3.1.2. Проверяемая гипотеза

В рамках исследования проверялась гипотеза о том, что качество научной статьи влияет на значения ранее определенных числовых критериев, а также то, что полученная выборка значений будет соответствовать нормальному распределению.

Исследование на выборке из 2500 прошедших рецензирование и опубликованных статей позволит получить математические параметры распределений, что позволит установить пороговые значения числовых критериев для статей хорошего качества.

3.1.3. Подчинение числовых критериев нормальному распределению

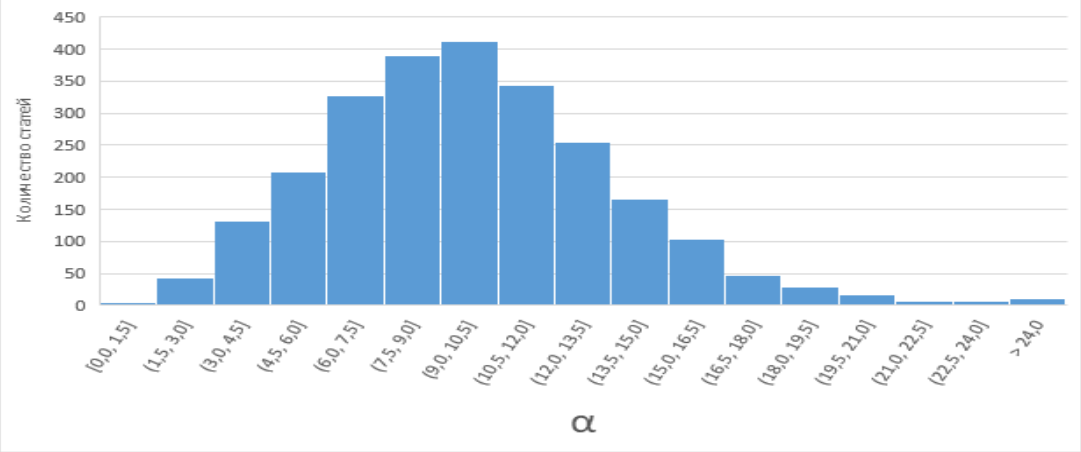

Рисунок 1 – Гистограмма распределения значений уровня ключевых слов в тексте статей из выборки

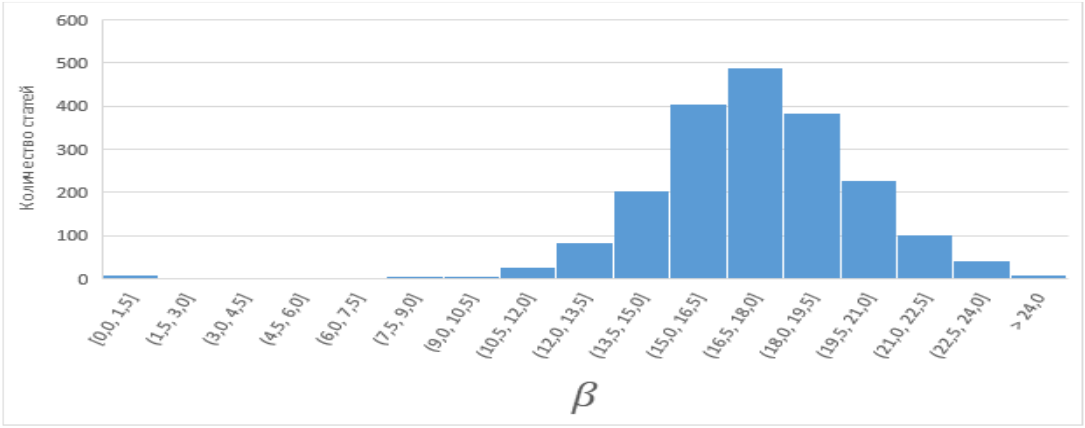

Рисунок 2 – Гистограмма распределения значений уровня водности текста статей из выборки

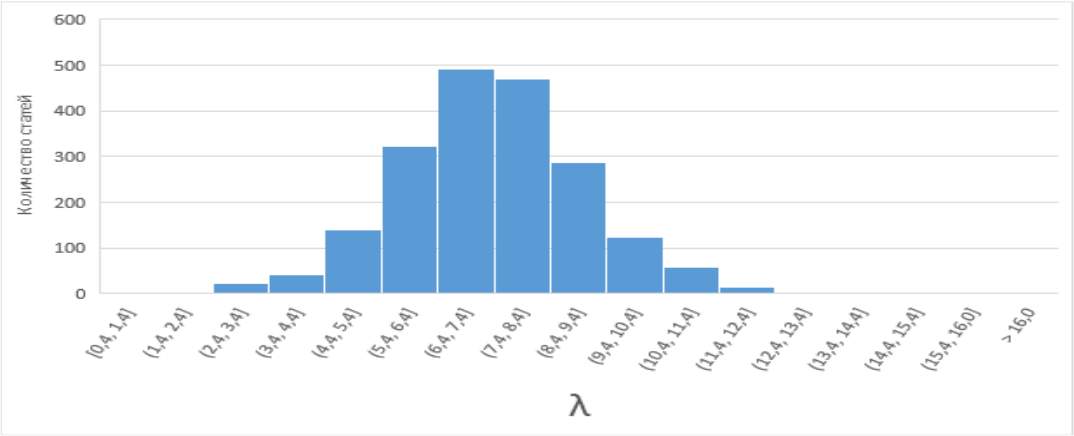

Рисунок 3 – Гистограмма распределения значений отклонения от идеальной кривой по Ципфу текста статей из выборки

Из рис. 1-3 видно, что у каждого из распределений наблюдается четкий пик и большинство значений сконцентрированы вокруг него симметрично, в связи с чем можно предположить, что распределения нормальные. Для доказательства воспользуемся тремя тестами нормальности: критерий Шапиро-Уилка [13], критерий Колмогорова-Смирнова [14], критерий Андерсона-Дарлинга [15]. В каждом из тестов проверяется нулевая гипотеза [16], о том, что каждая выборка получена из нормального распределения. Так, нулевая гипотеза считается верной до того момента, пока нельзя доказать обратное. Статистическая значимость [16] для тестов равна 0,05. P-значение [17] — величина, используемая при тестировании статистических гипотез. Фактически это вероятность ошибки при отклонении нулевой гипотезы.

Использовалась [18] реализация тестов из статистической библиотеки SciPy [19]. На выходе каждый тест выдает два значения – D (Статистика критерия для эмпирической функции распределения [14]) и P-значение. В случае, если значение P-значение близко к 0, или значительно меньше D – нулевая гипотеза не может быть отвергнута.

Результаты по каждому числовому критерию представлены в табл. 1-3:

Таблица 1 - результаты тестов для выборки значений уровня ключевых слов в тексте

Критерий	D	P-значение
Шапиро	0.967	1.407e-23
Колмогоров-Смирнов	0.309	0.0
Андерсон-Дарлинг	8.293	0.787

Таблица 2 - результаты тестов для выборки значений водности текста

Критерий	test-statistics	p-value
Шапиро	0.942	3.815e-30
Колмогоров-Смирнов	0.229	0.0
Андерсон-Дарлинг	14.957	0.787

Таблица 3 - результаты тестов для выборки значений отклонения текста от идеальной кривой по Ципфу

Критерий	D	P-значение
Шапиро	0.864	3.512e-42
Колмогоров-Смирнов	0.129	0.0
Андерсон-Дарлинг	28.732	0.787

Как видно из результатов тестов – нет поводов отклонить нулевую гипотезу для каждой выборки, то есть можно считать, что каждый числовой критерий подчиняется нормальному закону распределения.

В таблице 4 представлены математическое ожидание и дисперсия каждой из выборок:

Таблица 4 – Характеристики выборок

Выборка	Мат. ожидание	Дисперсия
α	9.822	3.902
β	17.145	3.082
λ	7.396	2.069

Так как распределения можно считать нормальными, то, согласно эмпирическому правилу [20], более 2/3 распределения будет содержаться в следующем интервале

$[\mu-\sigma, \mu+\sigma]$, где μ – среднее значение выборки, а σ – среднеквадратичное отклонение.

На основе этих данных были установлены интервалы для каждого из числовых критериев:

Таблица 5 – Установленные интервалы

Критерий	Интервал
α	~ [6, 14]
β	~ [14, 20]
λ	~ [5.5, 9.5]

3.1.4. Независимость числовых критериев

Независимость числовых критериев друг от друга показывает ценность каждого из них в отдельности – ни один из критериев не дублирует уже известную информацию. Для доказательства этого была вычислена матрица ковариации. Был использован линейный коэффициент корреляции (коэффициент корреляции Пирсона) для расчета корреляции числовых критериев на основе полученных выборок:

$$r_{XY} = \frac{\mathrm{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum (X-\bar{X})^2 (Y-\bar{Y})^2}} \quad (1)$$

где X и Y – значения критериев статьи, σ – среднеквадратичное отклонение, cov_{XY} - ковариация X и Y, $\bar{X}$ и $\bar{Y}$ - средние значения выборок.

Полученная матрица ковариации:

$$\begin{pmatrix} 1 & -0.07 & 0.22 \\ -0.07 & 1 & 0.01 \\ 0.22 & 0.01 & 1 \end{pmatrix} \quad (2)$$

Коэффициент корреляции Пирсона может принимать значения от -1 до 1, где 0 означает полную независимость переменных друг от друга. Полученный коэффициент корреляции между α и β равен -0.07, а между β и λ равен 0.01, что позволяет утверждать о независимости данных критериев. Между критериями α и λ наблюдается незначительная зависимость, что связано с учетом количества ключевых слов при вычислении обоих критериев.

3.1.5. Запуски на тестовой выборке и других текстах

Для проверки адекватности полученных интервалов и формулировки критерия принятия решения о соответствии научному стилю, было проведено оценивание 80 дипломных бакалаврских работ студентов СПбГЭТУ «ЛЭТИ» кафедры МОЭВМ 2016 и 2017 годов. Кафедрой были предоставлены оценки данных работ, что позволит сравнить их с результатами анализа критериев, и подсчитать количество ошибок 1 и 2 рода [21]. Примем допущение о том, что качество текста дипломной работы определяет ее оценку.

Перед сравнением примем следующие условия оценки работ с помощью анализа критериев:

Таблица 6 – Условия оценки работ

Оценка	Количество критериев, попадающих в интервал
5	$N \in [2;3]$
4	$N \in [1;2]$
3	$N \in [0;1]$

В ходе проверки статей было выявлено 28 ошибок 1 или 2 рода, то есть в 65% случаев оценка по анализу критериев совпала с оценкой, поставленной аттестационной комиссией. Таким образом можно сформулировать следующий критерий принятия решений о качестве статьи

$$\alpha \in [6;14] \wedge \beta \in [14,20] \wedge \lambda \in [5.5,\ 9.5] \quad (3),$$

то есть все три числовых критерия должны попадать в установленные интервалы. Данное условие нужно считать необходимым, но не достаточным, в связи отсутствием анализа полезности содержания статьи.

Для оценки корректности критерия, рассмотрим его работу на текстах других жанров. Результаты проверки данных текстов должны показать несоответствие текста научному стилю. Тексты, используемые для проверки:

- работа «Корчеватель» [22-23] – сгенерированная в научном стиле, не имеющая смысла статья, используемая как пример формально корректного, но бессмысленного научного текста;
- популярные статьи в it-сообществе Хабр [28]: «Моё разочарование в софте» [24], «Наши с вами персональные данные ничего не стоят» [25], «Рассказ о том, как я ворую номера кредиток и пароли у посетителей ваших сайтов» [26], «Трехмерный движок на формулах Excel для чайников» [27];
- первый том «Капитала» Карла Маркса;
- роман «Идиот» Фёдора Достоевского;
- роман-поэма «Мёртвые души» Николая Гоголя;
- роман «Путешествие к центру Земли» Жюля Верна.

Результаты оценки представлены в таблице 7:

Таблица 7 – Результаты оценки текстов

Текст	α	$\alpha \in [6;14]$	β	$\beta \in [14,20]$	λ	$\lambda \in [5.5,\ 9.5]$
Псевдонаучная статья «Корчеватель»	10.38	Да	18.50	Да	6.84	Да
Интернет-статья «Моё разочарование в софте»	3.66	Нет	31.68	Нет	5.35	Нет
Интернет-статья «Наши с вами	10.56	Да	32.10	Нет	6.84	Да

персональные данные ничего не стоят»						
Интернет-статья «Рассказ о том, как я ворую номера кредиток и пароли у посетителей ваших сайтов»	6.61	Да	36.46	Нет	6.82	Да
Интернет-статья «Трехмерный движок на формулах Excel для чайников»	11.61	Да	27.91	Нет	9.27	Да
«Капитал» Карла Маркса	5.84	Нет	28.94	Нет	138.22	Нет
«Идиот» Фёдора Достоевского	6.65	Да	45.65	Нет	53.12	Нет
«Мёртвые души» Николая Гоголя	7.14	Да	40.81	Нет	35.58	Нет
«Путешествие к центру Земли» Жюля Верна	5.03	Нет	35.19	Нет	21.56	Нет

По результатам проверки, значения всех трёх критериев статьи «Корчеватель» попали в установленные интервалы, т.е. работу можно считать соответствующей научному стилю, что показывает соответствие стиля данной статьи предъявляемым требованиям. Интернет-статьи и литературные произведения не написаны в научном стиле, и выделяются повышенным значением β. Поскольку, на всех примерах альтернативных жанров критерий

не показал ложных срабатываний, можно считать, что он корректно выполняет задачу определения соответствия научному стилю.

3.2. Ошибки несоответствия текста научному стилю

На основе обзора предметной области были выделены ошибки соответствия текста научному стилю для реализации, которые можно классифицировать:

- Стилистические ошибки и предупреждения – пренебрежение правилами написания научных работ;
- Структурные ошибки – ошибки соблюдения рекомендаций по структуре научной статьи, а также несоответствия в структуре статьи.

Реализована проверка следующих стилистических ошибок:

- Использование личных местоимений;
- Использование обобщений;
- Необъективная оценка;
- Использование усилителей;
- Использование риторических вопросов.

Реализована проверка следующих структурных ошибок:

- Отсутствие ссылки на указанный источник;
- Использование устаревшего источника;
- Отсутствие ссылки на рисунок;
- Отсутствие ссылки на таблицу;
- Наличие коротких разделов – разделов, состоящих менее чем из трёх предложений.
- Использование указанных ключевых слов в тексте.

3.3. Описание модели оценки соответствия научной статьи заданным требованиям

Наглядным способом оценки на соответствие идеалу является шкала, в связи с чем соответствие научной статьи заданным требованиям – числовое

значение в промежутке от 0 до 100. Для получения значения по шкале используются полученные значения числовых критериев, а также информация об ошибках в статье.

Назовем значение по шкале оценкой и обозначим как K. Основой для определения K является формула 3, в которой используются числовые критерии α, β и λ, определенные ранее. Попадание значений числовых критериев в ранее установленные дозволенные промежутки определяет базовое значение K. Попадание критерия в установленный промежуток обозначим как функцию E:

$$\mathrm{E}(\alpha)=\begin{cases}1, \alpha \in [6;14]\\ 0, \alpha \notin [6;14]\end{cases},\ \mathrm{E}(\beta)=\begin{cases}1, \beta \in [14;20]\\ 0, \beta \notin [14;20]\end{cases},\ \mathrm{E}(\lambda)=\begin{cases}1, \lambda \in [5.5;9.5]\\ 0, \lambda \notin [5.5;9.5]\end{cases}.$$

Обозначим базовое значение K как B, тогда:

$$\mathrm{B}=35\times\mathrm{E}(\alpha)+35\times\mathrm{E}(\beta)+30\times\mathrm{E}(\lambda)$$

Коэффициент при $\mathrm{E}(\lambda)$ меньше других коэффициентов в связи с тем, что λ отражает отклонение речи от естественной, что менее важно в контексте научной статьи, чем употребление ключевых слов и уровень «воды» в тексте.

Итоговое значение K получается при вычете штрафов за ошибки из B. Обозначим штраф как Φ. Штраф за каждую стилистическую и структурную ошибку равен двум баллам, то есть:

$$\Phi=2\times N,$$

где N – количество ошибок.

Рассмотрим пример анализа статьи. Допустим, в результате анализа, значения всех трёх числовых критериев попадают в заданные промежутки, значит B равно 100. В статье было найдено 7 ошибок, значит Φ равно 14. В итоге, оценка соответствия статьи научному стилю – 84 из 100.

3.4. Заключение

В результате исследования была сформулирована модель оценки соответствия статьи научному стилю, реализованная в решении.

4. ОПИСАНИЕ РЕШЕНИЯ

4.1. Общая архитектура решения

Выбранное решение подразумевает веб-приложение, взаимодействующее с сервером для анализа текста статьи и получения результатов, база данных необходима для сохранения результатов анализа статей. На рис. 4 представлена обобщенная архитектура решения:

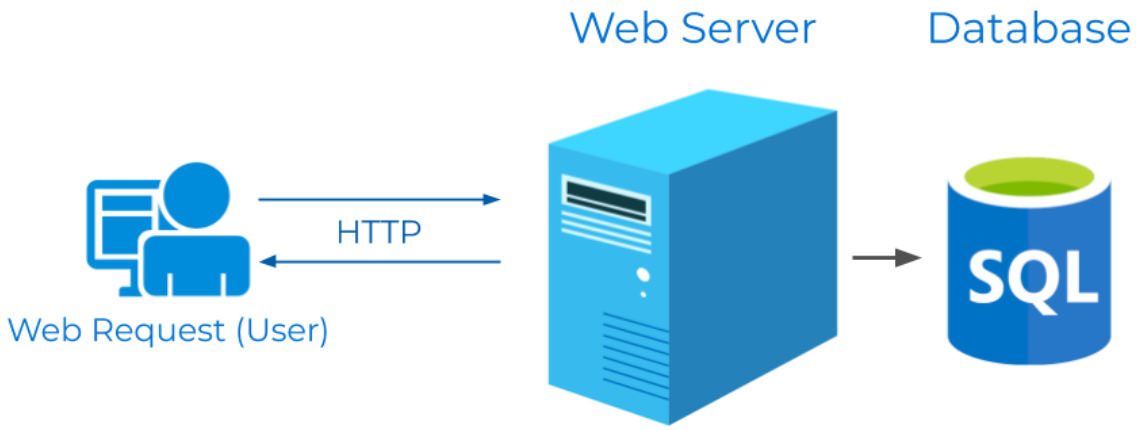

Рисунок 4 – Обобщенная архитектура решения

Более подробно архитектура будет рассмотрена после описания используемых технологий.

4.2. Сценарии использования

Практическая польза решения заключается в анализе статей на соответствие научному стилю и просмотре результатов анализа, поэтому существуют следующие сценарии использования: [СКРИНШОТЫ]!!!

Сценарий №1:

1. Пользователь открывает веб-приложение;
2. Пользователь выбирает файл со статьей для проверки;

3a. Пользователь заполняет настройки проверки;

3b. Пользователь импортирует настройки в формате json;

3*. Пользователь экспортирует настройки в формате json;

4. Пользователь нажимает на кнопку «Начать проверку»;

5. Пользователь попадает на страницу с результатом проверки статьи;

6. Пользователь видит оценку стиля статьи, значения числовых критериев, советы по улучшению значений критериев;

7. Пользователь видит количество ошибок в тексте, выделенные ошибки в тексте, советы по их исправлению;

8. Пользователь пользуется советами и улучшает статью.

Сценарий №2:

1. Пользователь открывает веб-приложение на странице с результатом проверки статьи;

2. Пользователь видит оценку стиля статьи, значения числовых критериев, советы по улучшению значений критериев;

3. Пользователь видит количество ошибок в тексте, выделенные ошибки в тексте, советы по их исправлению;

4. Пользователь пользуется советами и улучшает статью.

4.3. Использованные технологии

В качестве основной платформы разработки был выбран .Net Core – стремительно развивающаяся, универсальная платформа разработки с открытым кодом, которую поддерживает корпорация Майкрософт и сообщество .Net на сайте GitHub [https://github.com/dotnet/core]. Она является кроссплатформенной (поддерживает Windows, macOS и Linux) и может использоваться для создания приложений для устройств, облака и Интернета вещей. В качестве языка разработки выбран основной язык платформы .Net и .Net Core – C#.

Платформа .Net Core предоставляет фреймворк ASP.Net Core – версия ASP.Net с открытым исходным кодом, которую поддерживает корпорация Майкрософт и сообщество .NET на сайте GitHub [https://github.com/aspnet/AspNetCore]. ASP.Net – популярный фреймворк для веб-разработки для .Net платформы.

.Net Core решения, в том числе и решения ASP.Net Core, могут быть быстро развернуты и опубликованы в облачном сервисе Microsoft – Azure, а также Microsoft предоставляет официальные docker-контейнеры, что упрощает контейнеризацию .Net Core решений.

ASP.Net Core предоставляет несколько шаблонов разработки веб-приложений, рассмотрим некоторые из них:

1. ASP.Net Core MVC – MVC [ссылка] фреймворк для создания динамических веб-страниц с явным разделением ответственности [https://en.wikipedia.org/wiki/Separation_of_concerns], использующий Web API [ссылка] - RESTful интерфейсы [ссылка], и движок представлений Razor [ссылка].
2. ASP.Net Core Web Api + JS фреймворк – данный шаблон позволяет создать Web Api и использовать JS фреймворки, такие как Angular, React и Vue.
3. Blazor – экспериментальный фреймворк использующий Razor и C# как в бекенде так и на фронтенде, запускающийся в браузере, используя WebAssembly [ссылка].

Был использован ASP.Net Core MVC, в связи с тем, что этот фреймворк стабилен, в отличие от экспериментального Blazor, а также проект проще в сборке и публикации в облаке или в контейнере, чем стека Web API + JS фреймворк.

В связи с использованием стека .NET используется база данных того же разработчика – SQL Server. Так же используется ORM (Object-relational mapper) [ССЫЛКА] Entity Framework Core для простоты работы с базой данных, так как логика работы с ней не предусматривает сложных запросов.

4.4. Архитектура решения

На рис. 5 представлена архитектура решения:

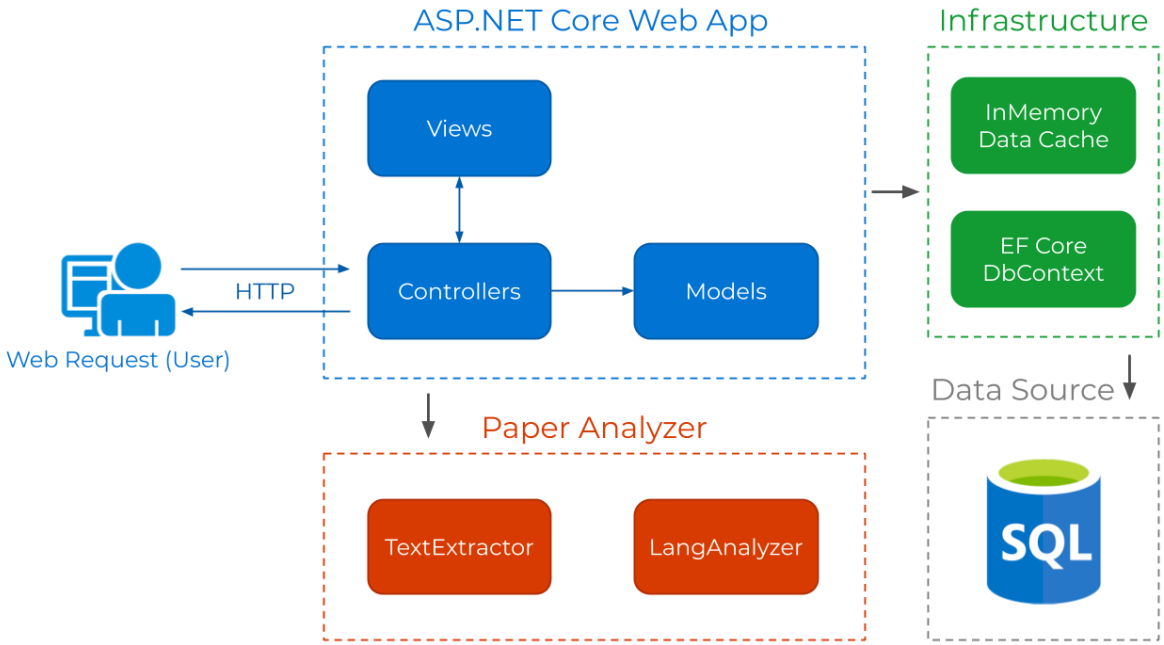

Рисунок 5 – Архитектура решения

Веб-приложение реализовано с помощью фреймворка ASP.NET Core MVC, и использует паттерн MVC (Model – View – Controller), позволяющий отделить логику приложения от данных и их представления пользователю. Взаимодействие с базой данных происходит через ORM фреймворк Entity Framework Core.

4.5. Описание алгоритмов работы

На рис. 5 изображен модуль PaperAnalyzer, который выполняет обработку pdf файла статьи – получение текста и его последующий анализ. Общий алгоритм обработки статьи представлен на рис. 6:

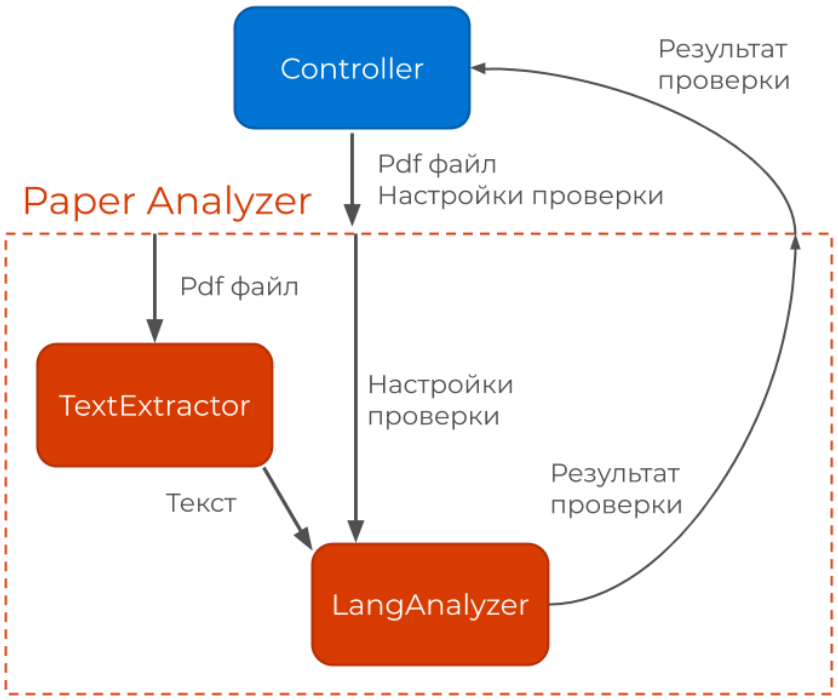

Рисунок 6 – Общий алгоритм обработки статьи

4.5.1. Получение текста из pdf

За извлечения текста из pdf файла отвечает модуль TextExtractor, изображенный на рис. 5. Работа с pdf форматом осуществляется с помощью бесплатной библиотеки iTextSharp [https://itextpdf.com/ru]. С помощью этой библиотеки pdf файл обрабатывает постранично, из каждой страницы извлекается весь текст.

4.5.2. Анализ текста

Полученный текст анализируется с помощью модуля LangAnalyzer, изображенного ни рис. 5. Модуль отвечает за нормализацию текста – приведение всех слов к словарной форме, определение морфологических признаков. Модуль реализует всю последовательность лингвистической обработки текста:

- Текст разбивается на предложения;
- Определяются части речи всех слов текста;
- Находятся морфологические характеристики всех слов;

- Снимается омонимия – выбирается одно слово из множества, предлагаемых морфословарем.

Используемая лингвистическая обработка теста реализована в библиотеке с открытым исходным кодом, расположенной на платформе GitHub [https://github.com/zamgi/lingvo--PosTagger-ru].

4.5.3. Вычисление числовых критериев

Формализуем числовые критерии для дальнейшего описания:

$$\alpha = \frac{k}{N} \times 100 \ , \beta = \frac{s}{N} \times 100 \ , \ \lambda = \sqrt{\frac{1}{N^*} \sum_{i=1}^{N^*} (x_i - \frac{x_1}{i})^2} \ , \text{ где}$$

N – количество слов в тексте, k – количество ключевых слов текста (два самых часто употребляемых слова), s – количество стоп-слов в тексте, N^* – количество слов в тексте, употреблённых минимум 5 раз, x – количество употреблений слова в тексте, x_1 – количество употреблений слова в тексте.

После лингвистического анализа, на выходе получен массив объектов, представляющих предложения – наборы объектов – слов с их морфологическими характеристиками. Для дальнейшей обработки массив обрабатывается и получается словарь, в котором ключом является слово, а значением – количество его употреблений в тексте.

Получив словарь, сразу получаем значение k. Стоп-слова, как было описано в 1 разделе, это слова, которые не несут никакой смысловой нагрузки, к ним относятся все предлоги, частицы, междометия, союзы, наречия, местоимения, а также вводные слова и выражения [3]. Так как морфологическая информация о каждом слове уже определена, словарь фильтруется по характеристикам ключей, а именно по части речи, или содержанию слова в специальном словаре стоп-слов [ССЫЛКА], таким образом вычисляется значение s.

4.5.4. Анализ стилистических ошибок в тексте

В разделе 3.2 были перечислены типы стилистических ошибок, проверка которых реализована:

1. Использование личных местоимений;
2. Использование обобщений;
3. Необъективная оценка;
4. Использование усилителей;
5. Использование риторических вопросов.

Проверка типов ошибок 1, 2 и 3 выполняется с помощью анализа морфологических признаков слов, полученных в результате лингвистического анализа текста. Проверка типа ошибок 4 выполняется с помощью словарей со словами усилителями (абсолютно, безусловно). Пятый тип ошибок является предупреждением, если предложение в статье является вопросом, создается предупреждение.

4.5.5. Анализ структурных ошибок в тексте

В разделе 3.2 были перечислены типы структурных ошибок, проверка которых реализована:

1. Отсутствие ссылки на указанный источник;
2. Использование устаревшего источника;
3. Отсутствие ссылки на рисунок;
4. Отсутствие ссылки на таблицу;
5. Наличие коротких разделов – разделов, состоящих менее чем из трёх предложений.
6. Использование указанных ключевых слов в тексте.

Ошибки типов 1-4 проверяются с помощью использования регулярных выражений на необработанном тексте. Используются регулярные выражения для обработки ссылок на источники, источников в списке литературы, ссылок на рисунки и таблицы, названий рисунков и таблиц.

Для проверки наличия ошибок пятого типа обрабатывается полученный после лингвистического анализа текста массив предложений. С помощью полученного на вход списка названий разделов, данный массив группируется по разделам – массивам меньшего размера, представляющими разделы статьи. Далее проверяется количество элементов в таких массивах.

Проверка наличия ошибок шестого типа выполняется на словаре с количеством употреблений слов в тексте. На вход был получен список ключевых слов текста, указываемых в статье для определения тем, к которым относится статья. Проверяется наличие слов из этого списка в словаре.

5. ИССЛЕДОВАНИЕ РЕШЕНИЯ

5.1. Исследование времени анализа статьи

Анализ статьи состоит из двух частей:

- Извлечение текста из pdf файла;
- Анализ текста.

Соответственно и время анализа статьи состоит из времени извлечения текста из pdf файла и анализа полученного текста.

В табл. 8 представлена информация о выборке из 6 файлов pdf содержащих текст – статьи, книги:

Таблица 8 – Информация о выборке файлов

Файл	Размер, Кбайт	Количество символов	Количество страниц	Время извлечения текста, мс	Время анализа текста, мс	Общее время обработки, мс
Файл 1	158,12	10156	4	80	38	118
Файл 2	293,79	20818	8	111	71	182
Файл 3	702,12	42706	37	206	144	350
Файл 4	1654,55	108521	76	841	295	1136
Файл 5	9738,32	4120625	2167	14437	37659	52096
Файл 6	42533,51	2717686	1047	35778	15842	51620

Так как на вход первому этапу обработки статьи попадает файл, необходимо исследовать зависимость времени извлечения текста от размера файла, а также от количества страниц в нем. На вход второму этапу обработки – анализу текста, поступает строка (текст). Необходимо исследовать зависимость времени обработки текста от количества символов в нём.

На рис. 7 представлен график зависимости времени извлечения текста из pdf файла от его размера на выборке из файлов 1 – 4, на рис. 8 представлен тот же график, но на выборке из файлов 1 – 6:

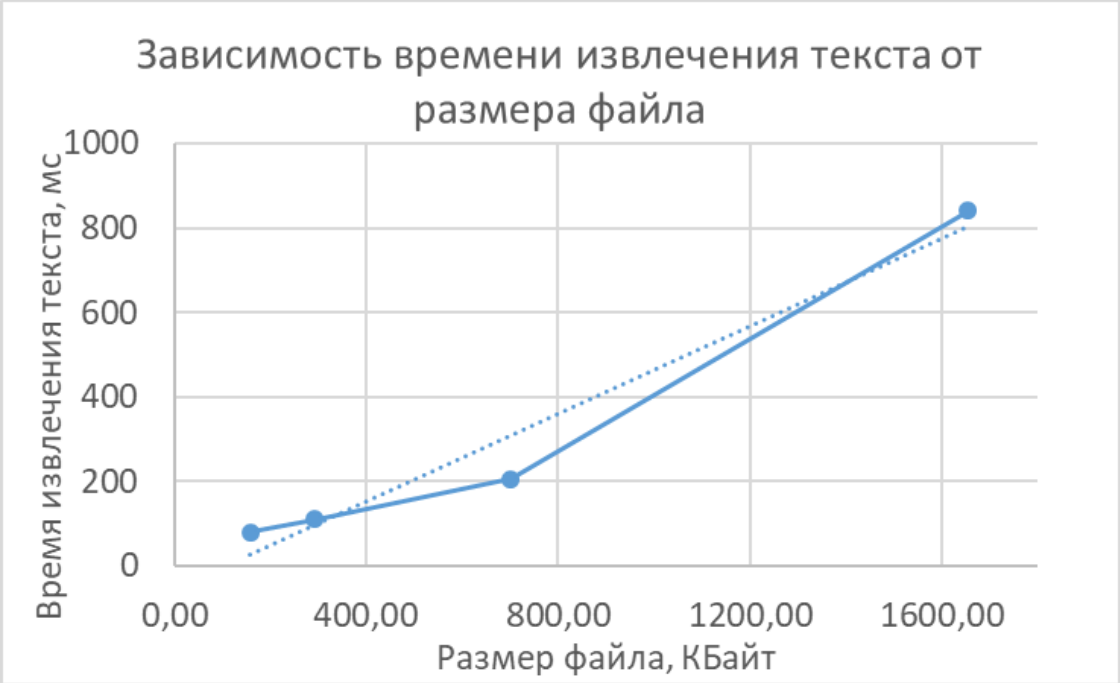

Рисунок 7 – График зависимости времени извлечения текста из pdf файла от его размера на выборке файлов 1 – 4

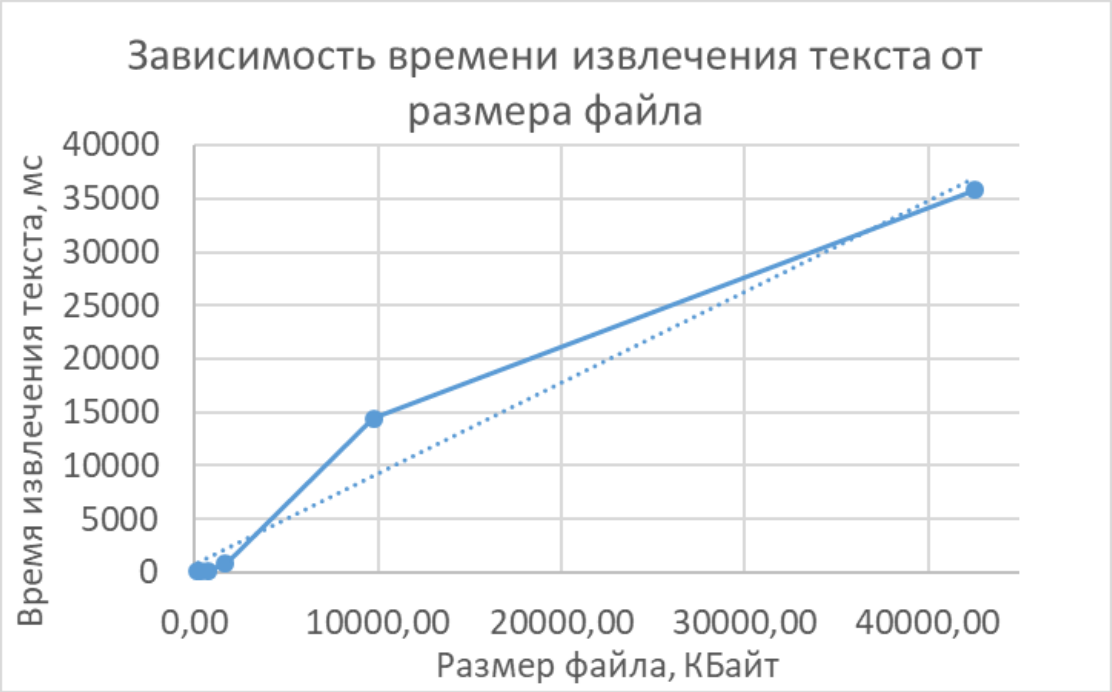

Рисунок 8 – График зависимости времени извлечения текста из pdf файла от его размера на выборке файлов 1 – 6

На рис. 7 наблюдается линейная зависимость времени извлечения текста из файла от его размера, на рис. 8, при добавлении файлов сильно большего размера, линейная зависимость так же прослеживается.

На рис. 9 представлен график зависимости времени извлечения текста из pdf файла от количества страниц в нем на выборке из файлов 1 – 6:

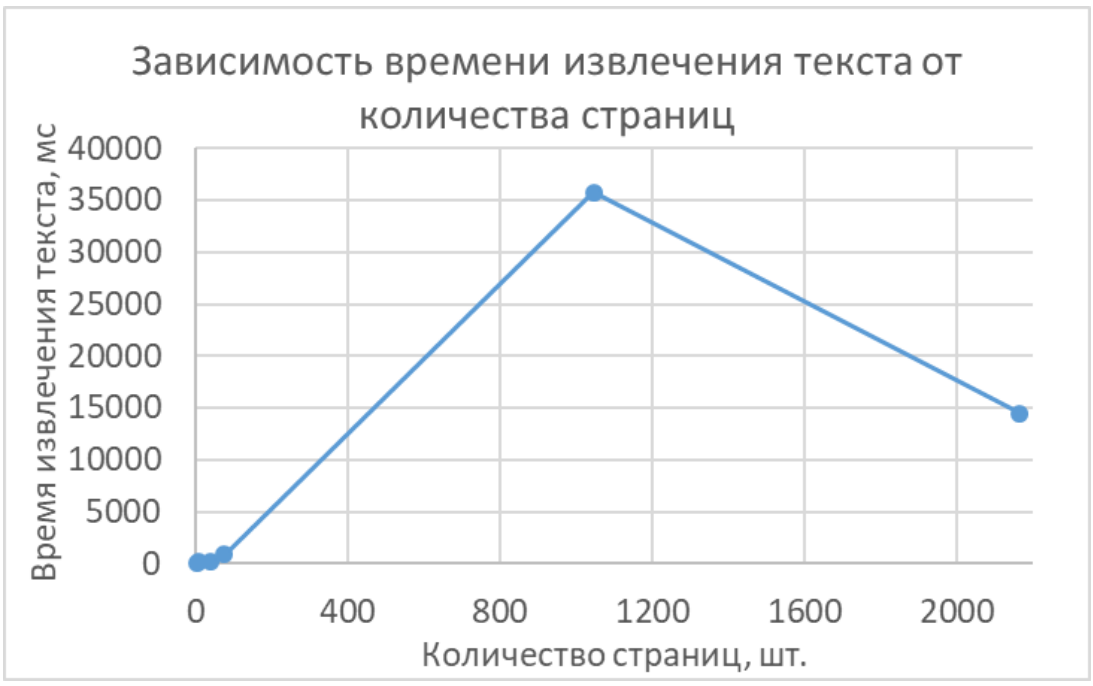

Рисунок 9 – График зависимости времени извлечения текста из pdf файла от количества страниц в нем на выборке файлов 1 – 6

Как видно из графиков на рис. 8 и рис.9 – пример файла 5 доказывает, что количество страниц не влияет на время извлечения текста из файла, но существует линейная зависимость между размером файла и временем извлечения текста. Важно отметить, что размер файла pdf зависит не только от содержимого текста в нем, но и от изображений, ссылок, шрифтов и других ресурсов.

На рис. 10 представлен график зависимости времени анализа текста от количества символов на выборке из файлов 1 – 4, на рис. 11 представлен тот же график, но на выборке из файлов 1 – 6:

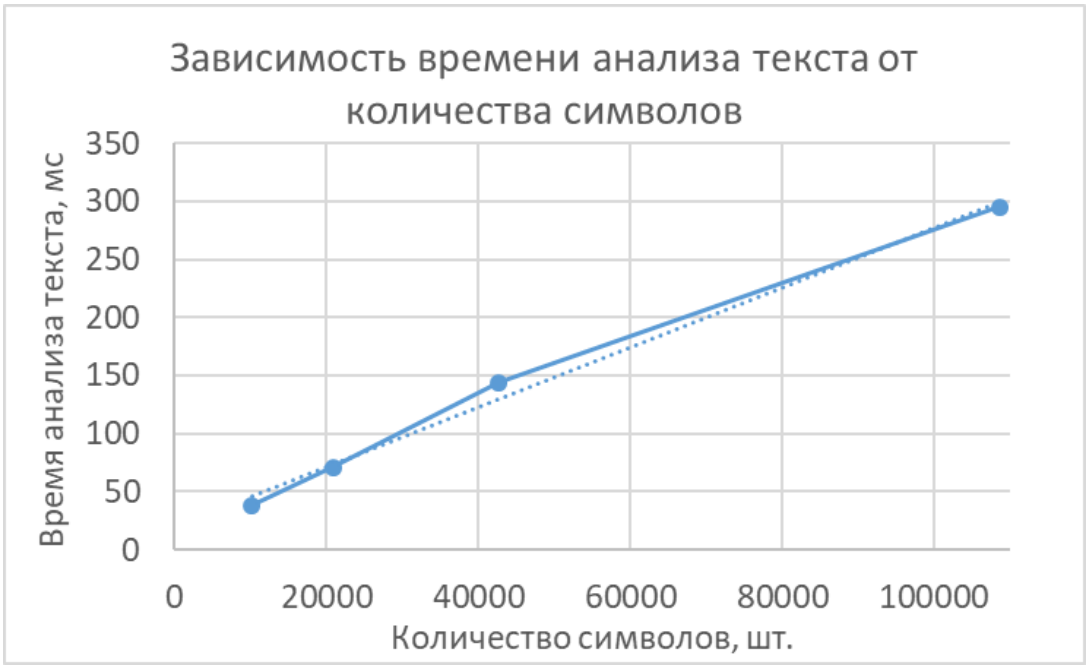

Рисунок 10 – График зависимости времени анализа текста от количества символов на выборке файлов 1 – 4

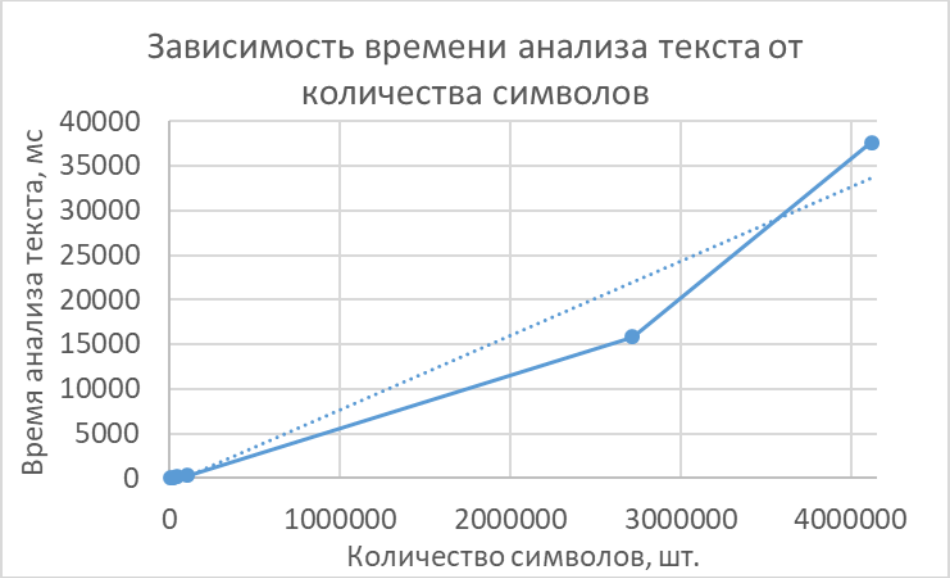

Рисунок 11 – График зависимости времени анализа текста от количества символов на выборке файлов 1 – 6

На рис. 10 наблюдается линейная зависимость времени анализа текста от количества символов, на рис. 11, при добавлении файлов с гораздо большим количеством текста, линейная зависимость так же прослеживается.

Средняя по выборке скорость обработки файла для извлечения текста составила 1977 КБайт в секунду. Средняя по выборке скорость анализа текста составила 251000 символов в секунду.

Учитывая направленность решения на анализ научных статей, файл статьи размером около 1.5 Мбайт и содержащий около 100000 символов (50 страниц) будет обработан и проанализирован за время, близкое к 1 секунде. Стоит заметить, что при использовании веб-сервиса, следует брать в расчет время загрузки файла на сервис, а также время загрузки ответа с сервера.

5.2. Пригодность использования приложения на кафедре

ЗАКЛЮЧЕНИЕ

По итогам работы были получены следующие результаты:

Поставленные задачи были решены, цель работы была достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ