«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

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.

Цель работы.

Предложить решение 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.

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. Характеризуется использованием научной терминологии, исключая жаргонизмы. Научный стиль не допускает личного изложения [1]. Проверяя текст на соответствие научному стилю, следует в первую очередь реализовать и базовую проверку на качество текста. К такого рода анализу можно отнести 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. 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]. 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. В предыдущей работе [7] был проведен более детальный обзор пригодности данных критериев к задачам автоматической проверки качества стиля статей. Помимо описанных числовых 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.

1.2. Обзор аналогов

Существуют веб сервисы, проверяющие текст по этим критериям - сервисы, позволяющие провести SEOанализ 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].

Так же 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.

2. ИССЛЕДОВАНИЕ

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.

2.1. 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]. 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 простое 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.

2.1.1. 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.

2.1.2. Подчинение числовых критериев нормальному распределению

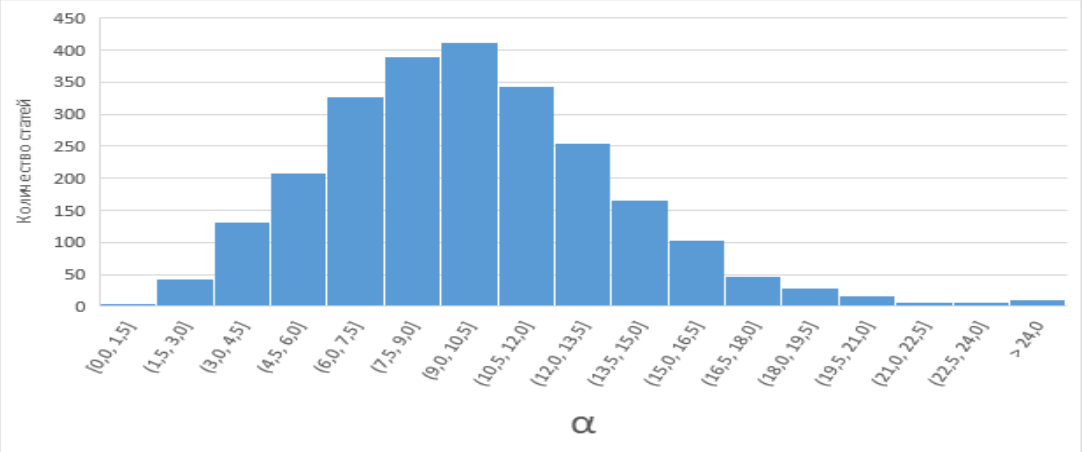

Рисунок 1 – Гистограмма распределения значений уровня ключевых слов в тексте статей из выборки

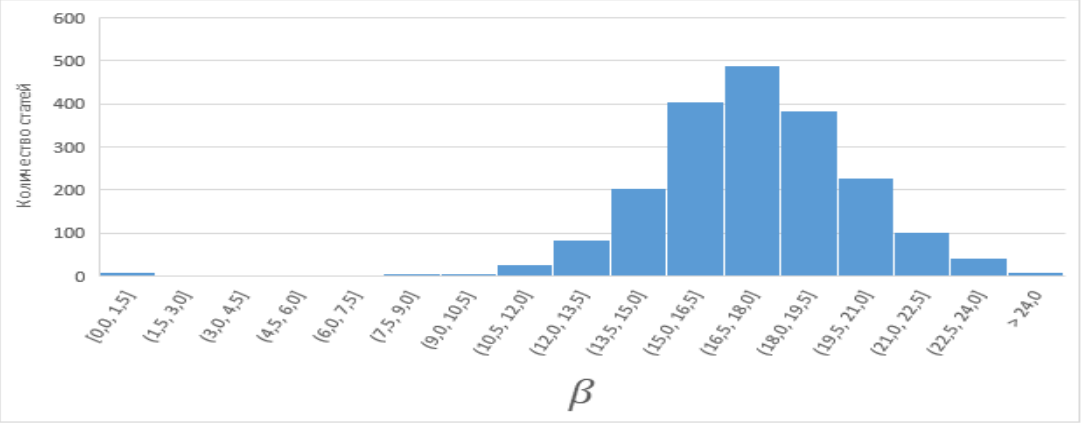

Рисунок 2 – Гистограмма распределения значений уровня водности текста статей из выборки

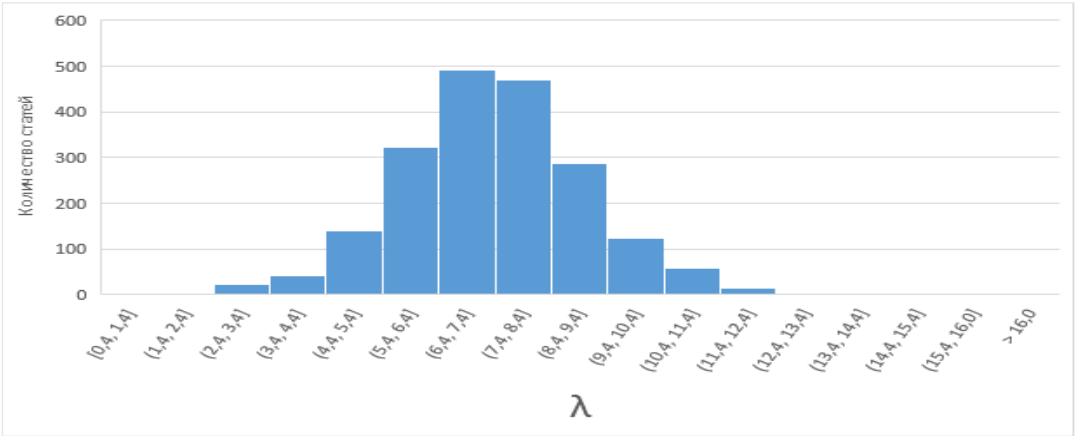

Рисунок 3 – Гистограмма распределения значений отклонения от идеальной кривой по Ципфу текста статей из выборки

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]

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.

2.1.3. **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:

- 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 персональные данные ничего не стоят»	10.56	Да	32.10	Нет	6.84	Да
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	11.61	Да	27.91	Нет	9.27	Да

движок на формулах Excel для чайников»						
«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.

2.2. Результаты исследования

В результате исследования были сформулированы три числовых критерия проверки статьи на соответствие научному стилю, были установлены пороговые значения данных критериев, позволяющие оценивать качество статей.

3. ОПИСАНИЕ РЕШЕНИЯ

3.1. Требования к решению и выбор метода решения

Были сформированы следующие требования к решению:

- Выполнение проверки на соответствие научному стилю и поиск наиболее частых ошибок;
- Простота использования решения;
- Возможность давать на вход решению файл;
- Наглядное представление результатов;
- Возможность контейнеризации решения для быстрого развертывания в любой среде.

Было принято решение реализовать веб-сервис, так как такой вид решения позволяет создать простой пользовательский интерфейс и наглядно предоставлять результаты.

3.2. 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 Blazor, в связи с тем, что это самая актуальная разработка Microsoft, активно поддерживаемая сообществом.

3.3. Архитектура решения

TO DO

3.4. Сценарии использования

3.5. Описание алгоритмов работы

ЗАКЛЮЧЕНИЕ

По итогам работы были получены следующие результаты:

Поставленные задачи были решены, цель работы была достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ