Программная реализация сравнения авторства текстов

Спецификация программных требований  
к программному средству

Версия 1.0

Журнал изменений

|  |  |  |  |
| --- | --- | --- | --- |
| **Дата** | **Версия** | **Описание** | **Автор** |
| 03.12.2024 | 1.0 | Создание спецификации программных требований | Михайлов Александр Сергеевич |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

Спецификация программных требований

# Введение

Данный документ представляет собой спецификацию программных требований (SRS) для разработки программного средства, предназначенного для определения авторства текстов. Программное средство будет анализировать тексты на основе различных стилистических и лексических признаков, таких как частота использования слов, синтаксические конструкции, пунктуация и т.д., для выявления сходства с известными авторами или текстами.

## Назначение

Цель данного документа — предоставить полное описание требований к функциональности и характеристикам программного средства определения авторства текстов. Этот документ охватывает все аспекты работы программного средства, включая:

1. Внешнее поведение: интерфейс пользователя, способы ввода и вывода данных, форматы поддерживаемых текстов.
2. Алгоритмы анализа: методы выявления стилистических и лексических признаков, алгоритмы сравнения текстов, способы оценки достоверности результатов.
3. База данных: структура хранения информации о известных авторах и текстах, методы добавления и обновления данных.
4. Нефункциональные требования: производительность, безопасность, масштабируемость, удобство использования.

Спецификация также описывает проектные ограничения и требования к интерфейсу, которые позволяют управлять базой данных, настраивать параметры анализа и интерпретировать результаты.

## Область применения

Данный документ описывает программное средство, которое предназначено для:

Определения авторства текстов: выявления вероятного автора неизвестного текста на основе анализа его стилистических и лексических особенностей.

Анализ плагиата: выявления сходства между текстами для выявления возможного плагиата.

Исследование стиля: изучения стилистических особенностей авторов и текстов для научных и образовательных целей.

Программное средство подключено к базе данных, которая хранит информацию о известных авторах и их текстах. База данных используется для сравнения анализируемых текстов с эталонными образцами.

## Определения, акронимы и сокращения

1. TF-IDF (Term Frequency-Inverse Document Frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, содержащегося в коллекции документов.
2. N-граммы — последовательности из N элементов (например, слов или символов) в тексте.
3. Стилистические признаки — характеристики текста, связанные с его формой, такие как частота использования слов, синтаксические конструкции, пунктуация и т.д.
4. Лексические признаки — характеристики текста, связанные с его содержанием, такие как используемые слова, их частота и контекст.
5. Эталонный образец — текст или набор текстов, используемых в качестве базы для сравнения при определении авторства.
6. Достоверность результата — вероятность того, что результат определения авторства является правильным.

## Обзор

Документ состоит из нескольких частей. В первой главе приводится общая информация о проекте и его области применения. Во второй главе будут представлены функциональные требования, включая описание интерфейса пользователя, способов ввода и вывода данных, а также основных функций программного средства. В третьей главе будет обсуждаться взаимодействие с базой данных и алгоритмы, используемые для анализа текстов. Четвёртая глава будет посвящена нефункциональным требованиям, таким как безопасность, производительность и удобство использования. Пятая глава будет содержать проектные ограничения.

# Общее описание

В данном разделе приводится общее описание программного средства для определения авторства текстов, описываются его основные функции, ограничения, а также предположения и зависимости, которые будут влиять на дальнейшую разработку и эксплуатацию продукта.

## Перспектива продукта

Программное средство для определения авторства текстов предназначено для автоматического анализа текстов на основе различных стилистических и лексических признаков с целью выявления сходства с известными авторами или текстами. Программное средство может использоваться для:

1. Определения авторства неизвестных текстов: выявления вероятного автора неизвестного текста на основе анализа его стилистических и лексических особенностей.
2. Анализ плагиата: выявления сходства между текстами для выявления возможного плагиата.
3. Исследование стиля: изучения стилистических особенностей авторов и текстов для научных и образовательных целей.

## Функции продукта

Программное средство предоставляет следующие основные функции:

1. Анализ текста: программное средство анализирует текст на основе различных стилистических и лексических признаков, таких как частота использования слов, синтаксические конструкции, пунктуация и т.д.
2. Сравнение с эталонными образцами: программное средство сравнивает анализируемый текст с эталонными образцами, хранящимися в базе данных, для выявления сходства.
3. Оценка достоверности результата: программное средство оценивает вероятность того, что результат определения авторства является правильным.
4. Управление базой данных: программное средство предоставляет инструменты для управления базой данных эталонных образцов, включая добавление, удаление и обновление данных.
5. Настройка параметров анализа: программное средство позволяет настраивать параметры анализа текста, такие как используемые стилистические и лексические признаки, пороговые значения для оценки достоверности результата и т.д.

## Пользовательские характеристики

Основными пользователями программного средства для определения авторства текстов будут:

1. Исследователи: научные сотрудники, занимающиеся изучением стилистических особенностей авторов и текстов, а также анализом плагиата.
2. Авторы: писатели, журналисты и другие авторы, которые хотят проверить оригинальность своих текстов.
3. Учителя и преподаватели: педагоги, использующие программное средство для проверки работ студентов на плагиат.
4. Администраторы баз данных: специалисты, ответственные за управление базой данных эталонных образцов.

Программное средство будет предоставлять разные уровни доступа и функциональности в зависимости от роли пользователя:

1. Исследователи и авторы: будут иметь доступ к базовому функционалу программного средства, включая загрузку текстов для анализа, настройку параметров анализа и просмотр результатов.
2. Учителя и преподаватели: помимо базового функционала, получат возможность создавать и управлять собственными базами данных эталонных образцов для проверки работ студентов.
3. Администраторы баз данных: будут иметь полный доступ к управлению базой данных, включая добавление, удаление и обновление эталонных образцов.

## Ограничения

1. Зависимость от API: Программное средство может зависеть от внешних API для доступа к базам данных эталонных образцов или для выполнения определенных функций анализа текста. Изменения в API или ограничения по запросам могут повлиять на работу программного средства.
2. Производительность: Производительность программного средства может быть ограничена объемом обрабатываемых данных и сложностью алгоритмов анализа текста. Для обработки больших объемов данных или сложных текстов может потребоваться высокопроизводительное оборудование.
3. Безопасность: Система анализа текста предполагает работу с конфиденциальными данными, такими как тексты, написанные известными авторами. Необходимо обеспечить высокий уровень безопасности при взаимодействии с базами данных, чтобы предотвратить несанкционированный доступ к конфиденциальной информации.
4. Масштабируемость: Программное средство должно быть спроектировано так, чтобы его можно было легко адаптировать под различные объемы данных и требования пользователей.

## Предположения и зависимости

1. Наличие подключения к интернету: Для функционирования программного средства требуется стабильное интернет-соединение для доступа к базам данных эталонных образцов и другим внешним ресурсам.
2. Существование баз данных: Предполагается, что для нормальной работы программного средства будет настроена база данных эталонных образцов, содержащая тексты, написанные известными авторами.
3. Права доступа: Для корректной работы программного средства потребуется предоставление соответствующих прав доступа к базам данных и другим ресурсам.

## Подмножества требований

Некоторые требования могут быть реализованы в нескольких этапах. Например, начальная версия программного средства будет реализовывать только основные функции анализа текста и сравнения с эталонными образцами, в то время как более поздние версии могут добавить поддержку дополнительных алгоритмов анализа, улучшение интерфейса пользователя и расширенные возможности управления базой данных.

# Конкретные требования

В этом разделе подробно описаны требования к системе, которые должны быть выполнены для успешной разработки и эксплуатации программного средства для определения авторства текстов.

## Функциональность

### **Модерация контента**

Cистема должна сканировать тексты пользователей в реальном времени.

Стилистические и лексические признаки должны извлекаться из текстов.

Система должна сохранять извлеченные признаки в базе данных.

Извлеченные признаки должны использоваться для сравнения с эталонными образцами.

### **Управление списком запрещённых слов**

Программное средство должно позволять добавлять и удалять эталонные образцы через интерфейс пользователя.

База данных должна быть доступна для просмотра и редактирования администраторам.

### **Управление ролями**

Администраторы должны иметь возможность назначать роли, которые могут выполнять команды программного средства.

Программное средство должно хранить список разрешенных ролей в базе данных.

### **Информация о сервере**

Программное средство должно предоставлять информацию о системе по запросу (например, количество обработанных текстов, доступные эталонные образцы и т.д.)

## Практичность

Этот раздел описывает требования к практичности системы, включая время обучения пользователей, скорость выполнения основных операций, соответствие общепринятым стандартам и интуитивность интерфейса. Требования направлены на обеспечение удобства использования как для конечных пользователей, так и для администраторов системы.

### **Время обучения**

Обычные пользователи смогут освоить основные функции программного средства за 15 минут.

Администраторы и исследователи, управляющие настройками и базами данных, смогут достичь высокой производительности при выполнении операций после 30 минут обучения.

### **Время выполнения операций**

Время обработки команды, введённой пользователем, должно составлять не более **1 секунды**. Время выполнения операций с базами данных (добавление/удаление запрещённых слов, управление ролями) не должно превышать **2 секунд**.

### **Интуитивность интерфейса**

Команды программного средства должны быть интуитивно понятны и следовать общепринятым стандартам пользовательского интерфейса.

## Надежность

Требования к надежности системы направлены на обеспечение ее стабильной и бесперебойной работы в различных эксплуатационных условиях. Надежность системы гарантирует минимизацию простоев, точность обработки данных и устойчивость к сбоям.

### **Доступность**

Система должна быть доступна не менее 99.9% времени, что соответствует допустимому простою не более 8.76 часов в год. Режим работы системы предусматривает круглосуточное использование с возможностью планового технического обслуживания не чаще одного раза в месяц. В случае пониженной производительности система должна обеспечивать минимальную функциональность для выполнения критически важных задач.

### **Точность**

Точность обработки данных должна соответствовать стандартам, установленным для реализации алгоритмов анализа текста.

Любые ошибки, возникающие в процессе работы системы, должны быть минимальными и не приводить к потере критической информации.

## Производительность

Данный раздел описывает основные показатели производительности системы, которые позволяют оценить ее эффективность при выполнении задач. Требования к производительности основываются на обеспечении быстрого времени реакции, поддержании высокой пропускной способности и оптимальном использовании ресурсов.

### **Время отклика**

Система должна обеспечивать среднее время реакции на команды не более 0.5 секунд и максимальное время реакции — не более 2 секунд.

Это включает обработку ввода, выполнение алгоритмов анализа текста, а также отправку результатов.

## Проектные ограничения

В данном разделе перечисляются и описываются все ограничения, которые необходимо учитывать при разработке системы. Эти ограничения определяют выбор инструментов, технологий, архитектуры и других аспектов разработки.

### **Языки программирования**

Основным языком разработки системы является **C#**, что обусловлено требованиями платформы .NET Framework/ASP.NET Core для обеспечения высокой производительности и интеграции с другими технологиями Microsoft.

### **Средства разработки**

Для разработки используется **Microsoft Visual Studio** версии 2022 и выше. Это решение основано на необходимости поддержки полного набора инструментов для разработки, отладки и тестирования в единой среде.

## Интерфейсы

### **Пользовательские интерфейсы**

Пользовательским интерфейсом является веб-интерфейс, который обеспечивает удобство использования и доступность с различных устройств.

Интерфейс должен быть интуитивно понятным и предоставлять все необходимые функции для работы с программным средством.