**Приложение 4**

**Теоретические положения**

**Атрибуция текста** – исследование текста с целью установления авторства или получения каких-либо сведений об авторе и условиях создания текстового документа. Задачи атрибуции можно разделить на идентификационные и диагностические

Методы атрибуции позволяют исследовать текст на пяти уровнях: пунктуационном, орфографическом, синтаксическом, лексико-фразеологическом, стилистическом.

1. **Пунктуационный** уровень помогает выявить особенности употребления автором знаков препинания, характерные ошибки.
2. Орфографический уровень выявляет характерные ошибки в написании слов.
3. Синтаксический уровень позволяет определить особенности построения предложений, предпочтение тех или иных языковых конструкций, употребление времен, активного или пассивного залога, порядок слов, характерные синтаксические ошибки.
4. **Лексико-фразеологический** уровень определяет словарный запас автора, особенности использования слов и выражений, склонность к употреблению редких и иностранных слов, диалектизмов, архаизмов, неологизмов, профессионализмов, арготизмов, навыки употребления фразеологизмов, пословиц, поговорок, «крылатых выражений» и т. д.
5. Стилистический уровень позволяет определить жанр, общую структуру текста, для литературных произведений – сюжет, характерные изобразительные средства (метафора, ирония, аллегория, гипербола, сравнение), стилистические фигуры (градация, антитеза, риторический вопрос и т. д.), другие характерные речевые приемы.

Под «авторским стилем» обычно понимаются последние три уровня. Анализ именно синтаксического, лексико-фразеологического и стилистического уровней представляет наибольший интерес и наибольшую сложностьA diagram of a company

Description automatically generatedВ данном программном комплексе будут рассматриваться пунктуационный и лексико-фразеологический уровни.

Одной из последних разработок в сфере атрибуции текста является методика, предложенная *О. Хрулевым*, определяющая автора текста на основе Евклидова расстояния. Суть метода заключается в следующем:

1. В распоряжении исследователя имеется текст неизвестного автора, и принимается гипотеза о некотором списке возможных претендентов на роль автора данного текста
2. На основе анализа всех известных произведений (содержащихся в базе данных) создается частотный словарь автора – словарь, содержащий лексемы, употребляемые в анализируемых текстах, с частотами употребления и частотный словарь текста, авторство которого необходимо определить;

Определяется расстояние между словарями авторов и словарем анализируемого

текста как

A black text on a white background

Description automatically generated

где xi – частота употребления i-ой лексемы в тексте, автор которого неизвестен , ai – частота употребления i-ой лексемы в частотном словаре предполагаемого автора, N – количество анализируемых лексем.

Учитываются только те лексемы, которые присутствуют в обоих словарях.

Рассматриваемые лексемы:

1. Основа слова
2. Пунктуационные и специальные знаки
3. Нечастые суффиксы, иностранные приставки

В качестве автора анализируемого текста выбирается тот автор, чей словарь находится на меньшем расстоянии до словаря анализируемого текста.

**Программная реализация:**

Программный комплекс реализован на языке javascript в среде компилирования Node.js и имеет следующую структуру:

project/

* authors/ - директория с текстами авторов
  + Гоголь.txt
  + Пушкин.txt
* input.txt – анализируемый текст
* index.js – анализатор

Анализатор имеет следующие функции:

* createAuthor(name, input) – создает частотный словарь основ слов автора на основе входного текста. Используется метод Стемминга для выделения основы слова
* stemText(input) - создает частотный словарь основ слов анализируемого текста. Используется метод Стемминга для выделения основы слова
* suffixAnalyze(input) – создает частотный словарь суффиксов для входного текста
* specSymbols(input) – создает частотный словарь пунктуационных и специальных символов во входном тексте

При запуске программного комплекса происходит следующее:

Программа считывает анализируемый текст, создает его частотные словари, считывает все текста авторов в папке authors и создает частотные словари рассматриваемых лексем для каждого автора. Перебирая каждую лексему в словарях анализируемого текста, если она встречается в обоих текстах, сравнивает евклидово расстояние часты использования данной лексемы и суммирует полученные расстояния. После чего выводится таблица полученных сумм евклидовых расстояний и в качестве автора выбирается автор с наименьшей суммой расстояний.