**Приложение**

**5.1 Теоретические положения**

**5.1.1 Атрибуция текста, задачи и методы**

**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 сложность. В данном программном комплексе будут рассматриваться пунктуационный и лексико-фразеологический уровни.

**5.1.2 Описание алгоритма. Сравнение мер расстояния.**

СОЗДАНИЕ ЧАСТОТНЫХ СЛОВАРЕЙ

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. Частоты нормируются по количеству слов в тексте.

ОПРЕДЕЛЕНИЯ РАССТОЯНИЯ МЕЖДУ ЧАСТОТНЫМИ СЛОВАРЯМИ(ВЕКТОРАМИ)

Геометрические меры расстояния:

1. **Евклидово расстояние**. Наиболее распространенная функция расстояния. Представляет собой геометрическим расстоянием в многомерном пространстве:  
   Изображение выглядит как текст, Шрифт, линия, рукописный текст

   Автоматически созданное описание (1)
2. **Квадрат евклидова расстояния**. Применяется для придания большего веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом:  
   Изображение выглядит как текст, Шрифт, рукописный текст, белый

   Автоматически созданное описание (2)
3. **Расстояние городских кварталов (манхэттенское расстояние)**. Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (т.к. они не возводятся в квадрат). Формула для расчета манхэттенского расстояния:  
   Изображение выглядит как текст, Шрифт, типография, белый

   Автоматически созданное описание(3)

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

Учитываются только те лексемы, которые присутствуют в обоих словарях.

Требования к мере расстояния:

* Мера должна являться метрикой (в математическом смысле)
  + Тождество: d(A, A) = 0 - Словарь находится от на расстоянии 0 от самого себя.
  + Симметрия: d(A, B) = d(B, A) - Расстояние от A до B равно расстоянию от B до A.
  + Неравенство треугольника: d(A, C) ≤ d(A, B) + d(B, C)
* Ограниченность:
* мера должна иметь фиксированный диапазон значений (тогда расстояние легко преобразуется в близость и наоборот).
* Устойчивость к окружению: d(A, B) не должно меняться в зависимости от того, какие корпуса есть в общем наборе для исследования.
* Соответствие реальности

Тестирование (20 прогонов × 45 пар словарей)

• Манхэттен 429 побед 48%

• Евклид 415 побед 46%

**Замечание**: В зависимости от анализируемых лексем будем использовать разные меры расстояний. Эксперимент показал, что при вычисления расстояния между частотными словарями суффиксов евклидово расстояние дает самый точный результат. При рассмотрении остальных лексем самым точным оказалось манхэттенское расстояние.

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.

**5.2 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:

1. Создание частотных словарей лексем введенного текста.
2. Создание частотных словарей рассматриваемых лексем для каждого автора, находящегося в базе данных.
3. Перебирая каждую лексему в словарях анализируемого текста, если она встречается в обоих текстах, сравнивается евклидово расстояние частоты использования данной лексемы, после чего полученные расстояния суммируются.
4. Вывод таблицы полученных сумм евклидовых расстояний и выбор автора с наименьшей суммой расстояний.

**5.3** **Инструкция по запуску программы**



* 1. **Статистический анализ частот слов на примере работы программного комплекса**
     1. **Построение вариационного ряда**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| слово | сист | Библиотек | исполня | функц | файл | программ | Процесс | unix | вызов |
| Ломако | 0,059 | 0 | 0,019 | 0,01 | 0 | 0,32 | 0,07 | 0 | 0 |
| Каливанов | 0,017 | 0 | 0 | 0,12 | 0,17 | 0 | 0,05 | 0 | 0 |
| Киселев | 0,106 | 0,24 | 0,106 | 0,03 | 2,97 | 0,31 | 0,14 | 1,06 | 0,1 |
| Исходный текст | 0,16 | 0,57 | 0,77 | 1,95 | 1,99 | 1,63 | 0,73 | 0,73 | 0,81 |

**Самые часто используемые слова – МатОжидание**

Ломако – язык (3,32%)

Каливанов – n (0,74%)

Киселев – файл (2,97%)

Исходный текст – файл (1,99%)

**Расстояние между словами - Дисперсия:**

**5.4.2 Построение графика**

