Автоматическое извлеченье ключевых слов и словосочетаний из электронных документов на русском языке

Студент: Барсуков Никита Михайлович

Руководитель: Барышников Марина Юрьевна

Москва 2022

Цель работы

Разработка программного обеспечения для извлечения ключевых слов (КС) и словосочетаний из электронного документа на русском языке.

Задачи:

1. Анализ существующих методов извлечения КС
2. Отбор и изучение выбранного алгоритма
3. Разработка архитектуры решения
4. Выбор инструментов
5. Реализация программного обеспечения
6. Проведение экспериментов

Классификация методов

Метод	По обучению	Лингвистические ресурсы	Матаппарат распозонования
Yake	Не требует обучения	Не использует	Гибридный
Rake	Не требует обучения	Не использует	Структурный / Графовый
Kea	Не требует обучения	На основе корпусов	Нейросетевой
TF-IDF	Не требует обучения	На основе корпусов	Статистический

Критерии

Справка: Подходит , не подходит по критерию

Метод	Не требует наличия корпусов текстов	Умеет извлекать многокомпонентные КС	Не привязан к предметной области применения
Yake	+	-	+
Rake	+	+	+
Kea	-	+	-
TF-IDF	-	-	+

Выбор алгоритма

Для реализации ПО был выбран метод **«Yake»**

Учитывает:

1. Расположение кандидата в документе;
2. Связь термина с контекстом;
3. Его форму написания;

До этого не использовался для извлечения КС из документов на русском языке

N-граммы

N-граммой на алфавите V называют произвольную цепочку длинной N.
На пример последовательность из слов или словосочетаний

Исходный текст: Автоматическое извлечение ключевых слов
Примеры н-грамм:

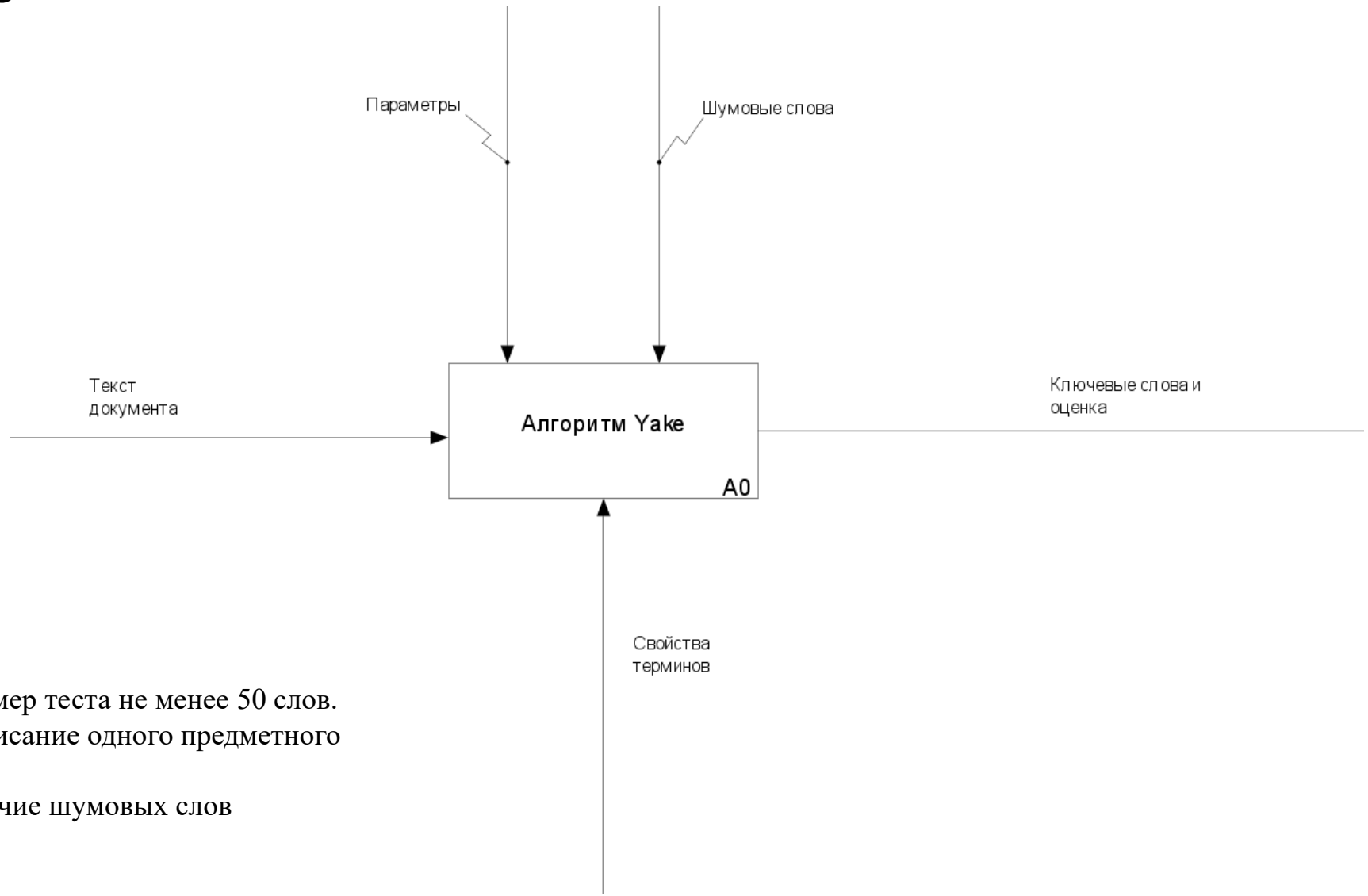
- Униграмма:
  - Автоматическое, извлечение, ключевых, слов;

- Биграмма:
  - Автоматическое извлечение, извлечение ключевых, ключевых слов;

- Триграммы:
  - Автоматическое извлечение ключевых, извлечение ключевых слов;

- N – грамма (н = 4)
  - Автоматическое извлечение ключевых слов.

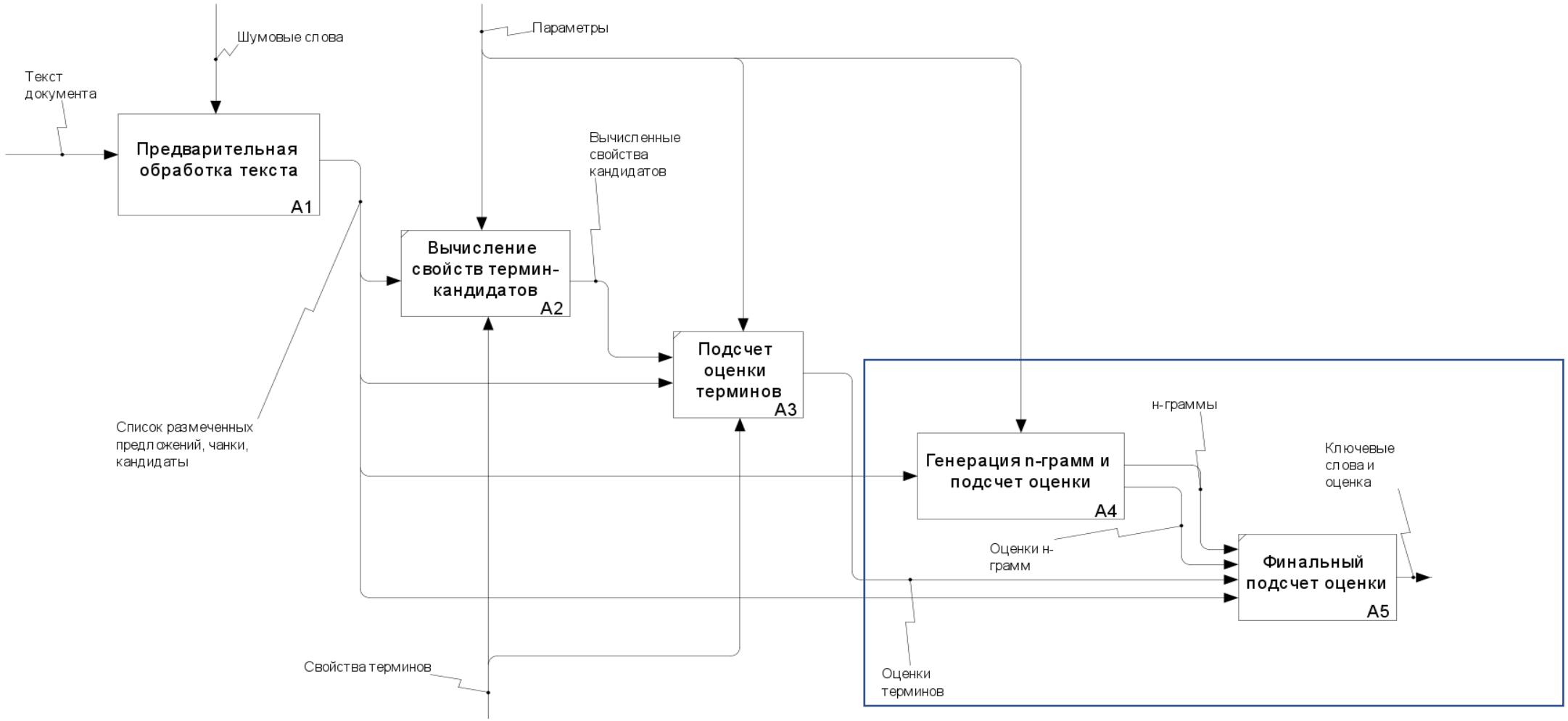
Yake



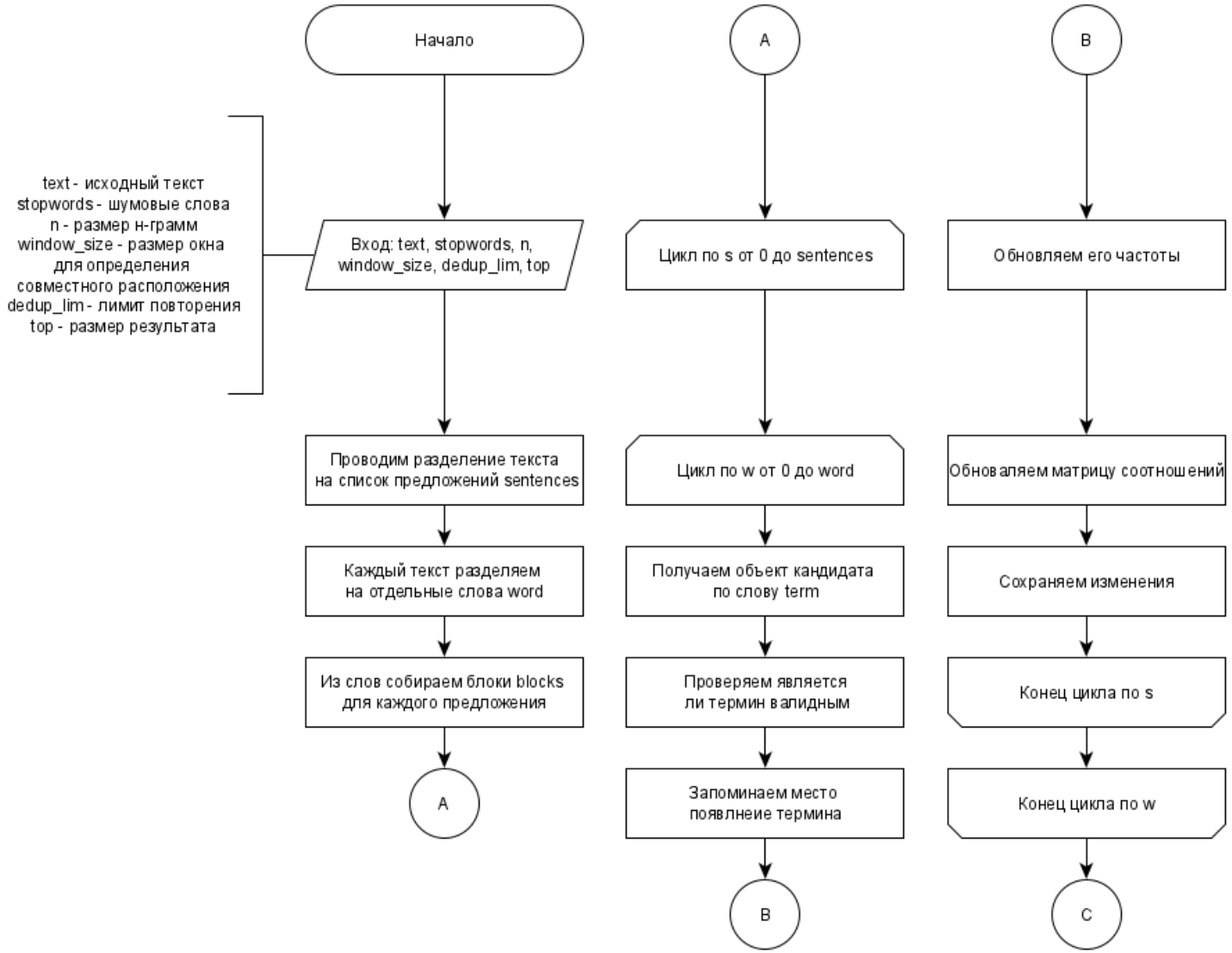
Ограничения:

- Минимальный размер теста не менее 50 слов.
- Текст содержит описание одного предметного объекта
- Обязательное наличие шумовых слов

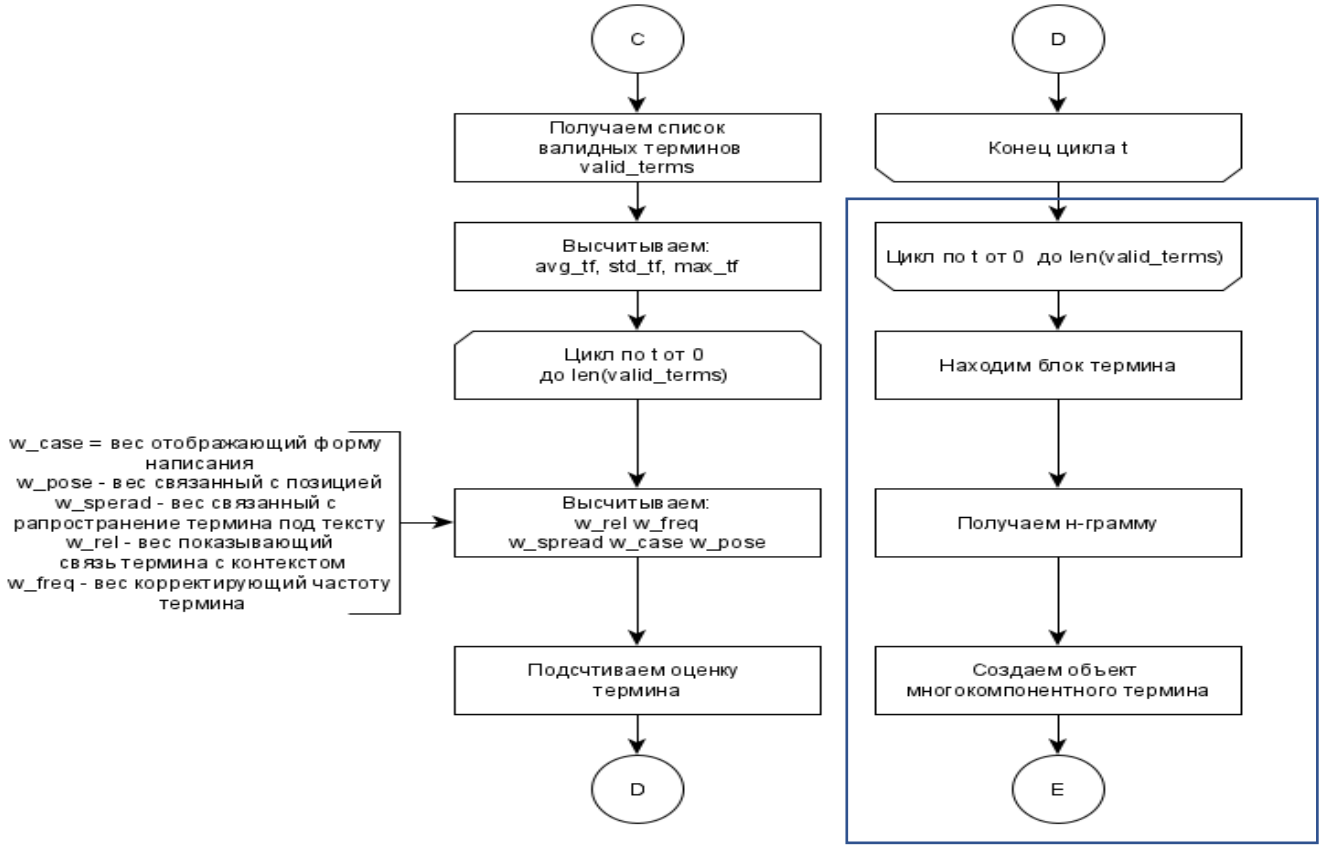
Метод извлечения КС



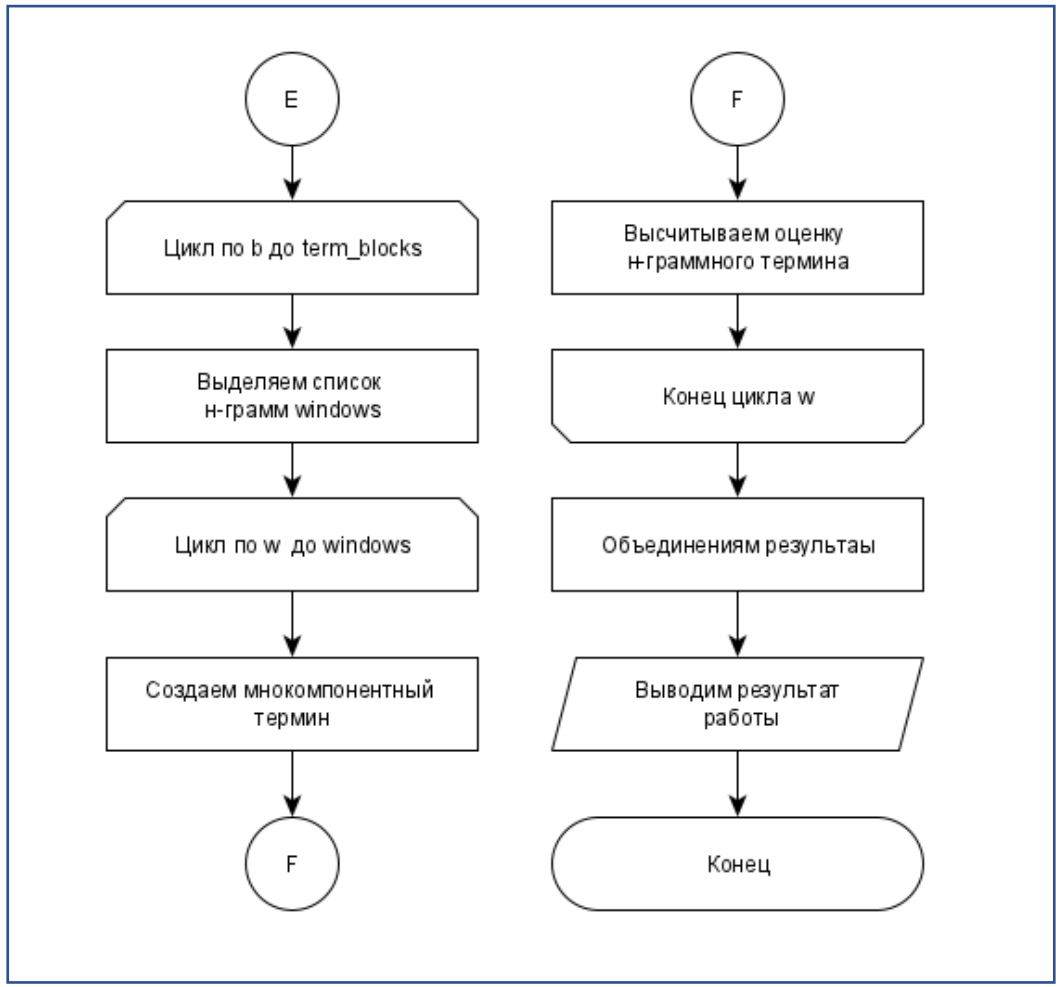
Предварительная обработка текста



Подсчет оценки термина



Вычисление н-грамм

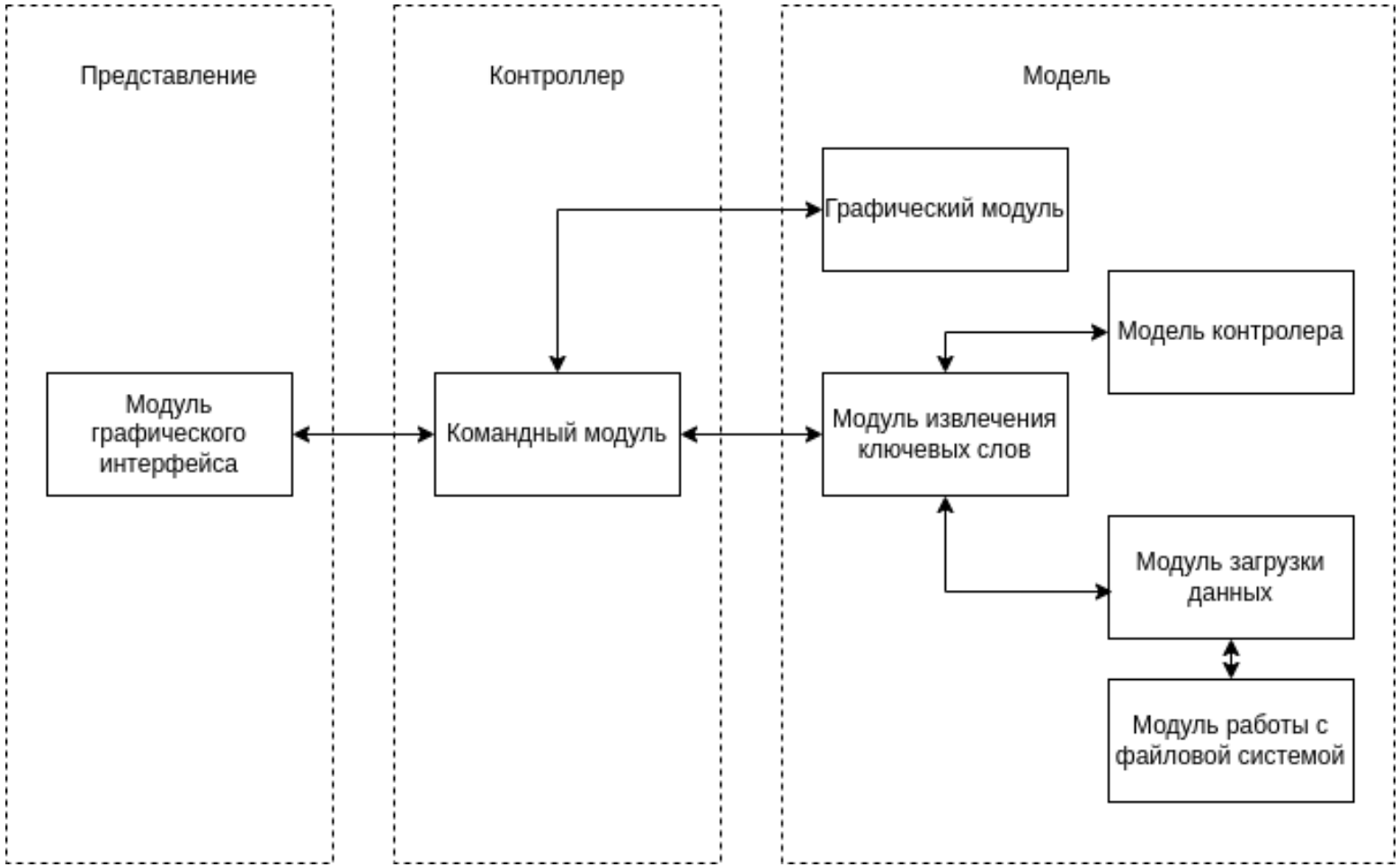


Архитектура ПО

При проектировании использовался шаблон MVC – Модель-представление-контролер

- Модель - это компонента отвечающая за предоставление данных конкретным элементам системы
- представление - это отображение состояния внутренний системы
- Контроллер - это связующее звено между представлением и моделью, обрабатывает действия пользователя, полученные от представления и отдает команды модели.

Схема архитектуры



Исследование

- Выборка:
  - 30 электронных документов

- Критерии оценки
  - Процент пересечения (1)
  - Средний процент пересечения (2)
  - Минимальный процент пересечения (3)
  - Максимальный процент пересечения (4)

- Условия:
  - Текс документа содержит в себе только одну тему
  - Документ написан на русском языке
  - Документы формата PDF
  - Должен содержать не менее 50 слов

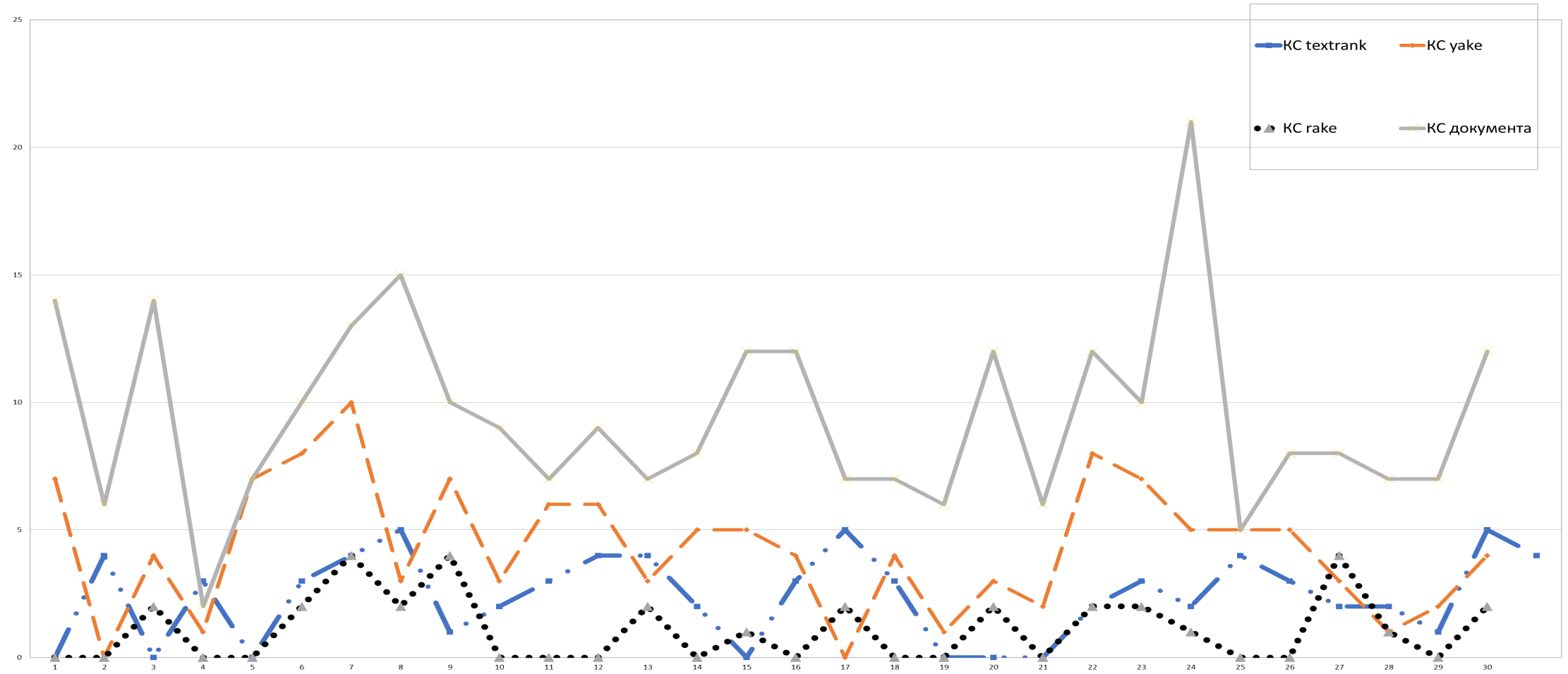
$$R = \frac{N_{doc}}{N_{cross}} \quad (1)$$

$$R_{mid} = \frac{\sum_0^N R}{N} \quad (2)$$

$$R_{min} = \min_{n \to N}(R_n) \quad (3)$$

$$R_{max} = \max_{n \to N}(R_n) \quad (4)$$

Сравнение с другими алгоритмами



Результаты сравнения

Метрики	Yake(mod)	Textrank	Rake
Максимальный % пересечения	100%	71%	50%
Средний % пересечения	42%	25%	2%
Минимальный % пересечения	0%	0%	0%

Исследование н-грамм

Документ: Идентификация личности по фрактальной размерности отпечатков пальцев и системы контроля и управления доступом

Ссылка на документ: https://cyberleninka.ru/article/n/identifikatsiya-lichnosti-po-fraktalnoy-razmernosti-otpechatkov-paltsev-i-sistemy-kontrolya-i-upravleniya-dostupom.pdf

Ключевые слова: биометрия, отпечаток пальца, фрактал, фрактальная размерность, идентификация и аутентификация личности, СКУД.

Результат работы алгоритма от N

```
Документ работы: identifikatsiya-lichnosti-po-fraktalnoy-razmernosti-otpechatkov-paltsev-i-sistemy
  -kontrolya-i-upravleniya-dostupom.pdf
Метод: yakemodified
размерности, личности, пальцев, Dcp, фрактальной, отпечатков, идентификации, значение, пользователь,
  системы, распознавания, For, СКУД, среднее, биометрические, log, Доклады, число, Lmax, часть
```

```
Документ работы: identifikatsiya-lichnosti-po-fraktalnoy-razmernosti-otpechatkov-paltsev-i-sistemy
  -kontrolya-i-upravleniya-dostupom.pdf
Метод: yakemodified
фрактальной размерности, размерности, личности, отпечатков пальцев, размерности отпечатков, идентификации
  личности, пальцев, распознавания личности, Dcp, фрактальной, отпечатков, идентификации, значение,
  пользователь, системы, распознавания, For, СКУД, значение фрактальной, Доклады ТУСУРа
```

```
Документ работы: identifikatsiya-lichnosti-po-fraktalnoy-razmernosti-otpechatkov-paltsev-i-sistemy
  -kontrolya-i-upravleniya-dostupom.pdf
Метод: yakemodified
гребней и впадин, размерности отпечатков пальцев, фрактальной размерности отпечатков, размерности,
  личности, отпечатков пальцев, идентификации личности, пальцев, распознавания личности, Dcp, фрактальной,
  отпечатков, идентификации, значение фрактальной размерности, значение, пользователь, системы,
  распознавания, For, размерности Минковского
```

Дальнейшее развитее

- Добавить процесс преобразование терминов к начальной форме
- Добавить автоматическое определение языка

Заключение

В результате выполнения работы поставленная цель была достигнута, а также были решены следующие задачи:

1. Проведен анализ методов извлечения ключевых слов;
2. Отобран метод по выбранные критерями;
3. Проработана модификация;
4. Разработана архитектура ПО;
5. Выбраны инструменты реализации
6. Реализовано программное обеспечение
7. Проведено тестирование работы
8. Проведены исследования