АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ И СЛОВОСОЧЕТАНИЙ ИЗ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ НА РУССКОМ ЯЗЫКЕ

Студент: Барсуков Никита Михайлович
Руководитель: Барышникова Марина Юрьевна

Москва, 2022

Цель и задачи

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

Задачи:

1. Анализ существующих методов извлечения ключевых слов.
2. Выбор основного алгоритма и определение направлений его модификации.
3. Проектирование и разработка программного обеспечения для реализации метода.
4. Экспериментальное исследование характеристик разработанного метода.

Классификация методов извлечения КС

Метод	По обучению	Лингвистические ресурсы	Мат. аппарат распознавания
Yake	Не требует обучения	Не использует	Гибридный
Rake	Не требует обучения	Не использует	Структурный / Графовый
Kea	Не требует обучения	На основе корпусов текстов	Нейросетевой
TF-IDF	Не требует обучения	На основе корпусов текстов	Статистический

Сравнение алгоритмов распознавания ключевых слов

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

Выбор алгоритма

Для решения поставленной задачи был выбран метод **«Yake»**

Преимущества: данный алгоритм позволяет учесть

1. расположение термина-кандидата в документе;
2. связь термина с контекстом;
3. его форму написания.

Примечание: до этого не использовался для извлечения КС из документов на русском языке

N-граммы

N-граммой на алфавите V называют произвольную цепочку длинной N, например последовательность из слов или словосочетаний

Исходный текст: Автоматическое извлечение ключевых слов

Примеры N-грамм:

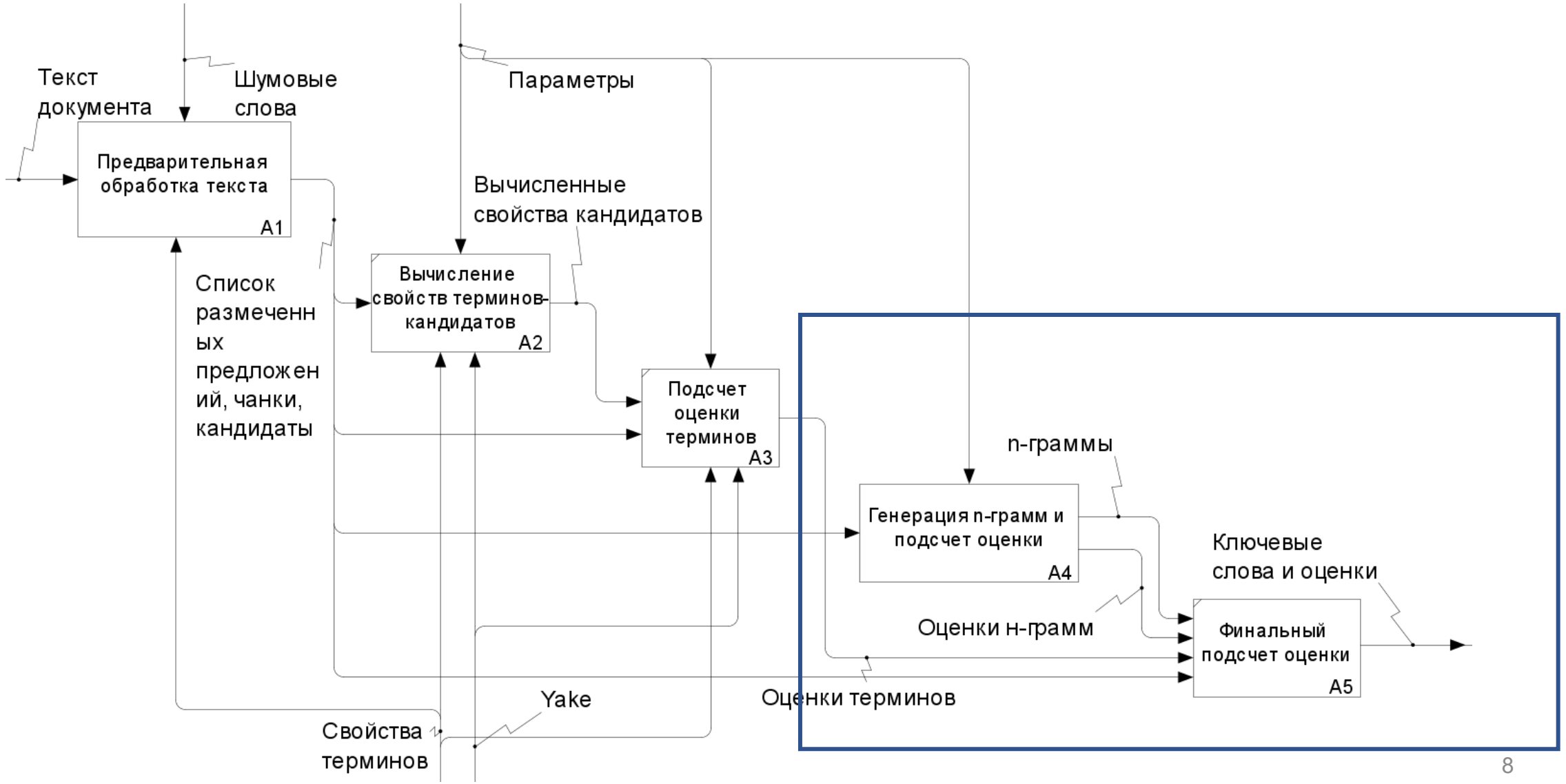
- Униграмма:
  Автоматическое, извлечение, ключевых, слов;
- Биграмма:
  Автоматическое извлечение, извлечение ключевых, ключевых слов;
- Триграммы:
  Автоматическое извлечение ключевых, извлечение ключевых слов;
- N – грамма (n = 4)
  Автоматическое извлечение ключевых слов.

Постановка задачи

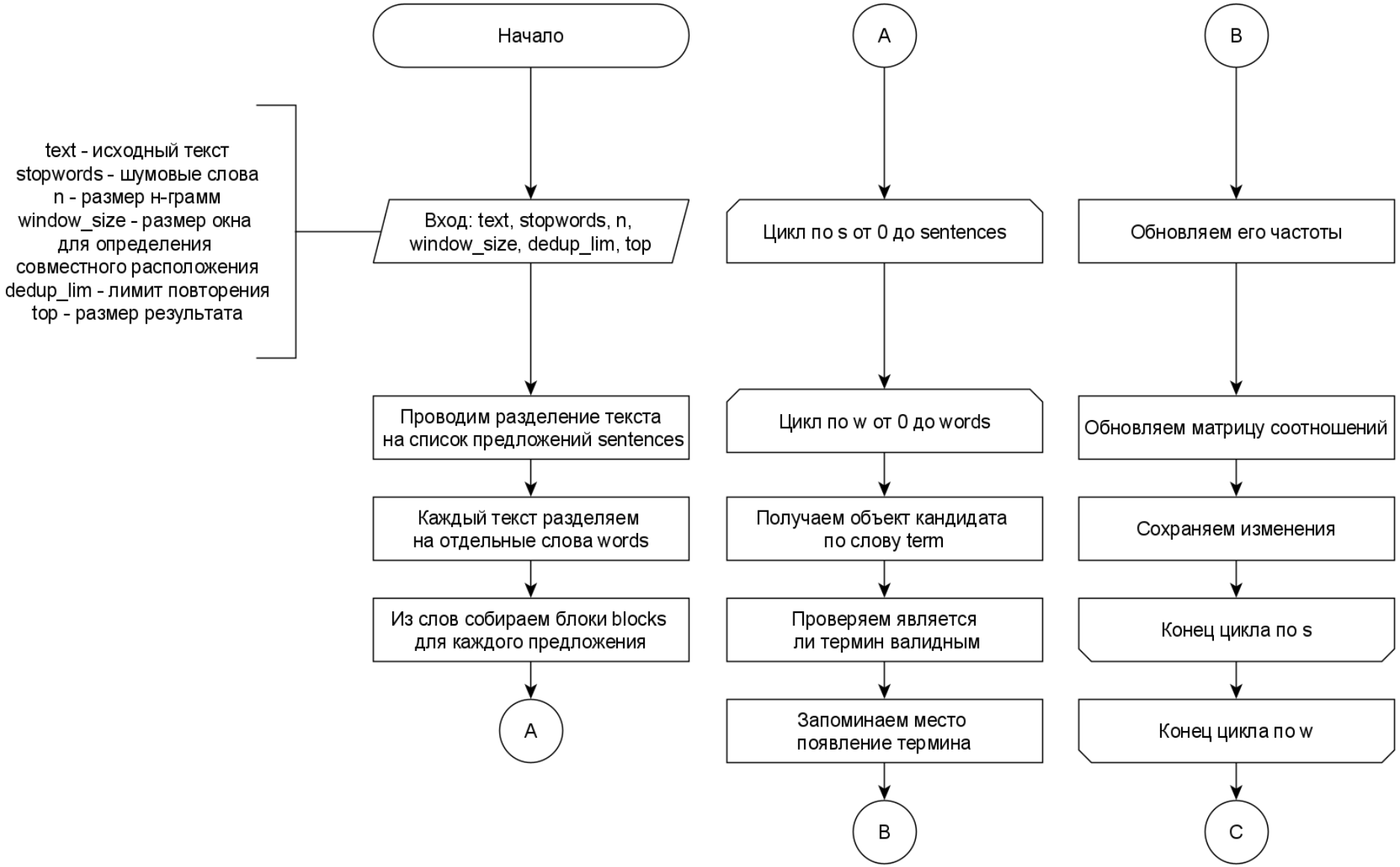
Ограничения:

- Минимальный размер теста не менее 50 слов.
- Текст принадлежит одному источнику (статье)

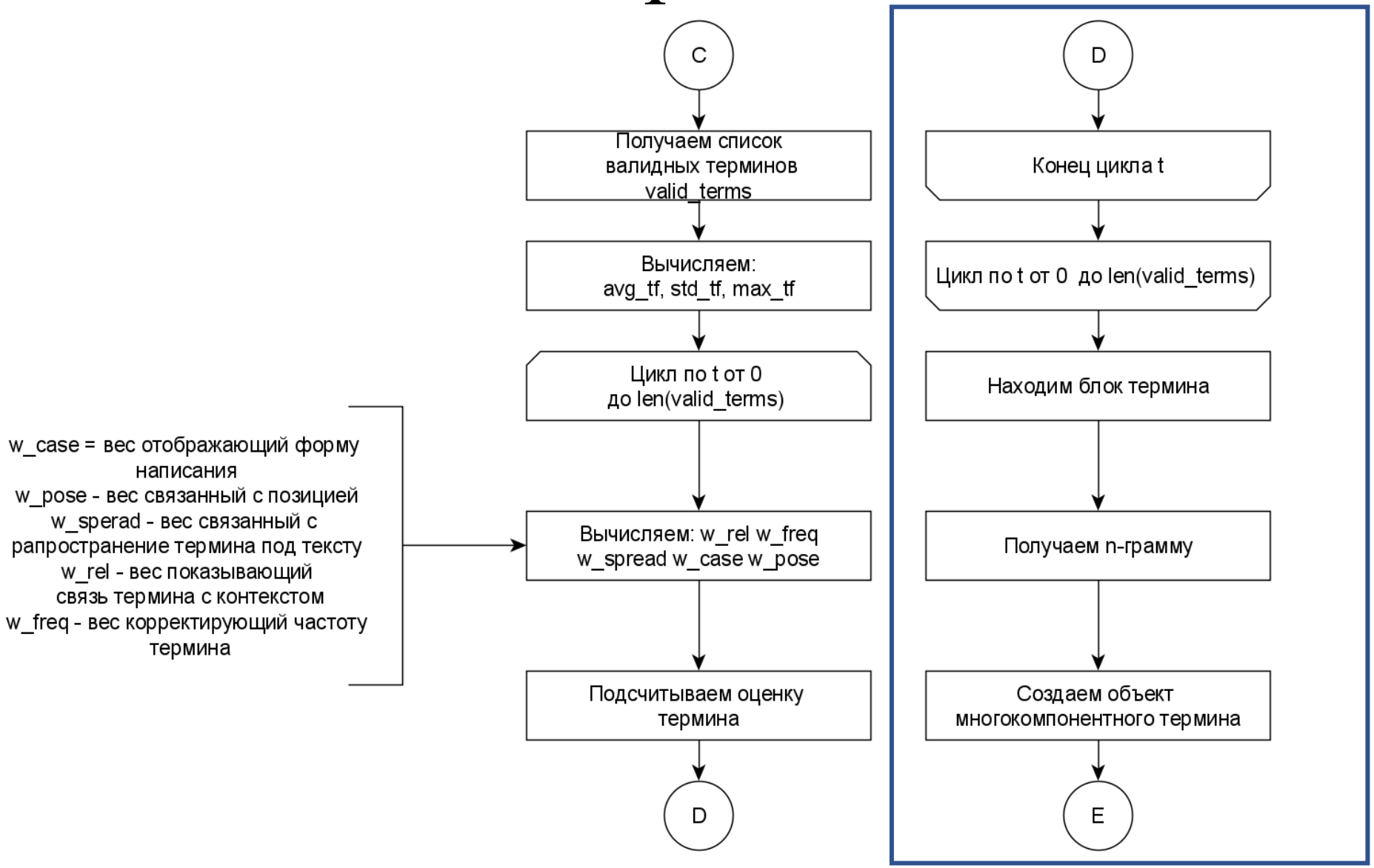
Метод извлечения КС



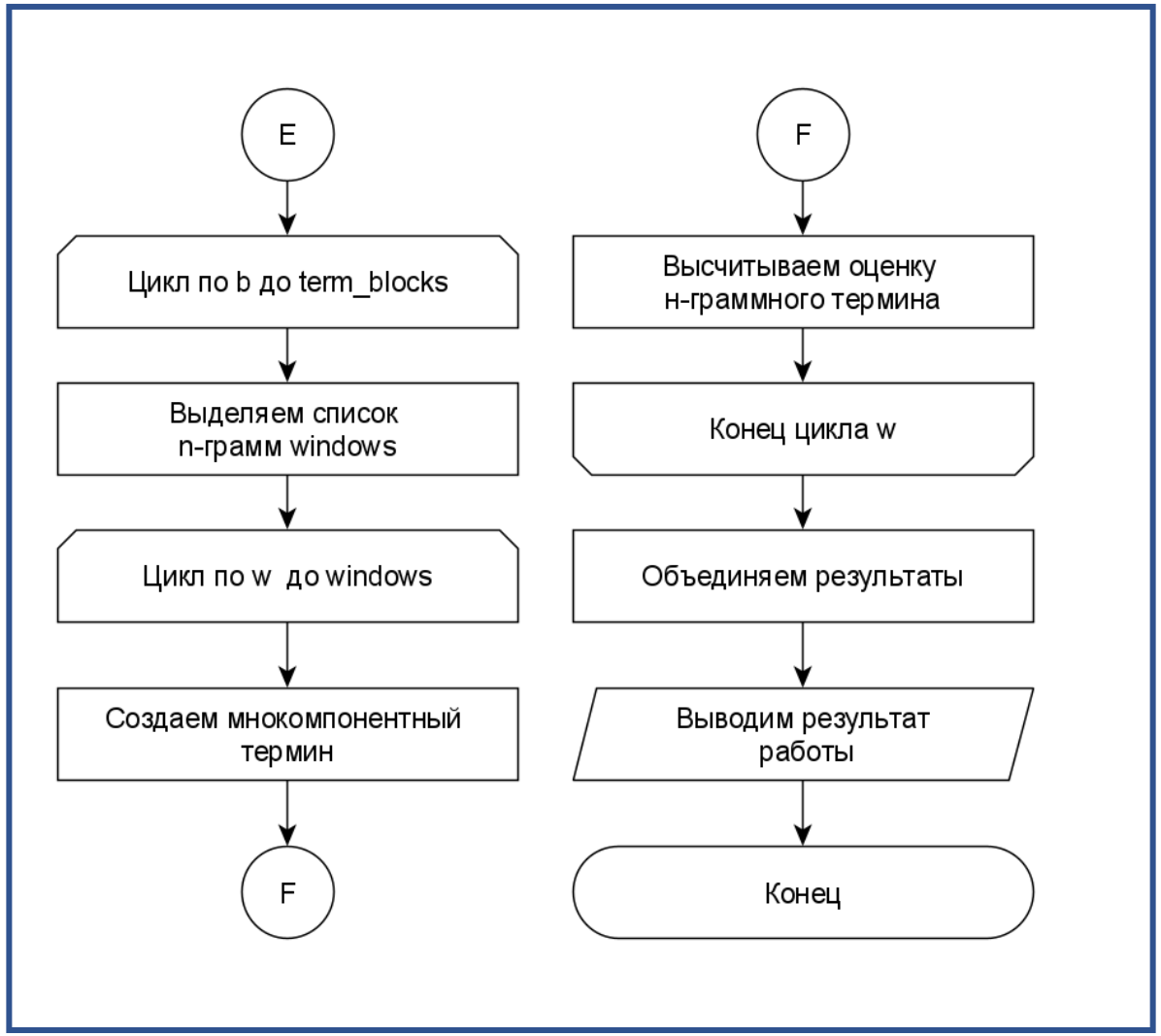
Предварительная обработка текста



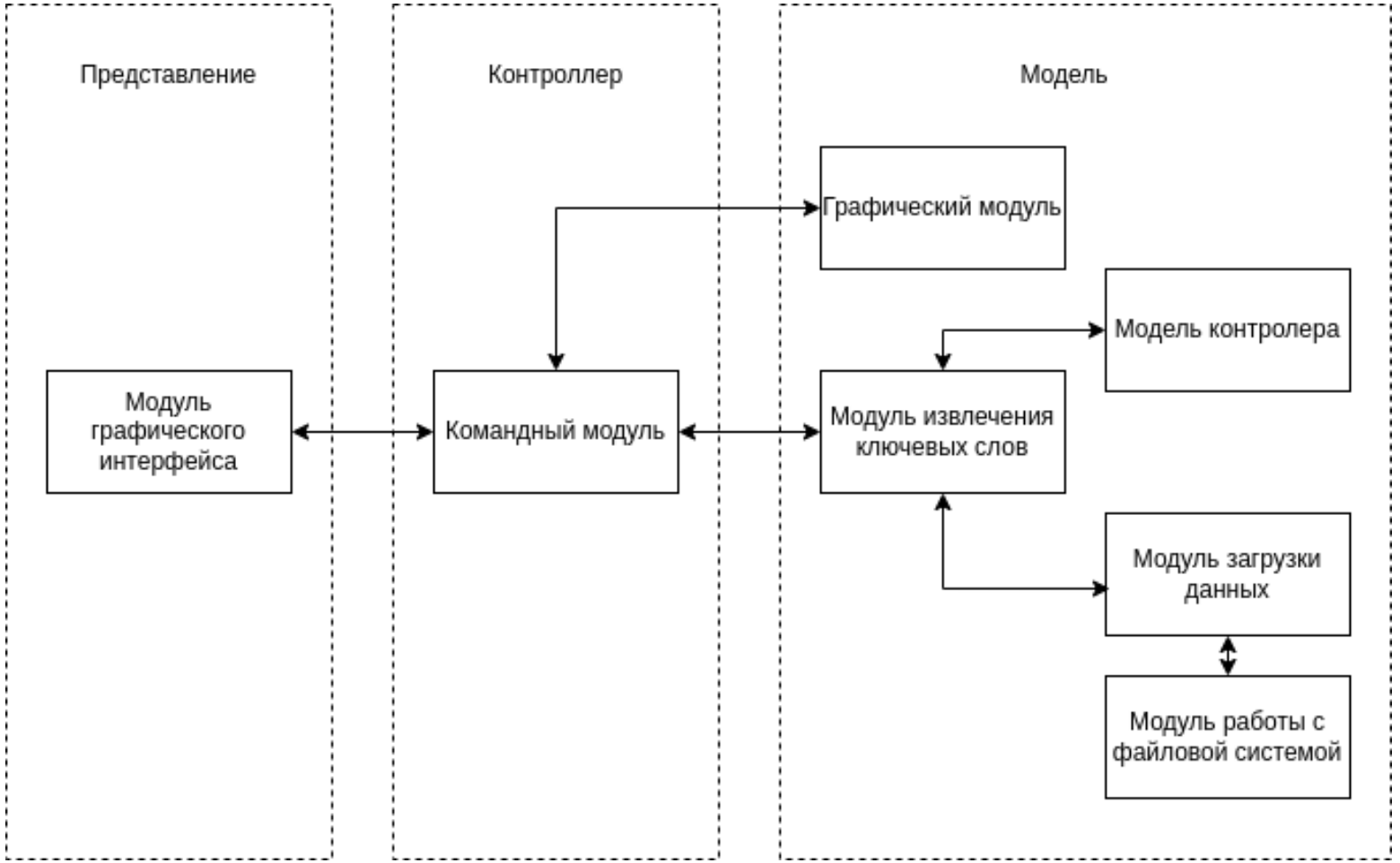
Подсчет оценки термина



Вычисление н-грамм



Архитектура ПО



Примечание: При проектировании использовался шаблон MVC – модель-представление-контролер

Исследование характеристик метода

- Выборка:
  - 30 электронных документов
- Критерии оценки
  - Процент пересечения авторских ключевых слов с КС полученными от методов
- Ограничения:
  - Текст документа содержит в себе только одну тему
  - Документ написан на русском языке
  - Документ формата PDF
  - Текст должен содержать не менее 50 слов

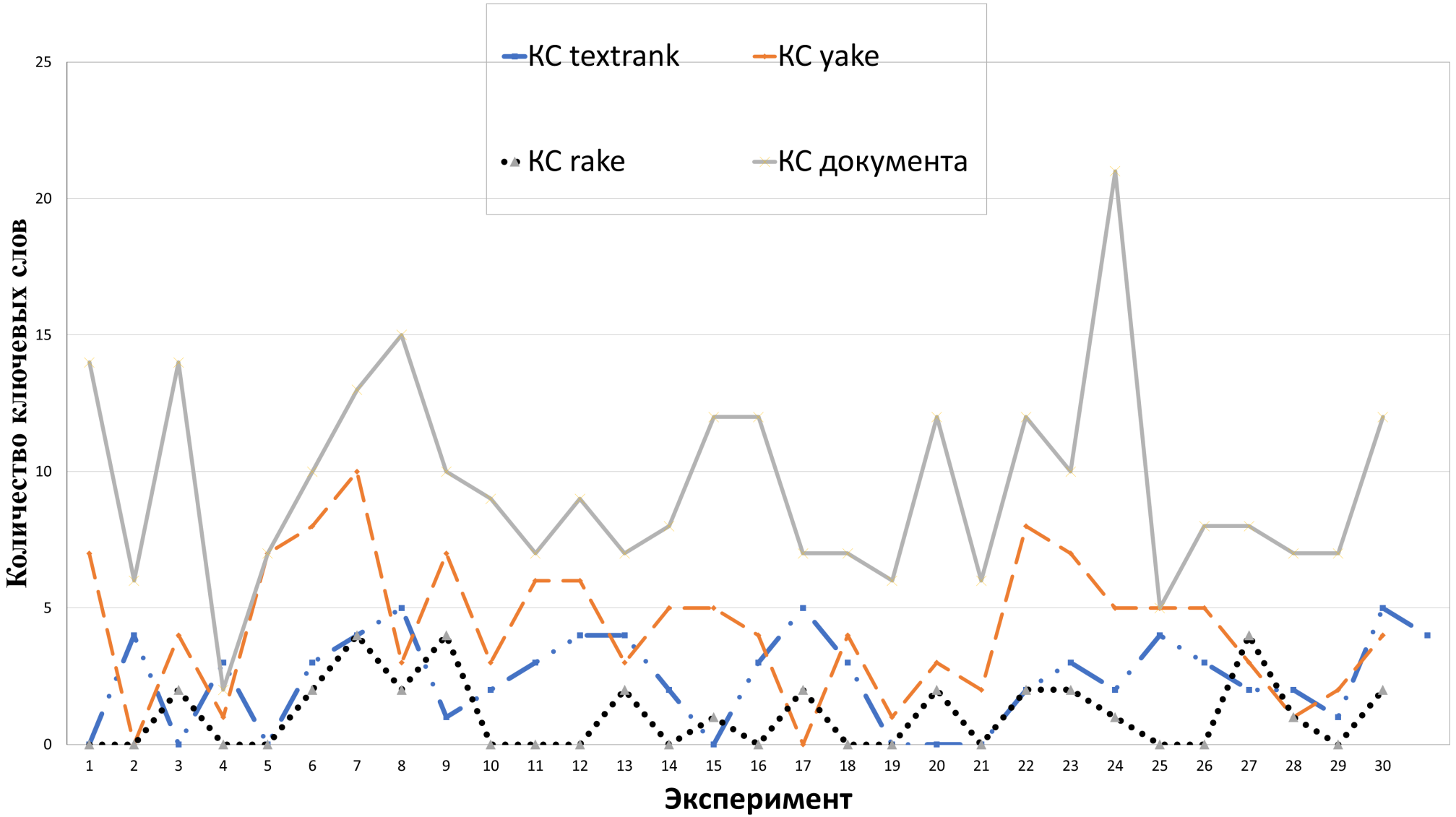
$$R = \frac{N_{doc}}{N_{cross}} \quad (1)$$

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

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

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

Сравнение с другими алгоритмами



Результаты сравнения методов

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

Исследование влияния n-грамм на результ работы метода

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 выделенные автором: биометрия, отпечаток пальца, фрактал, фрактальная размерность, идентификация и аутентификация личности, СКУД.

Выделенные КС при использовании различных n-грамм

```
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
```

Заключение

Разработан метод автоматического извлечения ключевых слов и словосочетаний из электронных документов на русском языке

1. Проведен анализ методов извлечения ключевых слов.
2. Отобран базовый алгоритм и выполнена его модификация.
3. Спроектировано и разработано программное обеспечение для реализации метода.
4. Проведено экспериментальное исследование характеристик разработанного метода

Направления дальнейшего развития

1. Добавить процесс преобразования терминов к начальной форме.
2. Улучшить поиск дублирующих терминов