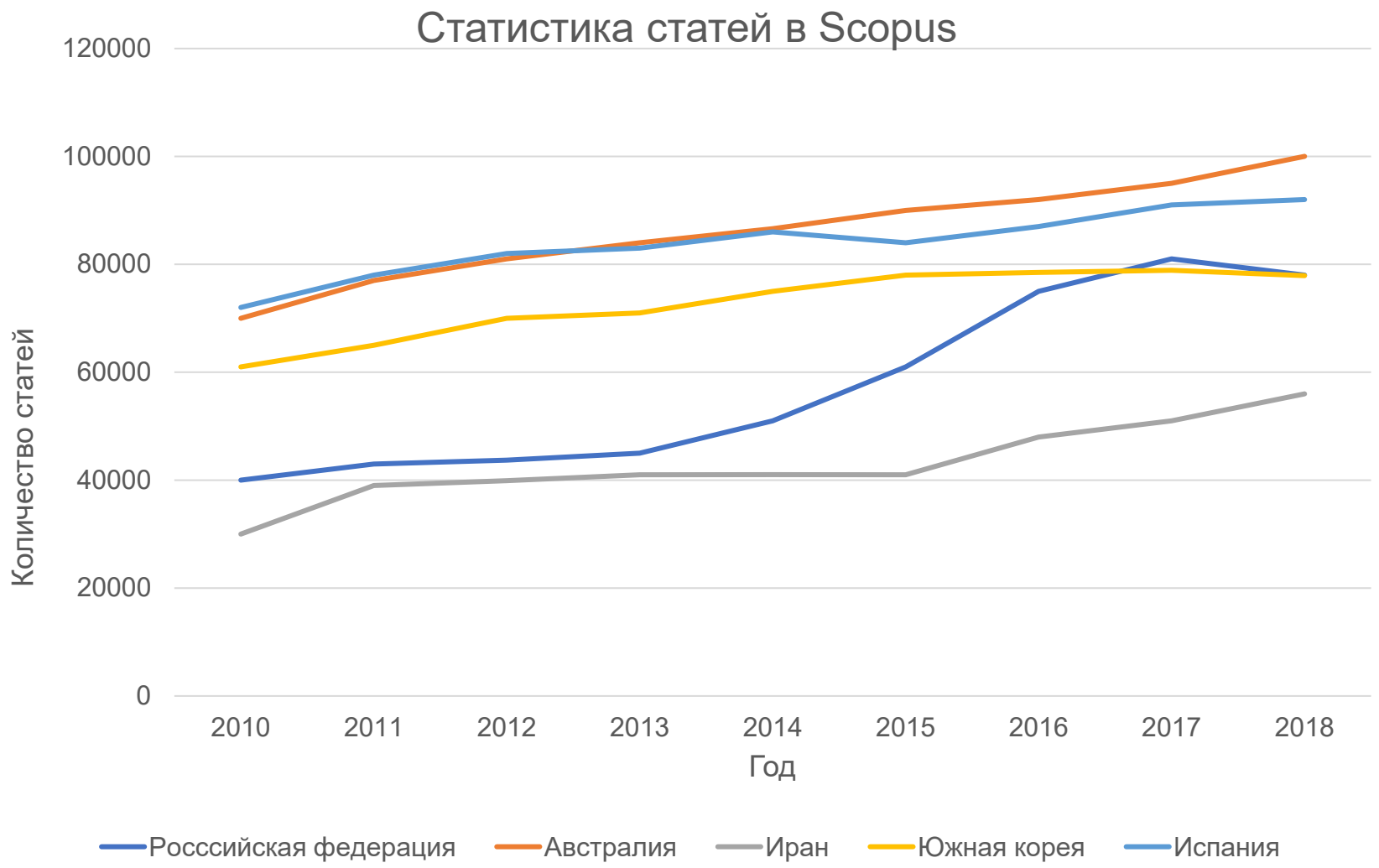
АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ КЛЮЧЕВЫХ СЛОВ И СЛОВОСОЧЕТАНИЙ ИЗ ЭЛЕКТРОННЫХ ДОКУМЕНТОВ НА РУССКОМ ЯЗЫКЕ

Студент: Барсуков Никита Михайлович
Руководитель: Барышникова Марина Юрьевна

Москва, 2022

Значимость темы



- С 2010 – 2018 количество статей издаваемых в год увеличилось в 2 раза.
- Растет сложность программных решений в области обработки текстов

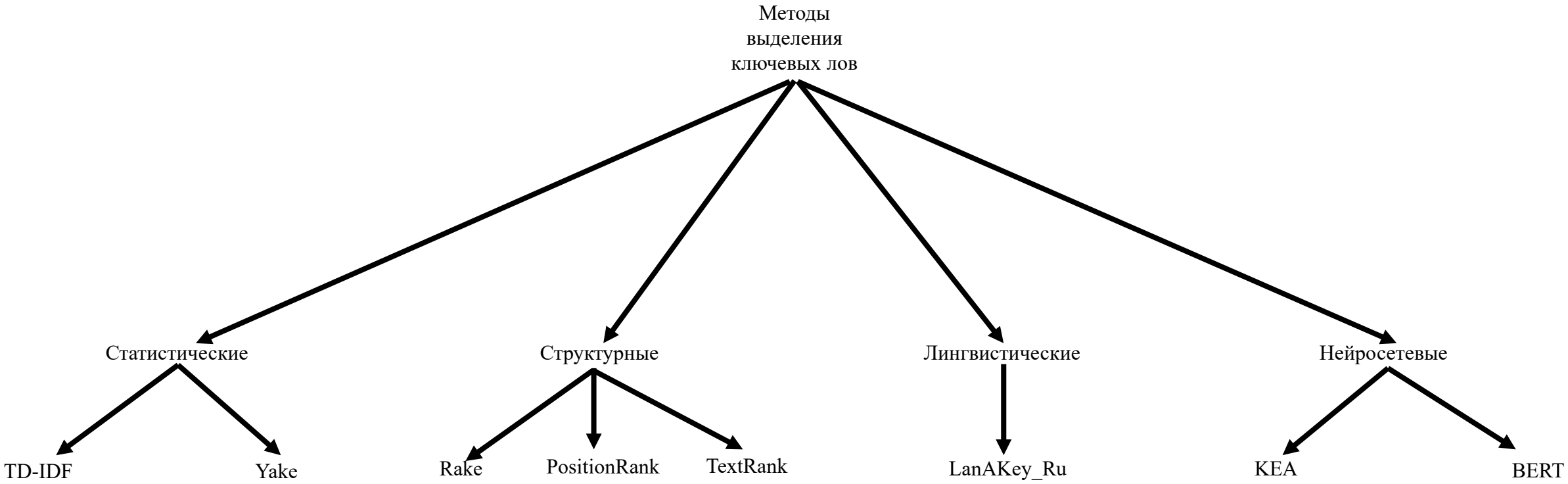
Цель и задачи

Разработка метода 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



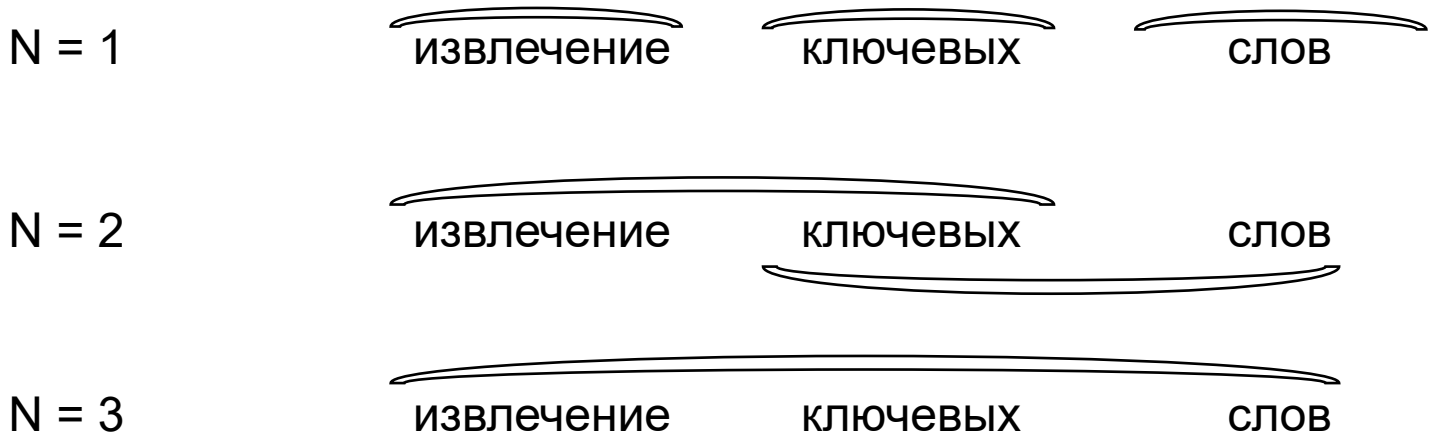
Статистические методы выделения КС

Метод	Не требует корпусов текстов	Не привязан к предметной области применения	Не использует словари, антологии	Учитывает схожесть терминов
Yake	+	+	+	+
Rake	+	+	+	-
textrank	+	+	+	-
TF-IDF	-	+	+	-

N-граммы

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

Исходный текст: Извлечение ключевых слов

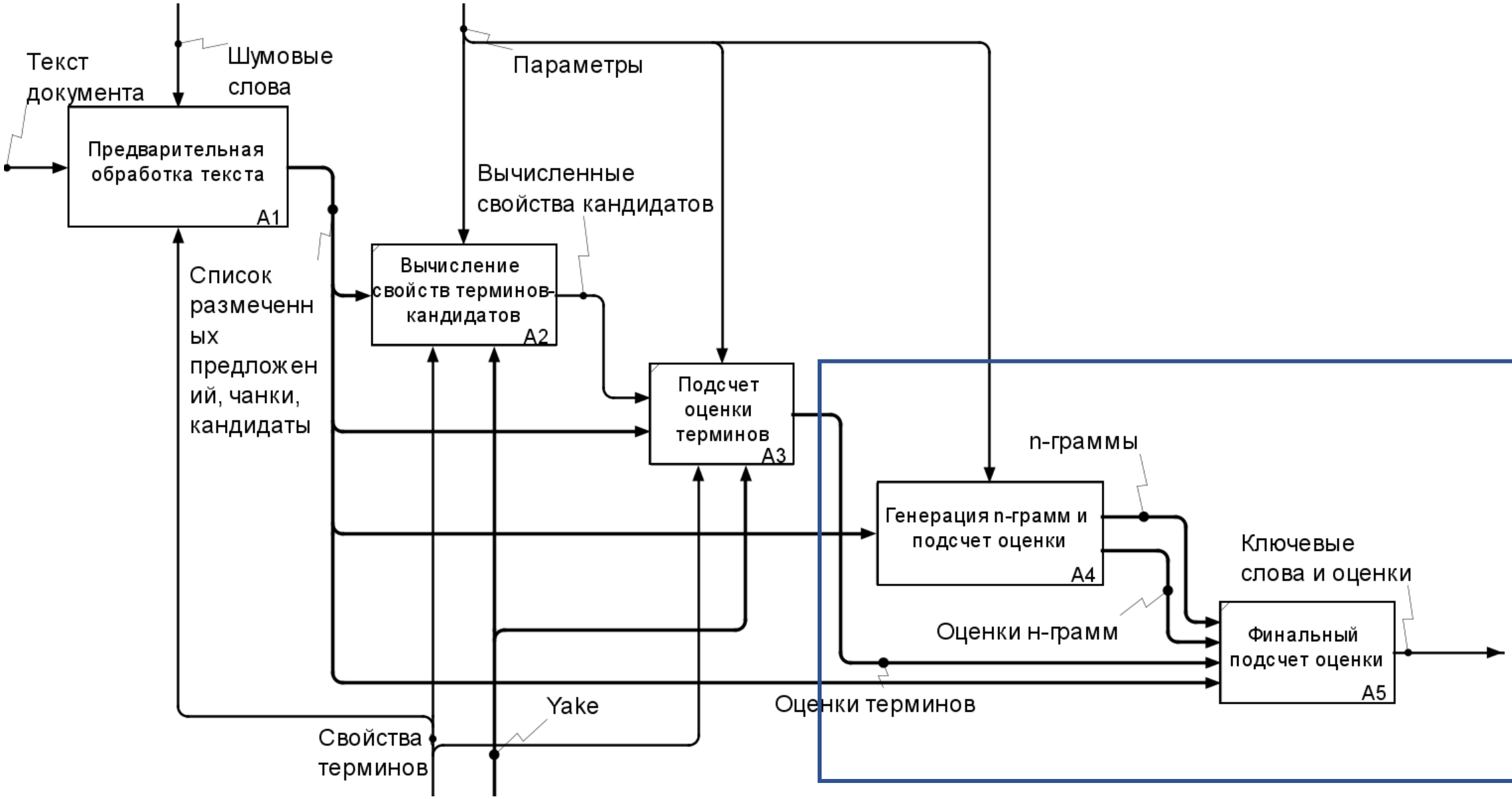


Постановка задачи

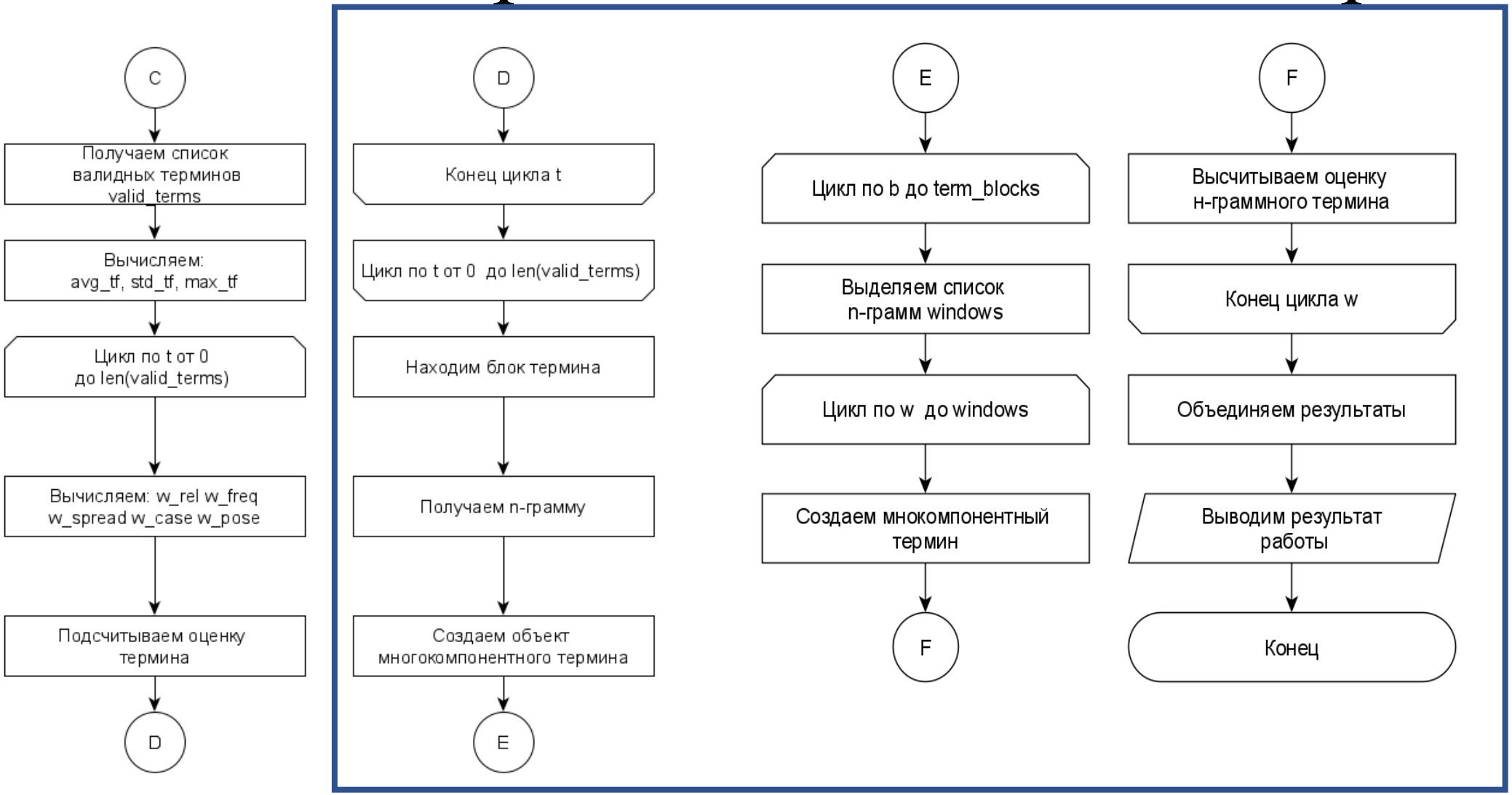
Ограничения:

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

Метод извлечения КС

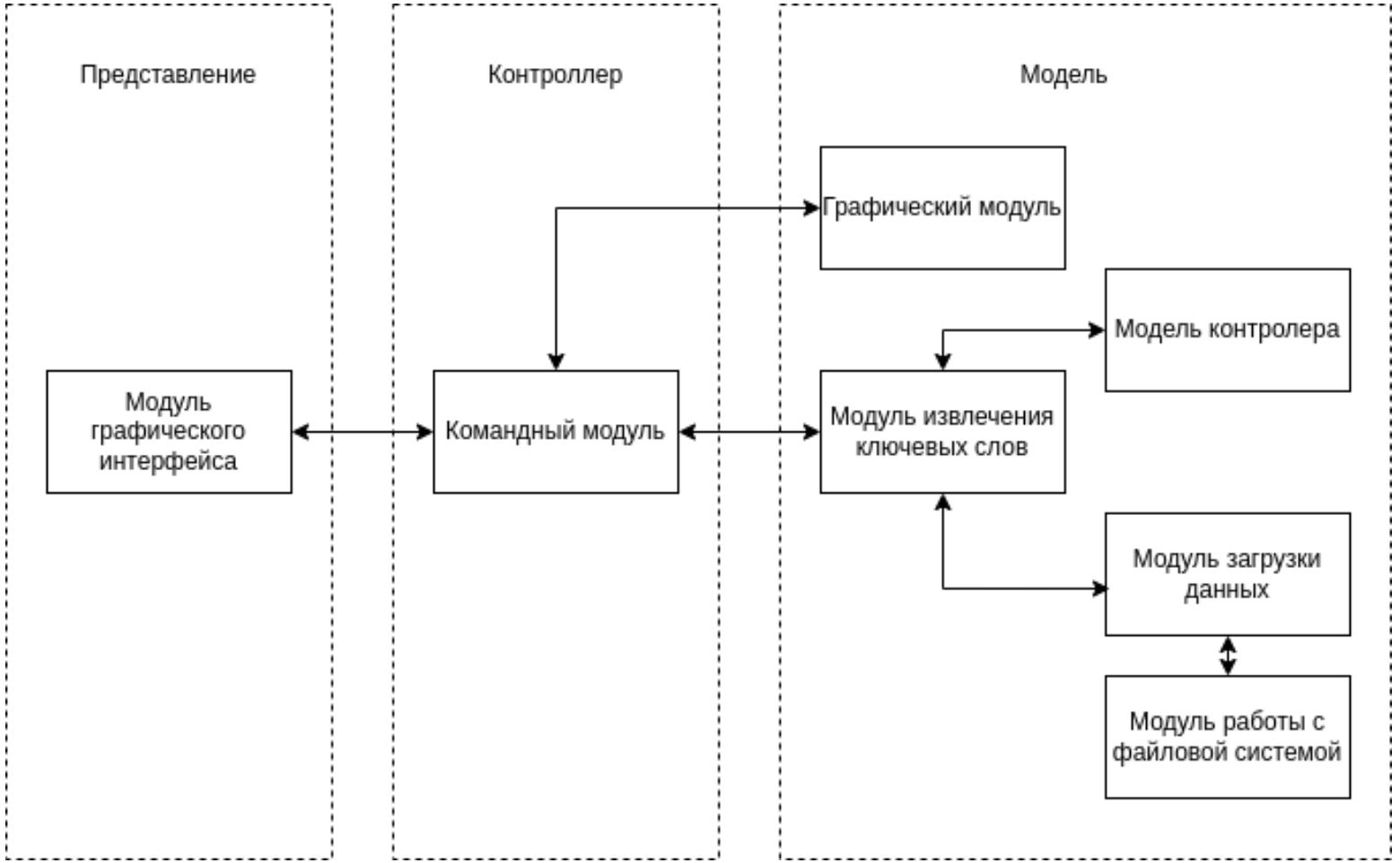


Подсчет оценки термина и выделение N-грамм



*w_case – вес связанный с регистром термина, w_pose – вес связанный с позицией в тексте, w_spread – вес связанный с рапспространением термина, w_rel – вес показывающий связь термина с контекстом, w_freq – вес корректирующий частоту

Структура ПО



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

Характеристики метода

- Выборка:
  - 30 электронных документов

- Критерии оценки
  - % ключевых слов попавших в пересечение с КС выделенными автором от количества выделенных (1)
  - % ключевых слов не попавших в пересечение от количества выделенных (2)

- Ограничения:
  - Текст документа содержит в себе только одну тему
  - Документ написан на русском языке
  - Документ формата PDF
  - Текст должен содержать не менее 50 слов

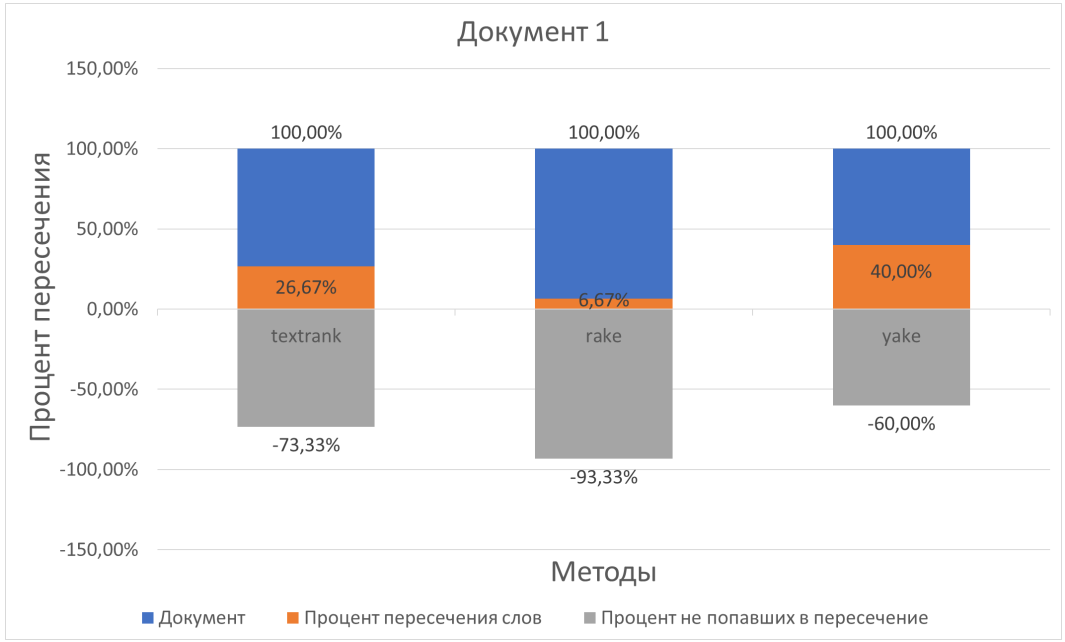
$$R_{true} = \frac{N_{res}}{N_{method}} \quad (1)$$

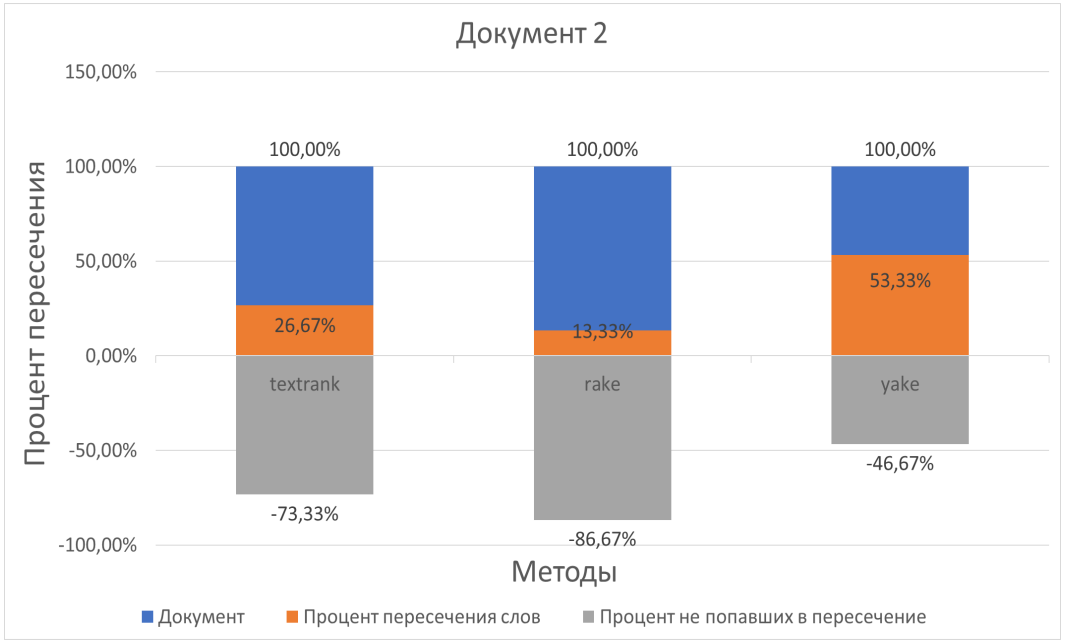
$$R_{false} = \frac{N_{method} - N_{res}}{N_{method}} \quad (2)$$

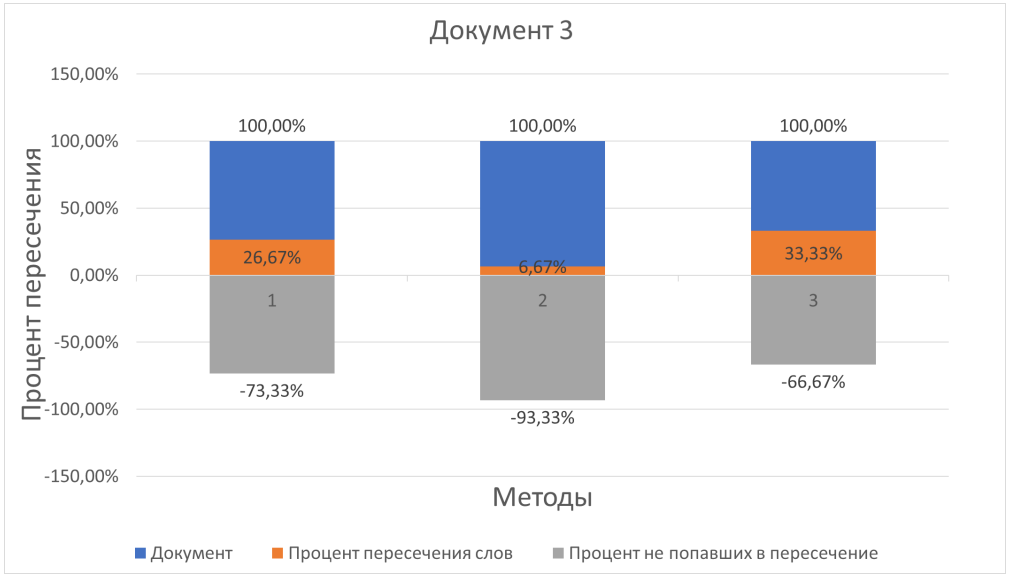
где N_{res} - количество слов из метода попавших в пересечение;

N_{method} - размеры выборки

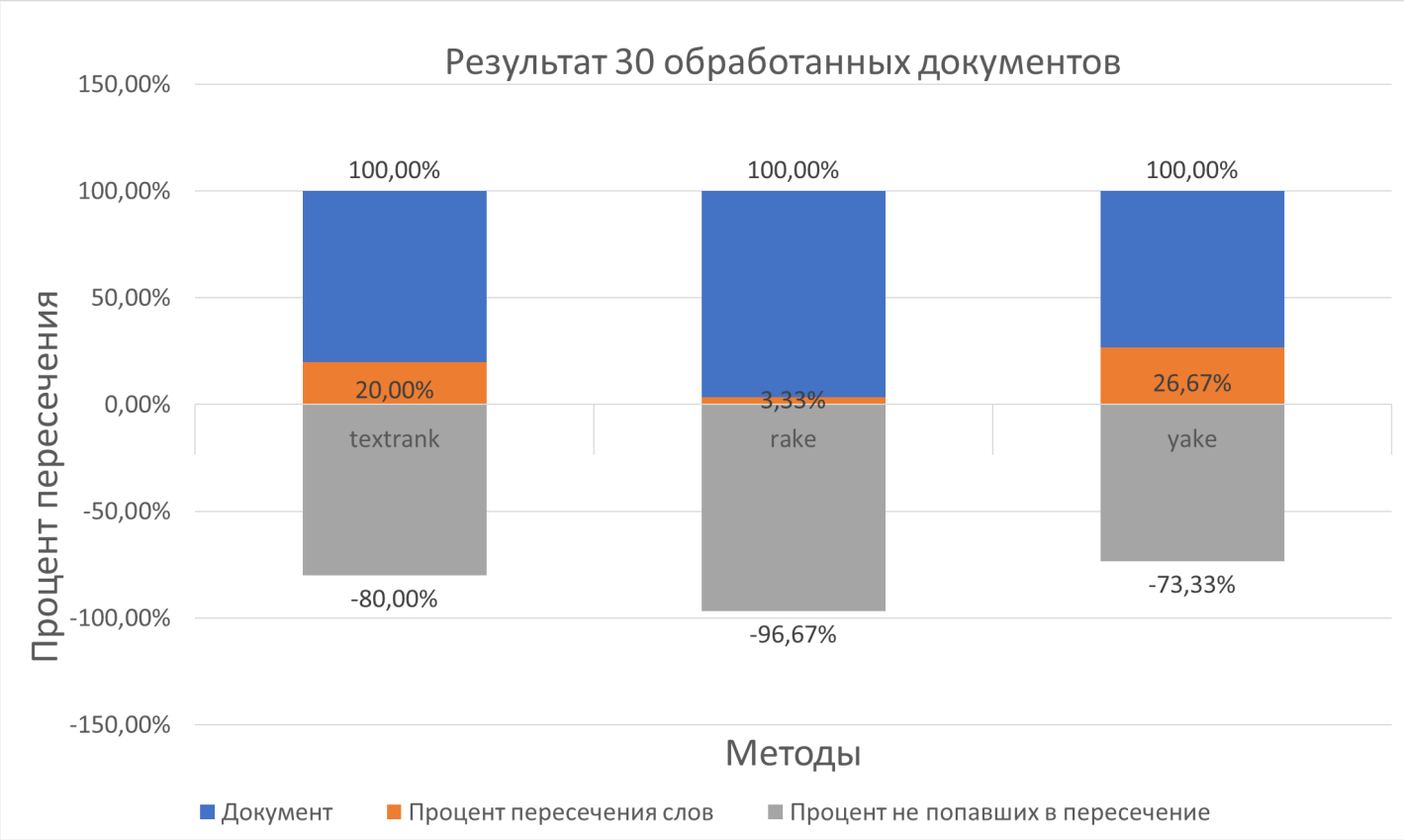
Визуализация эксперимента







Средние значения критериев



	+%	-%
Yake	26.67%	73.33%
Rake	3.33%	96.67%
Textrank	20%	80%

*+% - процент от количества слов попавших в пересечение; -% - процент слов не попавших в пересечение

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