Универсальный механизм первичного поиска повторов в тексте для пакета Duplicate Finder

Глазырин Антон Георгиевич, 21.M07-мм

Научный руководитель: доц. каф. СП, к.ф-м.н. Д. В. Луцив

Рецензент: gg

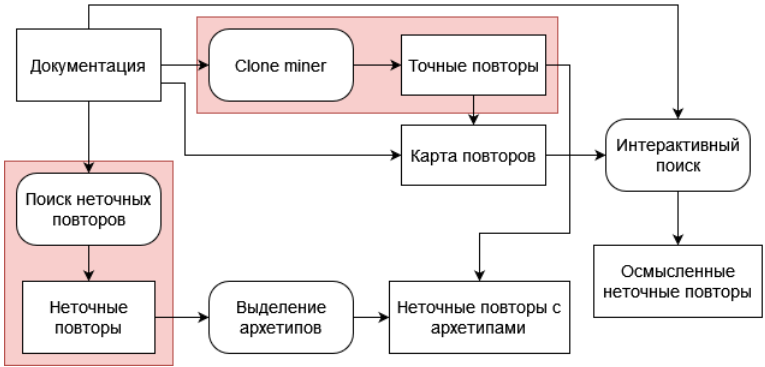
СПбГУ

18 мая 2023 г.

Мотивация

- E. Juergens, J. Porubän: ≈10% документации — дублированные фрагменты
- Негативное влияние повторов:
  - Раздувание объема
  - Усложнение модификации
- Управление повторами — улучшение документации

Мотивация: Duplicate Finder



Постановка задачи

Цель — разработка унифицированной подсистемы поиска точных и неточных повторов для Duplicate Finder Toolkit.

- Анализ предметной области
- Определение проблем поиска повторов в DuplicateFinder и требований к новому механизму
- Проектирование конвейера механизма поиска повторов
- Разработка алгоритмов точного и неточного поиска
- Реализация инструмента и его интеграция в DuplicateFinder
- Проведение тестирования разработанного инструмента

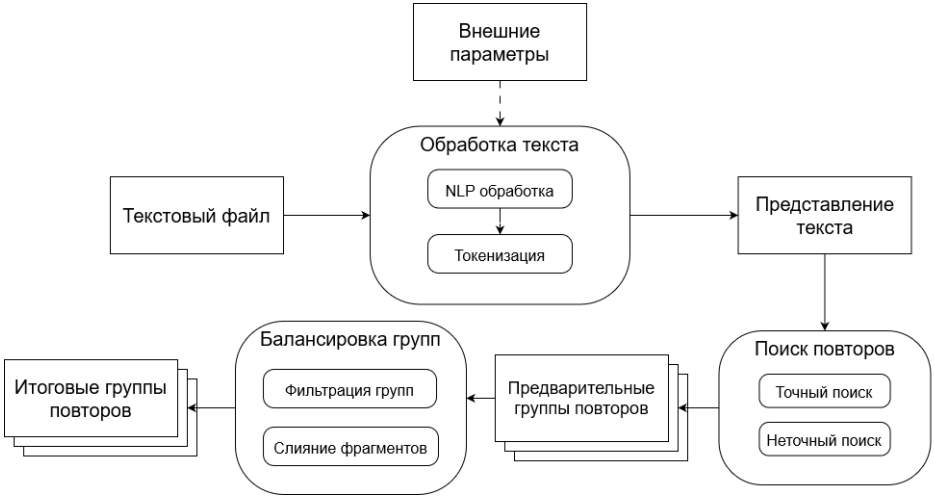
Существующие решения

- Поиск клонов в ПО:
  - CCFinder
  - CloneMiner
  - Klocwork inSight
  - cpdetector
- Сравнение текстовых документов:
  - Align
  - TxtAlign
- Поиск по образцу:
  - Duplicate Defect Detection
  - Apache Lucene
  - FactorLCS

Определение требований

1. Реализация на языке Python
2. Поиск точных и неточных повторов
3. Универсальность процесса поиска
4. Наличие API и CLI
5. Возможности конфигурации

Конвейер поиска повторов

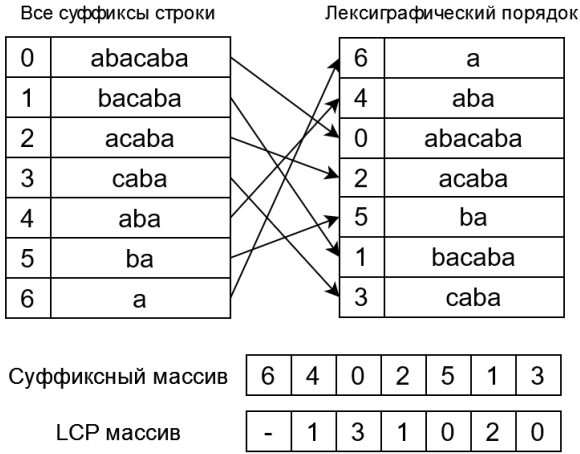


Предобработка текста

Методы NLP:

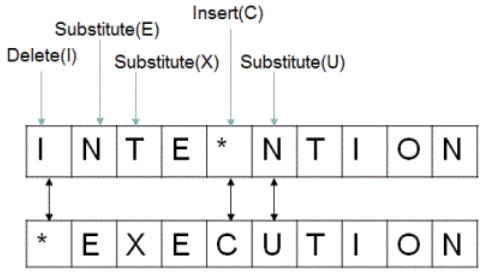
- Фильтрация спецсимволов
- Удаление стоп слов
- Лемматизация
- Стемминг

Поиск точных повторов



Поиск неточных повторов I

Расстояние Левенштейна:



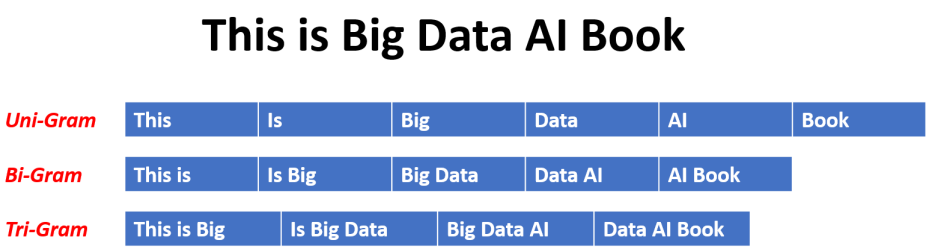
Поиск неточных повторов I

SimHash:

Hash 1	1	1	0	1	1	0	1	1
Hash 2	1	1	0	0	0	1	1	0
Hash 3	0	1	1	0	1	0	0	1
Result	1	1	0	0	1	0	1	1

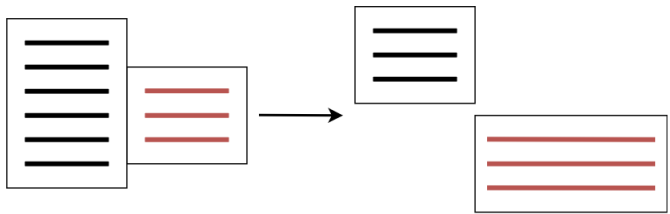
Поиск неточных повторов II

Множества N-грамм:

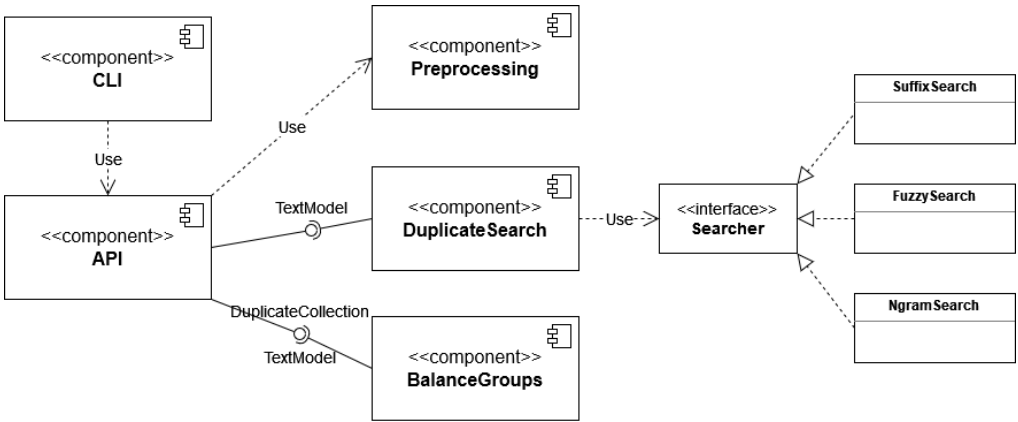


Балансировка групп повторов

- Фильтрация незначимых групп
- Слияние фрагментов



Архитектура



Тестирование: неточный поиск

Документ	Группы повторов	Средний размер группы	Средняя длина повтора	Покрытие документа
GIMP Manual	574	2.65	13.64	15%
PostgreSQL Manual	464	2.66	17.17	25%
Subversion book	282	2.27	18.93	10%
Zend Framework guide	522	2.32	22.96	16%
Blender Manual	1393	2.48	14.22	16%

Тестирование: точный поиск

Документ	Группы повторов	Средний размер группы	Средняя длина повтора	Покрытие документа
GIMP Manual	400	2.57	14.59	11%
PostgreSQL Manual	289	2.30	16.31	14%
Subversion book	218	2.17	17.27	7%
Zend Framework guide	557	2.44	16.58	13
Blender Manual	587	2.33	19.14	9%

Заключение

1. Проанализированы основные подходы и средства, которые используются в существующих инструментах для поиска повторов
2. Выявлены требования к новому механизму поиска
3. Спроектирован конвейер для механизма поиска: предобработка текста, применение алгоритмов поиска повторов, балансировка групп повторов
4. Разработаны алгоритмы для точного и неточного поиска повторов на основе использованных в Duplicate Finder инструментов
5. Выполнена реализация инструмента на языке Python с использованием пакета NLTK для предобработки текста, исходный код выложен на GitHub; проведена интеграция с Duplicate Finder
6. Проведено тестирование инструмента на корпусе документов, по результатам работы собрана статистика и проведен ее анализ