Универсальный механизм первичного поиска повторов в тексте для пакета Duplicate Finder

Глазырин Антон Георгиевич, 21.М07-мм

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

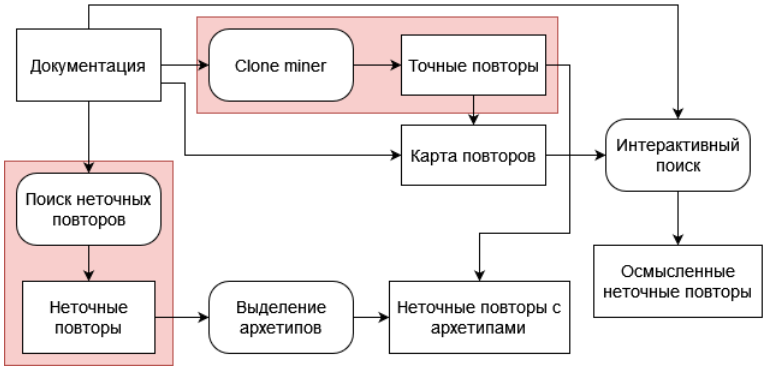
СПбГУ

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:
  - CloneMiner
  - CCFinder
  - 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.

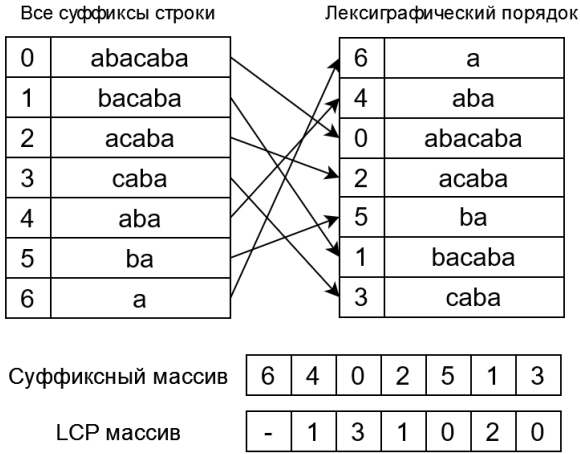
Конвейер поиска повторов

Предобработка текста

Методы NLP:

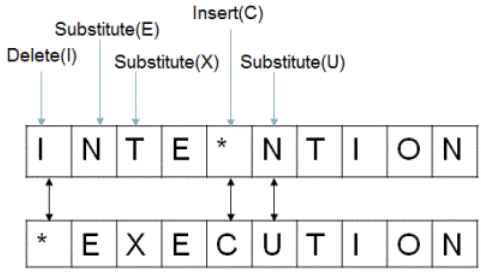
- Фильтрация спецсимволов
- Удаление стоп слов
- Лемматизация
- Стемминг

Поиск точных повторов



Поиск неточных повторов I

Расстояние Левенштейна:



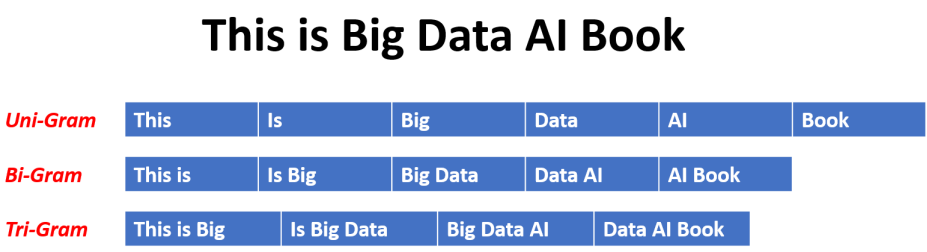
Поиск неточных повторов I

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

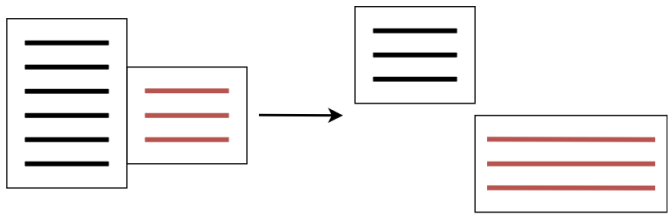
Поиск неточных повторов II

Множества N-грамм:

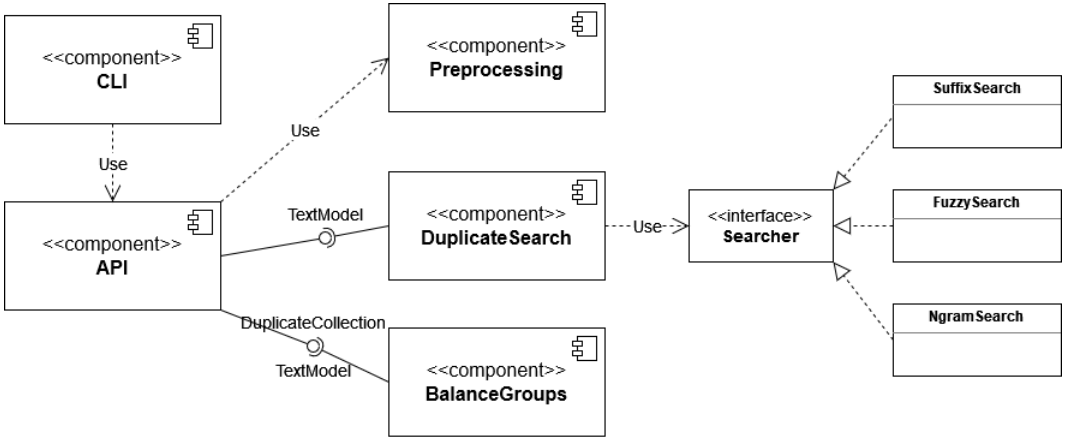


Балансировка групп повторов

- Фильтрация незначимых групп
- Слияние фрагментов



Архитектура



Тестирование: точный поиск

	GIMP	PostgreSQL	Subversion	Zend Framework
Токены	132554	72728	110270	164035
Группы повторов	400	289	218	557
Средний размер группы	2.57	2.30	2.17	2.44
Средняя длина повтора	14.59	16.31	17.27	16.58
Покрытие документа	11%	14%	7%	13%

Тестирование: неточный поиск

	GIMP	PostgreSQL	Subversion	Zend Framework
Токены	132554	72728	110270	164035
Группы повторов	574	464	282	522
Средний размер группы	2.65	2.66	2.27	2.32
Средняя длина повтора	13.64	17.17 2	18.93	22.96
Покрытие документа	15%	25%	10%	16%

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.