Факультет компьютерных наук
Департамент программной инженерии
Курсовая работа

Программа локального поиска документов по их имени и содержимому

Выполнил студент группы БПИ-143
Смилянский Александр Андреевич

Научный руководитель
Доцент департамента программной инженерии факультета компьютерных наук, к. т. н.
Дегтярёв К.Ю.

Высшая школа экономики, Москва, 2017
Hse.ru

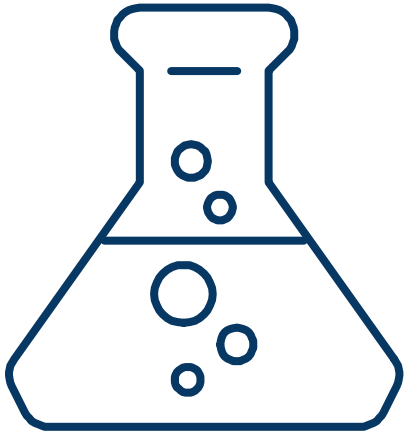
Вводная часть

Вопрос: данные?

Данные – зарегистрированная информация; представление фактов, понятий или инструкций в форме, приемлемой для общения, интерпретации, или обработки человеком или с помощью автоматических средств (ISO/IEC/IEEE 24765-2010).

Откуда приходят данные

- Делопроизводство
- Активность пользователей сети
- Работы творческого характера
- Новости
- Показатели датчиков
- Наблюдения
- …



Возникающие невольно вопросы

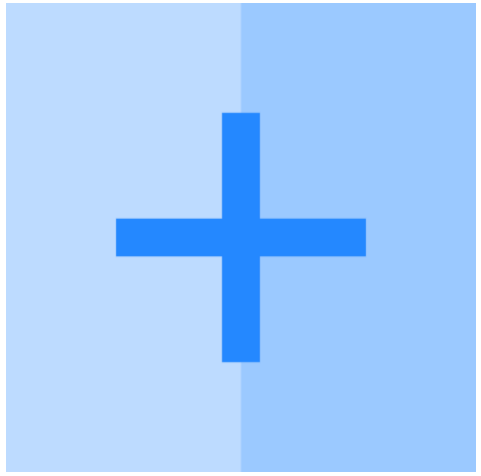
Хранение **Поиск** Обработка

Терминология

- Директория (каталог, папка) - объект в файловой системе, упрощающий организацию файлов.
- Индекс – указателем в общем смысле, в данной работе – указатель на место слов в файлах директорий под его управлением.
- Индексирование – процесс записи данных в индекс.
- Запрос на … – сущность (объект), к которой может быть применено действие и которая содержит всю необходимую для операции информацию.

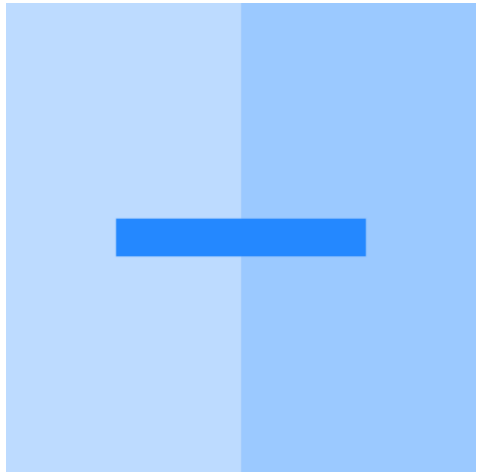
Самый простой пример

Поиск в Windows проводнике



Преимущества

+ Встроен в Windows
+ Не требует времени на создание индекса

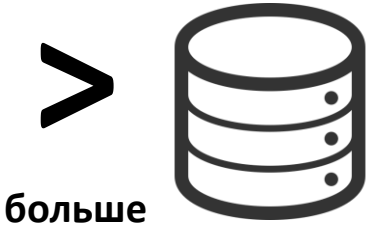


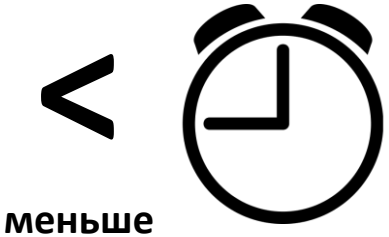
Недостатки

- Медленный
- Очень долгий поиск по содержимому

Цель разработки

Неформально: сделать программу, позволяющую производить поиск по файлам в директории, ищущую по содержимому и именам файлов **быстрее**, чем стандартные средства MS WINDOWS.





Существующие решения

- Copernic
- FileSearchy
- Архивариус 3000

Archivarius 3000

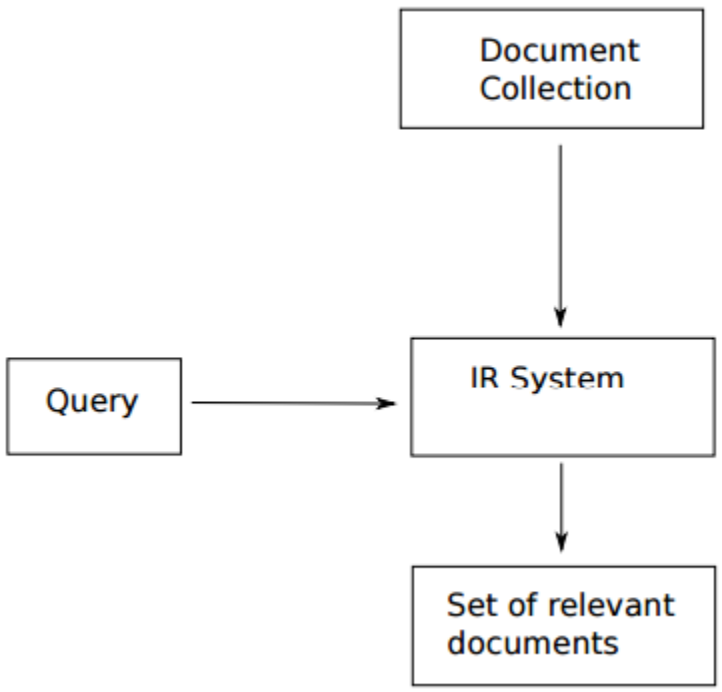
Проблемы существующих решений

- Долгая скорость индексирования
- Нет выбор расширений для индексирования
- Нет просмотра вхождения
- Нет поиска слова как подслова

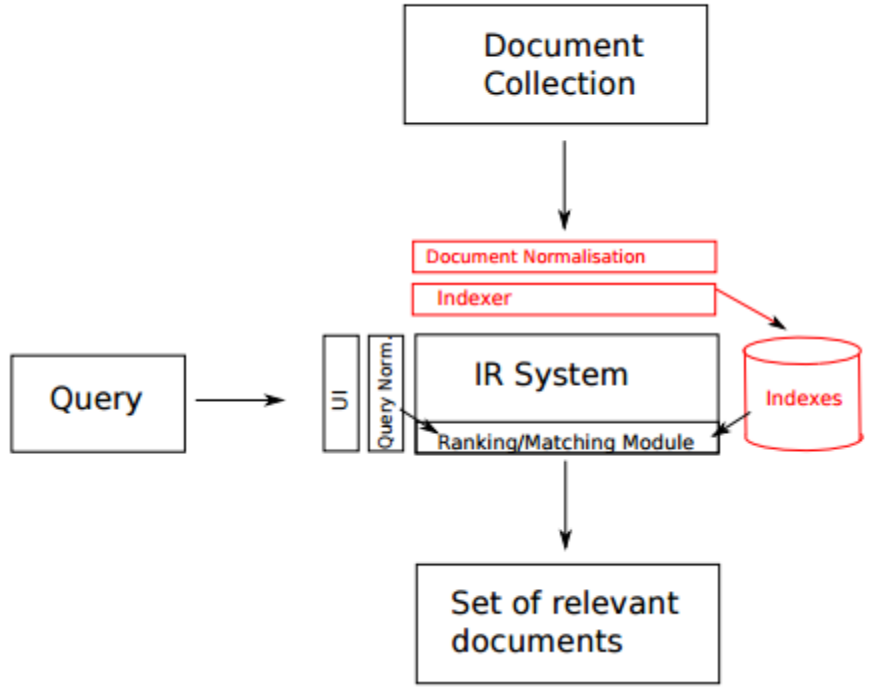
2 часть

Решение

Способ быстрого поиска

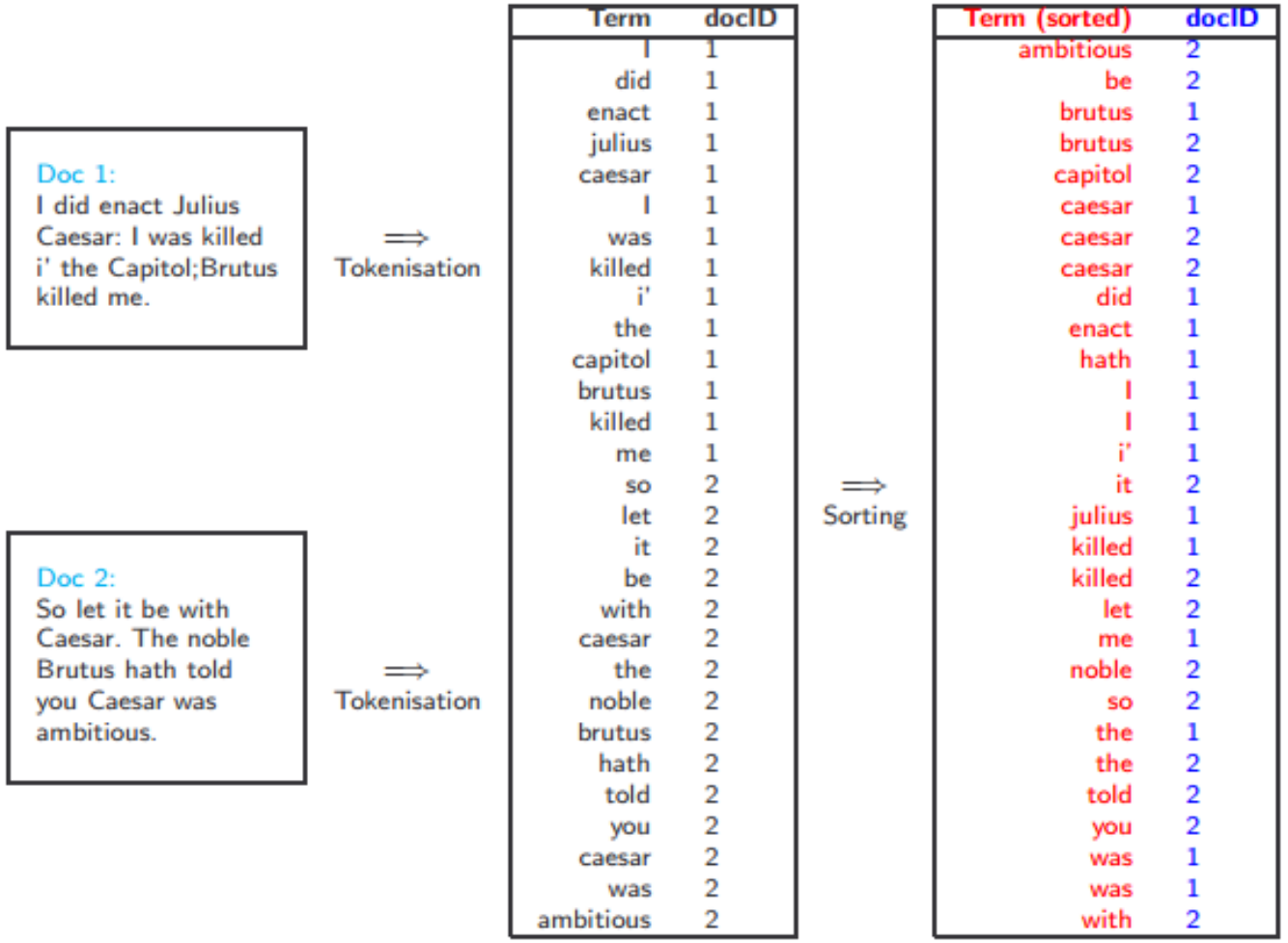


Самая простая схема задачи

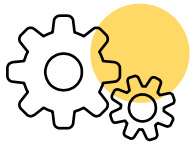


Раскрытие задачи

Способ быстрого поиска



Группы задач

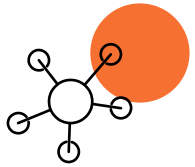


Алгоритмы построения индекса

- Нормализация
- Токенизация
- Лемматизация
- …

Алгоритмы нахождения слова

* Зависит от типа хранилища



Алгоритмы предоставляющие дополнительный функционал

- Поиск множеств
- Регулярные выражения
- …

Алгоритмы хранения

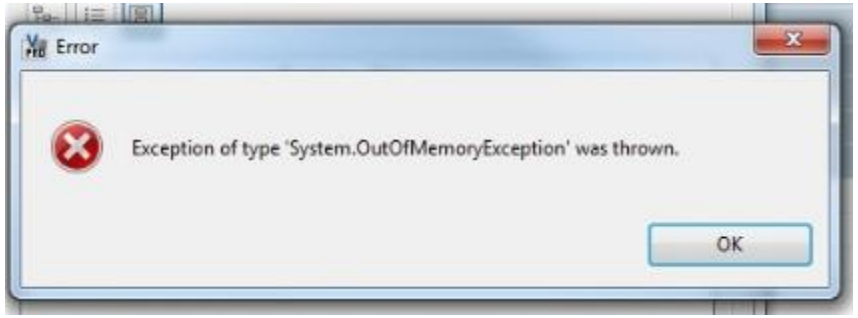
- SkipList
- LinkedList
- B-tree
- …

Построение индекса

Самый быстрый поиск – оперативная память

Размер индекса может быть даже больше размера индексируемых файлов, поэтому хранение всего в оперативной памяти – возможно только с небольшими данными, и без сохранения результатов.

Проблема:



В силу этого, хороший выход – БД, чтобы не загружать все ресурсы

Построение индекса

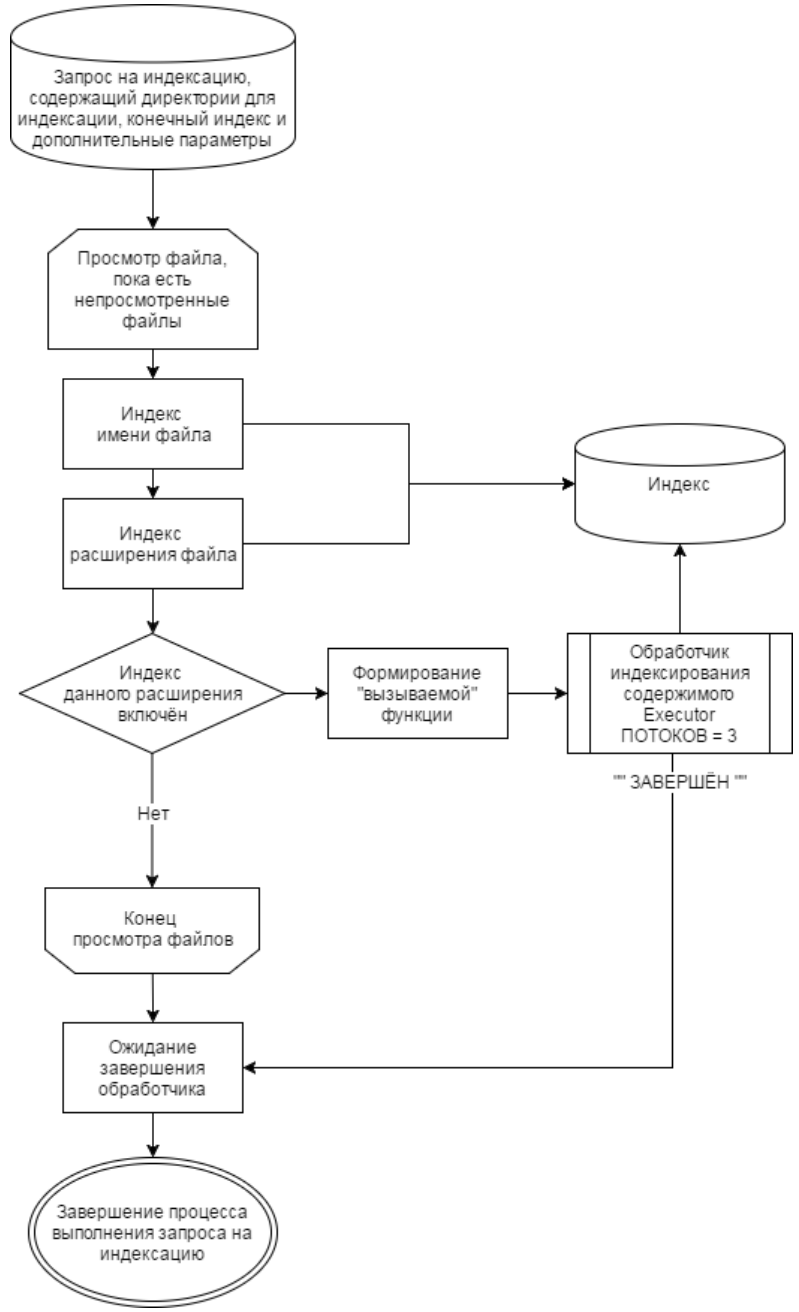
Решением стала Java EE и фреймворк записи в базу данных JPA. Он позволяет сохранять и поддерживать в up-to-date с БД состоянии объекты, находящиеся ещё в оперативной памяти.

Проблема индексирования

Проблема: быстрая скорость индексирования названия, долгая – содержания.
Названия – 1 слово на файл
Содержание – n слов в файле

Решение: многопоточность.

Результат: запись в индекс наиболее важной информации происходит за короткое время

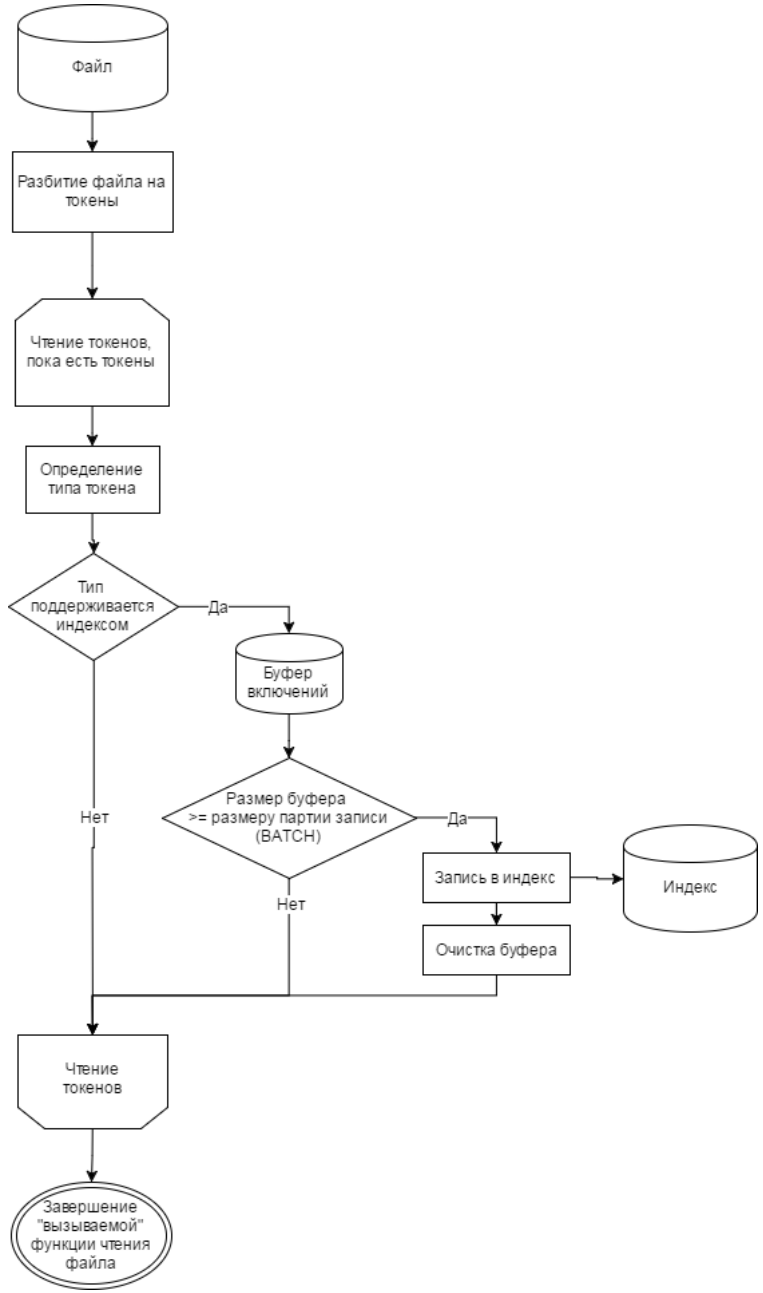


Проблема записи в БД

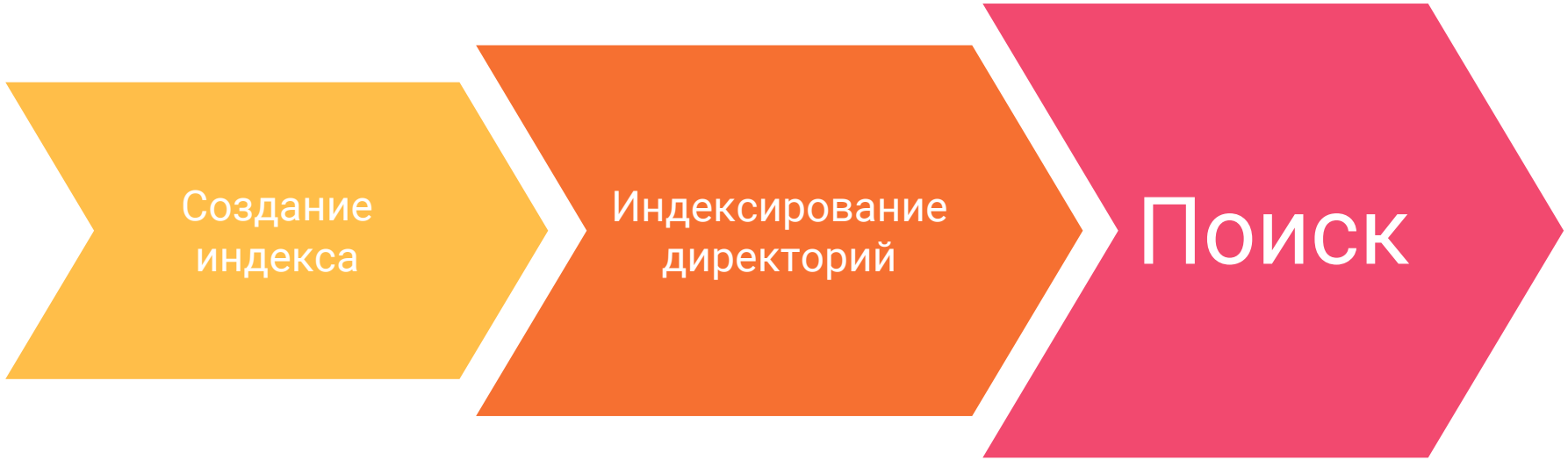
Проблема: создание транзакции для каждого сохранения каждого слова ведёт к низкой скорости записи.

Решение: Буфер слов и разделение записи на партии.

Результат: скорость записи увеличилась в разы за счёт снижения кол-ва транзакций



Работа программы

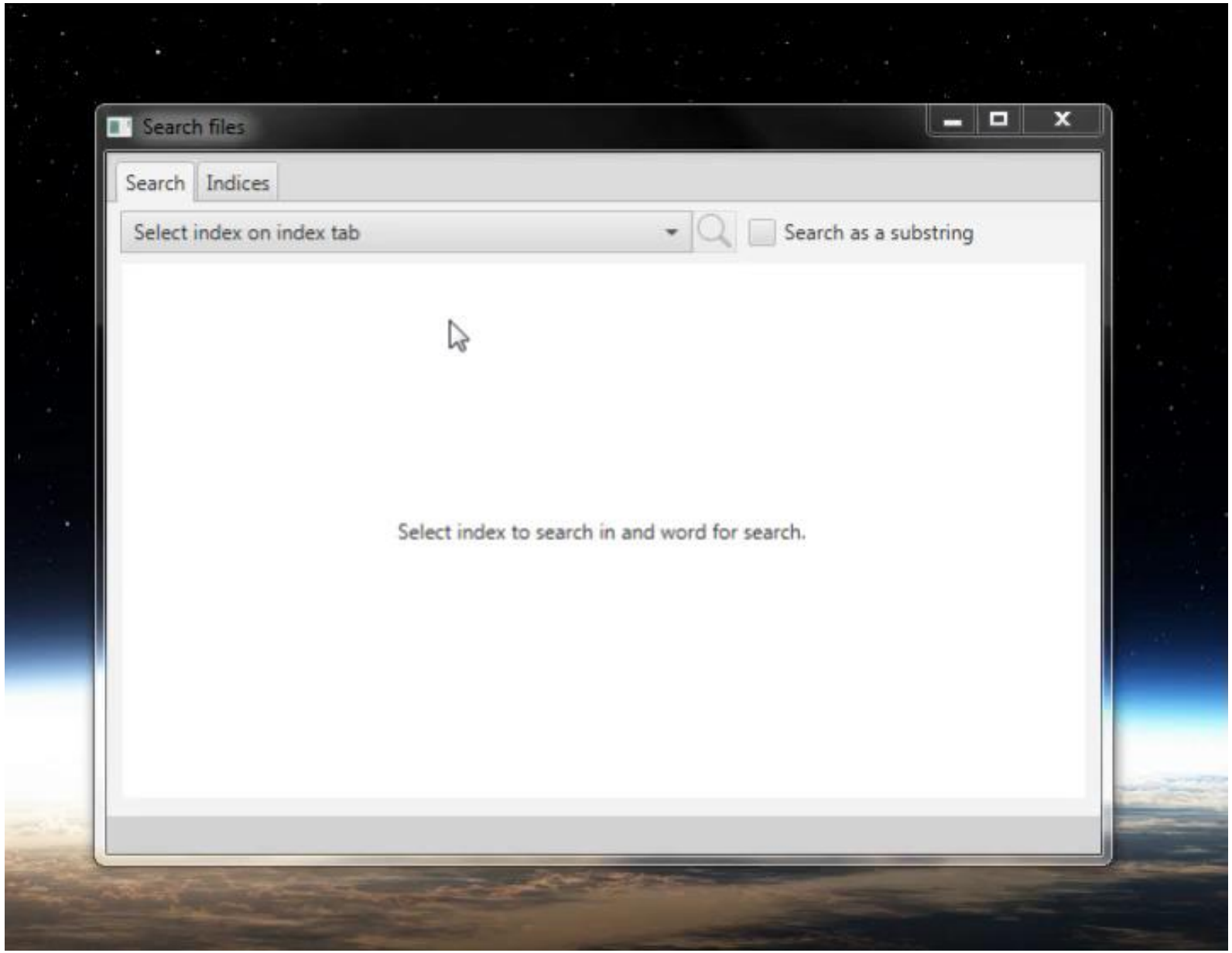


Технологии

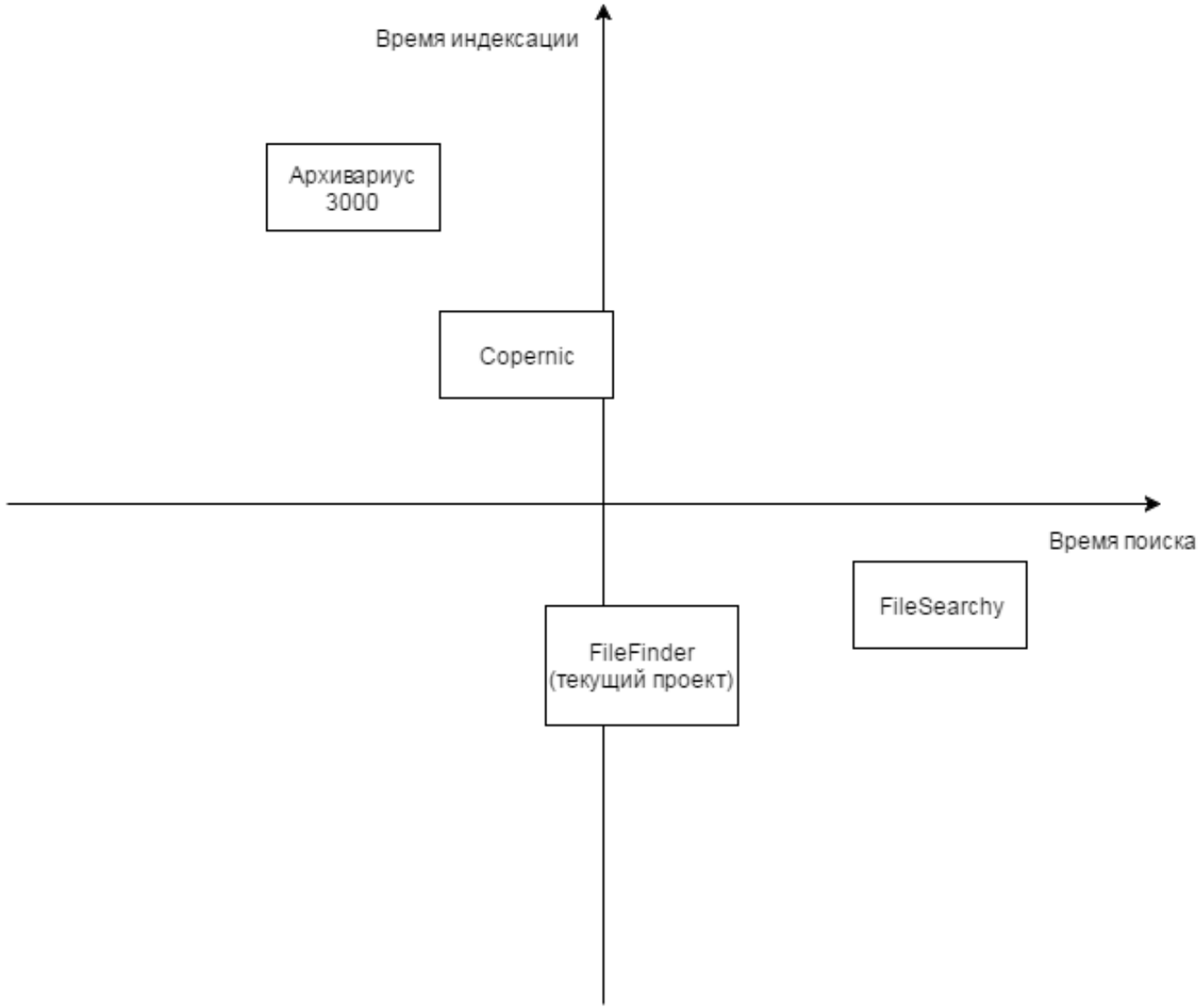
Интерфейс

Данные

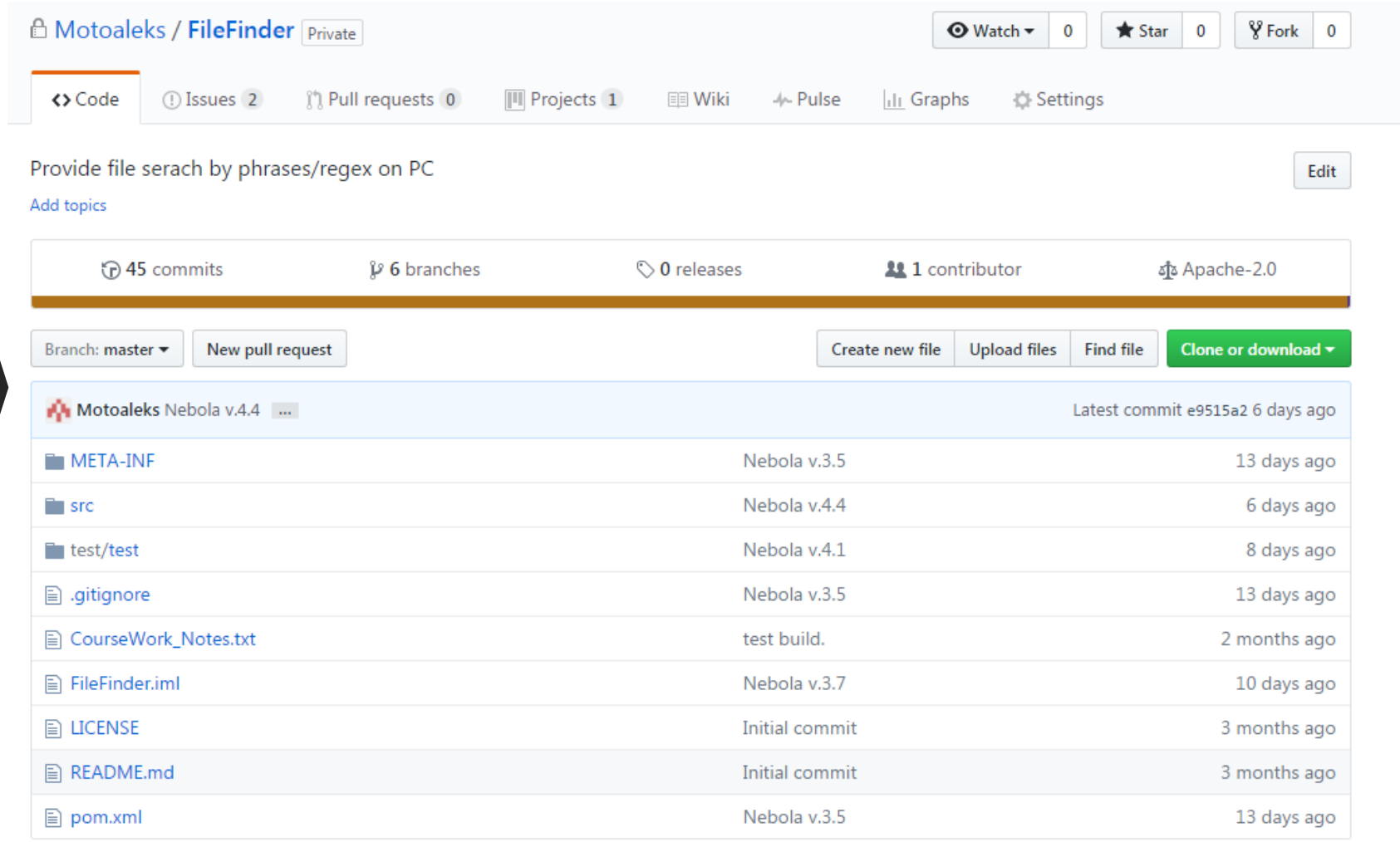
Результаты


Search files
Search
Indices
Select index on index tab
Search as a substring
Select index to search in and word for search.

Выводы


Время индексации
Архивариус 3000
Copernic
Время поиска
FileSearchy
FileFinder (текущий проект)

Github


Motoaleks / FileFinder Private
Watch 0
Star 0
Fork 0
Code
Issues 2
Pull requests 0
Projects 1
Wiki
Pulse
Graphs
Settings
Provide file serach by phrases/regex on PC
Edit
Add topics
45 commits
6 branches
0 releases
1 contributor
Apache-2.0
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
Motoaleks Nebola v.4.4 ...
Latest commit e9515a2 6 days ago
META-INF Nebola v.3.5 13 days ago
src Nebola v.4.4 6 days ago
test/test Nebola v.4.1 8 days ago
.gitignore Nebola v.3.5 13 days ago
CourseWork_Notes.txt test build. 2 months ago
FileFinder.iml Nebola v.3.7 10 days ago
LICENSE Initial commit 3 months ago
README.md Initial commit 3 months ago
pom.xml Nebola v.3.5 13 days ago

Источники

- Datastructures and Algorythms for Indexing. [Электронный ресурс] //URL: https://www.cl.cam.ac.uk/teaching/1314/InfoRtrv/lecture2.pdf (дата обращения: 21.12.2016, режим доступа: свободный)
- Токен, словарь терминов. [Электронный ресурс] // URL: https://ru.wikipedia.org/wiki/%D0%A2%D0%BE%D0%BA%D0%B5%D0%BD (дата обращения 18.04.17, режим доступа: свободный)
- Executor. [Электронный ресурс] // URL: https://docs.oracle.com/javase/7/docs/api/java/util/concurrent/Executor.html (дата обращения: 18.04.17, режим доступа: свободный)

Спасибо за внимание!

kerzkon@gmail.com

motoaleks

Студент 143 группы:
Смилянский А.А.