Zoekt

Codesearch for Git

Han-Wen Nienhuys (hanwen@google.com)

Codesearch inside google

Example session:

```
ssh

ssh f:gerrit

ssh f:gerrit -f:js (kex|key.*exchange)
```

What makes this work?

- Fast
- Complex queries: AND, OR, NOT
- Regular expressions

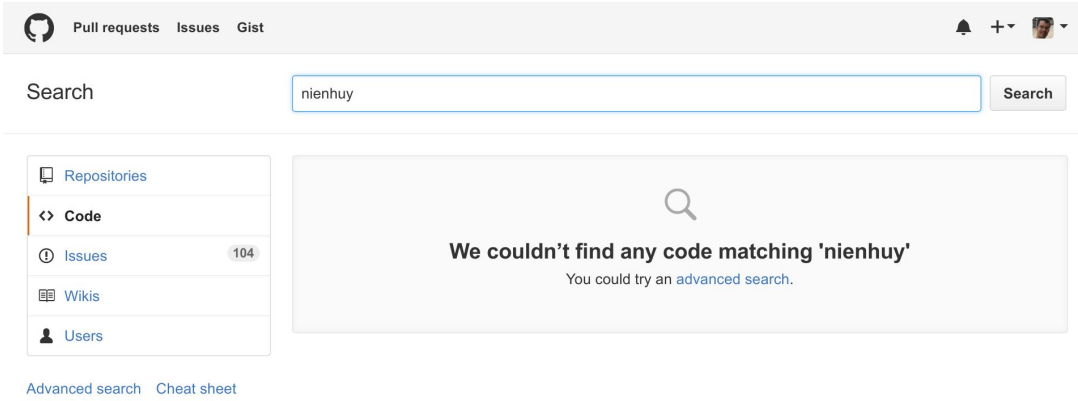
Alternatives outside?

Services:

- Limited coverage
- No boolean expressions
- No regexps; whole words only

Local (grep, IDE)

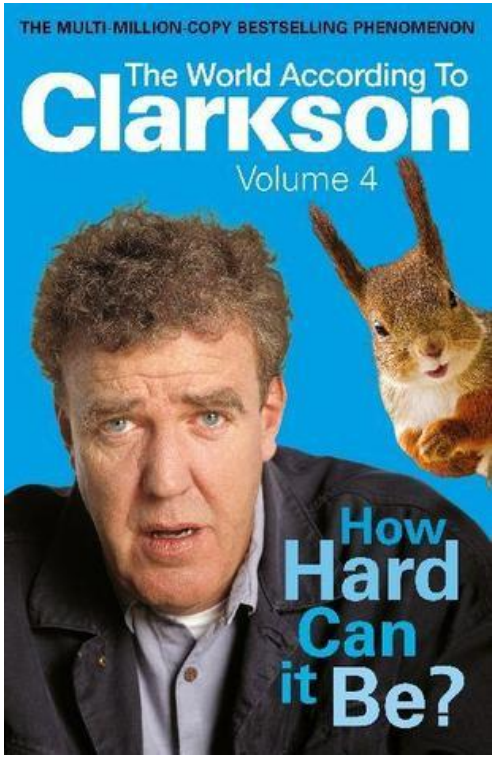
- Clumsy
- Slow
- No boolean expressions



I'll just build my own!

How do search engines work, anyway?

How does full-text search work?



Regular expression search

Regexps can be arbitrarily hard,

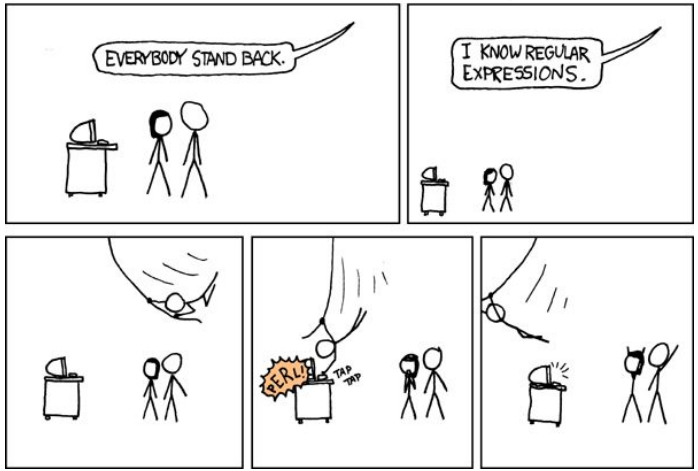
```
<^<()(?R){2}>\z|\1\Q^<()(?R){2}>\z|\1\Q>
```

But most aren't

```
extends.*Immutable(Map|List)
```

Use substring search to winnow

```
(AND "extends"
     "Immutable" (OR "Map" "List"))
```



Substring search: trigrams

Input	Index	Search
File 0: code	cod: 0	“needle”
File 1: model	ode: 0, 1	=>
	mod: 1	(AND “nee” “eed” “edl” “dle”)
	del: 1	

+ Space: 1.2 x RAM required
+ Fast indexing
- Slow searching

Substring search: suffix array

```
code model
0123 45678
```

0	codemodel
1	odemodel
2	demodel
3	emodel
4	model
5	odel
6	del
7	el
8	l

sort →

0	codemode
6	del
2	demode
7	el
3	emodel
8	l
4	model
5	odel
1	odemodel

\+ Fast search
\- ~3-5x RAM required
\- Slow indexing (sorting!)
\- Random result order (complicates nested queries)

Substring search: positional trigram

Input

```
code model
0123 45678
```

Index

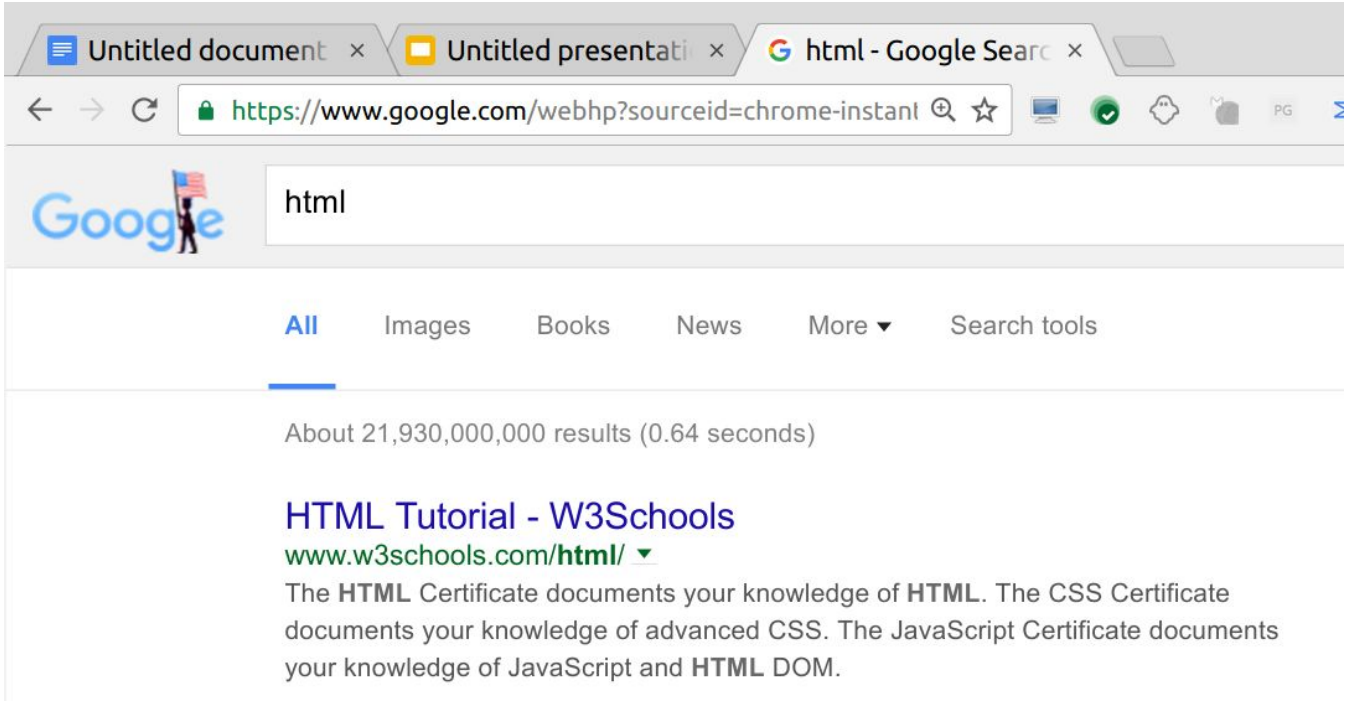
```
cod: 0
ode: 1,5
mod: 4
del: 6
```

Search

```
Query “keychain”
=>
find
  “key”, ”ain”
  2 chars apart
```

+ Space: 1x RAM, 2x SSD
+ Fast searching
+ Fast indexing
+ Results in document order

Ranking


Untitled document
Untitled presentati
html - Google Searc
https://www.google.com/webhp?sourceid=chrome-instant
Google
html
All
Images
Books
News
More
Search tools
About 21,930,000,000 results (0.64 seconds)
HTML Tutorial - W3Schools
www.w3schools.com/html/
The HTML Certificate documents your knowledge of HTML. The CSS Certificate documents your knowledge of advanced CSS. The JavaScript Certificate documents your knowledge of JavaScript and HTML DOM.

Ranking: too many results?

- Give up after enough matches
- Prefer important matches

```
// Licensed under the Apache License, Version 2.0 (the "License");

public class License implements Serializable {
```

- Start with important code

github.com/google/**guava**

Google Core Libraries for Java 6+

★ 12,726 Updated an hour ago

github.com/StarfruitStack/**guava**

Guava is a lightweight Python web framework written in C

★ 138 Updated on May 13, 2015

github.com/google/zoekt

Seek (and ye shall eat spinach)

- Code search using positional trigrams
- Regexps, AND/OR/NOT
- Uses ctags for symbol ranking
- Read git, git manifest XML and file system
- Web server with basic HTML interface
- Index server for googlesource.com and github.com
- Open source, written in Go
- Public instance to launch shortly

Git support

- Indexing bare Git repos
- Indexing manifest.xml repos (e.g. Android)
- Mirroring github & gitiles
- Query operators:

```
b:BRANCH
```

```
r:REPONAME
```

- Indexing multiple branches without space overhead

```
buggyFunction (b:master or b:stable)
```

Gitiles/Gerrit support (todo)

- Search engine can’t replicate Gerrit ACL logic.
- But Gerrit plugins know the rules:
  - put search UI in Gitiles.
  - Gitiles receives query
  - Gitiles queries *zoekt* with

```
(AND "query" r:repo (b:branch1 OR b:branch2 OR ... ))
```

- Proof of concept by Michael Ochmann & Saša Živkov (Sept '16 hackathon)

Summary

- Finally learned how search engines work!
- Code search = substring search
- https://cs.bazel.build to launch shortly
- Volunteers wanted for Gitiles integration