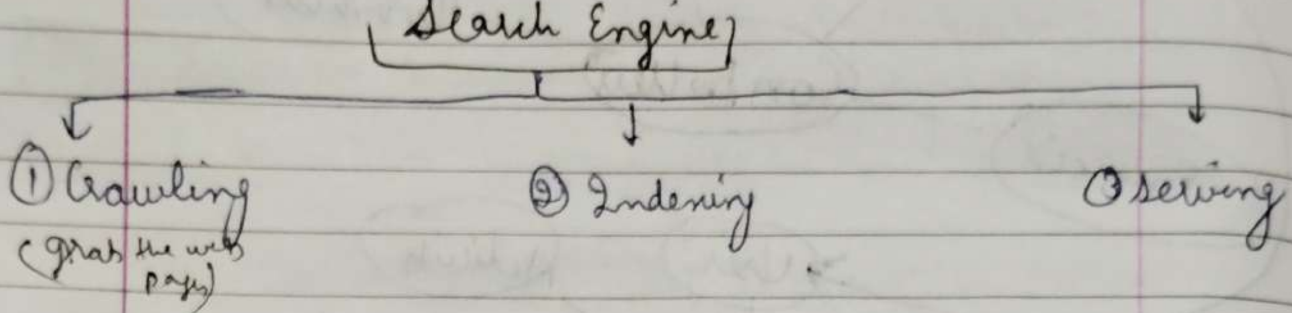
Date: 29/10/24

How Google Works:

— Coding sloth


Search Engine

1 Crawling (grab the web pages)

2 Indexing

3 Serving

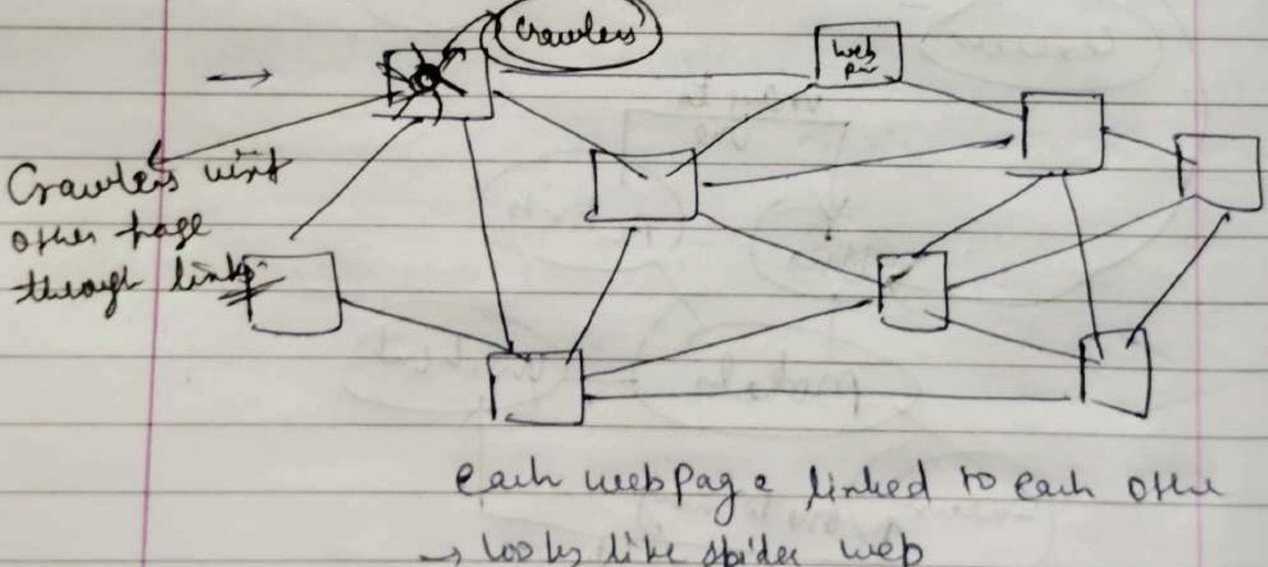
1 Crawling

→ through crawling google knows every website you visit.

→ automated programs: Crawlers / Spiders

⇒ why is internet called web?

→



each webpage linked to each other

→ looks like spider web

→ URL discovery

Sitemap

↓

file which keep infomation about pages

→ Google has bunch of crawler

Google bot Desktop

Google bot mobile

Google bot

→ (I) which page to crawl

→ (II) How often it should be crawled

→ (iii) How many pages to crawl

Algorithms

robots.txt

↳ this file tells which webpage bot should crawl.

Our Bot

→ it visits the url (our case: Wikipedia page of Google)

→ then it will grab all links in webpage

→ Then all the grabbed links will be put in list to be visited later

→ we will then move on to new url and repeat the process

time.sleep
Thread.sleep } make sure we donot overload the Server

↓

if it detects that it is bot

↓

Banned!

Spring Boot

Controller

service

① J soup

↳ API for extracting & manipulating data
↳ deal with real world HTML

→ Fetching url: Jsoup.connect("https://example.com").get();

→ Selecting elements:

```
Elements links = document.select("a[href]");
```

→ Extracting element attributes

```
for (Element link : links) {
    String href = link.attr("href");
    String text = link.text();
}
```

→ Manipulating

```
Element element = document.select("div.classname").
                                             first();
      element.text("New text");
```

→ CRAWL_LIMIT :- limit of pages to visit

Set
↳ track of URL visited
↳ avoid recrawling

Queues
↳ manage URL to be crawled
⇓
[poll] remove

→ Thread.sleep(1000 + random.nextInt(2000))
↳ delay b/w 1 & 3 sec

→ fetchDocument

→ formatUrl { null
invalid

DOES WHAT IT NEEDS TO DO!
WHO KNOWS WHETHER
CORRECT ?

② Indexing

→ google saved every info fromwebsite.

We are doing — Saving:

- → url
- → title
- → description
  - → if it has
  - ↳ if not
  - ↳ first 200 words of page
- → words.

- → URL
- → title
- → description
- → No. of words

★ Inversting Indexing

Words → webpage

Java Program 🔍

- → DB conting words "Java Prog"
- ↳ return the webites having word "Java Prog"

★ Stop words

the, is, at, which, on

→ google remove these words

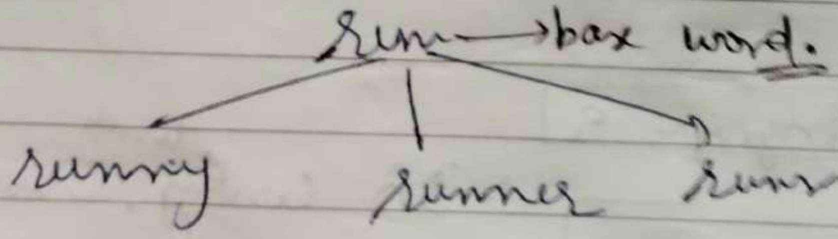
~~how~~ ~~to~~ ~~a~~ clear ~~my~~ internet browsing history?
↓
stop words.

↳ makes it faster

Nowadays

↳ stop words aren't removed
↳ using ML we can get better results through stop words

★ Stemming



search any thing will consider base word

Indexing Service

model

url
titele
description
words

gettersetter

① Collect title

document.title

② description

if it has ✓

it not : first 200 words

③ words

extract

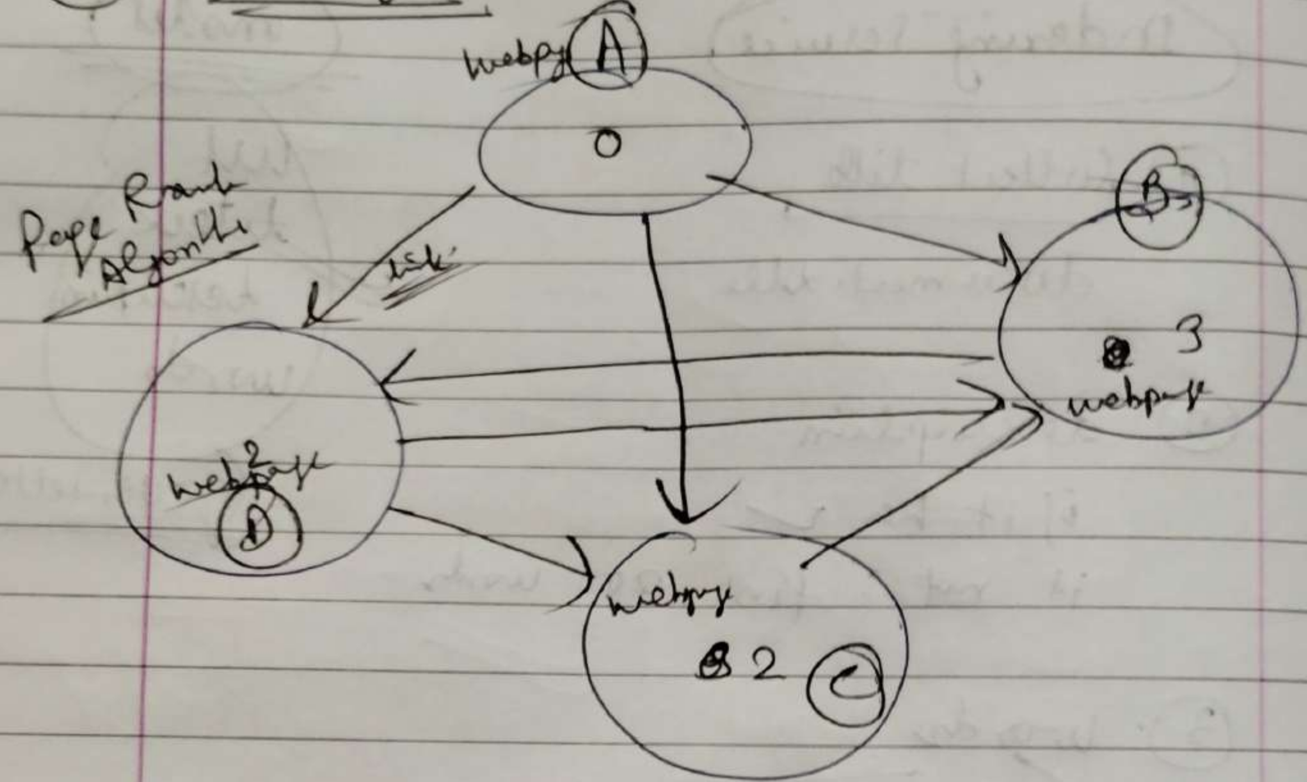
count

→ Pattern
→ Matches } List

WebCrawler Controller

map → indexing service

③ Serving:



→ Topological Sorting

Ranking: B > C = D > A

$$\boxed{PR(P_i) = \frac{1-d}{N} + d \cdot \sum_{P_j \in In(P_i)} \frac{PR(P_j)}{|Out(P_j)|}}$$

WTF?