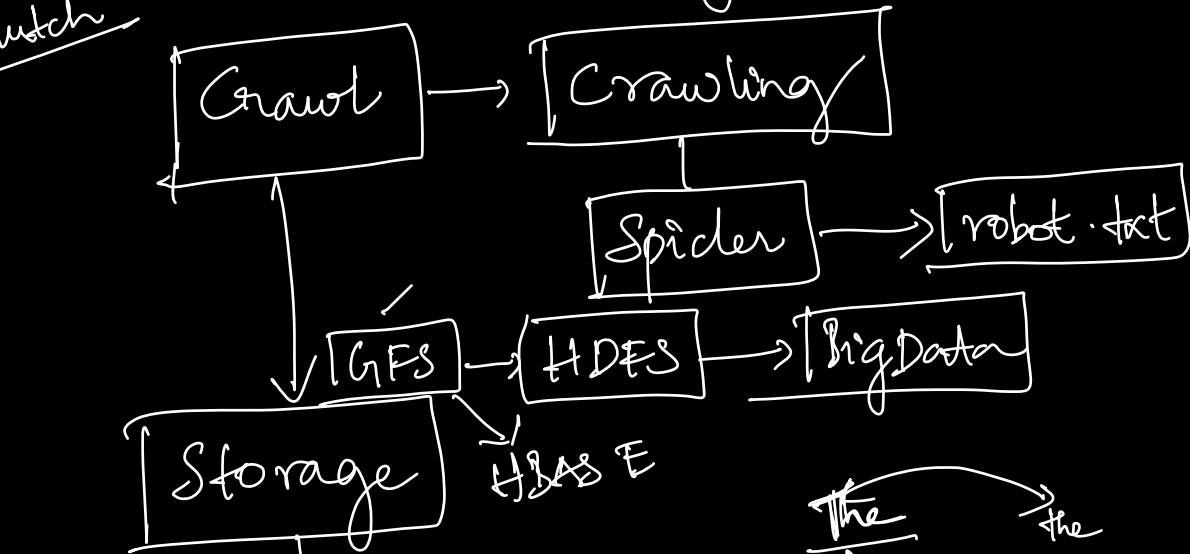


Search Engine

nutch



Pre processing

Indexing

Structured Representation

Query
↓
correct
↓
Relevant

↓ RDBS / NOSQL

- Parsing
- lowercase
- punctuation
- Special character
- Stop words
- Spell correction
- Count emoji
- Root word

porton stemming
rules

stemming
grammatical rules
Lemmatization

accurate
→
slow

run
site

running
sit

runs
sitting

stemming
stem

sit

Page Ranking (1000)
Larry Page
text → web
Video →
Images →
Ads →

→ Tokenization

I had enough breakfast \Rightarrow enough breakfast
breakfast was awesome \Rightarrow breakfast awesome
I had to take a trip \Rightarrow take trip
I went for a running \Rightarrow go run
he runs very fast \Rightarrow run very fast
he runs fast, while he runs he collapsed

run fast	run collapse
----------	--------------

n grams \Rightarrow 100 grams, 200 grams, 300 grams

Archives

enough breakfast

Chlorogram

→ [enough, breakfast]

bì gươm

[enough breaks)

r breakfast awesome

→ [breakfast, awesome]

[briefs an]

u take trip

$$\rightarrow [4 \text{ dec, 100}]$$

[fake trip]

✓ go run

$$\rightarrow [g, \frac{1}{\sqrt{2}}]$$

(g, n)

star ^{very} fast

→ [unzuf.]

[nun vry] [vry]

	enough	breakfast	running	at the	help	enough	breakfast	breakfast
1.	1	1	0	0	0	1	0	
2.	0	1	0	0	0	0	1	
3.								
4.								
5.								

Indexing

1. One hot encoding →

- a. Occurrences not there ^{for document} _{and corpus}
- b. orders are not preserved
- c. Context is not preserved

2. ~~at~~ Bag of Words

- a. Occurrences not there ^{for} _{corpus}
- b. orders are not preserved
- c. Context is not there

3. TF-IDF

4. Word embedding