Document structure:

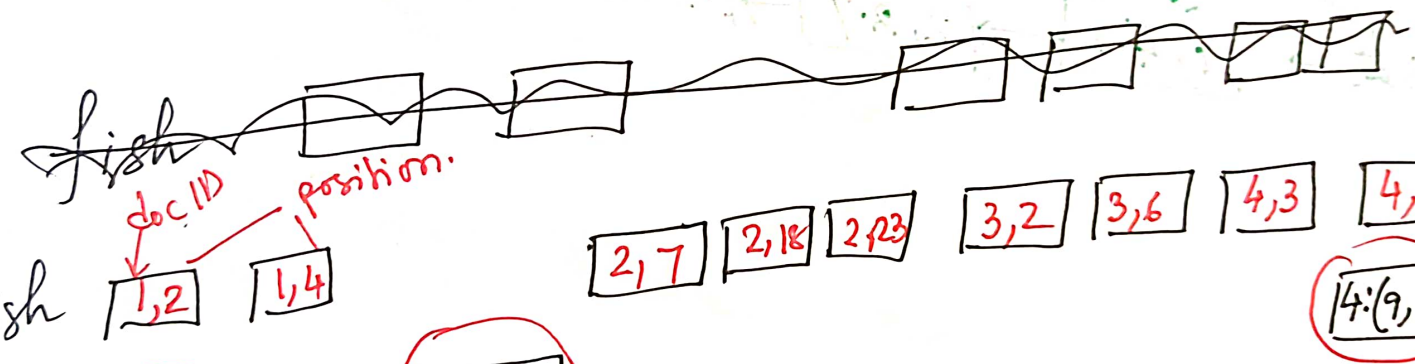
↳ fields (special)

↳ date.

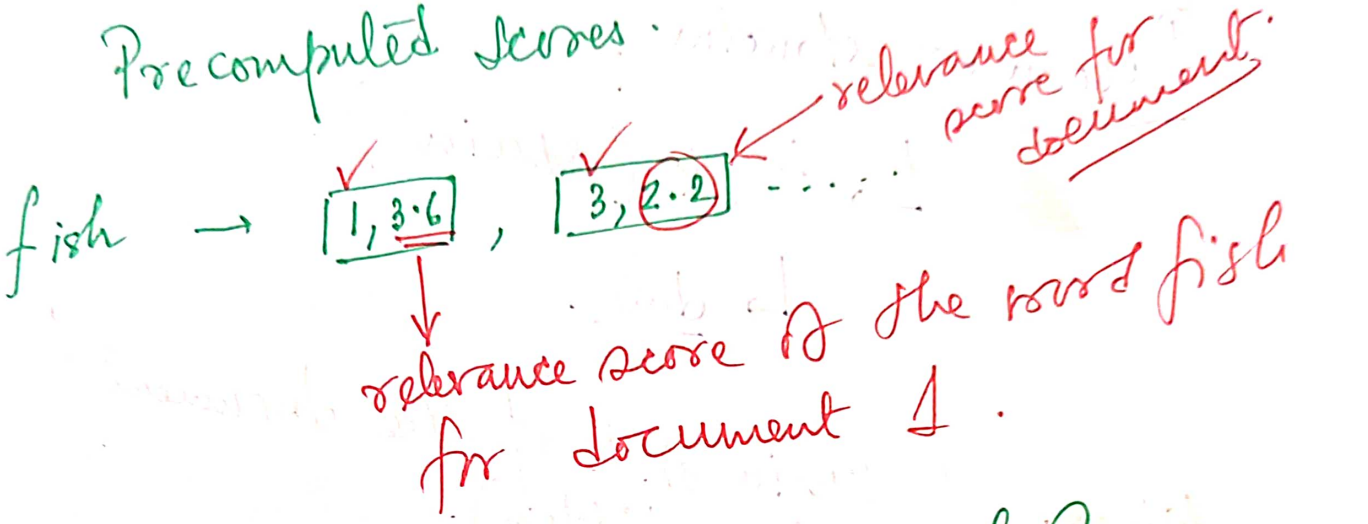
titles. (main title of the document / section titles).

Index these special fields separately.

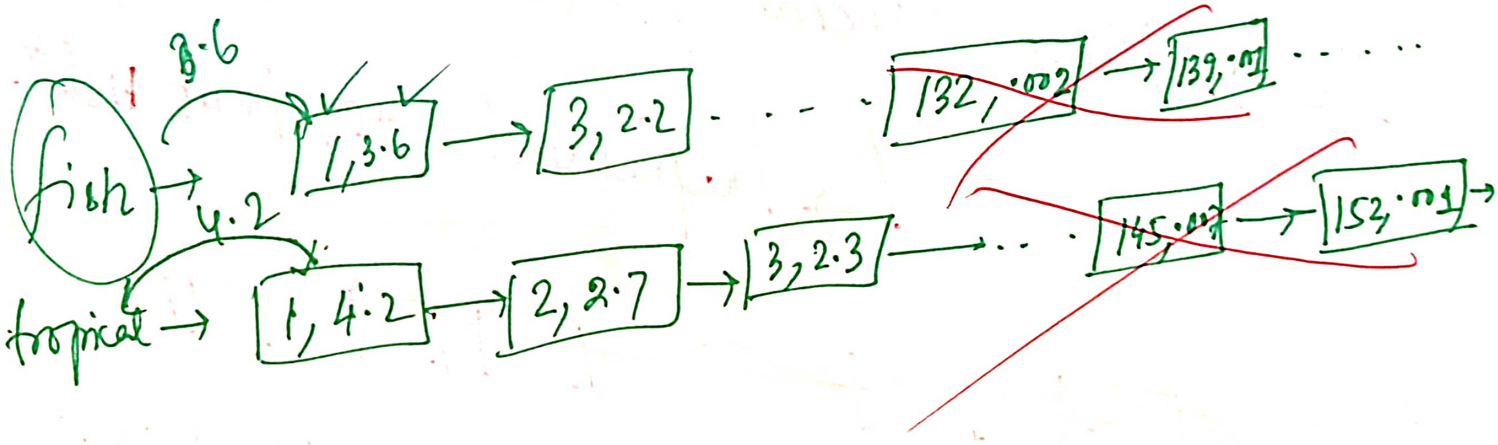
↳ Implement "extent lists".



Precomputed scores.



Merge → how does this help?



Abstract Model of Ranking for search / retrieval.

— → key words.

• Fred's Tropical Fish Shop is the best place to find tropical fish at low, low prices. Whether you're looking for a little fish or a big fish we've got what you need. We even have fake seaweed for your fish tank (and little surfboards too)

9.7 fish
4.2 tropical
22.1 tropical fish
8.2 seaweed
4.2 surfboards

Topical / keyword features.

Tropical fish
Query

5.2 fish
3.4 tropical
9.9 tropical fish
1.2 chichlids
0.7 barbs.

(R) Ranking function

14 Incoming links
3 days since last updated

Quality features.

1.2 incoming links.
0.2 update content

→ Doc → [9.7 4.2 22.1 ....] ← document feature (f_i)

Q → [5.2 3.4, 9.9 ....] ← query feature (g_i)

$$R(Q, D) = \sum_i g_i(Q) f_i(D)$$

1. tropical fish aquarium. [a]

Question(s): What is size of the result set for this query?

Query: "a b c" ⟶ how many pages will have all three terms.

f_a = frequency of the word 'a' in the corpus.

f_b = frequency of the word 'b' in the corpus.

f_c = frequency of the word 'c' in the corpus.

occurence of a, b, c are independent & N is the number of tokens in corpus what is the expected frequency of "a b c"

prob of a's occurrence = f_a/N

prob of b's " = f_b/N

prob of c's " = f_c/N

prob of occurrence of abc → ?

$$f_{a/N} \times f_{b/N} \times f_{c/N}$$

Expected frequency

$$f_{abc} = N \times \left(f_{a/N} \times f_{b/N} \times f_{c/N} \right)$$

$$= \left(f_a \cdot f_b \cdot f_c \right) / N^2$$

Independence assumption is bad.

Nodes	Doc freq.
tropical	120,900
fish	1,131,855
aquarium	26480
breeding	81885
tropical fish	18472
tropical aquarium	1921
fish breeding	~~6510~~ 36427
aquarium breeding	~~1722~~ 1848
tropical fish aquarium	~~3629~~ 1529
tropical fish breeding	3629

$$P(a \cap b \cap c) = P(a \cap b) \cdot P(c \mid a \cap b) \rightarrow \text{chain rule.}$$

$$\min \left[P(c \mid a), P(c \mid b) \right]$$

$$P(a \cap b \cap c) = P(a \cap b) \cdot \min \left[P(c \mid a), P(c \mid b) \right]$$

$f_{\underline{aq}\cap \overset{b}{\underline{fish}}\cap \overset{c}{tropical}}$

$= f_{aq\cap fish} \times \min\left[f_{aq\cap \overset{c}{tropical}}, f_{\overset{b}{fish}\cap \overset{c}{tropical}}\right] / f_{aq}$

$= \boxed{9722 \times 1921 / 26840}$

→ S documents
find out some C triplets. containing aquarium

Instead of processing 26480 docs

↳ 3000 docs.

↳ 258 cases where (tropical fish aquarium

$3000 \rightarrow 258$

$26480 \rightarrow \frac{258 \times 26480}{3000}$

$S = 3000$, $C = 258$ →