Document structure:
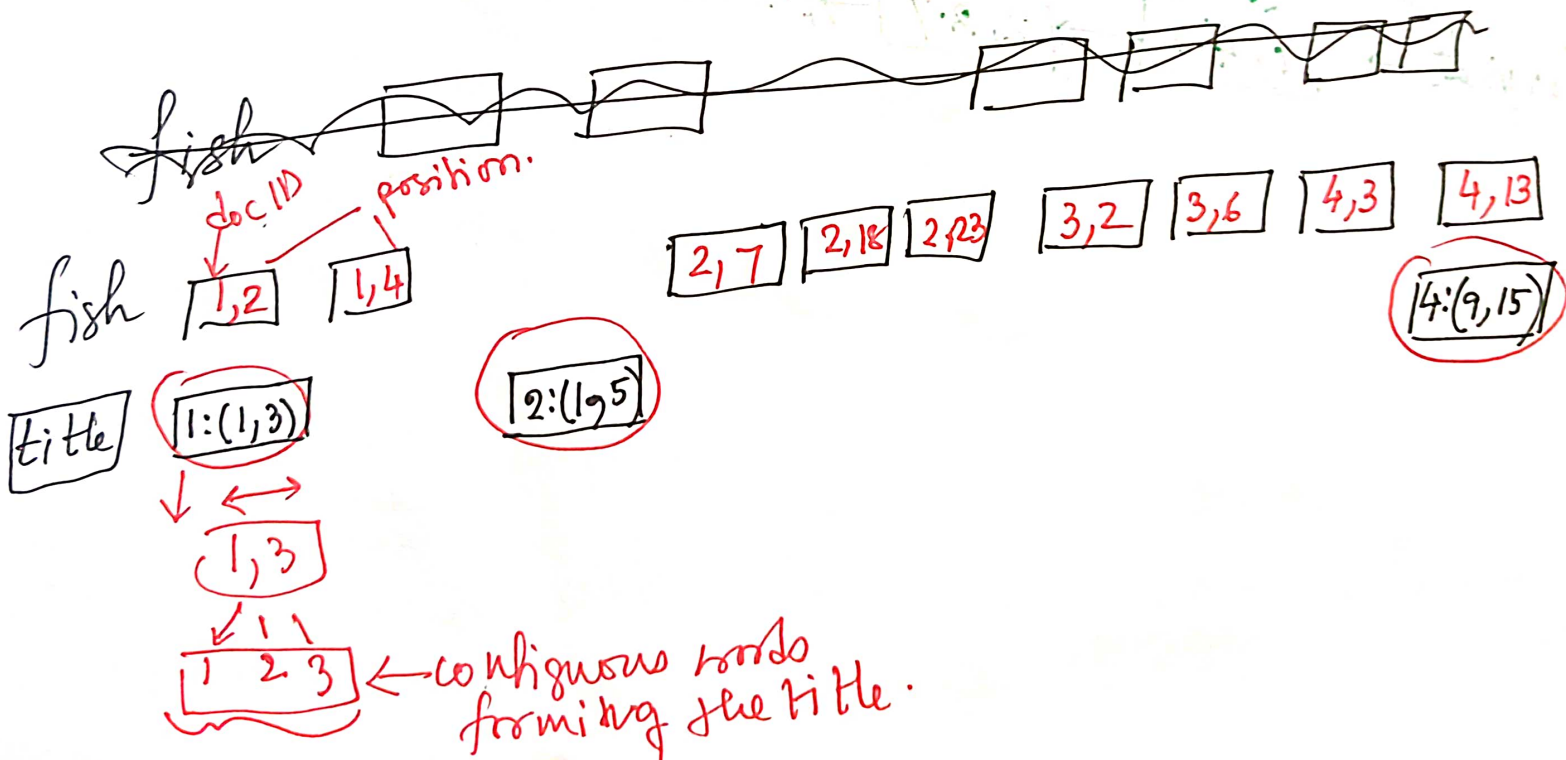↳ fields (special)
  ↳ date.
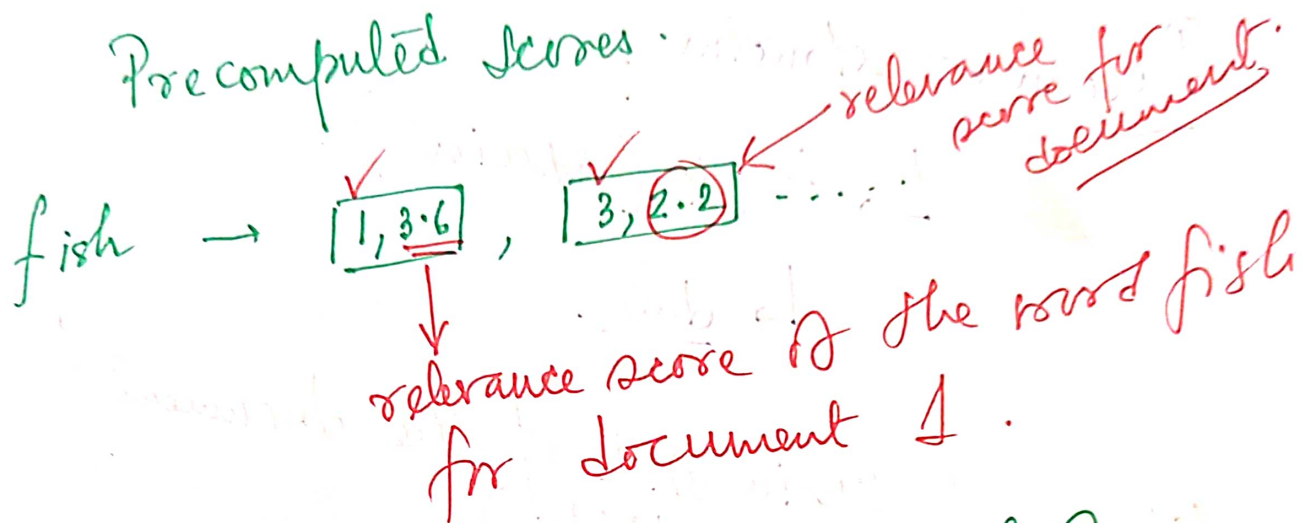
titles. (main title of the document /
          section titles).

Index these special fields separately

↳ Implement "extent lists".

fish

fish | 1,2 | | 1,4 | | 2,7 | | 2,18 | | 2,23 | | 3,2 | | 3,6 | | 4,3 | | 4,13 |
                                                                                                  | 4:(9,15) |

doc ID   position.

| 2:(1,5) |

| title | | 1:(1,3) |

| 1,3 |

| 1 | 2 | 3 | ←contiguous words
              forming the title.

Precomputed scores

fish $\rightarrow$ | 1, 3·6 | , | 3, 2·2 | · · · · · · $\leftarrow$ relevance score for document.

relevance score of the word fish for document 1.

Merge $\rightarrow$ how does this help?

fish $\rightarrow$ 3·6 | 1, 3·6 | $\rightarrow$ | 3, 2·2 | · · · · | 132, ·02 | $\rightarrow$ | 139, ·07 | · · · · · ·

fish $\rightarrow$ 4·2

tropical $\rightarrow$ | 1, 4·2 | $\rightarrow$ | 2, 2·7 | $\rightarrow$ | 3, 2·3 | $\rightarrow$ · · · | 145, ·07 | $\rightarrow$ | 153, ·04 | $\rightarrow$ · ·
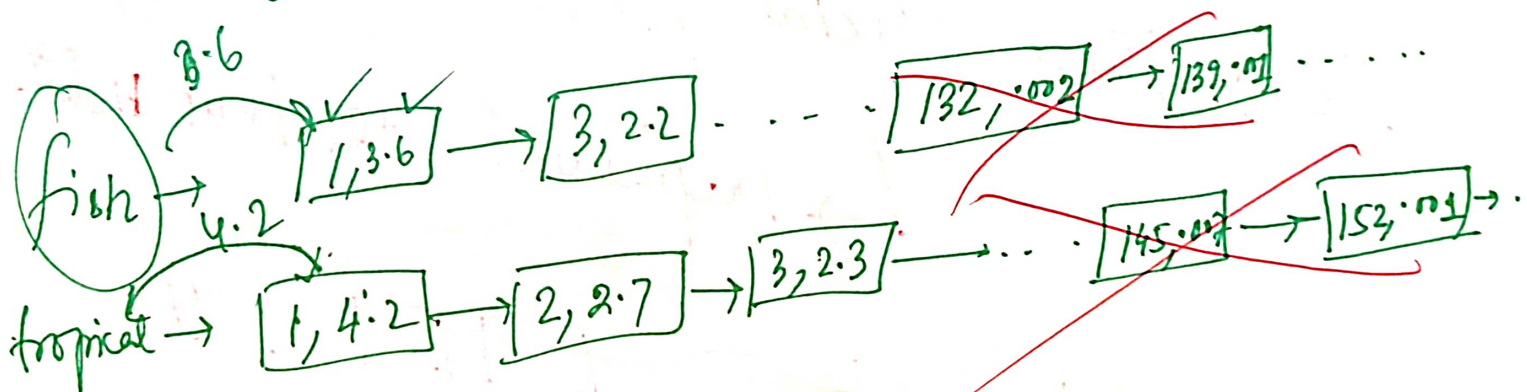
Abstract Model of Ranking for search / retrieval.

→ 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 surf boards too)

9.7 fish
4.2 tropical
22.1 tropical fish
8.2 seaweed
4.2 surf boards

tropical / seaweed features.

14 Incoming links
3 days since last updated

Quality features.

tropical fish Query

Ranking function (R)

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

1.2 incoming links.
0.2 update content

$$\text{Doc} \rightarrow [9.7 \ 4.2 \ 22.1 \dots] \leftarrow \text{document feature } (f_i)$$

$$Q \rightarrow [5.2 \ 3.4, \ 9.9 \dots] \leftarrow \text{query feature } (g_i)$$

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

1. tropical fish aquarium. $\boxed{a}$

Question(s): What is Size of the result set

for this query?

Query: "a b c" $\longrightarrow$ 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.

occurrence 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 a c's " $= f_c/N$

prob of occurrence of a bc $\rightarrow$?

$$f_a/N \times f_b/N \times f_c/N$$

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.

Nodo	Doc freq.
tropical	120,900
fish	1,131,855
aquarium	26480
breeding	81885
tropical fish	18472
tropical aquarium	1921
fish breeding	~~5510~~ 36427
acquarium breeding	~~8422~~ 1848
tropical fish aquarium	~~36429~~ 1529
tropical fish breeding	3629

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

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

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

$$f^{\frac{a}{2}}_{aq \cap fish \cap tropical}$$

$$= f_{aq \cap fish} \times \min \left[f^{\frac{a}{2}}_{aq \cap tropical}, \; f^{\frac{b}{2}}_{fish \cap tropical} \right] / f_{aq}$$

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

→ S documents
find out some C triplets. containing
aquarium

Instead of processing 26480 docs

↳ 3000 docs.

↳ 258 cases where
(tropical fish aquarium

$3000 \rightarrow 25$

$S = 3000$, $C = 258 \rightarrow$

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

20%