Web search basics (Chapter 19)

https://miro.com/app/board/o9J_lDUku7A=/

Definition 1 (Mark and Recapture)
Suppose that we could pick a random page from the index of E_1 and test whether it is in E_2's index and symmetrically, test whether a random page from E_2 is in E_1. These experiments give us the fraction x of the pages in E_1 that are also in E_2, and the fraction y of the pages in E_2 that are also in E_1. Let $|E_i|$ denote the size of the index E_i. Then

$$x|E_1| \approx y|E_2| \iff \frac{|E_1|}{|E_2|} \approx \frac{y}{x}.$$

Exercise 19/2

Two web search engines A and B each generate a large number of pages uniformly at random from their indexes. 30% of A's pages are present in B's index, while 50% of B's pages are present in A's index. What is the number of pages in A's index relative to B?

A .. $x = 3/10$

B .. $y = 5/10$

$$\frac{3}{10} \cdot |A| \approx \frac{5}{10} \cdot |B|$$

$$\frac{|A|}{|B|} = \frac{5}{3}$$

Exercise 19/3

Using shingling with k= 4 and the threshold 0.9 to decide whether the documents d_1 = "now is the time for all good men to come to the aid of their country", and d_2 = "now is the time for all good men to come to the aid of their party" are near-duplicates.

Definition 2 (Shingling)
By some estimates, as many as 40% of the pages on the Web are duplicates of other pages. Search engines try to avoid indexing multiple copies of the same content, to keep down storage and processing overheads.

Given a positive integer k and a sequence of terms in a document d, define the k-shingles of d to be the set of all consecutive sequences of k terms in d. As an example, consider the following text: "a rose is a rose is a rose". The 4-shingles for this text ($k = 4$ is a typical value used in the detection of near-duplicate web pages) are "a rose is a", "rose is a rose", and "is a rose is". Intuitively, two documents are near duplicates if the sets of shingles generated from them are nearly the same.

Let $S(d_j)$ denote the set of shingles of document d_j. The Jaccard coefficient measures the degree of overlap between the sets $S(d_1)$ and $S(d_2)$ as

$$J(S(d_1), S(d_2)) = \frac{|S(d_1) \cap S(d_2)|}{|S(d_1) \cup S(d_2)|}.$$

If the Jaccard index exceeds a preset threshold (say, 0.9), we declare documents d_1 and d_2 near-duplicates and eliminate one from indexing.

Since computing the Jaccard index between all pairs of documents is time-consuming, an estimate is often used, see Section 19.6 in the Manning book.

$S(d_1) = \{$"now is the time", "is the time for", ...$\}$

$S(d_2) = \{$"now is the time", ...$\}$

formula for no-duplicates

Set() data structure

no duplicates

$|S(d_2)| = |S(d_1)| = len(d_1) - k + 1 = 16 - 4 + 1 = 13$

d_1 = "now is the time for all good men to come to the aid of their country"

d_2 = "now is the time for all good men to come to the aid of their party"

$|S(d_1) \cap S(d_2)| = 12$

$|S(d_1) \cup S(d_2)| = 14$

$J(S(d_1), S(d_2)) = \frac{12}{14} = \frac{6}{7} \approx 0,86 < 0,9$