CS 124 — Lecture 13

Hashing → 64-bit hash ?

Duplicate / near-duplicates

≈ 30% of webpages

Dealing with near-duplicates

Sets of #'s

2 sets A, B, let each be of [redacted] size n

∈ 64-bit #'s

Resemblance (A, B) = $\frac{|A \cap B|}{|A \cup B|}$

1 if A = B

0 if disjoint

if ordered $O(n)$

$O(n \log n)$ for ordering

unordered $O(n^2)$

} too expensive for each document comparison

↓

change problem

↓

approximate resemblance

assume

A black box for random permutations

$\pi_3(x) \to y$

x	0	1	2	3	4	5	6	7
$\pi_3(x)$	2	4	3	1	7	6	5	0

$\pi_3(2) = 3$

permutation of a set, function applied to each element

$A = \{2, 4, 6\}$ $\pi_3(A) = \{3, 5, 7\}$ $\min \pi_3(A) = 3$

$\min(\pi_1(A)), \min(\pi_2(A)), \min(\pi_3(A)) \ldots \min(\pi_{100}(A))$

keep 100 #s associated with each set across all sets

→ calling card for each set

A calling card

$\min(\pi_1(A)), \min(\pi_2(A)), \min(\pi_3(A)) \ldots$

B calling card

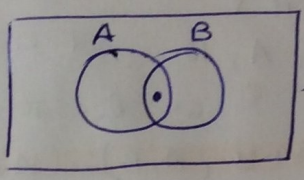
$\min(\pi_1(B)), \min(\pi_2(B)), \min(\pi_3(B)) \ldots$

Prob [$\min \pi_1(A) = \min \pi_1(B)$] = Resemblance (A, B)

$A = \{2, 4, 6\}$ $\pi_3(A) = \{3, 5, 7\}$ min $\pi_3(A) = 3$

$B = \{4, 5, 6\}$ $\pi_3(B) = \{5, 6, 7\}$ min $\pi_3(B) = 5$

every number has an equal chance to be the minimum of a permutation function. If a number is present in both documents (sets) and one of the "right" permutations is applied, the minima match



$$\frac{|A \cap B|}{|A \cup B|}$$

- can elt can be min in several π_3
- when minima match an elt is in both sets
- given a doc, every elt has equal chance to be min, πs are random

go through 100 πs to get an estimate of resemblance

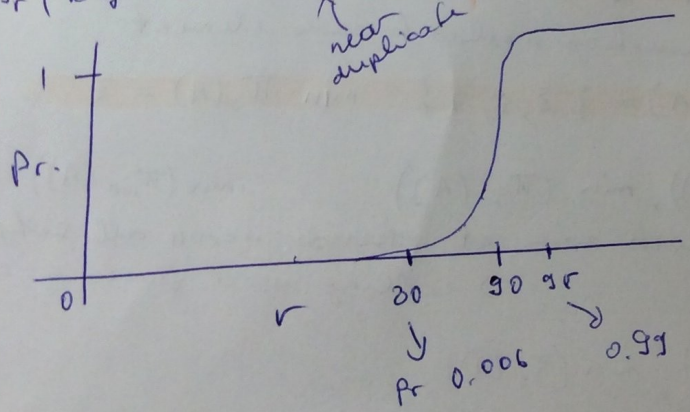
Suppose true resemblance is r

$\geq$ 90 matches → similar

$<$ 90 not

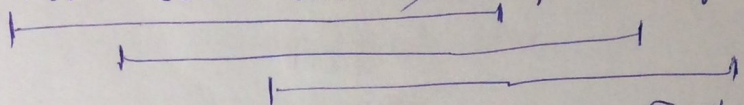
$Pr(\#\text{ of matches} = 90 \text{ when resemblance} = r) = \binom{100}{90} r^{90}(1-r)^{10}$

$Pr(\#\text{ of matches} \geq 90 \text{ when resemblance} = r) = \sum_{k=90}^{100} \binom{100}{k} r^k (1-r)^{100-k}$



Turn a document into a set of #s
# of semantic

Four score and seven years ago. ——



← takes some context into account

$k = 4$

then hash

shingling:

Syntactic vs. semantic