Radhika Patwari
18CS10062

Class Test - 3

11/4/2022

Q1. a) Use fingerprints to remove all but one copy of identical documents.

Remove common HTML tags and integers from the single computation, to eliminate shingles that occur very commonly in documents without telling us anything about duplication.

Use union-find algorithms to create clusters that contain documents that are similar.

Calculate d_i and d_j to check similarity.

Compute # of shingles in common for any pair of documents whose sketches have any members in common.

For list $\langle x_\pi, i, d_i \rangle$, sorted by (x_π, i) pairs, generate j for all i where (x_π, i) is present in both their sketches. For each pair i, j, get a non-zero sketch overlap, a count of # of (x_π, i) values in

Use a threshold to find pairs (i, j) that have overlapping sketches.

Run union find to group documents into near duplicate 'syntactic clusters'.

In this way, we can use single-link clustering algorithm.

. Radhika Patwari
18CS10062

Q5. a) No, we cannot say that documents are duplicates as jaccard similarity is bag of words based approach and doesn't consider word order.

eg. " I live in Delhi but I often visit Gurgaon. "
" I live in Gurgaon but I often visit Delhi. "

have jaccard similarity = 1 but are different in meaning

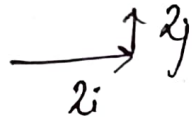
(b) Log function allows damping effect.
Any function that is monotonic and grow slowly compared to a linear function can be used.

eg. $x^{1/k}$ where $k = 1$

(c) If we normalise w_a, w_b, w_c before computing $(w_b - w_a + w_c)$ then the result will be unnormalized.

Adding → Normalising

$p = 2i$
$q = 2j$

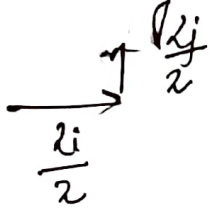


$$p+q = \frac{2i+2j}{2\sqrt{2}}$$

$$= \frac{i}{\sqrt{2}} + \frac{j}{\sqrt{2}}$$

$$|p+q| = 1$$

Normalising → Adding individually



$$\hat{p}+\hat{q} = \frac{i+j}{\sqrt{2}}$$

$$|\hat{p}+\hat{q}| = 1$$

Radhika Patwari
18CS10062

Q3. a) Heap's law = $(M = kT^b)$

= estimating vocabulary size given collection size

M = vocabulary size
T = # of tokens in collection
$b \approx 0.5$
$30 \leq k \leq 100$

Zipf's law = the ith most frequent term has frequency proportional to $1/i$.

= estimating collection frequency (cf) given $cf_i \propto \frac{1}{i}$ for term rank i

Both laws are power laws but Heap law talks about size of vocabulary and zipfs law talks about distribution of the vocabulary

---

Q3. c) (i) $\frac{1}{8}\left(\frac{1}{1} + \frac{1}{1} + \frac{3}{9} + \frac{4}{11} + \frac{5}{15} + \frac{6}{20} + \frac{7}{21} + \frac{8}{22}\right)$

↳ MAP ↳ common pair

(Assuming relevant docs are at rank 21, 22)

↓

we also disregard relative relevance score & just consider relevant / non relevant.

min MAP (relevant doc are at 9999 and 10,000)

$$\frac{1}{8}\left(\frac{2}{2} + \frac{3}{9} + \frac{4}{11} + \frac{1}{3} + \frac{3}{10}\right) + \frac{1}{8}\left(\frac{7}{9999} + \frac{8}{10,000}\right)$$

Radhika Patwari
18CS10062

Q3. c)(ii) Only NDCG can be computed as we can just determine ranking of documents and not the absolute relevance.

To compute it we can merge pairwise rankings (if possible, i.e, common elements exist) to recover total ordering of docs and compute NDCG

Q3.(b) Brute force approach

Keep a backup copy of every auxillary index I_i in the disk as $I_i$backup.

Now while retrieving, we will check if I_i is present. If not present (Ii is deleted), we can use I_i_backup. If it is not present, then I_i does not exist. If I_i_backup is present, then it implies I_i is deleted. We use I_i_backup.

This method has search overhead but requires 2 if conditions only.

Radhika Patwari
18CS10062

Q4.(a) $O(R|q, \vec{x}) = \frac{P(R=1 \mid q, \vec{x})}{P(R=0, q, \vec{x})}$

$= \frac{P(R=1|q) \cdot P(\vec{x} \mid R=1, q)}{P(R=0|q) \cdot P(\vec{x} \mid R=0, q)}$

Now if probability of term x_i is dependent only on previous term x_{i-1}, then this is a Markov assumption.

∴ Under Markov assumption,

$O(R|q, \vec{x}) = O(R|q) \cdot \frac{P(x_1 \mid R=1, q)}{P(x_1 \mid R=0, q)}$

$= \prod_{i=2}^{n} \frac{P(x_i \mid R=1, q, x_{i-1})}{P(x_i \mid R=0, q, x_{i-1})}$

Q4.(b) Let vocabulary size = $|V|$

⇒ BIM model

Every term / word x_i has 2 parameters

γ_i and ρ_i

∴ Total # of parameters = $2 = |V|$

→ Bigram Model

$P(x_i|x_{i-1})$ and $r(x_i|x_{i-1})$ are required for each term pair.

∴ Total # of parameters = $2 \times |V|^2$

Radhika Patwari
18CS10062

Q2.(b) Semantic relations are captured by word pairs.

Sharing common offset vectors (semantic relations are captured by certain direction in the vector space)

Another implicit assumption is that a word isn't ambiguous (no polysemy)

eg. Bank - money (place - product-analogy)

might not match with

BMW - car

because BMW is clearly an automobile company but bank can be a place or riverbank.