Hemin Patel - Hw 8

(1)

A) Consider table of term frequencies for 3 docs. Compute tf-idf weights for each terms. Corpus size = 100K documents.

term	df	idf	
car	18165	1.74	$1+\log(\frac{100K}{18165})$
auto	6723	2.17	$1+\log(\frac{100K}{6723})$
insurance	19241	1.72	$1+\log(\frac{100K}{19241})$
best	25235	1.60	$1+\log(\frac{100K}{25235})$

weights

	Doc1	Doc2	Doc3
car	46.98	6.96	41.76
auto	6.51	71.61	0
insurance	0	56.76	49.88
best	22.4	0	27.2

$\text{tf-idf} = tf_{t,d} \cdot idf_t$

$27 \cdot 1.74 \mid 4 \cdot 1.74 \mid 24 \cdot 1.74$

B) Compute normalized doc vectors for each doc

- Doc1: 46.98, 6.51, 0, 22.4

  $|doc1|: \sqrt{46.98^2 + 6.51^2 + 22.4^2} = \underline{52.45}$

  $Doc1/|doc1|: \boxed{0.90, 0.12, 0, 0.43}$

- Doc2: 6.96, 71.61, 56.76, 0

  $|doc2|: \sqrt{6.96^2 + 71.61^2 + 56.76^2} = \underline{91.64}$

  $Doc2/|doc2|: \boxed{0.076, 0.78, 0.62, 0}$

- Doc 3: 41.76, 0, 49.88, 27.2

  $|Doc3|: \sqrt{41.76^2 + 49.88^2 + 27.2^2} = \underline{70.51}$

  $Doc3/|Doc3|: \boxed{0.59, 0, 0.71, 0.40}$

	Doc1	Doc2	Doc3
car	0.90	0.076	0.59
auto	0.12	0.78	0
insur	0	0.62	0.71
best	0.43	0	0.40

2) normal: url 1: normal.org
2: merrian-webster.com
3: dictionary.com

form: url 1: google.com/forms/about
2: merrian-webster.com
3: form.com

cmd + F w/o space

Hemin - HW 8 - cont.

① c) compute cosine similarity between each pair of doc

- DOC1 · DOC2: (46.98 · 6.96) + (6.51 · 71.61) + / + / = 793.1619

$|DOC1| = 52.45$, $|DOC2| = 91.64$] $\cos(DOC1, DOC2) = \frac{793.1619}{(52.45 \cdot 91.64)} = \boxed{0.1650}$

- DOC2 · DOC3: (6.96 · 41.76) + / + (56.76 · 49.88) + / = 3121.8384

$|DOC2| = 91.64$, $|DOC3| = 70.51$] $\cos(DOC2, DOC3) = \frac{3121.8384}{(91.64 \cdot 70.51)} = \boxed{0.4831}$

- DOC1 · DOC3: (46.98 · 41.76) + / + / + (22.4 · 27.2) = 2571.1648

$|DOC1| = 52.45$, $|DOC3| = 70.51$] $\cos(DOC1, DOC3) = \frac{2571.1648}{(52.45 \cdot 70.51)} = \boxed{0.6952}$

② let google query Q = normal form | 3 URLs, Rank approx cos similarity

N = 1 trillion

term	df	idf
normal	2,440,000,000	5.61261
form	4,970,000,000	5.30364

$1 + \log \frac{1 \text{ tril}}{2{,}440 \text{ mil}}$

query Q

Norm(idf)
7.7220

Term Frequency

	url 1	url 2	url 3
normal	28	119	38
form	29	151	13

weights

	url 1	url 2	url 3
normal	157.1531	667.9006	213.2792
form	153.8056	800.8496	68.9473

$\text{tf-idf} = tf_{t,d} \cdot idf_t$

vectors + Normalization

	url 1	url 2	url 3
normal	17.9397	21.4666	18.6841
form	16.9026	20.7030	15.0545
NORM	24.6481	29.8233	23.9944

refers to weights

$(1 + \log(tf) \cdot (1 + \log(\frac{N}{df}))) \rightarrow$ or idf

$\rightarrow 1 + \log(157.1531) \cdot (1 + \log \frac{1 \text{ tril}}{2{,}440 \text{ mil}})$

Ex: $\sqrt{17.9397^2 + 16.9026^2}$

Hemin HW 8 - cont.

(2) Cosine similarity w/ query Q & 3 web pages

- URL 1 • Q: $(17.9397 \cdot 5.61261) + (16.9026 \cdot 5.30364) = 190.3338$

$$\left.\begin{array}{l} |URL\ 1| = 24.6481 \\ |Q| = 7.7220 \end{array}\right] \cos(URL\ 1, Q) = \frac{190.3338}{(24.6481 \cdot 7.7220)} = 1.000006$$

- URL 2 • Q: $(21.4666 \cdot 5.61261) + (20.7030 \cdot 5.30364) = 230.2849$

$$\left.\begin{array}{l} |URL\ 2| = 29.8233 \\ |Q| = 7.7220 \end{array}\right] \cos(URL2, Q) = \frac{230.2849}{(29.8233 \cdot 7.7220)} = 0.999954$$

- URL 3 • Q: $(18.6841 \cdot 5.61261) + (15.0545 \cdot 5.30364) = 184.7102$

$$\left.\begin{array}{l} |URL\ 3| = 23.9944 \\ |Q| = 7.7220 \end{array}\right] \cos(URL\ 3, Q) = \frac{184.7102}{(23.9944 \cdot 7.7220)} = 0.996899$$