CS336 HW8

1. a. IDF = $1 + \log_{10}$ (No. of documents / number of docs word occurs)

IDF(Covid) = $\log_{10}(4/4)+1 = 1$
IDF(Vaccine) = $\log_{10}(4/4)+1 = 1$
IDF(effective) = $\log_{10}(4/3)+1 = 1.125$
IDF(pfizer) = $\log_{10}(4/3)+1 = 1.125$
IDF(moderna) = $\log_{10}(4/2)+1 = 1.301$
IDF(trial) = $\log_{10}(4/2)+1 = 1.301$
IDF(astrazeneca) = $\log_{10}(4/1)+1 = 1.602$
IDF(halt) = $\log_{10}(4/1)+1 = 1.602$
IDF(race) = $\log_{10}(4/1)+1 = 1.602$

b.

TF:	Headline 1	2	3	4
CoVid	1	1	1	1
vaccine	1	1	1	1.3
effective	1	1	0	1
pfizer	1	1	0	1
moderna	0	1	0	1
trial	0	1	1	0
astrazeneca	0	0	1	0
halt	0	0	1	0
race	0	0	0	1

TF • IDF:	Headline 1	2	3	4
Covid	1	1	1	1
Vaccine	1	1	1	1.3
effective	1.125	1.125	0	1.125
pfizer	1.125	1.125	0	1.125
moderna	0	1.301	0	1.301
trial	0	1.301	1.301	0
astrazeneca	0	0	1.602	0
halt	0	0	1.602	0
race	0	0	0	1.602

C. Normalization $\|\vec{x}\| = \sqrt{\sum x_i^2}$ $\vec{x} = \frac{\vec{x}}{\|\vec{x}\|}$

	Norm 1	Norm 2	Norm 3	Norm 4
L2-Norm	2.128	2.92	2.971	3.08
Covid	0.47	0.34	0.337	0.32
vaccine	0.47	0.34	0.337	0.42
effective	0.53	0.385	0	0.37
pfizer	0.53	0.385	0	0.37
moderna	0	0.45	0	0.42
trial	0	0.45	0.438	0
astrazeneca	0	0	0.54	0
halt	0	0	0.54	0
race	0	0	0	0.52

d.

Term	Q	Norm Q
covid	1	0.55
vaccine	1	0.55
effective	1.125	0.622
pfizer	0	0
moderna	0	0
trial	0	0
astrazeneca	0	0
halt	0	0
race	0	0

e.

		similarity
Most similar:	Headline 1	0.847
	Headline 4	0.637
	Headline 2	0.613
Least similar:	Headline 3	0.371

2. A.

	Doc 1	Doc 2	Doc 3
car	4.23	2.79	4.14
auto	3.21	5.47	0
insurance	0	4.32	4.22
best	3.43	0	3.56

B.

Term	Norm 1	Norm 2	Norm 3
L2 Norm	4.86	7.51	6.9
car	0.87	0.372	0.6
auto	0.66	0.73	0
insurance	0	0.58	0.612
best	0.71	0	0.516

C.

	Similarity
Doc1 - Doc2	0.805
Doc2 - Doc3	0.578
Doc1 - Doc3	0.888

3

TF:	URL 1	URL 2	URL 3	URL 4	URL 5
TF-IDF	25	9	6	47	10
Cosine	31	34	2	14	14
similarity	31	41	7	13	31

term	DF
TF-IDF	3,930,000
Cosine	29,200,000
Similarity	183,000,000

$$TF \times IDF = \left(1 + \log(tf)\right) \times \left(1 + \log\left(\frac{100 \text{ trillion}}{df}\right)\right)$$

ex: $(1 + \log(25)) \times (1 + \log(1 \times 10^{14} / 3{,}930{,}000))$

TF x IDF	URL 1	URL 2	URL 3	URL 4	URL 5
TF-IDF	20.16	16.43	14.95	22.46	16.81
Cosine	18.77	19.07	9.8	16.17	16.17
Similarity	16.79	17.6	12.43	14.24	16.79

Normalize	Norm 1	Norm 2	Norm 3	Norm 4	Norm 5
L2 Norm	32.26	30.71	21.77	31.12	28.74
TF-IDF	0.625	0.535	0.687	0.722	0.585
Cosine	0.582	0.621	0.45	0.52	0.563
Similarity	0.52	0.573	0.571	0.458	0.584

X_i = TF x IDF

ex. $\sqrt{0.625^2 + 0.582^2 + 0.52^2} = 32.26 \Rightarrow \frac{20.16}{32.26} = 0.625$

L2 Norm = $\|X\| = \sqrt{\sum_{i=1} X_i^2}$

unit vector: $\hat{X} = \frac{\vec{X}}{\|X\|}$

Normalize Query	Q	Norm Q
TF-IDF	1	0.578
Cosine	1	0.578
Similarity	1	0.578

$\|X\| = \sqrt{1^2+1^2+1^2} = 1.73$ $\quad \bar{x} = 1 \quad \therefore \quad \frac{1}{1.73} = 0.578$

Cosine similarity is found by $\sum_{i=1}^{\wedge} X_i Y_i$

So we multiply the Normi * Norm Q

and add their summations

Cosine Similarity	Norm Q	Norm 1	Norm 2	Norm 3	Norm 4	Norm 5
TF-IDF	0.578	0.625	0.535	0.687	0.722	0.585
Cosine	0.578	0.582	0.621	0.45	0.52	0.563
Similarity	0.578	0.52	0.573	0.571	0.458	0.584

Q-URL1 Similarity : $\sum_{i=1}^{\wedge} X_i Y_i = (0.578 \cdot 0.625) + (0.578 \cdot 0.582) + (0.578 \cdot 0.52)$

Q-URL1 Similarity = 0.998206

Q-URL2 Similarity = 0.999362

Q-URL3 Similarity = 0.987224

Q-URL4 Similarity = 0.9826

Q-URL5 Similarity = 1.001096

		Similarity
Most similar URL to Query:	URL 5	1.001096
	URL 2	0.999362
	URL 1	0.998206
	URL 3	0.987224
least similar:	URL 4	0.9826