ANS 3

a) $tfidf = tf \cdot \log\left(\frac{N}{df}\right)$

Now, removing stop words such as

d_1 = dog ate homework

d_2 = cat ate homework

d_3 = dog cat ate hot dog

d_4 = dog cat wrote homework

d_5 = yesterday hot day

→ $idf = \log(N/df)$

here $N = 5$ where N = number of documents

N/df (for all words)	$idf = \log(N/df)$
ate - 5/3	0.51
cat - 5/2	0.92
day - 5/1	1.61
dog - 5/4	0.22
homework - 5/3	0.51
hot - 5/1	1.61
wrote - 5/1	1.61
yesterday - 5/1	1.61

TF:

	ate	cat	day	dog	homework	hot	wrote	yesterday
d1	1	0	0	1	1	0	0	0
d2	1	1	0	0	1	0	0	0
d3	1	1	0	2	0	0	0	0
d4	0	1	0	1	1	0	1	0
d5	0	0	1	0	0	1	0	1

- Now we multiply TF with IDF

	ate	cat	day	dog	homework	hot	wrote	yesterday
d1	0.51	0	0	0.22	0.51	0	0	0
d2	0.51	0.92	0	0	0.51	0	0	0
d3	0.51	0.92	0	0.44	0	0	0	0
d4	0	0.92	0	0.22	0.51	0	1.61	0
d5	0	0	1.61	0	0	1.61	0	1.61

NOW we can rewrite our results (terms vs documents)

terms \ documents					
	0.51	0.51	0.51	0	0
	0	0.92	0.92	0.92	0
	0	0	0	0	1.61
	0.22	0	0.44	0.22	0
	0.51	0.51	0	0.51	0
	0	0	0	0	1.61
	0	0	0	1.61	0
	0	0	0	0	1.61

b)

1. cosine similarity betⁿ D_1 & D_2

$$Sim(d_1, d_2) = \frac{\vec{d_1} \cdot \vec{d_2}}{|\vec{d_1}||d_2|}$$

$$= 2/3$$

$$= 0.67$$

here D_1 is more similar to D_2

2. cosine similarity (D_1 & D_5)

$$Sim(d_1, d_5) = \frac{\vec{d_1} \cdot \vec{d_5}}{|\vec{d_1}| \cdot |\vec{d_5}|}$$

$$= 0/3$$

$$= 0$$

here we can say that D_1 & D_5 are totally different.