③ (a)

(lexicographical order) $N = 5$

term	d_f	$idf = \ln(\frac{N}{df})$	$tf(d_1)$	$tf(d_2)$	$tf(d_3)$	$tf(d_4)$	$tf(d_5)$	$tf \cdot idf(d_1)$	$tf \cdot idf(d_2)$	$tf \cdot idf(d_3)$	$tf \cdot idf(d_4)$	$tf \cdot idf(d_5)$
and	1	$\ln(\frac{5}{1}) = 1.61$	0	0	1	0	0	$0 * \ln(\frac{5}{1}) = 0$	$0 * \ln(\frac{5}{1}) = 0$	$1 * \ln(\frac{5}{1}) = 1.61$	$0 * \ln(\frac{5}{1}) = 0$	$0 * \ln(\frac{5}{1}) = 0$
cat	4	$\ln(\frac{5}{4}) = 0.22$	0	1	1	1	2	$0 * \ln(\frac{5}{4}) = 0$	$1 * \ln(\frac{5}{4}) = 0.22$	$1 * \ln(\frac{5}{4}) = 0.22$	$1 * \ln(\frac{5}{4}) = 0.22$	$2 * \ln(\frac{5}{4}) = 0.45$
dog	3	$\ln(\frac{5}{3}) = 0.51$	1	0	2	1	0	$1 * \ln(\frac{5}{3}) = 0.51$	$0 * \ln(\frac{5}{3}) = 0$	$2 * \ln(\frac{5}{3}) = 1.02$	$1 * \ln(\frac{5}{3}) = 0.51$	$0 * \ln(\frac{5}{3}) = 0$
eat	3	$\ln(\frac{5}{3}) = 0.51$	1	1	1	0	0	$1 * \ln(\frac{5}{3}) = 0.51$	$1 * \ln(\frac{5}{3}) = 0.51$	$1 * \ln(\frac{5}{3}) = 0.51$	$0 * \ln(\frac{5}{3}) = 0$	$0 * \ln(\frac{5}{3}) = 0$
homework	2	$\ln(\frac{5}{2}) = 0.92$	1	1	0	0	0	$1 * \ln(\frac{5}{2}) = 0.92$	$1 * \ln(\frac{5}{2}) = 0.92$	$0 * \ln(\frac{5}{2}) = 0$	$0 * \ln(\frac{5}{2}) = 0$	$0 * \ln(\frac{5}{2}) = 0$
hot	1	$\ln(\frac{5}{1}) = 1.61$	0	0	1	0	0	$0 * \ln(\frac{5}{1}) = 0$	$0 * \ln(\frac{5}{1}) = 0$	$1 * \ln(\frac{5}{1}) = 1.61$	$0 * \ln(\frac{5}{1}) = 0$	$0 * \ln(\frac{5}{1}) = 0$
play	2	$\ln(\frac{5}{2}) = 0.92$	0	0	0	1	1	$0 * \ln(\frac{5}{2}) = 0$	$0 * \ln(\frac{5}{2}) = 0$	$0 * \ln(\frac{5}{2}) = 0$	$1 * \ln(\frac{5}{2}) = 0.92$	$1 * \ln(\frac{5}{2}) = 0.92$
the	5	$\ln(\frac{5}{5}) = 0$	1	1	3	2	2	$1 * \ln(\frac{5}{5}) = 0$	$1 * \ln(\frac{5}{5}) = 0$	$3 * \ln(\frac{5}{5}) = 0$	$2 * \ln(\frac{5}{5}) = 0$	$2 * \ln(\frac{5}{5}) = 0$
with	2	$\ln(\frac{5}{2}) = 0.92$	0	0	0	1	1	$0 * \ln(\frac{5}{2}) = 0$	$0 * \ln(\frac{5}{2}) = 0$	$0 * \ln(\frac{5}{2}) = 0$	$1 * \ln(\frac{5}{2}) = 0.92$	$1 * \ln(\frac{5}{2}) = 0.92$

* here $tf(d_n)$ signifies tf for a particular document n.
* All values rounded off to 2 decimals.
* log has been written as ln i.e. log to the base e.

All terms (words) are written in the Lexicographic order in the first column.

d_f = It is the document frequency i.e. no. of documents in the collection containing the term.

t_f = It is the frequency (count) of a term in document.

N = Total number of documents in the collection.

Since we have 5 documents d_1, d_2, d_3, d_4, d_5. Hence $N = 5$

$$t_f idf = t_f \cdot \ln\left(\frac{N}{d_f}\right)$$

Here ln( ) represent log to the base e.

All calculations & values are mentioned in the Table.

Eg. For the term "and", $d_f = 1$, as there is only 1 document i.e. d_3, in which this term is present. idf can be easily calculated using $\ln\left(\frac{N}{d_f}\right)$, if we know the value of d_f.

$t_f(d_1)$ represents t_f for document d_1. It is 0 as the term "and" is absent in document d_1 i.e. it is present 0 times in the document d_1. Similarly all calculations are done, as shown in table.

So final document vectors using the tfidf weights are —

$\vec{d_1} = [0, 0, 0.51, 0.51, 0.92, 0, 0, 0, 0]$

$\vec{d_2} = [0, 0.22, 0, 0.51, 0.92, 0, 0, 0, 0]$

$\vec{d_3} = [1.61, 0.22, 1.02, 0.51, 0, 1.61, 0, 0, 0]$

$\vec{d_4} = [0, 0.22, 0.51, 0, 0, 0, 0.92, 0, 0.92]$

$\vec{d_5} = [0, 0.45, 0, 0, 0, 0, 0.92, 0, 0.92]$

③ (b) Cosine Similarity between vectors d_x & d_y ⇒

$$sim(d_x, d_y) = \frac{d_x \cdot d_y}{\|d_x\| \, \|d_y\|}$$

Here $d_x \cdot d_y$ represents dot product of d_x & d_y.
$\|d_x\|$ represents square root of sum of squares of "tf-idf"s of different terms present in the document d_x. Similar meaning is for $\|d_y\|$.

So cosine similarity between between vectors for documents d_1 & d_2 ⇒

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

$d_1 \cdot d_2$ as per values taken Table

$= (0*0) + (0*0.22) + (0.51*0) + (0.51*0.51) + (0.92*0.92) + (0*0) + (0*0) + (0*0) + (0*0)$

$= 0 + 0 + 0 + 0.2601 + 0.8464 + 0 + 0 + 0 + 0$

$= 1.1065$

$\|d_1\| = \sqrt{0^2 + 0^2 + (0.51)^2 + (0.51)^2 + (0.92)^2 + 0^2 + 0^2 + 0^2 + 0^2}$

$= \sqrt{0 + 0 + 0.2601 + 0.2601 + 0.8464 + 0 + 0 + 0} = \sqrt{1.3666}$

$\|d_2\| = \sqrt{0^2 + (0.22)^2 + 0^2 + (0.51)^2 + (0.92)^2 + 0^2 + 0^2 + 0^2 + 0^2}$

$= \sqrt{0 + 0.0484 + 0 + 0.2601 + 0.8464 + 0 + 0 + 0 + 0} = \sqrt{1.1549}$

So $sim(d_1, d_2) = \frac{1.1065}{\sqrt{1.3666} \times \sqrt{1.1549}} = 0.88$

Similarity Cosine Similarity between vectors for documents d_1 & d_4 ⇒

$$sim(d_1, d_4) = \frac{d_1 \cdot d_4}{||d_4|| \cdot ||d_4||}$$

$d_1 \cdot d_4$ as per values taken from table ⇒

$= (0*0) + (0*0.22) + (0.51*0.51) + (0.51*0) + (0.92*0) + (0*0) + (0*0) + (0*0.92) + (0*0) + (0*0.92)$

$= 0 + 0 + 0.2601 + 0 + 0 + 0 + 0 + 0 + 0 + 0$

$= 0.2601$

$||d_1||$ as already calculated is $\sqrt{1.3666}$

$$||d_4|| = \sqrt{0^2 + (0.22)^2 + (0.51)^2 + 0^2 + 0^2 + 0^2 + (0.92)^2 + 0^2 + (0.92)^2}$$

$$= \sqrt{0 + 0.0484 + 0.2601 + 0 + 0 + 0 + 0.8464 + 0 + 0.8464}$$

$$= \sqrt{2.0013}$$

So $sim(d_1, d_4) = \frac{0.2601}{\sqrt{1.3666} \times \sqrt{2.0013}} = 0.16$

More the value of Similarity (Sim), lesser the angle between the two vectors & more is the similarity between them.

Since $sim(d_1, d_2) > sim(d_1, d_4)$

Hence d_1 is more similar to d_2.