

## Introduction to Data Science

### Assignment No. 04



**Name:** Mashail Jannat

**Roll No:** FA21-BSE-027

**Section:** C

**Submitted to:** Sir Muhammad Sharjeel

**Date:** 11th December 2023

COMSATS University Islamabad, Lahore Campus

## Question no. 01 :-

(1) Bow :-

Vocabulary = 'data', 'science', 'is', 'one',  
'of', 'the', 'most', 'important',  
'courses', 'in', 'computer',  
'this', 'best', 'scientists',  
'perform', 'analysis'.

data science is one of the most important.

S1	1	2	1	1	1	1	1	1
S2	1	1	1	1	1	0	0	
S3	2	0	0	0	0	1	0	0

courses in computer this best scientists perform

S1	1	1	1	0	0	0	0	0
S2	1	0	0	1	1	0	0	
S3	0	0	0	0	0	1	1	

analysis

S1	0							
S2	0							
S3	1							

Vector S1 = [ 1 2 1 1 1 1 1 1 1 0 ]

Vector S2 = [ 1 1 1 1 0 0 1 0 0 1 1 1 ]

Vector S3 = [ 2 0 0 1 0 0 0 0 0 0 0 0 ]

Vector S1 = [ 1 2 1 1 1 1 1 1 1 0 0 0 0 0 ]

Vector S2 = [ 1 1 1 1 0 0 1 0 0 1 1 1 0 0 0 ]

Vector S3 = [ 2 0 0 1 0 0 0 0 0 0 0 0 0 1 1 1 ]

(2) If :-

S1 :-

$$tf(\text{data}) = 1/12 = 0.0833$$

$$tf(\text{science}) = 2/12 = 0.1666$$

$$tf(\text{is}) = 1/12 = 0.0833$$

$$tf(\text{one}) = 1/12 = 0.0833$$

$$tf(\text{of}) = 1/12 = 0.0833$$

$$tf(\text{the}) = 1/12 = 0.0833$$

$$tf(\text{most}) = 1/12 = 0.0833$$

$$tf(\text{important}) = 1/12 = 0.0833$$

$$tf(\text{courses}) = 1/12 = 0.0833$$

$$tf(\text{in}) = 1/12 = 0.0833$$

$$tf(\text{computer}) = 1/12 = 0.0833$$

$$tf(this) = 0$$

$$tf(best) = 0$$

$$tf/scientists = 0$$

$$tf/perform = 0$$

$$tf/analysis = 0$$

82 :-

$$tf(data) = 1/9 = 0.111$$

$$tf/science = 1/9 = 0.111$$

$$tf/is = 1/9 = 0.111$$

$$tf/one = 1/9 = 0.111$$

$$tf/of = 1/9 = 0.111$$

$$tf/the = 1/9 = 0.111$$

$$tf/most = 0$$

$$tf/important = 0$$

$$tf/courses = 1/9 = 0.111$$

$$tf/in = 0$$

$$tf/computer = 0$$

$$tf(this) = 1/9 = 0.111$$

$$tf(best) = 1/9 = 0.111$$

$$tf/scientists = 0$$

$$tf/perform = 0$$

$$tf/analysis = 0$$

SS:-

$$tf(\text{data}) = 2/6 = 0.333$$

$$tf(\text{science}) = 0$$

$$tf(\text{is}) = 0$$

$$tf(\text{one}) = 0$$

$$tf(\text{of}) = 0$$

$$tf(\text{the}) = 1/6 = 0.1666$$

$$tf(\text{most}) = 0$$

$$tf(\text{important}) = 0$$

$$tf(\text{courses}) = 0$$

$$tf(\text{iu}) = 0$$

$$tf(\text{computer}) = 0$$

$$tf(\text{this}) = 0$$

$$tf(\text{best}) = 0$$

$$tf(\text{scientists}) = 1/6 = 0.1666$$

$$tf(\text{perform}) = 1/6 = 0.1666$$

$$tf(\text{qualysis}) = 1/6 = 0.1666$$

(idf) :-

(3) idf :-

$$idf(\text{data}) = \log(3/3) = 0$$

$$idf(\text{science}) = \log(3/2) = 0.17609$$

$$idf(\text{is}) = \log(3/2) = 0.17609$$

$$idf(\text{one}) = \log(3/2) = 0.17609$$

$$idf(\text{of}) = \log(3/2) = 0.17609$$

$$idf(\text{the}) = \log(3/3) = 0$$

$$idf(\text{most}) = \log(3/1) = 0.47712$$

$$idf(\text{important}) = \log(3/1) = 0.47712$$

$$idf(\text{courses}) = \log(3/2) = 0.17609$$

$$idf(\text{in}) = \log(3/1) = 0.47712$$

$$idf(\text{computer}) = \log(3/1) = 0.47712$$

$$idf(\text{this}) = \log(3/1) = 0.47712$$

$$idf(\text{best}) = \log(3/1) = 0.47712$$

$$idf(\text{scientists}) = \log(3/1) = 0.47712$$

$$idf(\text{perform}) = \log(3/1) = 0.47712$$

$$idf(\text{analysis}) = \log(3/1) = 0.47712$$

# (4) Tf. idf :-

$$tf \cdot idf = tf \times idf$$

S1

S2

S3

$tf \cdot idf(\text{data})$ :-	0	0	0
$tf \cdot idf(\text{science})$ :-	0.0293	0.0195	0
$tf \cdot idf(\text{is})$ :-	0.0146	0.0195	0
$tf \cdot idf(\text{one})$ :-	0.0146	0.0195	0
$tf \cdot idf(\text{of})$ :-	0.0146	0.0195	0
$tf \cdot idf(\text{the})$ :-	0	0	0
$tf \cdot idf(\text{most})$ :-	0.0397	0.030	0
$tf \cdot idf(\text{important})$ :-	0.0397	0	0
$tf \cdot idf(\text{courses})$ :-	0.0146	0.0195	0
$tf \cdot idf(\text{in})$ :-	0.0397	0	0
$tf \cdot idf(\text{computer})$ :-	0.0397	0	0
$tf \cdot idf(\text{this})$ :-	0	0.0530	0
$tf \cdot idf(\text{best})$ :-	0	0.0530	0
$tf \cdot idf(\text{scientists})$ :-	0	0	0.0794
$tf \cdot idf(\text{perform})$ :-	0	0	0.0794
$tf \cdot idf(\text{analysis})$ :-	0	0	0.0794

Question no. 02 :-

(1) cosine distances :-

→ Bow & S1, S2

$$\cos(S1, S2) = \frac{S1 \cdot S2}{\|S1\| \cdot \|S2\|}$$

$$S1 = [0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 2, 0, 1, 0]$$
$$S2 = [0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1]$$

$$S1 \cdot S2 = (0 \times 0 + 0 \times 1 + 1 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 1 + 1 \times 0 + 1 \times 1 + 1 \times 0 + 1 \times 1 + 0 \times 0 + 2 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 1)^{0.5}$$
$$= (8)^{0.5} = 2.8284$$

$$\|S1\| = (0 + 0 + 1 \times 1 + 0 + 2 \times 2 + 0 + 1 \times 1 + 0)^{0.5}$$

$$= (14)^{0.5} = 3.7416$$

$$\|S_2\| = (0 + |x| + 0 + |x| + |x| + 0 + 0 + |x| + 0 + |x| + |x| + 0 + |x| + 0 + |x| + |x|)^{0.5}$$
$$= (9)^{0.5} = 3$$

$$\cos(S_1, S_2) = \frac{2.8284}{(3.7416)(3)} = \frac{2.8284}{11.2248}$$
$$= 0.2519$$

→ If. idf of  $S_1, S_2$

$$\cos(S_1, S_2) = S_1 \cdot S_2$$

$$\|S_1\| \cdot \|S_2\|$$

$$S_1 \cdot S_2 = (0 + 0.0293 \times 0.0195 + 0.0146 \times 0.0195 + 0.0146 \times 0.0195 + 0.0146 \times 0.0195 + 0 + 0 + 0.0397 * 0.0530 + 0 + 0.0146 \times 0.0195 + 0 + 0 + 0 + 0 + 0 + 0)^{0.5}$$
$$= (0.003814)^{0.5} = 0.06175.$$

$$\|S_1\| = \sqrt{0 + 0.00087 + 0.000213 + 0.000213 + 0.000213 + 0 + 0.00157 + 0.00157 + 0.000213 + 0.00157}$$
$$= (0.008015)^{0.5} = 0.08952$$

$$\|S2\| = \sqrt{0 + 0.0038 + 0.0038 + 0.0038 + 0.0038 + 0.0028 + 0.0028 + 0.0028 + 0.0028 + 0.0028} \\ = (0.0103)^{0.5} = 0.10162.$$

$$\cos(S1, S2) = \frac{S1 \cdot S2}{\|S1\| \cdot \|S2\|} = \frac{0.06175}{0.08952 \times 0.10162} \\ = 6.7879$$

$\rightarrow$  BOW :-  $S2, S3$

$$\cos(S2, S3) = \frac{S2 \cdot S3}{\|S2\| \cdot \|S3\|}$$

$$S2 = [0, 1, 0, 1, 1, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1]$$

$$S3 = [1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0]$$

$$S2 \cdot S3 = (0 \times 1 + 1 \times 0 + 0 \times 0 + 1 \times 0 + 1 \times 2 + 0 \times 0 + 0 \times 0 + 1 \times 0 + 0 \times 0 + 1 \times 0 + 1 \times 0 + 0 \times 1 + 1 \times 0 + 0 \times 1 + 1 \times 0 + 0 \times 1 + 1 \times 1 + 1 \times 0) \\ = (3)^{0.5} = 1.7320$$

$$\|S2\| = (\sum (0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 0 + 1 \times 0 + 0 \times 1 + 1 \times 0 + 0 \times 1 + 1 \times 0 + 1 \times 1 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 1 \times 0 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 1 \times 1))^{0.5} \\ = (9)^{0.5} = 3$$

$$\|S3\| = (1 \times 1 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 2 \times 2 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 0 \times 0 + 1 \times 1 + 0 \times 0 + 1 \times 1 + 0 \times 0)^{0.5} = (8)^{0.5} = 2.8284$$

$$\cos(S2, S3) = \frac{S2 \cdot S3}{\|S2\| \cdot \|S3\|} = \frac{1.7320}{(3)(2.8284)} = \frac{1.7320}{8.4852} = 0.2041.$$

$\rightarrow$  If idf  $\vdash S2, S3$

$$\cos(S2, S3) = \frac{S2 \cdot S3}{\|S2\| \cdot \|S3\|}$$

$$S2 \cdot S3 = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0)^{0.5} = 0$$

$$\|S2\| = (0 + 0.00038 + 0.00038 + 0.00038 + 0.00038 + 0.0028 + 0.00038 + 0.0028 + 0.0028)^{0.5} = (0.0103)^{0.5} = 0.10162$$

$$\|S3\| = (0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0.006304 + 0.006304 + 0.006304)^{0.5} = (0.01891)^{0.5} = 0.1378$$

$$\cos(82.83) = \frac{0}{(0.10162)(0.1378)} = 0$$

$\rightarrow \text{BOW} \text{ of } S1, S3$

$$\cos(S1, S3) = \frac{S1 \cdot S3}{\|S1\| \cdot \|S3\|}$$

$$S1 = [0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 2, 0, 1, 0]$$

$$S2 = [1, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0]$$

$$\begin{aligned} S1 \cdot S3 &= (0 \times 1 + 0 \times 0 + 1 \times 0 + 1 \times 0 + 1 \times 2 + 1 \times 0 + 1 \times 0 + 1 \times 0 + \\ &\quad 1 \times 0 + 1 \times 0 + 0 \times 1 + 2 \times 0 + 0 \times 1 + 1 \times 1 + 0 \times 0)^{0.5} \\ &= (3)^{0.5} = 1.7320 \end{aligned}$$

$$\begin{aligned} \|S1\| &= (\sqrt{0+0+1+1+1+1+1+1+1+1+0+2+0+1+0})^{0.5} = (14)^{0.5} = 3.7416 \end{aligned}$$

$$\begin{aligned} \|S3\| &= (\sqrt{1+0+0+0+2+2+0+0+0+0+0+0+0+0})^{0.5} = (8)^{0.5} \\ &= 2.8284 \end{aligned}$$

$$\begin{aligned} \cos(S1, S3) &= \frac{(1.7320)}{(3.7416)(2.8284)} = \frac{1.7320}{10.582} \\ &= 0.16367 \end{aligned}$$

$\rightarrow \text{If } \text{idf} \propto S1, S3$

$$\cos(S1, S3) = S1 \cdot S3$$

$$\|S1\| \cdot \|S3\|$$

$$S1 \cdot S3 = (0+0+0+0+0+0+0+0+0+0+0+0+0+0+0)^{0.5} = 0$$

$$\begin{aligned}\|S1\| &= (0+0.0293 \times 0.0293 + 0.0146 \times 0.0146 + 0.0146 \times \\&\quad 0.0146 + 0.0146 \times 0.0146 + 0+0.0397 \times 0.0397 \\&\quad + 0.0397 \times 0.0397 + 0.0146 \times 0.0146 + 0.0397 \\&\quad \times 0.0397 + 0.0397 \times 0.0397 + 0+0+0+0+0)^{0.5} \\&= (0.008015)^{0.5} = 0.08952.\end{aligned}$$

$$\begin{aligned}\|S3\| &= (0+0+0+0+0+0+0+0+0+0+0+0+0+0+0)^{0.5} \\&\quad + 0+0.0794 \times 0.0794 + 0.0794 \times 0.0794 + \\&\quad 0.0794 \times 0.0794)^{0.5} = (0.01891)^{0.5} \\&= 0.1375.\end{aligned}$$

$$\cos(S1, S3) = S1 \cdot S3 = 0$$

$$\|S1\| \cdot \|S3\| = (0.08952)(0.1375)$$

$$= 0$$

## (2) Manhattan distances :-

→ Bow :- S<sub>1</sub>, S<sub>2</sub>

$$\text{manhattan distance}(S_1, S_2) = \sum_{i=1}^{16} |S_1 - S_2|$$

$$\begin{aligned}
 &= |0-0| + |0-1| + |1-0| + |1-1| + |1+1| + |1-0| + \\
 &\quad |1-0| + |1-1| + |1-0| + |1-1| + |1-1| + |0-0| + \\
 &\quad |2-1| + |0-0| + |1-1| + |0-1| \\
 &= 0 + 1 + 1 + 0 + 0 + 1 + 0 + 1 + 0 + 0 + 0 + 1 \\
 &\quad + 0 + 1 = 8
 \end{aligned}$$

→ Tf. idf :- S<sub>1</sub>, S<sub>2</sub>

$$\text{manhattan distance}(S_1, S_2) = \sum_{i=1}^{16} |S_1 - S_2|$$

$$\begin{aligned}
 &= |0-0| + |0.0293 - 0.0195| + |0.0146 - 0.0195| + \\
 &\quad |0.0146 - 0.0195| + |0.0146 - 0.0195| + |0-0| + \\
 &\quad |0.0397-0| + |0.0397-0| + |0.0146 - 0.0195| + \\
 &\quad |0.0397-0| + |0.0397-0| + |0-0.0530| + \\
 &\quad |0-0.0530| + |0-0| + |0-0| + |0-0| \\
 &= 0.0098 + 0.0049 + 0.0049 + 0.0049 + 0.0397 + \\
 &\quad 0.0397 + 0.0049 + 0.0397 + 0.0397 + 0.0530 + \\
 &\quad 0.0530 = 0.2942
 \end{aligned}$$

$\rightarrow$  Bow :- S2, S3

$$\text{manhattan distance}(S2, S3) = \sum_{i=1}^{16} |S2_i - S3_i|$$

$$\begin{aligned} &= |0-1| + |1-0| + |1-0| + |1-2| + |1-0| + |1-0| + \\ &\quad |1-0| + |0-1| + |1-0| + |0-1| + |1-1| + |1-0| \\ &= 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 1 \\ &= 11 \end{aligned}$$

$\rightarrow$  Tf. idf :- S2, S3

$$\text{manhattan distance}(S2, S3) = \sum_{i=1}^{16} |S2_i - S3_i|$$

$$\begin{aligned} &= 0 + |0.0195 - 0| + |0.0195 - 0| + |0.0195 - 0| + \\ &\quad |0.0195 - 0| + 0 + 0 + 0 + |0.0195 - 0| + 0 \\ &\quad + 0 + |0.0530 - 0| + |0.0530 - 0| + |0 - 0.0794| \\ &\quad + |0 - 0.0794| + |0 - 0.0794| \\ &= 0.0195 + 0.0195 + 0.0195 + 0.0195 + 0.0195 + \\ &\quad 0.0530 + 0.0530 + 0.0794 + 0.0794 + \\ &\quad 0.0794 \\ &= 0.4417 \end{aligned}$$

$\rightarrow$  Bow & S1, S3

$$\text{manhattan distance } (S1, S3) = \sum_{i=1}^{16} |S1 - S3|$$

$$\begin{aligned}&= |0-1| + |1-0| + |1-0| + |1-2| + |1-0| + |1-0| + \\&\quad |1-0| + |1-0| + |1-0| + |1-0| + |0-1| + |2-0| \\&\quad + |0-1| + |1-1| \\&= 1+1+1+1+1+1+1+1+1+2+1 \\&= 12+2 = 14\end{aligned}$$

$\rightarrow$  Tf. idf & S1, S3

$$\text{manhattan distance } (S1, S3) = \sum_{i=1}^{16} |S1 - S3|$$

$$\begin{aligned}&= |0.0293 - 0| + |0.0146 - 0| + |0.01616 - 0| + \\&\quad |0.0146 - 0| + |0.0397 - 0| + |0.0397 - 0| \\&\quad + |0.0146 - 0| + |0.0397 - 0| + |0.0397 - 0| \\&\quad + |0 - 0.0794| + |0 - 0.0794| + |0.0794| \\&= 0.0293 + 0.0146 + 0.0146 + 0.0146 + 0.0397 + \\&\quad 0.0397 + 0.0146 + 0.0397 + 0.0397 + 0.0794 \\&\quad + 0.0794 + 0.0794 \\&= 0.4847\end{aligned}$$

### (3) Euclidean distances &

→ Bow &  $S_1, S_2$

$$\text{euclidean distance } (S_1, S_2) = \sqrt{\sum_{i=1}^{16} (S_1 - S_2)^2}$$

$$\begin{aligned} &= \left[ (0-1)^2 + (1-0)^2 + (1-0)^2 + (1-2)^2 + (1-0)^2 + \right. \\ &\quad (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + \\ &\quad \left. (0-1)^2 + (2-0)^2 + (0-1)^2 + (1-1)^2 \right]^{0.5} \\ &= (1+1+1+1+1+1+1+1+1+1+1+1+1+1+1+0)^{0.5} \\ &= (16)^{0.5} = 3.7416 \end{aligned}$$

→ Tf. idf &  $S_1, S_2$

$$\text{euclidean distance } (S_1, S_2) = \sqrt{\sum_{i=1}^{16} (S_1 - S_2)^2}$$

$$\begin{aligned} &= \left[ (0.0293-0.0195)^2 + (0.0146-0.0195)^2 + (0.0146-0.0195)^2 \right. \\ &\quad + (0.0146-0.0195)^2 + (0.0397-0)^2 + (0.0397-0)^2 + \\ &\quad (0.0146-0.0195)^2 + (0.0397-0)^2 + (0.0397-0)^2 + \\ &\quad \left. (0-0.0530)^2 + (0-0.0530)^2 \right]^{0.5} \\ &= (0.00009604 + 0.00002401 + 0.00002401 + 0.00002401 \\ &\quad + 0.0015760 + 0.0015760 + 0.00002401 + 0.0015760 \\ &\quad + 0.0015760 + 0.002809 + 0.002809)^{0.5} \\ &= (0.01211)^{0.5} = 0.11006 \end{aligned}$$

$\rightarrow$  Bow :- S2, S3

$$\text{euclidean distance}(S2, S3) = \sqrt{\sum_{i=1}^{16} (S2_i - S3_i)^2}$$

$$\begin{aligned} &= \sqrt{(0-1)^2 + (1-0)^2 + (1-0)^2 + (1+2)^2 + (1-0)^2 + (1-0)^2 \\ &\quad + (1-0)^2 + (0-1)^2 + (1-0)^2 + (0-1)^2 + (1-1)^2 + (1-0)^2} \\ &\approx \sqrt{(1+1+1+1+1+1+1+1+1+0+1)^{0.5}} \\ &= \sqrt{11}^{0.5} = 3.3166 \end{aligned}$$

$\rightarrow$  Tf. idf :- S2, S3

$$\text{euclidean distance}(S2, S3) = \sqrt{\sum_{i=1}^{16} (S2_i - S3_i)^2}$$

$$\begin{aligned} &= \sqrt{(0.0195-0)^2 + (0.0195-0)^2 + (0.0195-0)^2 + (0.0195-0)^2 + \\ &\quad (0.0195-0)^2 + (0.0530-0)^2 + (0.0530-0)^2 + (0.0794)^2 \\ &\quad + (0-0.0794)^2 + (0.0794)^2}^2 \\ &= \sqrt{(0.00038025 + 0.00038025 + 0.00038025 + 0.00038025 + \\ &\quad 0.00038025 + 0.002809 + 0.002809 + 0.00630436 \\ &\quad + 0.00630436 + 0.00630436)^{0.5}} \\ &= \sqrt{(0.026432)}^{0.5} = 0.16257 \end{aligned}$$

$\rightarrow$  Bow  $\vdash S_1, S_3$

$$\text{euclidean distance } (S_1, S_3) = \sqrt{\sum_{i=1}^{16} (S_1 - S_3)^2}$$

$$\begin{aligned} &= [(0-1)^2 + (1-0)^2 + (1-0)^2 + (1-2)^2 + (1-0)^2 + (1-0)^2 \\ &\quad + (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2 + \\ &\quad (2-0)^2 + (0-1)^2 + (1-1)^2]^{0.5} \\ &= (1+1+1+1+1+1+1+1+1+1+2+1+0)^{0.5} \\ &= (16)^{0.5} = 4 \end{aligned}$$

$\rightarrow$  Tf. idf  $\vdash S_1, S_3$

$$\text{euclidean distance } (S_1, S_3) = \sqrt{\sum_{i=1}^{16} (S_1 - S_3)^2}$$

$$\begin{aligned} &= [(0.0293-0)^2 + (0.0146-0)^2 + (0.0146-0)^2 + (0.0146-0)^2 \\ &\quad + (0.0397-0)^2 + (0.0397-0)^2 + (0.0146-0)^2 \\ &\quad + (0.0397-0)^2 + (0.0397-0)^2 + (0-0.0794)^2 \\ &\quad + (0-0.0794)^2 + (0-0.0794)^2]^{0.5} \\ &= (0.00085849 + 0.00021316 + 0.00021316 + 0.00021316 \\ &\quad + 0.00157609 + 0.00157609 + 0.00021316 + \\ &\quad 0.00157609 + 0.00157609 + 0.00630436 + \\ &\quad 0.00630436 + 0.00630436)^{0.5} \\ &= (0.02692857)^{0.5} = (0.164099) \end{aligned}$$