

COMSATS UNIVERSITY ISLAMABAD

Lahore Campus

NAME:

ROLL NO:

TEACHER NAME:

ASSIGNMENT NO:

ASSIGNMENT TITLE:

M. Ammar Mukhtar

FA21-BSE-141

Prof. M. Sharjeel

BOW, TF, IDF, Similarity

Sun Mon Tue Wed Thu Fri Sat



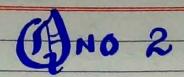
	BAG	OF	Wort	s:	(Word	occ	covance	in each	h Senler	(e)
ν	D.	ata	Science	is	one	of	the	most	important	Courde	in
S	1	1	2	1	1	1	1	1	1	1	1
5	2 1		1		1	1	1	0	0	1	0
5:									0		
	Corr	piter	this	bust	scie	entist	per	born	analysis	Tota	(= 16
.51		/									
52	()	1	1		0	(0	0	9	
\$3	0)	0	0		ı		1	L	6	
S1:[12											
82: [111110010011000]											
83: [2000 10000 000 111]											
TERM FREQUENCY: (TF = No. of Tems Total Terms in Doc)											
	Data	Scie	inence	is one	e of	the	me	st	important	cocurbes	in
51	1/16	2/1		46 VI				16	Y16	1/16	1/16
52	1,9	1/4		19 4	9 1/9	1/6	7 (0	0	1/9	0
53	2/6	0		0 0	0	y YE	5 (0	0	0
computer this pest scientist perform analysis.											
51	1/16			0	0		0		0		
32	0		9!	4	0		0		0		
53	0	(0	1/6		1/6		1/6		

					Sun I	Mon Tue Wed Thu	Fri Sat	
1	nc.	17	DF =	log 1	Vo. of D	reace in Doc)		
- J	Ur:			- 0 (1	erm occur	ence in Doc)		
THE STATE OF THE S	Data	science	is one	e of t	he mos	t inportar	nt courses	
idA	0	0.17		7 0.17	0.1	0.4	0.17	
0								
	in	computer	this	pest	scien	List perfo	rm analysis	
		0.4	0.4	0.4	0.4	0.4	0.9	
				w 25				
-	- 1	-	10 =1		10	TOF		
10	r x I	r:	TF-IL	JP =	I X	1DF		
	Of			,	41	at in a	I de sounder	
	Ihla	science	<u>es</u> o	n of	The	most ings	nted course	
f-idf.	0				0			
	-	a. atu	77.	T 10	right it	muchan	andesis-	
	m	computer	ems j	sest sc	ceru-gr	perform	aranger	
	Data	Science	is one	of the	most	important 0.14	Courses	
SI	0	0-34	0.17 6.17	017 0	044	0.14	0.17	
S2	0	0.17	0.17 0.17	0.17 0	0	0	0.17	
53	0	0	0 0	0 0	0	0	0	
	THE	- in- 17						
	in	computer	this	best	scientis	t perform	analysis.	
		,						
51	014	044	0	0	0	0	0	
52	0	0	0.4	0-4	0	0	0	
53	D	D	0	0.0	04	0.4	0.4	

Date:

Date:....

Sun Mon Tue Wed Thu Fri Sat



(i) Cosine Similarity
$$S_1 . S_2$$

 $(S_1, S_2) = \frac{1}{|S_1| . |S_2|}$

$$S_1.S_2 = 1+2+1+1+1+1+0+0+1+0+0+0+0+0+0+0=8$$

$$|S_1| = \sqrt{1+4+1+1+1+1+1+1+1+1+1} = \sqrt{19}$$

$$|5_2| = \sqrt{|+|+|+|+|+|+|+|} = \sqrt{9} = 3$$

cosine similarity
$$(S_1, S_2) - 8 = 0.713$$

$$(ii)$$
 Coine $(S_1, S_3) = \frac{S_1 \cdot S_3}{|S_1| |S_3|} = \frac{2+1+0}{\sqrt{14} \sqrt{8}} = 0.28$

$$60(5_1,5_3) = 0.28$$

$$(iii)$$
 (3) (5) $(5$

Sun Mon Tue Wed Thu Fri Sat

MONHATTON:

EUCLIDEAN:

$$=\sqrt{17} = 4.12$$

(iii) Distance =
$$\sqrt{|^2+|^2+|^2+|^2+|^2+|^2+|^2+|^2+|^2+|^2} = \sqrt{11}$$