Assignment # 4

Date: Muhammad Anas Day:

Registration No. :
Course :
Section :

Question # 01:-

S1: "data science is one of the most important courses in computer science"
S2: " this is one of the best data science courses"
S3: " the data scientists perform data analysis."

Vocabulary: {dat, analysis, best, computer, couses, data, important,
BoW: in, is, mest, of, one, perfoom science, sientist, the, this}

	1	2	3
analysis	0	0	1
best	0	1	0
computer	1	0	0
courses	1	1	0
data	1	1	2
important	1	0	0
in	1	0	0
is	1	1	0
most	1	0	0
of	1	1	0
one	1	1	0
perfoom	0	0	1
science	2	1	0

Date: ____________ Day: ____________

scientists	0	0	1
the	1	1	1
this	0	1	0

TF:-

$$T.F = \frac{\text{term occur in document}}{\text{total terms in document}}$$

	analysis	best	computer	courses
1	0·0000	0·0000	0·0833	0·0833
2	0·0000	0·1111	0·0000	0·1111
3	0·1666	0·0000	0·0000	0·0000

	data	important	in	is
1	0·0833	0·0833	0·0833	0·8888
2	0·1111	0·0000	0·0000	0·1111
3	0·3333	0·0000	0·0000	0·0000

	most	of	one	perform
1	0·0833	0·0833	0·0833	0·0000
2	0·0000	0·1111	0·1111	0·0000
3	0·0000	0·0000	0·0000	0·1666

	science	scientists	the	this
1	0·1666	0·0000	0·0833	0·0000
2	0·1111	0·0000	0·1111	0·1111
3	0·0000	0·1666	0·1666	0·0000

IDF:-

$$\text{IDF} = \log\left(\frac{\text{total number of docements}}{\text{number of document have term}}\right)$$

analysis	1·4771
best	1·4771
computer	1·4771
couses	1·1760
data	1·0000
important	1·4771
in	1·4771
is	1·1760
most	1·4771
of	1·1760
one	1·1760
perform	1·4771
science	1·1760
scientists	1·4771
the	1·0000
this	1·4771

TF·IDF:

$$\text{TF·IDF} = (\text{T·F}) \times (\text{IDF}$$

	analysis	best	computer	couses
1	0·0000	0·0000	0·3274	0·2490
2	0·0000	0·4229	0·0000	0·3216
3	0·4591	0·0000	0·0000	0·0000

	data	important	in	is
1	0·1934	0·3247	0·3274	0·2490
2	0·2498	0·0000	0·0000	0·3216
3	0·5423	0·0000	0·0000	0·0000

	most	of	one	perform
1	0·3274	0·2490	0·2490	0·0000
2	0·0000	0·3216	0·3216	0·0000
3	0·0000	0·0000	0·0000	0·4591

	science	sciencetists	the	this
1	0·4981	0·0000	0·1934	0·0000
2	0·3216	0·0000	0·2498	0·4229
3	0·0000	0·4591	0·2711	0·0000

QUESTION # 02 :-

Similarity using Cosine:

cosine(S1,S2) = dotproduct(T·F×IDF(S1) ·(TF·IDF(S2))/|S1| · |S2|

	1	2	3
1	1·0000	0·7126	0·2834
2	0·7126	1·0000	0·3535
3	0·2834	0·3535	1·0000

Day:

Similarity using Manhatan:

Manhatan (S1, S2) = sum(abs(abs(T·F·IDF(S1))·
·(TF·IDF(S2))

Manhatan (S1, S2) = 0·241223

Similarity using Eucliedan:-

Euclidean (S1, S2) = sqrt(sum(TF·IDF(S1))
· TF·IDF(S2))^2))

Euclidean (S1, S2) = 1·862364