ASSIGNMENT # 04

MUHAMMAD UMER AMIR FA21-BSE-114

S_1 = "data science is one of the most important courses in computer science".

S_2 = "this is the one of best data science course".

S_3 = "the data scientists perform data analysis".

BOW

Token	S_1	S_2	S_3	Total
data	1	1	2	4
science	2	1	0	3
is	1	1	0	2
one	1	1	0	2
of	1	1	0	2
the	1	1	1	3
most	1	0	0	1
important	1	0	0	1
courses	1	1	0	2
in	1	0	0	1
computer	1	0	0	1
scientist	0	0	1	1
perform	0	0	1	1

analysis	0	0	1	1
this	0	1	0	1
best	0	1	0	1
TOTAL	12	9	6	27

TF (TERM FREQUENCY)

	S_1	S_2	S_3
data	1/12 = 0.0833	1/9 = 0.1111	2/6 = 0.3333
Science	2/12 = 0.1667	1/9	0
is	1/12	1/9	0
one	1/12	1/9	0
of	1/12	1/9	0
the	1/12	1/9	1/6 = 0.167
most	1/12	0	0
important	1/12	0	0
courses	1/12	1/9	0
in	1/12	0	0
computer	1/12	0	0
Scientist	0	0	1/6
perform	0	0	1/6
analysis	0	0	1/6
this	0	1/9	0
best	0	1/9	0

IDF

TOKEN	IDF
data	$\log(3/3) = 0$
Scicne	$\log(3/2) = 0.176$
is	$\log(3/2) = 0.176$
one	$\log(3/2) = 0.176$
of	$\log(3/2) = 0.176$
the	$\log(3/2) = 0.176$
most	$\log(3/1) = 0.477$
important	$\log(3/1) = 0.477$
course	$\log(3/2) = 0.176$
in	$\log(3/1) = 0.477$
Computer	$\log(3/1) = 0.477$
Scientist	$\log(3/1) = 0.477$
perform	$\log(3/1) = 0.477$
analysis	$\log(3/1) = 0.477$
this	$\log(3/1) = 0.477$
best	$\log(3/1) = 0.477$

TF * IDF

	S1	S2	S3
data	0	0	0
Scicne	0.4981	0.3216	0
is	0.0146	0.3216	0
one	0.2490	0.32167	0
of	0.2490	0.3216	0
the	0.1934	0.2498	0
most	0.3274	0	0.3711
			0

important	0.3374	0	0
recurse	0.2490	0.326	0
in	0.3374	0	0
computer	0.3374	0	0
scientist	0	0	0.4591
perform	0	0	0.4591
analysis	0	0	0.4591
this	0	0.4229	0
best	0	0.4229	0

Cosine Similarity

$S_1 = [1,2,1,1,1,1,1,1,1,1,1,1,0,0,0,0,0,0]$

$S_2 = [1,1,1,1,1,1,1,0,1,0,0,0,0,1,1]$

$S_1 \cdot S_2 = [1+2+1+1+1+1+1+0+0+1+0+0+0+0+0+0+0]$

$S_1 \cdot S_2 = 8$

$\|S_1\| = [1+4+1+1+1+1+1+1+1+1+1+1+1+1]^{0.5} = 14^{0.5} = 3.7413$

$\|S_2\| = [1+1+1+1+1+1+1+1+1]^{0.5} = 9^{0.5} = 3$

$\text{Cosine}(S_1, S_2) = \frac{|S_1 \cdot S_2|}{\|S_1\| \times \|S_2\|} = \frac{8}{11.24} = 0.7114$

MANHATTAN (CITY BLOCK)

$$(S_1, S_2) = \Sigma |S_{1i} - S_{2i}|$$

(S_1, S_2) = 2.737
(S_2, S_3) = 4.1455
(S_3, S_1) = 4.60

EUCLIDEAN DISTANCE

$$(S_1, S_2) = \sqrt{\Sigma (S_1 - S_2)^2}$$

(S_2, S_3) = 1.262
(S_1, S_2) = 0.919
(S_1, S_3) = 1.0288