Name : Azeem Manzoor

Reg# : SP20-BCS-046

Group : 4

Semester : 6th

Section : B

Assignment# : 5th

Course : Introduction To Data Science

Submitted To.

Prof. Dr. Muhammad Sharjeel.

Question 1.

S1 → Sunshine state enjoy sunshine.

S2 → Brown fox jump high, brown fox run..

S3 → Sunshine state fox run fast.

Bag of Words:

Documents	**Vocabulary**									**Length**
	sunshine	state	enjoy	brown	fox	jump	high	run	fast	
S1	2	1	1	0	0	0	0	0	0	4
S2	0	0	0	2	2	1	1	1	0	7
S3	1	1	0	0	1	0	0	1	1	5

Term Frequency (TF)

Formula:

TF = (number of times term appears in document) / (total number of terms in document)

S1:

sunshine → 2/4

state → 1/4

enjoy → 1/4

S2:

brown → 2/7

fox → 2/7

jump → 1/7

high → 1/7

run → 1/7

S3:

sunshine → 1/5

state → 1/5

fox → 1/5

run → 1/5

fast → 1/5

Table of TF

Documents	**Vocabulary**								
	sunshine	state	enjoy	brown	fox	jump	high	run	fast
S1	2/4	1/4	1/4	0	0	0	0	0	0
S2	0	0	0	2/7	2/7	1/7	1/7	1/7	0
S3	1/5	1/5	0	0	1/5	0	0	1/5	1/5

Inverse Document Frequency (IDF)

Formula:

IDF = log (total number of documents) / (number of documents containing the term)

sunshine → log (3/2) = 0.176

state → log (3/2) = 0.176

enjoy → log (3/1) = 0.477

brown → log (3/1) = 0.477

fox → log (3/2) = 0.176

jump → log (3/1) = 0.477

high → log (3/1) = 0.477

run → log (3/2) = 0.176

fast → log (3/1) = 0.477

Term Frequency-Inverse Document Frequency (TF-IDF)

S1:

sunshine → 2/4 × 0.176 = 0.088

state → 1/4 × 0.176 = 0.044

enjoy → 1/4 × 0.477 = 0.1192

S2:

brown → 2/7 × 0.477 = 0.136

fox → 2/7 × 0.176 = 0.051

jump → 1/7 × 0.477 = 0.068

high → 1/7 × 0.477 = 0.068

run → 1/7 × 0.176 = 0.025

S3:

sunshine → 1/5 × 0.176 = 0.0352

state → 1/5 × 0.176 = 0.0352

fox → 1/5 × 0.176 = 0.0352

run → 1/5 × 0.176 = 0.0352

fast → 1/5 × 0.477 = 0.0954

TF-IDF Table.

Vocabulary	S1	S2	S3
sunshine	0.088	0	0.0352
state	0.044	0	0.0352
enjoy	0.11925	0	0
brown	0	0.136	0
fox	0	0.051	0.0352
jump	0	0.068	0
high	0	0.068	0
run	0	0.025	0.0352
fast	0	0	0.0954

Question 2.

Cosine similarity between S1 and S3.

Formula:

$$\text{Cos } \Theta = S1.S3 \div |S1|\,|S3|$$

Vector Representation of S1 and S3:

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

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

To Find S1.S3:

S1.S3 = (2*1)+(1*1) +(1*0)+(0*0)+(0*1)+(0*0)+(0*0)+(0*1)+(0*1)

S1.S3 = 2+1

S1.S3 = **3**

<u>To Find |S1| and |S2|:</u>

$|S1| = (2*2 + 1*1 + 1*1)^{0.5}$

$= (4+1+1)^{0.5}$

$= (6)^{0.5}$

= 2.45

$|S2| = (1*1 + 1*1 + 1*1 + 1*1 + 1*1)^{0.5}$

$= (1+1+1+1+1)^{0.5}$

$= (5)^{0.5}$

= 2.24

<u>Now put these values in Cos ϴ = S1.S3 ÷ |S1| |S3|</u>

Cos(S1,S3) = 3 **/** **(**2.45)(2.24)

= 3 / 5.47

= 0.547

<u>Cosine similarity of S1 and S3 is,</u>

Cos(S1,S3) = 0.547