SP20-BCS-033 Husnain Ahmed

(Group: 4)

(Introduction to Data Science)

<Assignment-5>

Q.1

S1 = "sunshine state enjoy sunshine"
S2 = "brown fox jump high, brown fox run"
S3 = "sunshine state fox run fast"

BOW Model:-

Unique words:

sunshine, state, enjoy, brown, fox, jump, high, run, fast

	sunshine	state	enjoy	brown	fox	jump	high	run	fast	Total length
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

vector S1 : [2,1,1,0,0,0,0,0,0]

vector S2 : [0,0,0,2,2,1,1,1,0]

vector S3 : [1,1,0,0,1,0,0,1,1]

TF Model :-

	sunshine	state	enjoy	brown	fox	jump	high	run	fast
tf-S1	1/2	1/4	1/4	0	0	0	0	0	0
tf-S2	0	0	0	2/7	2/7	1/7	1/7	1/7	0
tf-S3	1/5	1/5	0	0	1/5	0	0	1/5	1/5

IDF Model :-

IDF = log[(total number of documents)/(number of documents containing current term)]

	IDF
sunshine	0.17
state	0.17
enjoy	0.47
brown	0.47
fox	0.17
jump	0.47
high	0.47
run	0.17
fast	0.47

TF.IDF values :-

$$tf.idf = tf_{i,d} \times idf_i$$

	tf.idf -(S1)	tf.idf -(S2)	tf.idf -(S3)
sunshine	0.085	0	0.034
state	0.0425	0	0.034
enjoy	0.1175	0	0
brown	0	0.134	0
fox	0	0.048	0.034
jump	0	0.067	0
high	0	0.067	0
run	0	0.024	0.034
fast	0	0	0.094

Q.2

Cosine similarity between S1 & S3:-

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

S1 = [0.085, 0.042, 0.117, 0, 0, 0, 0, 0, 0]

S3 = [0.034, 0.034, 0, 0, 0.034, 0, 0, 0.034, 0.094]

S1.S3 = (0.085 × 0.034) + (0.042 × 0.034) + (0.117 × 0) + (0 × 0) + (0 × 0) + (0 × 0.034) + (0 × 0) + (0 × 0) + (0 × 0.034) + (0 × 0.094)

S1.S3 = 0.0043

|S1| = [(0.085 × 0.085) + (0.042 × 0.042) + (0.117 × 0.117) + (0 × 0) + (0 × 0) + (0 × 0) + (0 × 0) + (0 × 0) + (0 × 0)] ^ 0.5

|S1| = 0.15

|S3| = [(0.034 × 0.034) + (0.034 × 0.034) + (0 × 0) + (0 × 0) + (0.034 × 0.034) + (0 × 0) + (0 × 0) + (0.034 × 0.034) + (0.094 × 0.094)] ^ 0.5

|S3| = 0.11

$$\cos(S1, S3) = \frac{(S1.S3)}{|S1|\,|S3|} \Rightarrow \frac{0.0043}{(0.15)(0.11)}$$

$$\boxed{\cos(S1, S3) = 0.2606}$$