Introduction To Data Science

NAME: MUHAMMAD HUZAIFA JAWAD
Roll No: SP20-BCS-144
Group: 4 (IV)
Submitted To: SIR Muhammad Sharjeel

ASSIGNMENT # 05

QUESTION 01:

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

BoW Model

	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

TF Model

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

IDF Model

formula :
IDF ('word') = log(Total No. of Documents / No. of Documents containig 'word')

IDF(sunshine) = log(3/2) = 0.1761

IDF (state) = log(3/2) = 0.1761

IDF(enjoy) = log (3/1) = 0.4771

IDF (brown) = log(3/1) = 0.4771

IDF (fox) = log(3/2) = 0.1761

IDF (jump) = log(3/1) = 0.4771

IDF (high) = log(3/1) = 0.4771

IDF (run) = log (3/2) = 0.1761

IDF (fast) = log(3/1) = 0.4771 ··

	sunshine	state	enjoy	brown	fox	jump	high	run	fast
IDF	0.1761	0.1761	0.4771	0.4771	0.1761	0.4771	0.4771	0.1761	0.4771

TF.IDF values

for S1:

tf.idf (sunshine) = $\frac{2}{4} * 0.1761 = 0.0880$

tf.idf (state) = $\frac{1}{4} * 0.1761 = 0.0440$

tf.idf (enjoy) = $\frac{1}{4} * 0.4771 = 0.1192$

for S2:

tf.idf (brown) = $(\frac{2}{7})(0.4771) = 0.1363$

tf.idf (fox) = $(\frac{2}{7})(0.1761) = 0.0503$

tf.idf (jump) = $(\frac{1}{7})(0.4771) = 0.0681$

tf.idf (high) = $(\frac{1}{7})(0.4771) = 0.0681$

tf.idf (run) = $(\frac{1}{7})(0.1761) = 0.0251$

for S3:

tf.idf (sunshine) = $(\frac{1}{5})(0.1761) = 0.0352$

tf.idf (state) = $(\frac{1}{5})(0.1761) = 0.0352$

tf.idf (fox) = $(\frac{1}{5})(0.1761) = 0.0352$

tf.idf (run) = $(\frac{1}{5})(0.1761) = 0.0352$

tf.idf (fast) = $(\frac{1}{5})(0.4771) = 0.0954$

	s1	s2	s3
sunshine	0.0880	0	0.0352
state	0.0440	0	0.0352
enjoy	0.1192	0	0
brown	0	0.1363	0
fox	0	0.0503	0.0352
jump	0	0.0681	0
high	0	0.0681	0
run	0	0.0251	0.0352
fast	0	0	0.0954

QUESTION # 02:

cosine Similarity b/w S1 & S3.

using Bow model to generate vectors

$S1 = \langle 2 \; 1 \; 1 \; 0 \; 0 \; 0 \; 0 \; 0 \; 0 \rangle$

$S3 = \langle 1 \; 1 \; 0 \; 0 \; 1 \; 0 \; 0 \; 1 \; 1 \rangle$

formula :

$$\cos(\theta)_{S_1,S_3} = \frac{S_1 \cdot S3}{|S_1|\,|S3|}$$

$$S_1 \cdot S_3 = (2)(1) + (1)(1) + (1)(0) + (0)(0) + (0)(1) + (0)(0) + (0)(0) + (0)(1) + (0)(1)$$

$$= 2+1$$

$$= 3$$

$$|S_1| = \sqrt{2^2+1^2+1^2+0^2+0^2+0^2+0^2+0^2+0^2}$$

$$= \sqrt{4+1+1} = \sqrt{6}$$

$$|S_1| = 2.4494$$

$$|S_3| = \sqrt{1^2+1^2+0^2+0^2+1^2+0^2+0^2+1^2+1^2}$$

$$= \sqrt{5}$$

$$|S_3| = 2.2360$$

$$\cos(S_1,S_3) = \frac{3}{(2.4494)(2.2360)}$$

$$\boxed{\cos(S_1,S_3) = 0.5477}$$

$$S_1, S_3 = \cos^{-1}(0.5477)$$

$$\boxed{S_1, S_3 = 56.78}$$