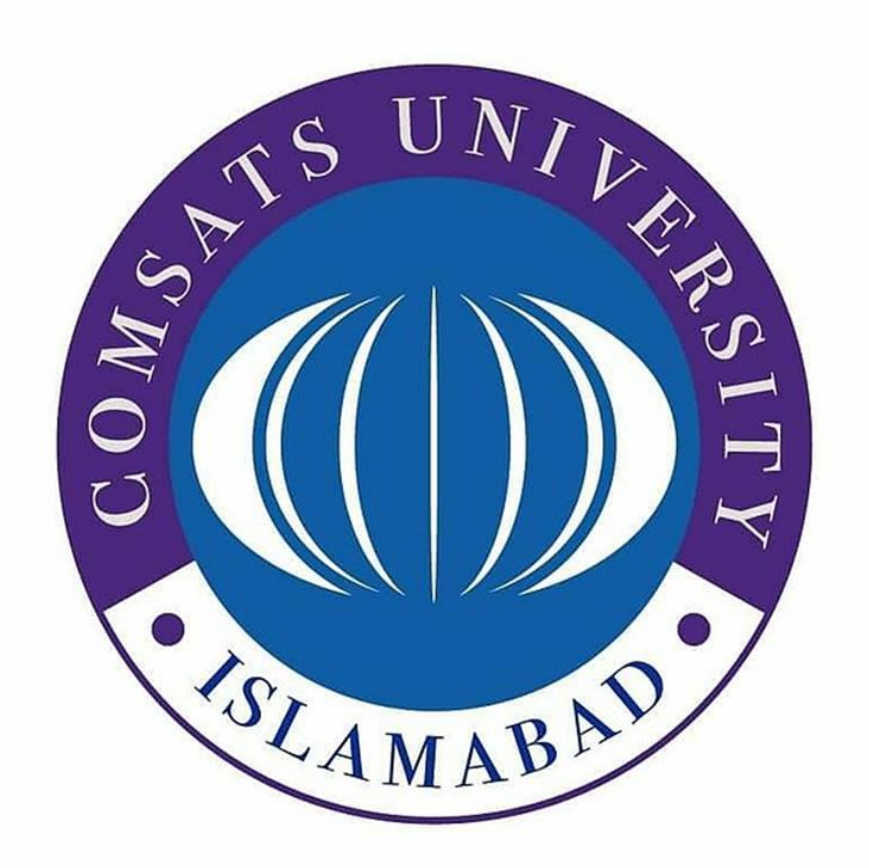

INTRODUCTION TO DATA SCIENCE

Muhammad Azeem Shoukat

SP20-BCS-077/B

Group-2

Assignment#5

Question 1

Compute the BoW model, TF Model, and IDF model for each of the terms in the following three sentences. Then calculate the TF.IDF values

S1 “sunshine state enjoy sunshine”
S2 “brown fox jump high, brown fox run”
S3 “sunshine state fox run fast”

Vocabulary

Sunshine, state, enjoy, brown, fox, jump, high, 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

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	2/4	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
Sunshine	$\log\left(\frac{3}{2}\right) = 0.18$
State	$\log\left(\frac{3}{2}\right) = 0.18$
Enjoy	$\log\left(\frac{3}{1}\right) = 0.48$
Brown	$\log\left(\frac{3}{1}\right) = 0.48$
Fox	$\log\left(\frac{3}{2}\right) = 0.18$
Jump	$\log\left(\frac{3}{1}\right) = 0.48$
High	$\log\left(\frac{3}{1}\right) = 0.48$
Run	$\log\left(\frac{3}{1}\right) = 0.48$
Fast	$\log\left(\frac{3}{1}\right) = 0.48$

TFIF Model

	tfidf - S1	tfidf – S2	tfidf- S3
Sunshine	$\frac{2}{4} * 0.18 =$ **0.09**	$0 * 0.18 =$ **0**	$\frac{1}{5} * 0.18 =$ **0.036**
State	$\frac{1}{4} * 0.18 =$ **0.045**	$0 * 0.18 =$ **0**	$\frac{1}{5} * 0.18 =$ **0.036**
Enjoy	$\frac{1}{4} * 0.48 =$ **0.12**	$0 * 0.48 =$ **0**	$0 * 0.48 =$ **0**
Brown	$0 * 0.48 =$ **0**	$\frac{2}{7} * 0.48 =$ **0.137**	$0 * 0.48 =$ **0**
Fox	$0 * 0.18 =$ **0**	$\frac{2}{7} * 0.18 =$ **0.051**	$\frac{1}{5} * 0.18 =$ **0.036**
Jump	$0 * 0.48 =$ **0**	$\frac{1}{7} * 0.48 =$ **0.068**	$0 * 0.48 =$ **0**
High	$0 * 0.48 =$ **0**	$\frac{1}{7} * 0.48 =$ **0.068**	$0 * 0.48 =$ **0**
Run	$0 * 0.48 =$ **0**	$\frac{1}{7} * 0.48 =$ **0.068**	$\frac{1}{5} * 0.48 =$ **0.096**
Fast	$0 * 0.48 =$ **0**	$0 * 0.48 =$ **0**	$\frac{1}{5} * 0.48 =$ **0.096**

Question 2

Compute the cosine similarity between S1 and S3.

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

Vector S3: [1, 1, 0, 0, 1, 0, 0, 1, 1]

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

$(S1\ .\ S3)$ = (2*1 + 1*1 + 1*0 + 0*0 +0*1 + 0*0 + 0*0 + 0*1 + 0*1) = 3

$|\text{S1}| = \sqrt{2*2 + 1*1 + 1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0}$ = 2.45

$|\text{S3}| = \sqrt{1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 0*0 + 0*0 + 1*1 + 1*1}$ = 2.24

$$\cos(\text{S1, S3}) = \frac{3}{2.45*2.24} = \mathbf{0.5466}$$

Hence, the cosine similarity between S1 and S2 is **0.55**