COMSATS University Islamabad (Lahore Campus)

Assignment <5> FALL 2022

Course Title:	Introduction to Data Science			Course Code:		CSC461	Credit Hours: 3
Course Instructor:	Muhammad Sharjeel			Programme Name:		BS Computer Science	
Semester:	6th	Batch:	SP20	Section:	C	Date:	25-12-2022
Due Date:	**30-12-2022**			**Maximum Marks:**			**10**
Student Name:	**YALMAZ ALI**			**Registration No.**			**SP20-BCS-093**

Solve the following two questions manually as well as implement the solution using Python. Submit both solutions.

Q1. 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"

Answer:

MANUALLY

∴(Q#1):-

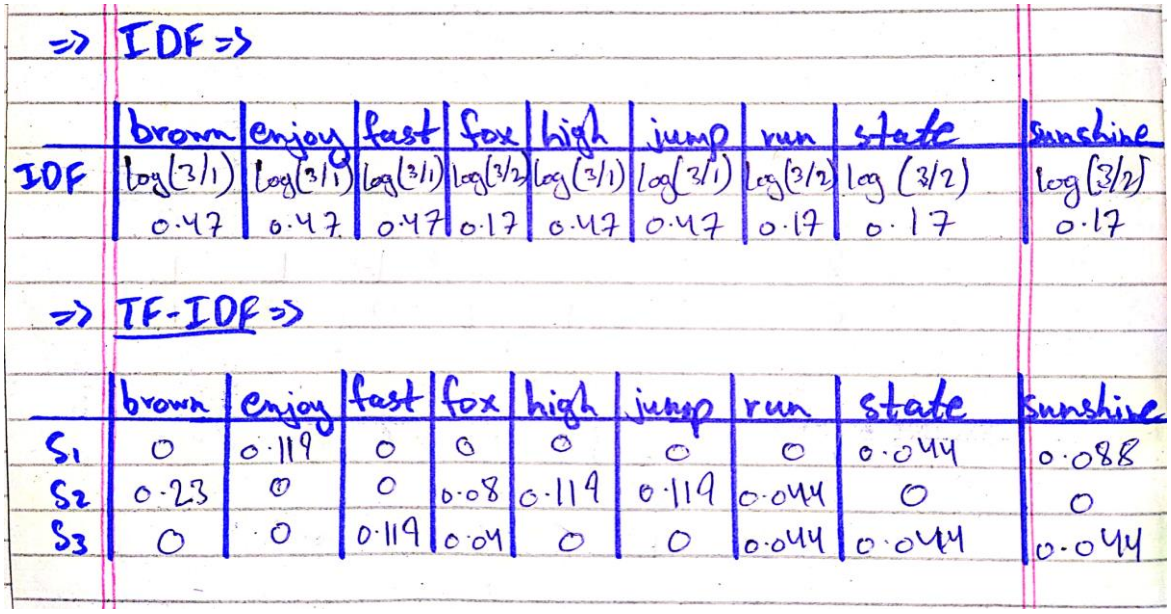
=> BOW =>

	brown	enjoy	fast	fox	high	jump	run
S_1	0	1	0	0	0	0	0
S_2	2	0	0	2	1	1	1
S_3	0	0	1	1	0	0	1

	state	sunshine
S_1	1	2
S_2	0	0
S_3	1	1

=> Term-Frequency =>

	brown	enjoy	fast	fox	high	jump	run	state	sunshine
S_1	0	1/4	0	0	0	0	0	1/4	1/2
S_2	2/7	0	0	2/7	1/7	1/7	1/7	0	0
S_3	0	0	1/5	1/5	0	0	1/5	1/5	1/5



BY CODE

BOW:

	brown	enjoy	fast	fox	high	jump	run	state	sunshine
s1	0	1	0	0	0	0	0	1	2
s2	2	0	0	2	1	1	1	0	0
s3	0	0	1	1	0	0	1	1	1

TF:

	brown	enjoy	fast	fox	high	jump	run	state	sunshine
s1	0	1/4	0	0	0	0	0	1/4	1/2
s2	2/7	0	0	2/7	1/7	1/7	1/7	0	0
s3	0	0	1/5	1/5	0	0	1/5	1/5	1/5

IDF:

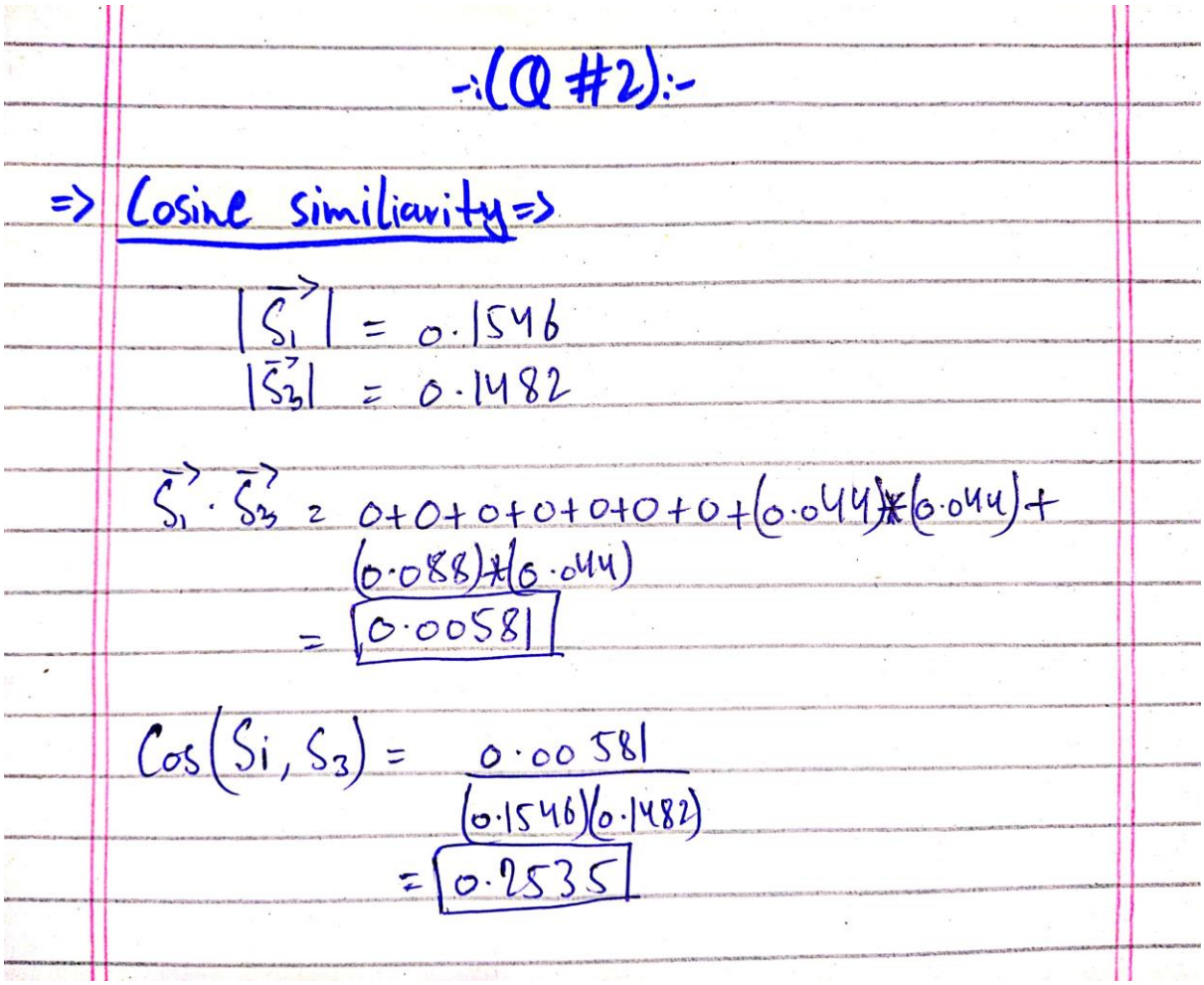
	brown	enjoy	fast	fox	high	jump	run	state	sunshine
IDF's	0.477121	0.477121	0.477121	0.176091	0.477121	0.477121	0.176091	0.176091	0.176091

TF-IDF:

	brown	enjoy	fast	fox	high	jump	run	state	sunshine
0	0.000000	0.50689	0.000000	0.000000	0.000000	0.000000	0.000000	0.385503	0.771006
1	0.670703	0.00000	0.000000	0.510087	0.335352	0.335352	0.255044	0.000000	0.000000
2	0.000000	0.00000	0.549351	0.417796	0.000000	0.000000	0.417796	0.417796	0.417796

Q2. ***Compute the cosine similarity between S1 and S3.***
Answer:

MANUALLY



BY CODE

COSINE:

```
0.4832
```