COMSATS University Islamabad, Lahore Campus

Assignment-4

Course Title: Introduction to Data Science
Course Code: CSC461

Name: Ayesha Tariq
Roll Number: SP20-BCS-020-B
Group: IV
Semester: 6th
Program Name: Bachelor of Computer Science
Submitted To: Dr. Muhammad Sharjeel

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”

For Bag Of words Model:

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

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

Vector S2: [0,0,0,2,2,1,1,1,1,0] (Including comma “,” as a separate token)

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

Tf Model:

	sunshine	state	enjoy	brown	fox	jump	high	,	run	fast
S1	1/2	1/4	1/4	0	0	0	0	0	0	0
S2	0	0	0	1/4	1/4	1/8	1/8	1/8	1/8	0
S3	1/5	1/5	0	0	1/5	0	0	0	1/5	1/5

IDF Model:

	sunshine	state	enjoy	brown	fox	jump	,	high	run	fast
idf	0.176	0.176	0.477	0.477	0.176	0.477	0.477	0.477	0.176	0.477

TF-IDF:

	sunshine	state	enjoy	brown	fox	jump	,	high	run	fast
tf-idf(S1)	0.088	0.044	0.119	0	0	0	0	0	0	0
tf-idf (S2)	0	0	0	0.119	0.044	0.059	0.059	0.059	0.022	0
tf-idf (S3)	0.035	0.035	0	0	0.035	0	0	0	0.035	0.095

Q2. Compute the cosine similarity between S1 and S3.

S1 “sunshine state enjoy sunshine”

S3 “sunshine state fox run fast”

V_S1: [2,1,1,0,0,0,0,0,0,0]

V_S3: [1,1,0,0,1,0,0,0,1,1]

V_S1. V_S3 = 2*1 + 1*1 + 1*0 + 0*0 + 0*1 + 0*0 + 0*0 + 0*0 + 0*1 + 0*1 = 3

| V_S1 | = (2*2 + 1*1 + 1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0) * 0.5 = 2.449

| V_S3 | = (1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 0*0 + 0*0 + 0*0 + 1*1 + 1*1) * 0.5 = 2.236

cos (V_S1, V_S3) = $\frac{(V\_S1.\ V\_S3)}{|V\_S1|\ |V\_S3|} = \frac{3}{2.449 \times 2.236}$ = 0.548