ASSIGNMENT # 5

INTRODUCTION TO DATA SCIENCE

ABDULLAH ADNAN
SP20-BCS-027

IDS-FA22-Assignment **Due Date: 30-12-2022**

Submission: Please upload the PDF report and Python code (preferably iPython notebook) to GitHub.

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

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”

Q2. Compute the cosine similarity between S1 and S3

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

Bag of Words

	sunshine	state	enjoy	brown	fox	jump	high	run	fast
S1	2	1	1	0	0	0	0	0	0
S2	0	0	0	2	2	1	1	1	0
S3	1	1	0	0	1	0	0	1	1

TF Table

	brown	enjoy	fast	fox	high	jump	run	state	sunshine
S1	0.000000	0.25	0.0	0.000000	0.000000	0.000000	0.000000	0.25	0.5
S2	0.285714	0.00	0.0	0.285714	0.142857	0.142857	0.142857	0.00	0.0
S3	0.000000	0.00	0.2	0.200000	0.000000	0.000000	0.200000	0.20	0.2

IDF Table

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

TF-IDF Table

	brown	**enjoy**	**fast**	**fox**	**high**	**jump**	**run**	**state**	**sunshine**
S1	0.00000	0.11928	0.000000	0.000000	0.00000	0.00000	0.000000	0.044023	0.088046
S2	0.13632	0.00000	0.000000	0.050312	0.06816	0.06816	0.025156	0.000000	0.000000
S3	0.00000	0.00000	0.095424	0.035218	0.00000	0.00000	0.035218	0.035218	0.035218

Cosine Similarity

Q2. Compute the cosine similarity between S1 and S3

S1 = “sunshine state enjoy sunshine”
S3 = “sunshine state fox run fast”

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

S1.S3 = 2x1 + 1x1 + 1x0 + 0x0 + 0x1 + 0x0 + 0x0 + 0x1 + 0x1
= 3

| S1 | = $\sqrt{4+1+1+0+0+0+0+0+0}$ = $\sqrt{6}$ = 2.449

| S3 | = $\sqrt{1+1+0+0+1+0+0+1+1}$ = $\sqrt{5}$ = 0.236

$$\cos(\theta) = \frac{3}{(2.449)*(2.238)} = 0.54735$$