INTRODUCTION TO DATA SCIENCE

ASSIGNMENT # 5

DECEMBER 30, 2022

SUBMITTED BY:

NAME: MINAHIL SADIQ
Reg # SP20-BCS-023
Section: 'B'

SUBMITTED TO:

DR. MUHAMMAD SHARJEEL

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"

ANSWER:

Bag Of Words:

Documents	Vocabulary									Total Length
	sunshine	state	enjoy	brown	fox	jump	high	run	fast	
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]

Calculating Term Frequency:

Denoted $tf_{i,d}$ = frequency of (term) word **i** in document **d**

$tf_{i,d}$ = number of times word (term) **i** appears in a document **d** / total number of words (terms) in document **d**

- Term frequencies of all the words in all three sentences:

Documents	**Vocabulary**								
	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

Calculating Inverse Document Frequency:

idf_i = log (total number of documents / number of documents with word (term) i)

- Inverse Document Frequency of all the words:

sunshine	$\log 3/2 = 0.176$
state	$\log 3/2 = 0.176$
enjoy	$\log 3/1 = 0.477$
brown	$\log 3/1 = 0.477$
fox	$\log 3/2 = 0.176$
jump	$\log 3/1 = 0.477$
high	$\log 3/1 = 0.477$
run	$\log 3/2 = 0.176$
fast	$\log 3/1 = 0.477$

Calculating Term Frequency- Inverse Document Frequency:

Tf-Idf = Tf * Idf

Words	**Tf-Idf ---S1**	**Tf-Idf ---S2**	**Tf-Idf ---S3**
sunshine	0.088	0	0.0352
state	0.044	0	0.0352
enjoy	0.11925	0	0
brown	0	0.136	0
fox	0	0.051	0.0352
jump	0	0.068	0
high	0	0.068	0
run	0	0.025	0.0352
fast	0	0	0.0954

QUESTION 2

Compute the cosine similarity between S1 and S3.

ANSWER:

Formula: Cos θ = S1.S3 ÷ |S1| |S3|

Vector Representation of S1 and S3:

S1 = [2,1,1,0,0,0,0,0,0]

S3 = [1,1,0,0,1,0,0,1,1]

S1.S3 = (2*1) + (1*1) +(1*0) + (0*0) + (0*1) + (0*0) + (0*0) + (0*1) + (0*1)

S1.S3 = 2+1 = 3

- **S1.S3 = 3**

|S1| = (2*2 + 1*1 + 1*1) 0.5 = (4+1+1) 0.5 = (6) 0.5 = 2.45

|S3| = (1*1 + 1*1 + 1*1 + 1*1 + 1*1) 0.5 = (1+1+1+1+1) 0.5 = (5) 0.5 = 2.24

- **|S1| = 2.45**
- **|S3| = 2.24**

Putting values in formula:

Cos θ = S1.S3 ÷ |S1| |S3|

Cos(S1,S3) = 3 / (2.45)(2.24) = 3 / 5.47 = 0.547

Cosine similarity between S1 and S3 is:

- **Cos(S1 , S3) = 0.547**