12/30/2022

INTRODUCTION TO DATA SCIENCE

ASSIGNMENT # 05

NAME:

MUHAMMAD SUBHAN KHAN

REGISTRATION NUMBER:

SP20-BCS-098

SECTION & GROUP:

(B)-4

SEMESTER:

6th

ASSIGNMENT:

Introduction to Data Science

QUESTION

Compute the BoW, TF, and IDF model for each of the terms in the following sentences then calculate the tf.idf value

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

Bag of Words (BoW) :-

Documents	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

Term Frequency (tf) :- = $\frac{\text{No. of times word appear in document}}{\text{Total Words}}$

S1 :-

Sunshine = 2/4 ⇒ 1/2
State = 1/4
enjoy = 1/4

S2 :-

brown = 2/7
fox = 2/7
jump = 1/7
high = 1/7
Run = 1/7

S_3 :-

Sunshine = 1/5
state = 1/5
fox = 1/5
run = 1/5
fast = 1/5

Invers Document Frequency :- $\log_{10}\left(\frac{\text{Total Documents}}{\text{Word appear in Documents}}\right)$

sunshine = $\log_{10}(3/2)$ = 0.1762
state = $\log_{10}(3/2)$ = 0.1762
enjoy = $\log_{10}(3/1)$ = 0.477
brown = $\log_{10}(3/1)$ = 0.477
fox = $\log_{10}(3/2)$ = 0.1762
jump = $\log_{10}(3/1)$ = 0.477
high = $\log_{10}(3/1)$ = 0.477
run = $\log_{10}(3/2)$ = 0.1762
fast = $\log_{10}(3/1)$ = 0.477

Term Frequency * Inverse Document Frequency

S1 :-

Sunshine = 2/4 * 0.1762 = 0.0881

state = 1/4 * 0.176 = 0.044

enjoy = 1/4 * 0.477 = 0.11925

S2 :-

braun = 2/7 * 0.477 = 0.136

fox = 2/7 * 0.176 = 0.051

jump = 1/7 * 0.477 = 0.068

high = 1/7 * 0.477 = 0.068

run = 1/7 * 0.176 = 0.025

S3 :-

Sunshine = 1/5 * 0.176 = 0.0352

state = 1/5 * 0.176 = 0.0352

fox = 1/5 * 0.176 = 0.0352

run = 1/5 * 0.176 = 0.0352

fast = 1/5 * 0.477 = 0.0954

TF * IDF Table

Vocabulary	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
braun	0	0.136	0
fox	0	0.051	0
jump	0	0.068	0.0352

high	0	0.068	0
run	0	0.025	0.0352
fast	0	0	0.0954

QUESTION-NO-02

Cosine Similarity between S1 and S3 :-

$$\text{Cosine Similarity} = \frac{\vec{S1} \cdot \vec{S3}}{|\vec{S1}|\,|\vec{S3}|}$$

Document Vectors :

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

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

$$\vec{S1} \cdot \vec{S3} = (2\times1) + (1\times1) + (1\times0) + (0\times0) + (0\times1) + (0\times0) + (0\times0) + (0\times1) + (0\times1)$$

$$= 2 + 1$$

$$\vec{S1} \cdot \vec{S3} = 3$$

$$|\vec{S1}| = \sqrt{(2\times2) + (1\times1) + (1\times1)}$$

$$= \sqrt{4+1+1}$$

$$|\vec{S1}| = 2.45$$

$$|\vec{S3}| = \sqrt{1\times1 + 1\times1 + 1\times1 + 1\times1 + 1\times1}$$

$$= \sqrt{5}$$

$$= 2.24$$

putting values in formula :—

$$\cos\theta = \frac{\vec{S1} \cdot \vec{S3}}{|\vec{S1}|\,|\vec{S3}|}$$

$$= \frac{3}{5.47}$$

$$\boxed{\cos(S1, S3) = 0.547}$$