FA19-BCS-076 (G2)

Fatima Safdar

Intro to Data Science

Assignment 05

Q1:-

⇒ Bag of Words (BoW)

S1 " sunshine state enjoy sunshine"

S2 " brown fox jump high, brown fox run"

S3 " sunshine state fox run fast"

Unique Words = 09

sunshine, state, enjoy, brown,

fox, jump, high, run, fast

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

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

vector S2 : [0 ,0 ,1 ,0 , 1 ,0 ,1 ,2 ,2]

vector S3 : [1 ,1 ,1 ,1 ,0 ,0 ,0 ,0 ,1]

⇒ TF model

Total no. of words in s1 = 4

Total no. of words in s2 = 7

Total no. of words in s3 = 5

In S1

tf ("sunshine") = 2/4 ⇒ 1/2

tf ("state") = 1/4

tf ("enjoy") = 1/4

In S2

tf ("brown") = 2/7	tf("high") = 1/7
tf("fox") = 2/7	tf("run") = 1/7
tf("jump") = 1/7	

In S3

tf("sunshine") = 1/5	tf("run") = 1/5
tf(" state") = 1/5	tf("fast") = 1/5
tf ("fox") = 1/5	

	TF		
	S1	S2	S3
fast	0	0	1/5
sunshine	1/2	0	1/5
run	0	1/7	1/5
state	1/4	0	1/5
high	0	1/7	0
enjoy	1/4	0	0
jump	0	1/7	0
brown	0	2/7	0
fox	0	2/7	1/5

⇒ IDF model

$$idf = \log\Big(\frac{\text{Total no. of documents}}{\text{no. of doc. containing term}}$$

In S1

idf("sunshine") = log (3/2) = 0.18

idf("state") = log (3/1)= 0.48

idf("enjoy") = log (3/1) = 0.48)

In S2

idf("brown") = log (3/2)=0.18

idf("fox") = log (3/2) = 0·18

idf ("jump") = log (3/1) = 0·48

idf ("high") = log (3/1) = 0·48

idf ("run") = log (3/1) = 0·48

In S3

idf ("sunshine") = log (3/1) = 0·48

idf ("state") = log (3/1) = 0·48

idf ("fox") = log (3/1) = 0·48

idf ("run") = log (3/1) = 0·48

idf (" fast") = log (3/1) = 0·48

	IDF		
	S1	S2	S3
fast	0	0	0·48
sunshine	0·18	0	0·48
run	0	0·48	0·48
state	0·48	0	0·48
high	0	0·48	0
enjoy	0·48	0	0
jump	0	0·48	0
brown	0·18	0	0
fox	0	0·18	0·48

⇒ TF x 9DF

	TF x 9DF		
	S1	S2	S3
fast	0	0	0.096
sunshine	0.09	0	0.096
run	0	0.068	0.096
state	0.12	0	0.096
high	0	0.068	0
enjoy	0.12	0	0
jump	0	0.068	0
brown	0	0	0
fox	0	0.051	0.096

Q2:- Cosine Similarity

$$\cos(S1, S3) = \frac{(S1 . S3)}{|S1| \, |S3|}$$

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

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

$(S1 . S3) = (0+2+0+1+0+0+0+0+0) \Rightarrow 3$

$|S1| = (2\times2 + 1\times1 + 1\times1)^{1/2} = 2.44$

$|S3| = (1\times1 + 1\times1 + 1\times1 + 1\times1)^{1/2} = 2.23$

$$\cos(S1, S3) = \frac{3}{(2.44)(2.23)} = \frac{3}{5.4412}$$

$$\boxed{\cos(S1, S3) = 0.5513}$$