M Saim Jamil

Fa19-bcs-125

G II

Assig#5

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.

Solution

Words:

S1 “sunshine state enjoy sunshine”

S2 “brown fox jump high, brown fox run”

S3 “sunshine state fox run fast”

Vocabulary

'brown', 'enjoy', 'fast', 'fox', 'high', 'jump', 'run', 'state', 'sunshine'

BOW model

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

Term frequencies

Tf = Val/total

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

IDF

Idf (word)= log(total/value of word)

S1: “sunshine state enjoy sunshine”

Idf(“sunshine”) = log(3/2) = 0.176

Idf(“state”) = log(3/2) = 0.176

Idf(“enjoy”) = log(3/1) = 0.477

S2: “brown fox jump high, brown fox run”

Idf(“brown”) = log(3/1) = 0.477

Idf(“fox”) = log(3/2) = 0.176

Idf(“jump”) = log(3/1) = 0.477

Idf(“high”) = log(3/1) = 0.477

Idf(“run”) = log(3/2) = 0.176

S3: “sunshine state fox run fast”

Idf(“sunshine”) = log(3/2) = 0.176

Idf(“state”) = log(3/2) = 0.176

Idf(“fox”) = log(3/2) = 0.176

Idf(“run”) = log(3/2) = 0.176

Idf(“fast”) = log(3/1) = 0.477

	'brown'	'enjoy'	'fast'	'fox'	'high'	'jump'	'run'	'state'	'sunshine'	Total
S1	0	0. 477	0	0	0	0	0	0.176	0.176	4
S2	0.477	0	0	0.176	0. 477	0. 477	0.176	0	0	7
S3	0	0	0.477	0.176	0	0	0.176	0.176	0.176	5

Tf-idf

	'brown'	'enjoy'	'fast'	'fox'	'high'	'jump'	'run'	'state'	'sunshine'	Total
S1	0	0.119	0	0	0	0	0	0.044	0.088	4
S2	0.136	0	0	0.050	0.068	0.068	0.025	0	0	7
S3	0	0	0.095	0.035	0	0	0.035	0.035	0.035	5

Q2

Cosine Similarity between S1 and S3
TF Vector:
S1= [2/4, 1/4, 1/4, 0, 0, 0, 0, 0, 0]
S3 = = [1/5, 1/5, 0, 0, 1/5, 0, 0, 1/5, 1/5]
S1 . S3 = 2/4 * 1/5 + 1/4 * 1/5 + 1/4 * 0 + 0 * 0 + 0 * 1/5 +0 * 0 + 0 * 0 + 0 * 1/5 + 0 * 1/5
S1.S3 = 0.15000
|S1| = (2/4 * 2/4 + 1/4 * 1/4 + 1/4 * 1/4 + 0 * 0 +0 * 0 + 0 * 0 + 0 * 0 + 0 * 0 + 0 * 0) ^1/2
|S1| = 0.61237
|S3|=(1/5 * 1/5 + 1/5 * 1/5 + 0 * 0 + 0 * 0 + 1/5 *1/5 + 0 * 0 + 0 * 0 + 1/5 * 1/5 + 1/5 * 1/5) ^1/2
|S3| = 0.44721

The Cosine similarity between S1 and S3 are as below:

COS(S1,S3) = 0.15/0.61237*0.44721
COS(S1,S3) = 0.54773