FA19-BCS-001

Muhammad Nouman Tahir

Introduction to data Science

Assignment 5

Vocabulary (all the unique words) in the three sentences is:

There are 9 unique words:

“sunshine”, “state”, “enjoy”, “brown”, “fox”, “jump”, “high”, “run”, “fast”

Bag of Words (BoW):

Sentences	sunshine	state	enjoy	brown	fox	jump	high	run	fast	Total Length
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**

Term Frequencies (TF):

TF for term ‘word’ = (number of times ‘word’ appears in sentence) / (total number of terms in sentence)

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

Inverse Document Frequency (Idf):

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

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

Idf-Sentences	**sunshine**	**state**	**enjoy**	**brown**	**fox**	**jump**	**high**	**run**	**fast**	**Total Length**
idf-S1	0.176	0.176	**0.477**	**0**	**0**	**0**	**0**	**0**	**0**	**4**
idf-S2	0	**0**	**0**	0.477	0.176	**0.477**	**0.477**	**0.176**	**0**	**7**
idf-S3	0.176	0.176	**0**	**0**	0.176	**0**	**0**	**0.176**	**0.477**	**5**

Term Frequency Inverse Document Frequency (Tf-Idf):

Term Frequency inverse document frequency = tf * idf

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

Question No. 2

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) = 0.15000/0.61237*0.44721

COS(S1,S3) = 0.54773