Assignment 05

NLP

Muhammad Tayyab

FA19-BCS-010

* S1 “sunshine state enjoy sunshine”
* S2 “brown fox jump high, brown fox run”
* S3 “sunshine state fox run fast”

**Vocabulary:**

[sunshine, state, enjoy, brown, fox, jump, high, run, fast]

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  | BoW |  |  |  |  |  |
|  |  | 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 |

# 

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

# IDF

Idf = log (total number of documents / number of documents with word that term)

* 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 |  |  |  |  |  |  |
|  | sunshine | state | enjoy | brown | fox | jump | high | run | fast | Total length |
| S1 | 0.176 | 0.176 | 0.477 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| S2 | 0 | 0 | 0 | 0.477 | 0.176 | 0.477 | 0.477 | 0.176 | 0 | 7 |
| S3 | 0.176 | 0.176 | 0 | 0 | 0.176 | 0 | 0 | 0.176 | 0.477 | 5 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | Tf-idf |  |  |  |  |  |
|  | sunshine | state | enjoy | brown | fox | jump | high | run | fast | Total length |
| TfidfS1 | 2/4\*0.176 | 1/4\*0.176 | 1/4\*0.477 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| TfidfS2 | 0 | 0 | 0 | 2/7\*0.477 | 2/7\*0.176 | 1/7\*0.477 | 1/7\*0.477 | 1/7\*0.176 | 0 | 7 |
| TfidfS3 | 1/5\*0.176 | 1/5\*0.176 | 0 | 0 | 1/5\*0.176 | 0 | 0 | 1/5\*0.176 | 1/5\*0.477 | 5 |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | sunshine | state | enjoy | brown | fox | jump | high | run | fast | Total length |
| TfidfS1 | 0.088 | 0.044 | 0.119 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| TfidfS2 | 0 | 0 | 0 | 0.136 | 0.050 | 0.068 | 0.068 | 0.025 | 0 | 7 |
| TfidfS3 | 0.035 | 0.035 | 0 | 0 | 0.035 | 0 | 0 | 0.035 | 0.095 | 5 |

**Question:02**

## Cosine Similarity

cos (S1, S3) = S1.S3/|S1||S3|

### Taking 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 =

|S1| =

|S3| =

**cos (S1, S3)** =