**Introduction to Data Science**

**Hunain Javed**

**FA18-BCS-061**

**Assignment # 5**

**­­S1:** “sunshine state enjoy sunshine”

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

**S3:** “sunshine state fox 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 |
| **S2** | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 5 |

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}

**TF:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **sunshine** | **state** | **enjoy** | **brown** | **fox** | **jump** | **high** | **run** | **fast** |
| **tf - S1** | 1/2 | 1/4 | 1/4 | 0 | 0 | 0 | 0 | 0 | 0 |
| **tf - S2** | 0 | 0 | 0 | 2/7 | 2/7 | 1/7 | 1/7 | 1/7 | 0 |
| **tf - S2** | 1/5 | 1/5 | 0 | 0 | 1/5 | 0 | 0 | 1/5 | 1/5 |

tf - S1 = {1/2, 1/4, 1/4, 0, 0, 0, 0, 0, 0}

tf - S2 = {0, 0, 0, 2/7, 2/7, 1/7, 1/7, 1/7, 0}

tf - S3 = {1/5, 1/5, 0, 0, 1/5, 0, 0, 1/5, 1/5}

**IDF:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **sunshine** | **state** | **enjoy** | **brown** | **fox** | **jump** | **high** | **run** | **fast** |
| **idf** | 0.176 | 0.176 | 0.477 | 0.477 | 0.176 | 0.477 | 0.477 | 0.176 | 0.477 |

**TF.IDF:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **sunshine** | **state** | **enjoy** | **brown** | **fox** | **jump** | **high** | **run** | **fast** |
| **tf.idf - S1** | 0.088 | 0.044 | 0.119 | 0 | 0 | 0 | 0 | 0 | 0 |
| **tf.idf - S2** | 0 | 0 | 0 | 0.136 | 0.050 | 0.068 | 0.068 | 0.025 | 0 |
| **tf.idf - S2** | 0.0352 | 0.0352 | 0 | 0 | 0.0352 | 0 | 0 | 0.0352 | 0.0954 |

S1 = {0.088, 0.044, 0.119, 0 0 0 0 0 0}

S2 = {0, 0, 0, 0.136, 0.050, 0.068, 0.068, 0.025, 0}

S3 = {0.0352, 0.0352, 0, 0, 0.0352, 0, 0, 0.0352, 0.0954}

**Cosine Similarity between S1 and S3:**

cos = S1.S3/|S1|\*|S3|

S1.S3 = (0.088\*0.0352) + (0.044\*0.0352) + 0 + 0 + 0 + 0 + 0 + 0 + 0

S1.S3 = 0.0030+0.00154

**S1.S3 = 0.00454**

|S1| = (0.088\*0.088 + 0.044\*0.044 + 0.119\*0.119 + 0 + 0 + 0 + 0 + 0 + 0) sq.root

|S1| = (0.0077 + 0.0019 + 0.00014) sq.root

|S1| = (0.0097) sq.root

**|S1| = 0.098**

|S3| = (0.0352\*0.0352 + 0.0352\*0.0352 + 0 + 0 + 0.0352\*0.0352 + 0 + 0 + 0.0352\*0.0352 + 0.0954\*0.0954) sq.root

|S3| = (0.0012 + 0.0012 + 0.0012 + 0.0012 + 0.0091) sq.root

|S3| = (0.0139) sq.root

**|S3| = 0.1179**

cos = S1.S3/|S1|\*|S3|

cos = 0.00454/0.098\*0.1179

**cos = 0.3929**