**Assignment # 04**

**Introduction to Data Science**



**Name:** Hamna Shahbaz

**Reg. #:** FA21-BSE-048

**Section:** BSE-C

**Question 01:**

**Sentences:**

**S1.** “data science is one of the most important courses in computer science”

**S2.** “this is one of the best data science courses”

**S3.** “the data scientists perform data analysis”

* **BoW**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **this** | **the** | **is** | **one** | **of** | **most** | **important** | **best** | **perform** | **data** | **computer** | **analysis** | **courses** | **science** | **in** | **scientists** | **Total Words** |
| **S1** | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 2 | 1 | 0 | **12** |
| **S2** | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | **9** |
| **S3** | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 1 | 0 | 0 | 0 | 1 | **6** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Vectors:** |  |  |  |
| **Vector S1** | [0 1 1 1 1 1 1 0 0 1 1 0 1 2 1 0] | | |
| **Vector S2** | [1 1 1 1 1 0 0 1 0 1 0 0 1 1 0 0] | | |
| **Vector S3** | [0 1 0 0 0 0 0 0 1 2 0 1 0 0 0 1] | | |

* **TF**

Total Number of words in **S1**: 12

TF(the) = 1/12

TF(this) = 0/12 = 0

TF(is) = 1/12

TF(one) = 1/12

TF(of) = 1/12

TF(most) = 1/12

TF(important) = 1/12

TF(best) = 0/12 = 0

TF(perform) = 0/12 = 0

TF(data) = 1/12

TF(computer) = 1/12

TF(analysis) = 0/12 = 0

TF(courses) = 1/12

TF(science) = 2/12 = 1/6

TF(in) = 1/12

TF(scientists) = 0/12 = 0

Total Number of words in **S2**: 9

TF(the) = 1/12

TF(this) = 1/12

TF(is) = 1/12

TF(one) = 1/12

TF(of) = 1/12

TF(most) = 0/12 = 0

TF(important) = 0/12 = 0

TF(best) = 1/12

TF(perform) = 0/12 = 0

TF(data) = 1/12

TF(computer) = 0/12 = 0

TF(analysis) = 0/12 = 0

TF(courses) = 1/12

TF(science) = 1/12

TF(in) = 0/12 = 0

TF(scientists) = 0/12 = 0

Total Number of words in **S3**: 6

TF(the) = 1/12

TF(this) = 0/12 = 0

TF(is) = 0/12 = 0

TF(one) = 0/12 = 0

TF(of) = 0/12 = 0

TF(most) = 0/12 = 0

TF(important) = 0/12 = 0

TF(best) = 0/12 = 0

TF(perform) = 1/12

TF(data) = 2/12 = 1/6

TF(computer) = 0/12 = 0

TF(analysis) = 1/12

TF(courses) = 0/12 = 0

TF(science) = 0/12 = 0

TF(in) = 0/12 = 0

TF(scientists) = 1/12

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **this** | **the** | **is** | **one** | **of** | **most** | **important** | **best** | **perform** | **data** | **computer** |  |  | **Analysis** | **courses** |  | **science** | **in** | **scientists** | **Total Words** |
| **S1** | 0 | 1/12 | 1/12 | 1/12 | 1/12 | 1  /  12 | 1  /  1  2 | 0 | 0 | 1/12 | 1  /  1  2 |  |  | 0 | 1  /  1  2 |  | 1  /  6 | 1/12 | 0 | **12** |
| **S2** | 1/12 | 1/12 | 1/12 | 1/12 | 1/12 | 0 | 0 | 1/12 | 0 | 1/12 | 0 |  |  | 0 | 1  /  1  2 |  | 1  /  1  2 | 0 | 0 | **9** |
| **S3** | 0 | 1/12 | 0 | 0 | 0 | 0 | 0 | 0 | 1  /  1  2 | 1/6 | 0 |  |  | 1  /  1  2 | 0 |  | 0 | 0 | 1  /  1  2 | **6** |

* **IDF**

IDF for **S1**:

IDF(data) = log(3/3) = log(1) = 0

IDF(science) = log(3/2) = log(1.5) = 0.176

IDF (is) = log(3/2) = log(1.5) = 0.176

IDF (one) = log(3/2) = log(1.5) = 0.176

IDF(of) = log(3/2) = log(1.5) = 0.176

IDF (the) = log(3/3) = log(1) = 0

IDF(most) = log(3/1) = log(3) = 0.477

IDF(important) = log(3/1) = log(3) = 0.477

IDF(courses) = log(3/2) = log(1.5) = 0.176

IDF(in) = log(3/1) = log(3) = 0.477

IDF(computer) = log(3/2) = log(1.5) = 0.176

IDF for **S2**:

IDF(this) = log(3/1) = log(3) = 0.477

IDF(is) = log((3/2) = log(1.5) = 0.176

IDF(one) = log(3/2) = log(1.5) = 0.176

IDF(of) = log(3/2) = log(1.5) = 0.176

IDF (the) = log(3/3) = log(1) = 0

IDF(best) = = log(3/1) = log(3) = 0.477

IDF(data) = log(3/3) = log(1) = 0

IDF(science) = log(3/2) = log(1.5) = 0.176

IDF(courses) = log(3/2) = log(1.5) = 0.176

IDF for **S3:**

IDF (the) = log(3/3) = log(1) = 0

IDF(data) = log(3/3) = log(1) = 0

IDF(scientists) = log(3/1) = log(3) = 0.477

IDF(perform) = log(3/1) = log(3) = 0.477

IDF(analysis) = log(3/1) = log(3) = 0.477

|  |  |
| --- | --- |
|  | IDF |
| **this** | 0.477 |
| **the** | 0 |
| **is** | 0.176 |
| **one** | 0.176 |
| **of** | 0.176 |
| **most** | 0.477 |
| **important** | 0.477 |
| **best** | 0.477 |
| **perform** | 0.477 |
| **data** | 0 |
| **computer** | 0.176 |
| **analysis** | 0.477 |
| **courses** | 0.176 |
| **science** | 0.176 |
| **in** | 0.477 |
| **scientists** | 0.477 |

* **TF-IDF**

TF-IDF for **S1:**

Tf-idf(data) = 1/12 \* 0 = 0

Tf-idf(science) = 1/6 \* 0.176 = 0.0293

Tf-idf(is) = 1/12 \* 0.176 = 0.014

Tf-idf(one) = 1/12 \* 0.176 = 0.014

Tf-idf(of) = 1/12 \* 0.176 = 0.014

Tf-idf(the) = 1/12 \* 0 = 0

Tf-idf(most) = 1/12 \* 0.477 = 0.039

Tf-idf(important) = 1/12 \* 0.477 = 0.039

Tf-idf(courses) = 1/12 \* 0.176 = 0.014

Tf-idf(in) = 1/12 \* 0.477 = 0.039

Tf-idf(computer) = 1/12 \* 0.176 = 0.014

TF-IDF for **S2:**

Tf-idf(this) = 1/9 \* 0.477 = 0.053

Tf-idf(is) = 1/9 \* 0.176 = 0.019

Tf-idf(one) = 1/9 \* 0.176 = 0.019

Tf-idf(of) = 1/9 \* 0.176 = 0.019

Tf-idf(the) = 1/9 \* 0 = 0

Tf-idf(best) = 1/9 \* 0.477 = 0.053

Tf-idf(data) = 1/9 \* 0 = 0

Tf-idf(science) = 1/9 \* 0.176 = 0.019

Tf-idf(courses) = 1/9 \* 0.176 = 0.019

TF-IDF for **S3: “the data scientists perform data analysis”**

Tf-idf(the) = 1/6 \* 0 = 0

Tf-idf(data) = 1/3 \* 0 = 0

Tf-idf(scientists) = 1/6 \* 0.477 = 0.079

Tf-idf(perform) = 1/6 \* 0.477 = 0.079

Tf-idf(analysis) = 1/6 \* 0.477 = 0.079

|  |  |  |  |
| --- | --- | --- | --- |
|  | TF-IDF S1 | TF-IDF S2 | TF-IDF S3 |
| **this** | 0 | 0.053 | 0 |
| **the** | 0 | 0 | 0 |
| **is** | 0.014 | 0.019 | 0 |
| **one** | 0.014 | 0.019 | 0 |
| **of** | 0.014 | 0.019 | 0 |
| **most** | 0.039 | 0 | 0 |
| **important** | 0.039 | 0 | 0 |
| **best** | 0 | 0.053 | 0 |
| **perform** | 0 | 0 | 0.079 |
| **data** | 0 | 0 | 0 |
| **computer** | 0.014 | 0 | 0 |
| **analysis** | 0 | 0 | 0.079 |
| **courses** | 0.176 | 0.019 | 0 |
| **science** | 0.029 | 0.019 | 0 |
| **in** | 0.039 | 0 | 0 |
| **scientists** | 0 | 0 | 0.079 |

**Question 02:**

* **Cosine Similarity**

|  |  |
| --- | --- |
| **Vector S1** | [0 1 1 1 1 1 1 0 0 1 1 0 1 2 1 0] |
| **Vector S2** | [1 1 1 1 1 0 0 1 0 1 0 0 1 1 0 0] |
| **Vector S3** | [0 1 0 0 0 0 0 0 1 2 0 1 0 0 0 1] |

Vector Length of S1 = **(∑ (S1i)2 ) ^ 1.5** = 0 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 1 + 1 + 0 + 1 + 4 + 1 + 0 = **14**

Vector Length of S2 = **(∑ (S2i)2 ) ^ 1.5 =** 1 + 1 + 1 + 1 + 1 + 0 + 0 + 1 + 0 + 1+ 0 + 0 + 1 + 1 + 0 + 0 = **9**

Vector Length of S3 = **(∑ (S3i)2 ) ^ 1.5 =**  0 + 1 + 1 + 4 + 1 + 1 = **8**

**Dot Product of S1 and S2 =** 0\*1 + 1\*1 + 1\*1 + 1\*1 + 1\*1 + 1\*0 + 1\*0 + 0\*1 + 0\*0 + 1\*1 + 1\*0 + 0\*0 + 1\*1 + 2\*1 + 1\*0 + 0\*0 = 1+1+1+1+1+1+2 = 8

**Dot Product of S1 and S3 =** 0\*0 + 1\*1 + 1\*0 + 1\*0 + 1\*0 + 1\*0 + 1\*0 + 0\*0 + 0\*1 + 1\*2 + 1\*0 + 0\*1 + 1\*0 + 2\*0 + 1\*0 = 1+2 = 3

**Dot product of S2 and S3 =** 1\*0 + 1\*1 + 1\*0 + 1\*0 + 1\*0 + 0\*0 +0\*0 + 1\*0 + 0\*1+ 1\*2+ 0\*0 + 1\*1+ 1\*0 + 1\*0 + 0\*0 + 0\*1 = 1+2+1 = 3

**Cosine S1-S2 =** 8/ 14\* 9 = 0.0634

**Cosine S2-S3 =** 3/ 9\*8 = 0.04166

**Cosine S1-S3 =** 3/ 14\* 8 = 0.0267

* **Manhattan Similarity**
* **Euclidean Similarity**