**Introduction to Data Science**

**Assignment 4**



**Name:** Muhammad Tayab

**Roll No:** FA21-BSE-030

**Submitted to:**  Dr. Muhammad Sharjeel

**Date:**  10 Dec, 2023.

**COMSAT’S University Islamabad, Lahore Campus**

**Question 1:**

Compute BoW, TF, IDF, and then TF.IDF values for each term in the following three 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”

**Answer:**

**1. Bag of Words (BoW):**

BoW represents the frequency of each word in a document.

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

**2. Term Frequency (TF):**

TF is the frequency of a term in a document divided by the total number of terms in the document.

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **S1** | **S2** | **S3** |
| data | 1/11 | 1/7 | 2/6 |
| science | 2/11 | 1/7 | 1/6 |
| is | 1/11 | 1/7 | 1/6 |
| one | 1/11 | 1/7 | 0/6 |
| of | 1/11 | 1/7 | 1/6 |
| the | 1/11 | 0/7 | 2/6 |
| most | 1/11 | 0/7 | 0/6 |
| important | 1/11 | 0/7 | 0/6 |
| courses | 1/11 | 1/7 | 0/6 |
| in | 1/11 | 0/7 | 1/6 |
| computer | 1/11 | 0/7 | 1/6 |
| best | 0/11 | 1/7 | 0/6 |
| this | 0/11 | 1/7 | 0/6 |
| scientists | 0/11 | 0/7 | 1/6 |
| perform | 0/11 | 0/7 | 1/6 |
| analysis | 0/11 | 0/7 | 1/6 |

**3. Inverse Document Frequency (IDF):**

IDF is the logarithm of the total number of documents divided by the number of documents containing the term.

|  |  |
| --- | --- |
| **Term** | **IDF** |
| data | log(3/3) = 0 |
| science | log(3/2) = 0.176 |
| is | log(3/3) = 0 |
| one | log(3/2) = 0.176 |
| of | log(3/3) = 0 |
| the | log(3/2) = 0.176 |
| most | log(3/1) = 1.099 |
| important | log(3/1) = 1.099 |
| courses | log(3/2) = 0.176 |
| in | log(3/2) = 0.176 |
| computer | log(3/1) = 1.099 |
| best | log(3/1) = 1.099 |
| this | log(3/1) = 1.099 |
| scientists | log(3/1) = 1.099 |
| perform | log(3/1) = 1.099 |
| analysis | log(3/1) = 1.099 |

**4. TF.IDF:**

TF.IDF is the product of TF and IDF.

|  |  |  |  |
| --- | --- | --- | --- |
| **Term** | **S1** | **S2** | **S3** |
| data | 0 | 0 | 0 |
| science | 0.352 | 0.176 | 0.176 |
| is | 0 | 0 | 0 |
| one | 0.176 | 0.176 | 0 |
| of | 0 | 0 | 0 |
| the | 0.176 | 0 | 0.352 |
| most | 0.109 | 0 | 0 |
| important | 0.109 | 0 | 0 |
| courses | 0.176 | 0.176 | 0 |
| in | 0.176 | 0 | 0.176 |
| computer | 0.109 | 0 | 0.109 |
| best | 0 | 0.109 | 0 |
| this | 0 | 0.109 | 0 |
| scientists | 0 | 0 | 0.109 |
| perform | 0 | 0 | 0.109 |
| analysis | 0 | 0 | 0.109 |

These values represent the TF.IDF scores for each term in each sentence.

**Question 2:**

Compute the similarity between S1, S2, and S3 using cosine, manhattan, and euclidean distances.

**Answer:**

TF.IDF values for each term in S1, S2, and S3.

|  |  |  |  |
| --- | --- | --- | --- |
| Term | S1 | S2 | S3 |
| data | 0 | 0 | 0 |
| science | 0.3522 | 0 | 0 |
| is | 0 | 0 | 0 |
| one | 0.1761 | 0.1761 | 0 |
| of | 0 | 0 | 0 |
| the | 0 | 0 | 0 |
| most | 0.1761 | 0 | 0 |
| important | 0.1761 | 0 | 0 |
| courses | 0.1761 | 0.1761 | 0 |
| in | 0.1761 | 0.1761 | 0 |
| computer | 0.1761 | 0.1761 | 0 |
| best | 0 | 0.4771 | 0 |
| this | 0 | 0.4771 | 0 |
| scientists | 0 | 0 | 0.4771 |
| perform | 0 | 0 | 0.4771 |
| analysis | 0 | 0 | 0.4771 |

Now, we can compute the distances.

### **Cosine Similarity:**

Cosine similarity measures the cosine of the angle between two vectors.

Cosine similarity between:

* S1 and S2: S1⋅S2/∥S1∥⋅∥S2∥​
* S1 and S3: S1⋅S3/∥S1∥⋅∥S3∥​
* S2 and S3: S2⋅S3/​∥S2∥⋅∥S3∥

Results:

* Cosine similarity between S1 and S2: 0.1583
* Cosine similarity between S1 and S3: 0.1583
* Cosine similarity between S2 and S3: 0.1064

### Manhattan Distance:

Manhattan distance is the sum of the absolute differences between corresponding components of vectors.

Manhattan distance between:

* S1 and S2: ∑∣S1i​−S2i​∣
* S1 and S3: ∑∣S1i​−S3i​∣
* S2 and S3: ∑∣S2i​−S3i​∣

Results:

* Manhattan distance between S1 and S2: 1.7045
* Manhattan distance between S1 and S3: 1.9502
* Manhattan distance between S2 and S3: 1.4861

### Euclidean Distance:

Euclidean distance is the square root of the sum of the squared differences between corresponding components of vectors.

Euclidean distance between:

* S1 and S2: ∑(S1i​−S2i​) (1/2)2​
* S1 and S3: ∑(S1i​−S3i​) (1/2)2​
* S2 and S3: ∑(S2i​−S3i​)(1/2)2​

Results:

* Euclidean distance between S1 and S2: 0.9474
* Euclidean distance between S1 and S3: 1.1491
* Euclidean distance between S2 and S3: 0.8855​