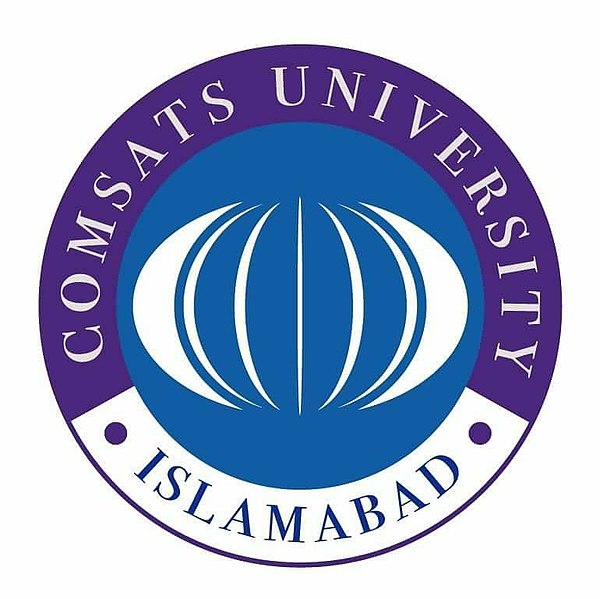
****

**INTRODUCTION TO DATA SCIENCE**

**-ASSIGNMENT # 5-**

**Registration Number:** SP20-BCS-095

**Student Name:** Rohaan Farooq

**Course:** Introduction to Data Science (CSC461)

**Group:** BCS 6 - G2-A

**Assignment #:** 4

**Submitted to:** Prof. Muhammad Sharjeel

**Date:** 30th December 2022

Question 1

**Compute the BoW model, TF Model, and IDF model for each of the terms in the following three sentences. Then calculate the TF.IDF values**

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 Model**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **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 |

Vector S1: [2, 1, 1, 0, 0, 0, 0, 0, 0]

Vector S2: [0, 0, 0, 2, 2, 1, 1, 1, 0]

Vector S3: [1, 1, 0, 0, 1, 0, 0, 1, 1]

**TF Model**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sunshine** | **State** | **Enjoy** | **Brown** | **Fox** | **Jump** | **High** | **Run** | **Fast** |
| **tf - S1** | 2/4 | 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 Model**

|  |  |
| --- | --- |
|  | **Idf** |
| **Sunshine** | 0.18 |
| **State** | 0.18 |
| **Enjoy** | 0.48 |
| **Brown** | 0.48 |
| **Fox** | 0.18 |
| **Jump** | 0.48 |
| **High** | 0.48 |
| **Run** | 0.18 |
| **Fast** | 0.48 |

**TFIF Model**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **tfidf - S1** | **tfidf – S2** | **tfidf- S3** |
| **Sunshine** | **0.09** | **0** | **0.036** |
| **State** | **0.045** | **0** | **0.036** |
| **Enjoy** | **0.12** | **0** | **0** |
| **Brown** | **0** | **0.137** | **0** |
| **Fox** | **0** | **0.051** | **0.036** |
| **Jump** | **0** | **0.068** | **0** |
| **High** | **0** | **0.068** | **0** |
| **Run** | **0** | **0.026** | **0.036** |
| **Fast** | **0** | **0** | **0.096** |

Question 2

**Compute the cosine similarity between S1 and S3.**

Vector S1: [2, 1, 1, 0, 0, 0, 0, 0, 0]

Vector S3: [1, 1, 0, 0, 1, 0, 0, 1, 1]

cos(S1, S3) =

= (2\*1 + 1\*1 + 1\*0 + 0\*0 +0\*1 + 0\*0 + 0\*0 + 0\*1 + 0\*1) = 3

|S1| = = 2.45

|S3| = = 2.24

cos(S1, S3) = = **0.5466**

Hence, the cosine similarity between S1 and S2 is **0.55**

The End