# COMSATS UNIVERSITY ISLAMABAD

## Lahore Campus

**NAME:** M. Ammar Mukhtar

**ROLL NO:** FA21-BSE-141

**TEACHER NAME:** Prof. M. Sharjeel

**ASSIGNMENT NO:** 04

**ASSIGNMENT TITLE:** BoW, TF, IDF, Similarity

Date:

Sun Mon Tue Wed Thu Fri Sat

# Q No 1

## BAG OF WORDS: (Word occurance in each sentence)

| | Data | Science | is | one | of | the | most | important | courses | in |
|---|---|---|---|---|---|---|---|---|---|---|
| S1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| S2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 |
| S3 | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

| | compiter | this | best | scientist | perform | analysis | Total = 16 |
|---|---|---|---|---|---|---|---|
| S1 | 1 | 0 | 0 | 0 | 0 | 0 | 12 |
| S2 | 0 | 1 | 1 | 0 | 0 | 0 | 9 |
| S3 | 0 | 0 | 0 | 1 | 1 | 1 | 6 |

S1 : [1 2 1 1 1 1 1 1 1 1 0 0 0 0]

S2 : [1 1 1 1 1 1 0 0 1 0 0 1 1 0 0 0]

S3 : [2 0 0 0 1 0 0 0 0 0 0 0 1 1 1]

## TERM FREQUENCY: ( $TF = \frac{\text{No. of Terms}}{\text{Total Terms in Doc}}$ )

| | Data | Scienence | is | one | of | the | most | important | courses | in |
|---|---|---|---|---|---|---|---|---|---|---|
| S1 | 1/16 | 2/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 | 1/16 |
| S2 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 0 | 0 | 1/9 | 0 |
| S3 | 2/6 | 0 | 0 | 0 | 0 | 1/6 | 0 | 0 | 0 | 0 |

| | computer | this | best | scientist | perform | analysis. |
|---|---|---|---|---|---|---|
| S1 | 1/16 | 0 | 0 | 0 | 0 | 0 |
| S2 | 0 | 1/9 | 1/9 | 0 | 0 | 0 |
| S3 | 0 | 0 | 0 | 1/6 | 1/6 | 1/6 |

Date:

Sun Mon Tue Wed Thu Fri Sat

## IDF: $\left( IDF = \log \left( \frac{\text{No. of Doc}}{\text{Term occurance in Doc}} \right) \right)$

| | Data | science | is | one | of | the | most | important | courses |
|---|---|---|---|---|---|---|---|---|---|
| idf | 0 | 0.17 | 0.17 | 0.17 | 0.17 | 0 | 0.4 | 0.4 | 0.17 |

| | in | computer | this | best | scientist | perform | analysis. |
|---|---|---|---|---|---|---|---|
| | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 | 0.4 |

## IDF × TF: TF-IDF = TF × IDF

| | Data | science | is | on | of | the | most | important | courses |
|---|---|---|---|---|---|---|---|---|---|
| ~~Tf-idf~~ | 0 | | | | | | 0 | | |

| | in | computer | this | best | scientist | perform | analysis. |
|---|---|---|---|---|---|---|---|

| | Data | Science | is | one | of | the | most | important | courses |
|---|---|---|---|---|---|---|---|---|---|
| S1 | 0 | 0.34 | 0.17 | 0.17 | 0.17 | 0 | 0.14 | 0.14 | 0.17 |
| S2 | 0 | 0.17 | 0.17 | 0.17 | 0.17 | 0 | 0 | 0 | 0.17 |
| S3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

| | in | computer | this | best | scientist | perform | analysis. |
|---|---|---|---|---|---|---|---|
| S1 | 0.14 | 0.14 | 0 | 0 | 0 | 0 | 0 |
| S2 | 0 | 0 | 0.4 | 0.4 | 0 | 0 | 0 |
| S3 | 0 | 0 | 0 | 0.4 | 0.4 | 0.4 | 0.4 |

Date:

Sun Mon Tue Wed Thu Fri Sat

# Q NO 2

**COSINE:** Cosine Similarity $(A, B) = \frac{A \cdot B}{|A| \cdot |B|}$

(i) Cosine Similarity $(S1, S2) = \frac{S_1 \cdot S_2}{|S_1| \cdot |S_2|}$

$S_1 : [1\ 2\ 1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0]$

$S_2 : [1\ 1\ 1\ 1\ 1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 0]$

$S_1 \cdot S_2 = 1+2+1+1+1+1+0+0+1+0+0+0+0+0+0+0 = 8$

$|S_1| = \sqrt{1+4+1+1+1+1+1+1+1+1+1+1+1} = \sqrt{14}$

$|S_2| = \sqrt{1+1+1+1+1+1+1+1+1} = \sqrt{9} = 3$

cosine similarity $(S_1, S_2) = \frac{8}{\sqrt{14} \cdot 3} = 0.713$

(ii) $\text{Cosine}(S_1, S_3) = \frac{S_1 \cdot S_3}{|S_1|\ |S_3|} = \frac{2+1+0}{\sqrt{14}\ \sqrt{8}} = 0.28$

$\cos(S_1, S_3) = 0.28$

(iii) $\cos(S_2, S_3) = \frac{S_2 \cdot S_3}{|S_2| \cdot (S_3)} = \frac{2+1+0}{3 \times \sqrt{8}} = \frac{1}{\sqrt{8}}$

$\cos(S_2, S_3) = 0.35$

Date:.....

Sun Mon Tue Wed Thu Fri Sat

## MANHATTAN:

$$\text{Distance}(A,B) = \sum_{i=1}^{n} |A_i - B_i|$$

(i) $\text{Distance}(S_1, S_2) = 1+1+1+1+1+1+1+0 = 7$

(ii) $\text{Distance}(S_1, S_3) = 1+2+1+1+1+1+1+1+1+1+1+1 = 14$

(iii) $\text{Distance}(S_2, S_3) = 1+1+1+1+1+1+1+1+1+1+1+1 = 11$

## EUCLIDEAN:

$$\text{Distance}(A,B) = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2}$$

(i) $\text{Distance}(S_1, S_2) = \sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2} = \sqrt{7} = 2.64$

(ii) $\text{Distance}(S_1, S_3) = \sqrt{1^2+4^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2}$

$= \sqrt{17} = 4.12$

(iii) $\text{Distance} = \sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2} = \sqrt{11}$

$= 3.3$