FA20-BSE-057(A)
Esha Naeem.

"Assignment #04"
(IDS)

Question #02:-

- **Bag of words:-**

Vocabulary:-

'data', 'science', 'is', 'one', 'of', 'the', 'most', 'important', 'courses', 'in', 'computer', 'this', 'best', 'scientists', 'perform', analysis'   Total length - 16

Bag of word Vectors:-

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

Total length of S1 = 12
Total length of S2 = 9
Total length of S3 = 6

- **Term Frequency:-**

S1 tf:-

tf = 'data'
$\boxed{tf = 1/12}$
tf = 'science'
$\boxed{tf = 2/12 = 1/6}$

Inverse Document Frequency.

Idf = 'data'
Idf = log 3/3
Idf = 0

tf= 'is'

$\boxed{tf = 1/12}$

tf= 'one'

$\boxed{tf = 1/12}$

tf= 'of'

$\boxed{tf = 1/12}$

tf = 'the'

$\boxed{tf = 1/12}$

tf = 'most'

= 1/12

tf = 'important'

= 1/12

tf = 'courses'

= 1/12

tf = 'in'

= 1/12

tf = 'computer'

= 1/12

tf = 'this'

= 0

tf = 'best'

= 0

tf = 'scientists'

= 0

tf = 'perform'

= 0

tf = 'analysis'

=

Idf = 'science'

Idf = log(3/2)

Idf = 0.176

Idf = 'is'

Idf = log(3/2)

Idf = 0.176

Idf = 'one'

Idf = log(3/2)

Idf = 0.176

Idf = 'of'

Idf = log(3/2)

Idf = 0.176

Idf = 'the'

Idf = log(3/3)

Idf = 0

Idf = 'most'

Idf = log(3/1)

Idf = 0.477

Idf = 'important'

Idf = log(3/1)

Idf = 0.477

Idf = 'courses'

Idf = log(3/2)

Idf = 0.176

S2 tf:-

tf ='data'
tf = 1/9

tf ='science'
tf = 1/9

tf ='is'
= 1/9

tf ='one'
tf = 1/9

tf ='of'
tf = 1/9

tf ='the'
tf = 1/9

tf ='most'
tf = 0/9

tf ='important'
tf = 0/9

tf ='courses'
tf = 1/9

tf ='in'
tf = 0/9

tf ='computer'
tf = 0/9

tf ='this'
tf = 1/9

Idf= 'in'
Idf = log(3/1)
Idf = 0.477

Idf ='computer'
Idf = log(3/1)
Idf = 0.477

Idf ='this'
Idf = log(3/1)
Idf = 0.477

Idf ='best'
Idf = log(3/1)
Idf = 0.477

Idf ='scientists'
Idf = log(3/1)
Idf = 0.477

Idf = 'perform'
Idf = log(3/1)
Idf = 0.477

Idf = 'analysis'
Idf = log(3/1)
Idf = 0.477

tf = 'best'

tf = 1/9

tf = 'scientist'

tf = 0/9

tf = 'perform'

tf = 0/9

tf = 'analysis'

tf = 0/9

S3 tf:-

tf = 'data'

tf = 2/6 = 1/3

tf = 'science'

tf = 0/6

tf = 'is'

tf = 0/6

tf = 'one'

tf = 0/6

tf = 'f'

tf = 0/6

tf = 'the'

tf = 1/6

tf = 'most'

tf = 0/6

tf = 'important'

tf = 0/6

tf = 'courses'

tf = 0/6

tf='in'
tf= 0/6
tf='computer'
tf= 0/6
tf='this'
tf= 0/6
tf='best'
tf= 0/6
tf='scientists
tf= 1/6
tf='perform'
tf= 1/6
tf='analysis'
tf= 1.6

- Tf·IDf = ?

Vocabulary	S1 (tf*Idf)	S2 (tf*Idf)	S3 (tf*Idf)
'data'	=(1/12)(0) tfIdf = 0	=(1/9)(0) tfIdf=0	= (1/3)(0) tfIdf = 0
'science'	=(1/6)(0.176) tfIdf=0.0293	=(1/9)(0.176) tfIdf =0.019	=(0/6)(0.176) tfIdf = 0
'is'	=(1/12)(0.176) tfIdf=0.014	= (1/9)(0.176) tfIdf=0.019	= (0/6)(0.176) tfIdf= 0
'one'	=(1/12)(0.176) tfIdf=0.014	=(1/9)(0.176) tfIdf=0.019	=(0/6)(0.176) tfIdf = 0

'of'	=(1/12)(0·176) tfIdf=0·014	=(1/9)(0·176) tfIdf=0·019	=(0/6)(0·176) tfIdf=0
'the'	=(1/12)(0) tfIdf=0	=(1/9)(0) tfIdf=0	=(1/6)(0) tfIdf=0
'most'	=(1/12)(0·477) tfIdf=0.0397	=(0/9)(0·477) tfIdf=0	=(0/6)(0·477) tfIdf=0
'important'	=(1/12)(0·477) tfIdf=0.0397	=(0/9)(0·477) tfIdf=0	=(0/6)(0·477) tfIdf=0
'courses'	=(1/12)(0·176) tfIdf=0·014	=(1/9)(0·176) tfIdf=0·019	=(0/6)(0·176) tfIdf=0
'in'	=(1/12)(0·477) tfIdf=0.0397	=(0/9)(0·477) tfIdf=0	=(0/6)(0·477) tfIdf=0
'computer'	=(1/12)(0·477) tfIdf=0.0397	=(0/9)(0·477) tfIdf=0	=(0/6)(0·477) tfIdf=0
'this'	=(0)(0·477) tfIdf=0	=(1/9)(0·477) tfIdf=0·053	=(0/6)(0·477) tfIdf=0

'best'	=(0)(0·477) tfIdf=0	=(1/9)(0·477) tfIdf=0·053	=(0/6)(0·477) tfIdf=0
'scientists'	=(0)(0·477) tfIdf=0	=(0/9)(0·477) tfIdf=0	=(1/6)(0·477) tfIdf=0·0795
'perform'	=(0)(0·477) tfIdf=0	=(0/9)(0·477) tfIdf=0	=(1/6)(0·477) tfIdf=0·0795
'analysis'	=(0)(0·477) tfIdf=0	=(0/9)(0·477) tfIdf=0	=(1/6)(0·477) tfIdf=0·0795

Q#02:- Cosine Similarity:- (BOW)

(S1, S2)

$$\frac{0+0+0+1+1+0+0+1+0+1+1+0+2+0+1+0}{\left(\sqrt{0^2+0^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+0+2^2+0^2}\right)\left(\sqrt{0^2+0^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+1^2+0+2^2}\right)}$$

+

$$= \frac{7}{(\sqrt{12})(\sqrt{9})}$$

$$\cos \boxed{(S1, S2) = 0.712}$$

(S1, S3)

$$\frac{0+0+0+0+2+0+0+0+0+0+0+0+0+0+0+0+1+0}{\left(\sqrt{S1^2}\right)\left(\sqrt{1^2+0+0^2+0^2+2^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+2^2+0}\right)}$$

$$= \frac{3}{\sqrt{12}\times\sqrt{7}}$$

$$\boxed{\cos(S1, S3) = 0.2835}$$

(S2, S3)

$$\frac{0+0+0+0+2+0+0+0+0+0+0+0+0+0+1+0}{(\sqrt{S_2})^2\,(\sqrt{S_3})^2}$$

$$= \frac{3}{\sqrt{9}\sqrt{7}}$$

$$\boxed{\cos(S2, S3) = 0.353}$$

Manhatten Distance

$$|S1-S2| = |-0| + |0-1| + |1-0| + |1-1| + |1-1| + |1-0|$$
$$+ |1-1| + |1-0| + |1-1| + |(-1| + |0-0 +$$
$$|2-1| + |0-1| + |-1 + |0-1|$$

$$\boxed{S1, S2 = 7}$$

(S1, S3)

$$|S1-S3| = |0-1| + |0-0| + |1-0| + |-0| + |1-2| + |-0|$$
$$+ |1-0| + |1-0| + |1-0| + |-0| + |0-1$$
$$+ |1-2| + |1-0| + |1-0| + |1-1| + |0-1$$
$$+ |2-0| + |0-1|$$

$$\boxed{(S1, S3) = 14}$$

(S2, S3)

$$|S2-S3| = |0-1| + |1-0| + |0-0| + |1-0| + |1-2| + |0-0|$$
$$+ |1-0| + |0-1| + |1-0|$$
$$+ |0-0| + |1-0| + |0-1| + |1-0| + |0-0| + |1-0| +$$
$$+ |0-0| + |1-1| + |1-0|$$
$$+ |0-1| + |1-1|$$
$$+ |0-0| + |0-0|$$
$$|-2| + |0-0|$$

$$\boxed{S2, S3 = 11}$$

Euclidean Distance :-

$$\sqrt{(S1, S2)^2} = \sqrt{0 + (-1)^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2 + 0^2 + (-1)^2}$$

$$\boxed{\sqrt{(S1, S2)^2} = 2.6458}$$

(S1, S3)

$$\sqrt{(S1,S3)^2} = \sqrt{(-1)^2+0^2+1^2+1^2+(-1)^2+1^2+1^2+1^2+1}$$
$$1^2+1^2+1^2+(-1)^2+0^2+0$$

$$\boxed{\sqrt{(S1,S3)^2}=4.0}$$

(S2, S3)

$$\sqrt{(-1)^2+1^2+0^2+1^2+(-1)^2+(0)^2+0^2+1^2+0^2+1^2}$$
$$+1^2+1^2+(-1)^2+1^2+(-1)^2+0^2+1^2$$

$$\boxed{\sqrt{(S2\,S3)^2}=3.316}$$

Q#02 TfIdf (Similarities)

TfIdf vectors = S1 [0, 0.0293, 0.014, 0.014, 0.014, 0, 0.039, 0.039, 0.014, 0.039, 0.039, 0, 0, 0, 0, 0]

S2 = [0, 0.019, 0.019, 0.019, 0.019, 0, 0, 0, 0.019, 0, 0, 0.053, 0.053, 0, 0, 0]

S3 = [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.0795, 0.0795, 0.0795]

$$\text{cos\_similarity}(S1, S2) = \frac{S1 \cdot S2}{\|S1\|\|S2\|}$$

$$\text{cos\_similarity}(S1, S3) = \frac{S1 \cdot S3}{\|S1\|\|S3\|}$$

$$\text{cos\_similarity}(S2, S3) = \frac{S2 \cdot S3}{\|S2\|\|S3\|}$$

$$\text{cos-similarity}(S1, S2) = \frac{\begin{gathered}(0.0 + (0.0293)(0.019) + (0.014)(0.019) \\ + (0.014)(0.019) + (0.014)(0.019) + \\ (0)(0) + (0.039)(0) + (0.039)(0) + \\ (0.014)(0.019) + (0.039)(0) + \\ (0.039)(0) + 0 + 0 + 0 + 0 + 0)\end{gathered}}{\begin{gathered}\left(\sqrt{0^2 + (0.0293)^2 + (0.014)^2 + (0.014)^2 + (0.014)^2 + \\ 0^2 + (0.039)^2 + (0.039)^2 + (0.014)^2 + (0.0 \\ + (0.039)^2 + 0^2 + \cdots}\right)\left(\sqrt{0^2 + (0.019)^2 + (0.019) \\ + (0.019)^2 + (0.019^2 + 0^2 + \cdots + (0.019)^2 + 0^2 \\ + (0.053)^2 + (0.053)^2 + 0^2 + 0^2 + 0^2}\right)\end{gathered}}$$

$$= \frac{1.62 \times 10^{-3}}{(0.087)(0.0861)}$$

$\boxed{\text{Cos-Similarity of } (S1, S2) = 0.2162}$

→ Cos-Similarity (S1, S3) $0+0+0+0+0+0+0+$
$0+0+0+0+0+0+0+0$
$+0$

$$= \frac{0+0+0+0+0+0+0+0+0+0+0+0+0+0+0+0}{(\sqrt{S1^2})\cdot(\sqrt{S3^2}}$$

$$= \frac{0}{(0.087)(\sqrt{0^2+0^2+''+(0.079)^2+(0.0795)^2+(0.09}}$$

$\boxed{\text{Cos-Similarity } (S1, S3) = 0}$

Cos-Similarity (S2, S3)

$$= \frac{0+''+0}{(\sqrt{S2^2})(\sqrt{S3})^2} = \frac{0}{0}$$

$\boxed{\text{Cos-Similarity}(S2, S3) = 0}$

Manhatten Distance:-

$(S_1, S_2) = \Sigma|S_1 - S_2|$

$|0.014 - 0.019| +$

$= |0 - 0| + |0.0293 - 0.019| + |0 - 0| + |0.0390|$

$|0.014 - 0.019| + |0.014 - 0.019| + |0.014 - 0| +$

$++|0.039 - 0| + |0.039 - 0.019| + |0.014 - 0| +$

$|0.039 - 0| + |0.039 - 0.053| + '' + ''$

$= 0 + 0.0103 + 0.005 + 0.005 + 0.005 + 0 +$
$0.039 + 0.039 + 0.005 + 0.039 + 0.039 + 0$
$+0+0+0+0$

$\boxed{(S1, S2) = }$ 0.296

$(S1, S3) = |0-0| + |0.0293-0| + |0.014-0|| +$
$|0.014-0| + |0.014-0| + |0-0| + |0.039-0| +$
$|0.039-0| + |0.014-0| + |0.039-0| + |0.039-0|$
$+ |0-0| + \ldots$

$(S1, S3) = 0.49$

$(S2, S3) = |0-0| + |0.019-0| + |0.019-0| +$
$|0.019-0| + |0.019-0| + |0-0| + |0-0| +$
$|0.019-0| + |0-0| + |0-0| + |0.0531-0| +$
$|0.053-0| + |0-0| + |0.53-0| + |0.53-0| + \ldots$

$\boxed{(S2, S3) = 0.446}$

Eucidean Distance:-

$(S1, S2) = \sqrt{(0.01)^2 + (-5\times10^{-3})^2 + (-5\times10^{-3})^2 + (0.04)^2 + (-5\times10^{-3})^2 + (0.04)^2 + (0.04)^2 + (0.04)^2 + (-0.053)^2 + (-0.053)^2}$

$\boxed{(S1, S2) = 0.110}$

$(S1, S3) = \sqrt{(0.03)^2 + (0.015)^2 + (0.015)^2 + (0.04) + (0.015)^2 + (0.04) + (0.04) + (0.07)^2 + (-0.08)^2 + (-0.08)^2 + (-0.08)^2}$

$\boxed{(S1, S3) = 0.165}$

$$(S2,S3) = \sqrt{(0.02)^2+(0.02)^2+(0.02)^2+(0.02)^2 + (0.053)^2+(0.04)+(-0.08)+ (-0.08)^2+(-0.08)^2}$$

$$\boxed{(S2,S3) = 0.163} \text{ Ans}$$