Q1. 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"

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

TF table:

	analysis	best	computer	courses	data	important	in	is	most	of	one	perform	science	scientists	the	this
S1	0	0	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{2}{12}$	0	$\frac{1}{12}$	0
S2	0	$\frac{1}{9}$	0	$\frac{1}{9}$	$\frac{1}{9}$	0	0	$\frac{1}{9}$	0	$\frac{1}{9}$	$\frac{1}{9}$	0	$\frac{1}{9}$	0	$\frac{1}{9}$	$\frac{1}{9}$
S3	$\frac{1}{6}$	0	0	0	$\frac{2}{6}$	0	0	0	0	0	0	$\frac{1}{6}$	0	$\frac{1}{6}$	$\frac{1}{6}$	0

TERMS	IDF
analysis	$log(\frac{3}{1})$
best	$log(\frac{3}{1})$
computer	$log(\frac{3}{1})$
courses	$log(\frac{3}{1})$
data	$log(\frac{3}{13})$
important	$log(\frac{3}{1})$
in	$log(\frac{3}{1})$
is	$log(\frac{3}{2})$
most	$log(\frac{3}{1})$

of	$log(\frac{3}{2})$
one	$log(\frac{3}{2})$
perform	$log(\frac{3}{1})$
science	$log(\frac{3}{2})$
scientists	$log(\frac{3}{1})$
the	$log(\frac{3}{3})$
this	$log(\frac{3}{1})$

Term	S1	S2	S3
analysis	0.000000	0.000000	0.079520
best	0.000000	0.053013	0.000000
computer	0.039760	0.000000	0.000000
courses	0.014674	0.019566	0.000000
data	0.000000	0.000000	0.000000
important	0.039760	0.000000	0.000000
in	0.039760	0.000000	0.000000
is	0.014674	0.019566	0.000000
most	0.039760	0.000000	0.000000
of	0.014674	0.019566	0.000000
one	0.014674	0.019566	0.000000
perform	0.000000	0.000000	0.079520
science	0.029349	0.019566	0.000000

scientists	0.000000	0.000000	0.079520
the	0.000000	0.000000	0.000000
this	0.000000	0.053013	0.000000

Q2. Compute the similarity between S1, S2, and S3 using cosine, manhattan, and euclidean distances.

Cosine similarity :(using BOW vectors)

Cosine Similarity between S1 and S2:

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

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

Dot Product = 0x0 + 0x1 + 1x0 + 1x1 + 1x1 + 1x0 + 1x0 + 1x1 + 1x0 + 1x1 + 1x1 + 0x0 + 2x1 + 0x0 + 1x1 + 0x1

= 7

Magnitude of S1 = $\sqrt{0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 2^2 + 0^2}$

=$\sqrt{12}$

Magnitude of S2 = $\sqrt{0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 2^2 + 0^2}$

=$\sqrt{9}$

cosine (S1,S2) =$\frac{7}{\sqrt{12}\, X \sqrt{9}}$

= 0.712

Cosine Similarity between S2 and S3:

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

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

= $0X1 + 0X0 + 1X0 + 1X0 + 1X2 + 1X0 \ + 1X0 + 1X0 + 1X0 + 1X0 + 1X0 + 0X1 + 2X0 + 0X1 + 1X1 + 0X0$

= 3

Magnitude of S3 = $\sqrt{1^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 0^2 + 2^2 + 0^2}$

=$\sqrt{7}$

cosine (S1,S3) $= \frac{3}{\sqrt{12}\, X \sqrt{7}}$

= 0.2835

Cosine Similarity between S2 and S3:

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

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

$= 0X1 + 1X0 + 0X0 + 1X0 + 1X2 + 0X0 \ + 0X0 + 1X0 + 0X0 + 1X0 + 1X0 + 0X1 + 1X0 + 0X1 + 1X1 + 1X0$

= 3

cosine (S2,S3) $= \frac{3}{\sqrt{9}\, X \sqrt{7}}$

= 0.3536

Manhattan Distance:

Manhattan Distance between S1 and S2:

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

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

= |0−0|+|0−1|+|1−0|+|1−1|+|1−1|+|1−0|+|1−0|+|1−1|+|1−0|+|1−1|+|1−1|+|0−0|+|2−1|+|0−0|+|1−1|+|0−1|

= 0+1+1+0+0+1+1+0+1+0+0+0+1+0+0+1=70+1+1+0+0+1+1+0+1+0+0+0+1+0+0+1

=7

Manhattan Distance between S1 and S3:

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

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

= |0−1|+|0−0|+|1−0|+|1−0|+|1−2|+|1−0|+|1−0|+|1−0|+|1−0|+|1−0|+|1−0|+|0−1|+|2−0|+|0−1|+|1−1|+|0−0|

= 1+0+1+1+1+1+1+1+1+1+1+1+2+1+0+0=141+0+1+1+1+1+1+1+1+1+1+1+2+1+0+0

= 14

Manhattan Distance between S2 and S3:

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

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

=
|0−1|+|1−0|+|0−0|+|1−0|+|1−2|+|0−0|+|0−0|+|1−0|+|0−0|+|1−0|+|1−0|+|0−1|+|1−0|+|0−1|+|1−1|+|1−0||0−1|+|1−0|+|0−0|+|1−0|+|1−2|+|0−0|+|0−0|+|1−0|+|0−0|+|1−0|+|1−0|+|0−1|+|1−0|+|0−1|+|1−1|+|1−0|

= 1+1+0+1+1+0+0+1+0+1+1+1+1+1+1+0+1=111+1+0+1+1+0+0+1+0+1+1+1+1+1+1+0+1

= 11

Euclidean Distance:

Euclidean Distance between S1 and S2:

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

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

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

=2.6458

Euclidean Distance between S1 and S3:

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

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

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

= 4.0

Euclidean Distance between S2 and S3:

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

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

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

= 3.3166

cosine similarity : (using TF vectors)

Between S_1 & S_2 :

S_1 : [0, 0, 0.083, 0.083, 0.083, 0.083, 0.083, 0.083
0.083, 0.083, 0.083, 0, 0.1667, 0, 0.083, 0]

S_2 : [0, 0.11, 0, 0.11, 0.11, 0, 0, 0.11, 0, 0.11,
0.11, 0, 0.11, 0, 0.11, 0.11]

$S_1 \cdot S_2 = 0\times0 + 0\times0.11 + 0.083\times0 + 0.083\times0.11 + 0.083\times0.11 + 0.083\times0 + 0.083\times0 + 0.083\times0.11 + 0.083\times0 + 0.083\times0.11 + 0.083\times0.11 + 0 + 0.1667\times0.11 + 0\times0 + 0.083\times0.11 + 0$

$= 0.065073$

$|S_1| = \sqrt{(0.083)^2 + (0.083)^2 + (0.083)^2 + (0.083)^2 + (0.083)^2 + (0.083)^2 + (0.083)^2 + (0.083)^2 + (0.083)^2 + (0.1667)^2}$

$|S_2| = \sqrt{(0.11)^2 + (0.11)^2 + (0.11)^2 + (0.11)^2 + (0.11)^2 + (0.11)^2 + (0.11)^2 + (0.11)^2 + (0.11)^2 + (0.11)^2}$

$|S_1| = 0.2996$

$|S_2| = 0.34785$

$|S_1||S_2| = 0.1042$

$\cos(S_1, S_2) = \dfrac{0.065073}{|0.2996| \times |0.34785|}$

$= 0.71217$

Between S_1 & S_3:

$$S_1 \cdot S_3 = 0 + 0 + 0 + 0 + 0.0278 + 0.013889$$

$$= 0.041666$$

$$|S_1| = \sqrt{0.041667}$$

$$|S_3| = \sqrt{0.1111}$$

$$\cos(S_1, S_3) = 0.2835$$

Between S_2 & S_3:

$$S_2 \cdot S_3 = 0.012346 + 0.037037 + 0.12346 + 0.012345$$

$$= 0.074074$$

$$|S_2| = \sqrt{0.074074} *$$

$$|S_3| = \sqrt{0.1111}$$

$$\cos(S_2, S_3) = \frac{0.074074}{\sqrt{0.074074} \times \sqrt{0.111}}$$

$$= 0.3536$$

Manhattan Distance b/w S_1 & S_2:

$$\text{distance}(S_1, S_2) = 0 + 0.11 + 0.083 + 0.0277 + 0.027 + 0.277 + 0.0277 + 0.0277 + 0.55 + 0.0277 + 0.111$$

Sum = 0.7778

Between S_1 & S_3:

= 0.1667 + 0.833 + 0.833 + 0.25 + 0.083 + 0.083 + 0.083 + 0.083 + 0.0833 + 0.1666 + 0.1666 + 0.1666 + 0.083

= 1.6667

Between S_2 & S_3:

= 0.1667 + 0.11 + 0.111 + 0.222 + 0.11 + 0.11 + 0.11 + 0.1667 + 0.11 + 0.1667 + 0.556 + 0.11

= 1.556

Euclidean Distance

Between S_1 & S_2:

= 0.0123 + 0.0069 + 0.000772 + 0.00073 + ... 0.000772

Sqrt = $\sqrt{0.060109}$

= 0.2453

Between S_1 & S_3:

= $\sqrt{\text{sum of squared differences}}$

= 0.4859

Between S_2 & S_3 :

$= \sqrt{\text{sum of squared differences}}$

$= 0.4714$

Cosine similarity (using TF-IDF):

Between S_1 & S_2:

$S_1 = [0, 0, 0.03976, 0.0146, 0.03976, 0.039, 0.014674, 0.0397, 0.01467, 0.0146, 0, 0.0293, 0, 0, 0]$

$S_2 = [0, 0.053013, 0, 0.01956, 0, 0, 0, 0.019, 0, 0.019, 0.019, 0, 0.019, 0, 0, 0.053]$

$S_3 = [0.079, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0.079, 0, 0.079, 0, 0]$

$S_1 . S_2 = 0.00028 + 0 + 0 + 0 + 0.00028 + 0 + 0 + 0.00028 + 0 + 0.000574 + 0$

$= 0.001438$

$|S_1| = \sqrt{0.0059}$

$|S_2| = \sqrt{0.007417}$

$\cos(S_1, S_2) = 0.2126$

Between S_1 & S_3:

$S_1 \cdot S_3 = 0$

$\cos(S_1, S_3) = 0$

Between S_2 & S_3:

$S_2 \cdot S_3 = 0$

$\cos(S_2, S_3) = 0$

Manhattan Distance (using TF-IDF)

Between S_1 & S_2:

$= 0 + 0.053013 + 0.039 + 0.004 + \dots + 0.053$

$= 0.778$

Between S_1 & S_3:

$= 0.079 + 0.03976 + 0.0146 + 0 + 0 + 0$

$= 1.667$

Between S_2 & S_3:

$= 0.0795 + 0.53013 + 0.019566 + \dots 0053$

$= 1.555$

Eculidean Distance (using TF-IDF)

Between S_1 & S_2:

$= 0 + 0.0028 + 0.00151 + 0.00024 + \dots + 0.0628$

$= \sqrt{0.060109}$

$= 0.2453$

Between S_1 & S_3

$= 0.0063 + 0.001579 + 0.000215 + 0.001579$
$+ 0.001579 + 0.000215 + 0.001579 + 0.000215$
$+ 0.000215 + 0.006327 + 0.000862 + 0.006323$
$= 0.026176$
$= \sqrt{0.026176}$
$= 0.4859$

Between S_2 & S_3:

$= 0.0063 + 0.0028 + 0.00038 + 0.00038 +$
$0.000382 + 0.00038 + 0.0063 + 0.000382$
$+ 0.0063 + 0.0028$
$= 0.026176$
$= \sqrt{0.026176}$
$= 0.1618$