① Q1 Bag of Words.

data science is one of the most important courses in computer this best scientists perform analysis

S_1	1	2	1	1	1	1	1	1	1	1	1	0	0	0	0	0	
S_2	1	1	1	1	1	1	0	0	1	0	0	1	1	0	0	0	
S_3	2	0	0	0	0	1	0	0	0	0	0	0	0	1	1	1	

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

$S_2 = [1\ 1\ 1\ 1\ 1\ 1\ 0\ 0\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 0]$

$S_3 = [2\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1\ 1\ 1]$

Term frequency.

data science is one of the most important courses in computer this best scientists perform analysis.

S_1	$\frac{1}{12}$	$\frac{2}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{12}$	0	0	0	0	0	
S_2	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	$\frac{1}{9}$	0	0	$\frac{1}{9}$	0	0	$\frac{1}{9}$	$\frac{1}{9}$	0	0	0	
S_3	$\frac{2}{6}$	0	0	0	0	$\frac{1}{6}$	0	0	0	0	0	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	

IDf & tf IDF

IDF	idf	tf-Idf(S1)	tf-Idf(S2)	tf-idf(S3)
data	log(3/3) = 0	0	1/9 x 0 = 0	2/6 x 0 = 0
Science	log(3/2) = 0.17609	2/12 x 0.17609 = 0.02935	1/9 x 0.17609 = 0.0196	0
is	log(3/2) = 0.17609	1/12 x 0.17609 = 0.015	1/9 x 0.17609 = 0.0196	0
one	log(3/2) = 0.17609	1/12 x 0.17609 = 0.015	1/9 x 0.17609 = 0.0196	0
of	log(3/2) = 0.17609	1/12 x 0.17609 = 0.015	1/9 x 0.17609 = 0.0196	0
the	log(3/3) = 0	1/12 x 0 = 0	1/9 x 0 = 0	1/6 x 0 = 0
most	log(3/1) = 0.4771	1/12 x 0.4771 = 0.03976	0 x 0.4771 = 0	0
important	log(3/1) = 0.4771	1/12 x 0.4771 = 0.03976	0 x 0.4771 = 0	0
cources	log(3/2) = 0.17609	1/12 x 0.17609 = 0.015	1/9 x 0.17609 = 0.0196	0
in	log(3/1) = 0.4771	1/12 x 0.4771 = 0.03976	0 x 0.4771 = 0	0
computer	log(3/1) = 0.4771	1/12 x 0.4771 = 0.03976	0 x 0.4771 = 0	0
this	log(3/1) = 0.4771	0	1/9 x 0.4771 = 0.0530	0
best	log(3/1) = 0.4771	0	1/9 x 0.4771 = 0.053	0
scientist	log(3/1) = 0.4771	0	0 x 0.4771 = 0	1/6 x 0.4771 = 0.07952
perform	log(3/1) = 0.4771	0	0 x 0.4771 = 0	1/6 x 0.4771 = 0.07952
analysis	log(3/1) = 0.4771	0	0 x 0.4771 = 0	1/6 x 0.4771 = 0.07952

Q2) Cosine:

Bag of words:

$$\cos\theta = \frac{\bar{S_1} \cdot \bar{S_2} \cdot \bar{S_3}}{|S_1||S_2||S_3|}$$

$$\cos(S_1, S_2) = \frac{S_1 \cdot S_2}{|S_1||S_2|}$$

$$S_1 \cdot S_2 = 1.1 + 2.1 + 1.1 + 1.1 + 1.1 + 1.1 + 1.0 + 1.0 + 1.1 + 1.0 + 1.0 + 1.1 + 1.0 + 1.0 + 0.1 + 0.1 + 0.0 + 0.0 + 0.0$$

$$S_1 \cdot S_2 = 9$$

$$|S_1| = (1 + 4 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 0 + 0 + 0 + 0 + 0)^{10.5}$$

$$= 14^{10.5} = 3.7417$$

$$|S_2| = 1 + 1 + 1 + 1 + 1 + 0 + 0 + 1 + 0 + 0 + 1 + 1 + 0 + 0 + 0$$

$$= \sqrt{9} = 3.$$

$$|S_1||S_2| = 11.2251$$

- $\cos(S_1, S_2) = 0.8017$.

$$\cos(S_2, S_3) = \frac{S_2 \cdot S_3}{|S_2||S_3|}$$

$$S_2 \cdot S_3 = 2 + 1 + 0 = 3.$$

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

$$|S_3| = \sqrt{6} = 2.4495$$

- $\cos(S_2, S_3) = 0.4082$

• $\cos(S_1, S_3) = \frac{S_1 S_3}{|S_1||S_3|}$

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

$S_1 S_3 = 2+1+0 = 3$

$|S_1| = 3.7417 \quad |S_3| = \sqrt{6} = 2.4495$

$\cos(S_1, S_2) = 0.327.$

Euclidean (Bag of words) distance.

$$e\text{-}d(S_1, S_2) = \sqrt{\sum_{i=1}^{n} (S1_i - S2_i)^2}$$

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

$$= \sqrt{7} = 2.646.$$

$$e\text{-}d(S1, S3) = \sqrt{\sum_{i=1}^{n} (S1_i - S3_i)^2}$$

$$= [(-1)^2 + (2)^2 + (1)^2 + (1)^2 + (1)^2 + (0)^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + (-1)^2 + (-1)^2 + (-1)^2$$

$$= \sqrt{16} = 4$$

$$ed(S_2, S_3) = \sqrt{\sum_{i=1}^{n} (S2_i - S3_i)^2}$$

$$[(-1)^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 0^2 + 0^2 + 1^2 + 1^2 + (-1)^2 + (-1)^2 + (-1)^2]^{1/2}$$

$$= \sqrt{11} = 3.32.$$

Manhattan distances (Bag of words)

$$m\text{-}d(S_1, S_2) = \boxed{\sum_{i=1}^{n} |S_{1i} - S_{2i}|}$$

$$= [\,1+1+0+0+0+0+1+1+0+1+1+1+1+0+0+0\,] = 8$$

$$m\text{-}d(S_1, S_3) = \sum_{i=1}^{n} |S_{1i} - S_{3i}|$$

$$[\,1+2+1+1+1+0+1+1+1+1+1+0+0+1+1+1\,] = 14$$

$$md(S_2, S_3) = \sum_{i=1}^{n} |S_{2i} - S_{3i}|$$

$$[\,1+1+1+1+1+0+0+0+1+0+0+1+1+1+1+1\,]$$

$$= 11.$$