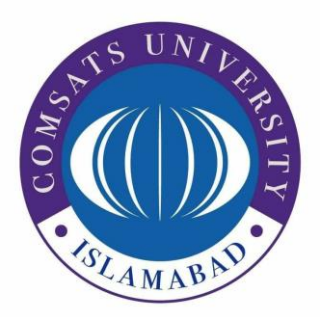
INTRODUCTION TO DATA SCIENCE

ASSIGNMENT # 04

Submitted By:

Arooba Masood

FA20-BSE-092



COMSATS Institute of Information Technology

Department of Computer Science

Question: 1

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"

Answer:

1. Bag of Words:

First we generate vocabulary of words:

Vocabulary: [analysis, best, computer, courses, data, important, perform, science, scientists, the, this, one, is, most, in, of]

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

BoW Vectors:

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

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

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

2. **TF**
3. **IDF**
4. **TF.IDF**

Question: 2

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

Answer:

Cosine Similarities:

Cosine Similarity between S1 and S2:

Cosine similarity = $\frac{S1.S2}{|S1|.|S2|}$

$= \frac{(0\times1)+(0\times0)+(1\times0)+(1\times1)+(1\times1)+(1\times0)+(0\times0)+(2\times1)+(0\times0)+(1\times1)+(0\times1)+(1\times1)+(1\times1)+(1\times0)+(1\times0)+(1\times1)}{\sqrt{14}\times\sqrt{9}}$

$= \frac{0+0+0+1+1+0+0+2+0+1+0+1+1+0+0+1}{\sqrt{14}\times\sqrt{9}}$

$= \frac{8}{\sqrt{14}\times\sqrt{9}}$

$= \frac{8}{11.224972}$

$= 0.71269$

Cosine Similarity between S1 and S3:

Cosine similarity = $\frac{S1.S3}{|S1|.|S3|}$

$= \frac{(0\times1)+(0\times0)+(1\times0)+(1\times0)+(1\times2)+(1\times0)+(0\times1)+(2\times0)+(0\times1)+(1\times1)+(0\times0)+(1\times0)+(1\times0)+(1\times0)+(1\times0)+(1\times0)}{\sqrt{14}\times2\sqrt{2}}$

$= \frac{0+0+0+0+2+0+0+0+0+1+0+0+0+0+0+0}{\sqrt{14}\times2\sqrt{2}}$

$= \frac{3}{\sqrt{14}\times2\sqrt{2}}$

$= \frac{3}{10.58300}$

$= 0.28347$

Cosine Similarity between S3 and S2:

Cosine similarity = $\frac{S2.S3}{|S2|.|S3|}$

$= \frac{(0\times1)+(1\times0)+(0\times0)+(1\times0)+(1\times2)+(0\times0)+(0\times1)+(1\times0)+(0\times1)+(1\times1)+(1\times0)+(1\times0)+(1\times0)+(0\times0)+(0\times0)+(1\times0)}{\sqrt{9}\times2\sqrt{2}}$

$= \frac{0+0+0+0+2+0+0+0+0+1+0+0+0+0+0+0}{\sqrt{9}\times2\sqrt{2}}$

$= \frac{3}{\sqrt{9}\times2\sqrt{2}}$

$= \frac{3}{8.48528}$

$= 0.35355$

Manhattan Distance

Manhattan distance between S1 and S2:

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

$= 1 + 1 + 1 + 1 + 1 + 1 + 1$

$= 7$

Manhattan distance between S1 and S3:

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

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

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

$= 1 + 1 + 1 + 1 + 1 + 1 + 2 + 1 + 1 + 1 + 1 + 1 + 1$

$= 14$

Manhattan distance between S2 and S3:

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

$= 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1$

$= 11$

Euclidean Distance

Euclidean distance between S1 and S2:

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

$= \sqrt{1 + 1 + 1 + 1 + 1 + 1 + 1}$

$= \sqrt{7} = 2.6457$

Euclidean distance between S1 and S3:

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

$= \sqrt{1 + 1 + 1 + 1 + 1 + 1 + 4 + 1 + 1 + 1 + 1 + 1 + 1}$

$= \sqrt{16}$

$= 4$

Euclidean distance between S2 and S3:

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

$$= \sqrt{1+1+1+1+1+1+1+1+1+1+1}$$

$$= \sqrt{11} = 3.31662$$

2. Term Frequency (tf):

$$Term\ frequency = \frac{Number\ of\ times\ the\ term\ appears\ in\ a\ sentence}{Total\ number\ of\ terms\ in\ the\ sentence}$$

Sentence 1:

"data science is one of the most important courses in computer science"

Tf value of ***'data'*** $= \frac{1}{12} = 0.083333$
Tf value of '***science'*** $= \frac{2}{12} = \frac{1}{6} = 0.16666$
Tf value of ***'is'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'one'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'of'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'the'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'most'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'important'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'courses'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'in'*** $= \frac{1}{12} = 0.083333$
Tf value of ***'computer'*** $= \frac{1}{12} = 0.083333$

Sentence 2:

"this is one of the best data science courses"

Tf value of ***'this'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'is'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'one'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'of'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'the'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'best'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'data'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'science'*** $= \frac{1}{9} = 0.11111111$

Tf value of ***'courses'*** $= \frac{1}{9} = 0.11111111$

Sentence 3:

"the data scientists perform data analysis"

Tf value of ***'the'*** $= \frac{1}{6} = 0.1666666$

Tf value of ***'data'*** $= \frac{2}{6} = \frac{1}{3} = 0.333333333$

Tf value of ***'scientists'*** $= \frac{1}{6} = 0.1666666$

Tf value of ***'perform'*** $= \frac{1}{6} = 0.1666666$

Tf value of ***'analysis'*** $= \frac{1}{6} = 0.1666666$

Calculating Vectors:

	analysis	best	computer	courses	data	important	perform	science	scientists	the	this	one	is	most	in	of
S1	0	0	0.083333	0.083333	0.083333	0.083333	0	0.16666	0	0.083333	0	0.083333	0.083333	0.083333	0.083333	0.083333
S2	0	0.111111	0	0.111111	0.111111	0	0	0.111111	0	0.111111	0.111111	0.111111	0.111111	0	0	0.111111
S3	0.166666	0	0	0	0.3333333	0	0.166666	0	0.166666	0.166666	0	0	0	0	0	0

S1 = [0, 0, 0.083333, 0.083333, 0.083333, 0.083333, 0, 0.16666, 0, 0.083333, 0, 0.083333, 0.083333, 0.083333, 0.083333, 0.083333]

S2 = [0, 0.11111, 0, 0.11111, 0.11111, 0, 0, 0.11111, 0, 0.11111, 0.11111, 0.11111, 0.11111, 0, 0, 0.11111]

S3 = [0.166666, 0, 0, 0, 0.33333, 0, 0.16666, 0, 0.166666, 0.166666, 0, 0, 0, 0, 0, 0]

Cosine Similarities:

Cosine Similarity between S1 and S2:

$$\text{Cosine similarity} = \frac{S1.S2}{|S1|.|S2|}$$

$$= \frac{(0\times 0)+(0\times 0.1111)+(0.083333\times 0)+(0.083333\times 0.1111)+(0.083333\times 0.1111)+(0.083333\times 0)+(0\times 0)+(0.16666\times 0.1111)+(0\times 0)+(0.083333\times 0.1111)+(0\times 0.1111)+(0.083333\times 0.1111)+(0.083333\times 0.1111)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0.1111)}{0.31180\times 0.33333}$$

$$= 0.833333$$

Cosine Similarity between S1 and S3:

$$\text{Cosine similarity} = \frac{S1.S3}{|S1|.|S3|}$$

$$= \frac{(0\times 0.166666)+(0\times 0)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0.33333)+(0.083333\times 0)+(0\times 0.166666)+(0.16666\times 0)+(0\times 0.166666)+(0.083333\times 0.166666)+(0\times 0)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0)}{0.31180\times 0.471399}$$

$$= 0.36514$$

Cosine Similarity between S3 and S2:

$$\text{Cosine similarity} = \frac{S2.S3}{|S2|.|S3|}$$

$$= \frac{(0.166666\times 0)+(0\times 0.1111)+(0\times 0)+(0\times 0.1111)+(0.33333\times 0.1111)+(0.083333\times 0)+(0\times 0.166666)+(0.16666\times 0)+(0\times 0.166666)+(0.083333\times 0.166666)+(0\times 0)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0)+(0.083333\times 0)}{0.471399\times 0.33333}$$

$$= 0.182574$$

Manhattan Distance

Manhattan distance between S1 and S2:

$$= |0-0|+|0-0.1111|+|0.083333-0|+|0.083333-0.1111|+|0.083333-0.1111|+|0.083333-0|+|0-0|+|0.16666-0.1111|$$
$$+|0-0|+|0.083333-0.1111|+|0-0.1111|+|0.083333-0.1111|+|0.083333-0.1111|+|0.083333-0|$$
$$+|0.083333-0|+|0.083333-0.1111|$$

$$= 0.333333$$

Manhattan distance between S1 and S3:

$$= |0-0.166666|+|0-0|+|0.083333-0|+|0.083333-0|+|0.083333-0.33333|+|0.083333-0|+|0-0.166666|+|0.16666-0|$$
$$+|0-0.166666|+|0.083333-0.166666|+|0-0|+|0.083333-0|+|0.083333-0|+|0.083333-0|$$
$$+|0.083333-0|+|0.083333-0|$$

$$= 1.333333$$

Manhattan distance between S2 and S3:

$$= |0.166666 - 0| + |0 - 0.1111| + |0 - 0| + |0 - 0.1111| + |0.33333 - 0.1111| + |0.083333 - 0| + |0 - 0.166666| + |0.16666 - 0| + |0 - 0.166666| + |0.083333 - 0.166666| + |0 - 0| + |0.083333 - 0| + |0.083333 - 0| + |0.083333 - 0| + |0.083333 - 0| + |0.083333 - 0|$$

$$= 1.66666667$$

Euclidean Distance

Euclidean distance between S1 and S2:

$$= \sqrt{(0-0)^2 + (0-0.1111)^2 + (0.083333-0)^2 + (0.083333-0.1111)^2 + (0.083333-0.1111)^2 + (0.083333-0)^2 + (0-0)^2 + (0.16666-0.1111)^2 + (0-0)^2 + (0.083333-0.1111)^2 + (0-0.1111)^2 + (0.083333-0.1111)^2 + (0.083333-0.1111)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0.1111)^2}$$

$$= 0.23570$$

Euclidean distance between S1 and S3:

$$= \sqrt{(0-0.166666)^2 + (0-0)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0.33333)^2 + (0.083333-0)^2 + (0-0.166666)^2 + (0.16666-0)^2 + (0-0.166666)^2 + (0.083333-0.166666)^2 + (0-0)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0)^2}$$

$$= 0.48304589$$

Euclidean distance between S2 and S3:

$$= \sqrt{(0.166666-0)^2 + (0-0.1111)^2 + (0-0)^2 + (0-0.1111)^2 + (0.33333-0.1111)^2 + (0.083333-0)^2 + (0-0.166666)^2 + (0.16666-0)^2 + (0-0.166666)^2 + (0.083333-0.166666)^2 + (0-0)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0)^2 + (0.083333-0)^2}$$

$$= 0.547722$$

3. IDF

$$IDF = \log\left(\frac{Total\ Number\ of\ documents}{Number\ of\ documents\ containing\ term}\right)$$

Analysis	$= \log\left(\frac{3}{1}\right) = 1.0986122886681098$
Best	$= \log\left(\frac{3}{1}\right) = 1.0986122886681098$
Computer	$= \log\left(\frac{3}{1}\right) = 1.0986122886681098$
Courses	$= \log\left(\frac{3}{2}\right) = 0.4054651081081644$
data	$= \log\left(\frac{3}{3}\right) = 0.0$
important	$= \log\left(\frac{3}{1}\right) = 1.0986122886681098$
in	$= \log\left(\frac{3}{1}\right) = 1.0986122886681098$
is	$= \log\left(\frac{3}{2}\right) = 0.4054651081081644$
most	$= \log\left(\frac{3}{1}\right) = 1.0986122886681098$
of	$= \log\left(\frac{3}{2}\right) = 0.4054651081081644$
one	$= \log\left(\frac{3}{2}\right) = 0.4054651081081644$
perform	$= \log\left(\frac{3}{1}\right) = 1.0986122886681098$

science $= \log\left(\frac{3}{2}\right) = 0.4054651081081644$

scientists $= \log\left(\frac{3}{1}\right) = 1.0986122886681098$

the $= \log\left(\frac{3}{3}\right) = 0.0$

this $= \log\left(\frac{3}{1}\right) = 1.0986122886681098$

4. Tf.IDF

For Sentence 1:

{'data': 0.0, 'science': 0.06757751801802739, 'is': 0.06757751801802739, 'one': 0.06757751801802739, 'of': 0.06757751801802739, 'the': 0.0, 'most': 0.0, 'important': 0.0, 'courses': 0.0, 'in': 0.0, 'computer': 0.0}

For Sentence 2:

{'this': 0.1831020481113516, 'is': 0.06757751801802739, 'one': 0.06757751801802739, 'of': 0.06757751801802739, 'the': 0.0, 'best': 0.0, 'data': 0.0, 'science': 0.06757751801802739, 'courses': 0.0}

For Sentence 3:

{'the': 0.0, 'data': 0.0, 'scientists': 0.21972245773362198, 'perform': 0.21972245773362198, 'analysis': 0.21972245773362

Calculating Vectors:

	analysis	best	computer	courses	data	important	perform	science	scientists	the	this	one	is	most	in	of
S1	0	0	0	0	0	0	0	0.067577	0	0	0	0.067577	0.067577	0	0	0.067577
S2	0	0	0	0	0	0	0	0.067577	0	0	0.183102	0	0.067577	0	0	0.067577
S3	0.219722	0	0	0	0	0	0.219722	0	0.219722	0	0	0	0	0	0	0

S1 = [0, 0, 0, 0, 0, 0, 0, 0.067577, 0, 0, 0, 0.067577, 0.067577, 0, 0, 0.067577]

S2 = [0, 0, 0, 0, 0, 0, 0, 0.067577, 0, 0, 0.183102, 0, 0.067577, 0, 0, 0.067577]

S3 = [0.219722, 0, 0, 0, 0, 0, 0.219722, 0, 0.219722, 0, 0, 0, 0, 0, 0, 0]

Cosine Similarities:

Cosine Similarity between S1 and S2:

Cosine similarity = $\frac{S1.S2}{|S1|.|S2|}$

$$= \frac{(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0.0667577\times0.0667577)+(0\times0)+(0\times0)+(0\times0.183102)+(0.0667577\times0)+(0.0667577\times0.0667577)+(0\times0)+(0\times0)+(0.0667577\times0.0667577)}{0.135154\times0.2173}$$

$= 0.57732$

Cosine Similarity between S1 and S3:

Cosine similarity = $\frac{S1.S3}{|S1|.|S3|}$

$$= \frac{(0\times0.219722)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0.219722)+(0.0667577\times0)+(0\times0.219722)+(0\times0)+(0\times0)+(0.0667577\times0)+(0.0667577\times0)+(0\times0)+(0\times0)+(0.0667577\times0)}{0.135154\times0.3805}$$

$= 0.15733$

Cosine Similarity between S3 and S2:

Cosine similarity = $\frac{S2.S3}{|S2|.|S3|}$

$$= \frac{(0\times0.219722)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0.219722)+(0.0667577\times0)+(0\times0.219722)+(0\times0)+(0.183102\times0)+(0\times0)+(0.0667577\times0)+(0\times0)+(0\times0)+(0.0667577\times0)}{0.2173\times0.3805}$$

$= 0.20321$

Manhattan Distance

Manhattan distance between S1 and S2:

S1 = [0, 0, 0, 0, 0, 0, 0, 0.067577, 0, 0, 0, 0.067577, 0.067577, 0, 0, 0.067577]

S2 = [0, 0, 0, 0, 0, 0, 0, 0.067577, 0, 0, 0, 0.183102, 0, 0.067577, 0, 0, 0.067577]

S3 = [0.219722, 0, 0, 0, 0, 0, 0.219722, 0, 0.219722, 0, 0, 0, 0, 0, 0, 0]

$$= |0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0.0667577-0.0667577|+|0-0|+|0-0|+|0-0.183102|$$
$$+|0.0667577-0|+|0.0667577-0.0667577|+|0-0|+|0-0|+|0.0667577-0.0667577|$$

$= 2.735564$

Manhattan distance between S1 and S3:

$$= |0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0.0667577-0.0667577|+|0-0|+|0-0|+|0-0.183102|$$
$$+|0.0667577-0|+|0.0667577-0.0667577|+|0-0|+|0-0|+|0.0667577-0.0667577|$$

$= 4.60822$

Manhattan distance between S2 and S3:

$$= |0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0-0|+|0.0667577-0.0667577|+|0-0|+|0-0|+|0-0.183102|$$
$$+|0.0667577-0|+|0.0667577-0.0667577|+|0-0|+|0-0|+|0.0667577-0.0667577|$$

$= 4.1455$

Euclidean Distance

Euclidean distance between S1 and S2:

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0.0667577 - 0.0667577)^2 + (0-0)^2 + (0-0)^2 + (0-0.183102)^2 + (0.0667577-0)^2 + (0.0667577 - 0.0667577)^2 + (0-0)^2 + (0-0)^2 + (0.0667577 - 0.0667577)^2}$$

$= 091943005$

Euclidean distance between S1 and S3:

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0.0667577 - 0.0667577)^2 + (0-0)^2 + (0-0)^2 + (0-0.183102)^2 + (0.0667577-0)^2 + (0.0667577 - 0.0667577)^2 + (0-0)^2 + (0-0)^2 + (0.0667577 - 0.0667577)^2}$$

$= 1.2982$

Euclidean distance between S2 and S3:

$$= \sqrt{(0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2 + (0.0667577 - 0.0667577)^2 + (0-0)^2 + (0-0)^2 + (0-0.183102)^2 + (0.0667577-0)^2 + (0.0667577 - 0.0667577)^2 + (0-0)^2 + (0-0)^2 + (0.0667577 - 0.0667577)^2}$$

$= 1.26236$