Assignment 4
FA21-BSE-077

Date: 11-12-2023
Maryam Yousaf

Question 1

S_1: "data science is one of the most important courses in computer science"
S_2: "This is one of the best data science courses"
S_3: "The data scientists perform data analysis"

Compute Bag of words vocabulary: 16 unique word

	data	science	is	one	of	the	most	importan
S_1	1	2	1	1	1	1	1	
S_2	1	1	1	1	1	0	0	
S_3	2	0	0	0	1	0	0	

	courses	in	computer	This	best	scientists
S_1	1	1	1	0	0	0
S_2	1	0	0	1	1	0
S_3	0	0	0	0	0	1

	perform	analysis
S_1	0	0
S_2	0	0
S_3	1	1

vector S_1 [121111111110000 0]
vector S_2 [111111001001100 0]
vector S_3 [200001000000000 111]

Term frequency:

for S_1 length: 12

data = $\frac{1}{12}$ = 0.0833
science = $\frac{2}{12}$ = 0.1667
is = $\frac{1}{12}$ = 0.0833
one = $\frac{1}{12}$ = 0.0833
of = $\frac{1}{12}$ = 0.0833
the = $\frac{1}{12}$ = 0.0833
most = $\frac{1}{12}$ = 0.0833
important = $\frac{1}{12}$ = 0.0833
courses = $\frac{1}{12}$ = 0.0833
in = $\frac{1}{12}$ = 0.0833
computer = $\frac{1}{12}$ = 0.0833

this = 0
best = 0
scientists = 0
perform = 0
analysis = 0

for S_2 length: 9

This: $\frac{1}{9}$ = 0.1111
is: $\frac{1}{9}$ = 0.1111
one: $\frac{1}{9}$ = 0.1111
of: $\frac{1}{9}$ = 0.111
the: $\frac{1}{9}$ = 0.11
best: $\frac{1}{9}$ = 0.11
data: $\frac{1}{9}$ = 0.11
science: $\frac{1}{9}$ = 0.11

courses: $\frac{1}{9}$ = 0.11
scientists = 0
perform = 0
analysis = 0
important = 0
computer = 0
most = 0
in = 0

idf

Total documents = 3

∴ one of the

S3 length = 6

The = 1/6 = 0.167

data = 2/6 = 0.333

Scientists = 1/6 = 0.167

perform = 1/6 = 0.167

analysis = 1/6 = 0.167

	data	Science	is	one	of	the	most	imp	courses	in	computer	this
S_1	0.833	0.1667	0.083	0.083	0.08	0.08	0.083	0.083	0.083	0.083	0.083	0
S_2	0.11	0.111	0.11	0.11	0.11	0.11	0	0	0.11	0	0	0.11
S_3	0.33	0	0	0	0	0.167	0	0	0	0	0	0

	best	scientists	perform	analysis
S_1	0	0	0	0
S_2	0.11	0	0	0
S_3	0	0.1667	0.1667	0.1667

idf

Total documents = 3

	data	science	is	one	of	the
S_1	$\log(\frac{3}{3})$ 0	$\log(\frac{3}{3})$ 0	$\log(\frac{3}{2})$ = 0.176	0.176	0.176	$\log(\frac{3}{3})$ 0
S_2	0	0	0.176	0.176	0.176	0
S_3	0	0	0.176	0.176	0.176	0

	most	important	courses	in	computer
S_1	$\log(\frac{3}{1})$ 0.477	0.477	0.176	0.477	0.477
S_2	0.477	0.477	0.176	0.477	0.477
S_3	0.477	0.477	0.176	0.477	0.477

	this	best	scientists	perform	analysis
S_1	0.477	0.477	0.477	0.477	0.477
S_2	0.477	0.477	0.477	0.477	0.477
S_3	0.477	0.477	0.477	0.477	0.477

Tf.idf for S_1

data : 0.0833 X 0 = 0

science : 0.1667 X 0 = 0

is : 0.0833 X 0.176 = 0.0147

one : 0.0833 X 0.176 = 0.0147

of : 0.0833 X 0.176 = 0.0147

the : 0.0833 X 0 = 0

most : 0.0833 X 0.477 = 0.0397

important : 0.0833 X 0.477 = 0.0397

courses : 0.0833 X 0.176 = 0

in : 0.0833 X 0.477 = 0

computer : 0.0833 X 0.477 = 0

this : 0

best : 0

scientists : 0

perform : 0

analysis : 0

Tfidf for S_2

This: 0.11 * 0.477 = 0.05247
is: 0.11 * 0.176 = 0.01936
one: 0.11 * 0.176 = 0.01936
of: 0.11 * 0.176 = 0.01936
the: 0.11 * 0 = 0
best: 0.11 * 0.477 = 0.05247
data: 0.11 * 0 = 0
Science: 0.11 * 0 = 0
analysis: 0
course: 0.11 * 0.176 = 0.01936
most: 0
important: 0
in: 0
computer: 0
scientists: 0
perform. 0

Tfidf for S_3

The : 0.167 * 0 = 0
data : 0
Scientists : 0.167 * 0.477 = 0.079659
perform : 0.167 * 0.477 = 0.079659
data : 0
analysis. 0.167 * 0.477 = 0.079659

Tfidf Table

	Data	Science	is	one	of	the	most	important	courses	in	computer
S_1	0	0	0147	0.0147	0.0147	0	0.0397	0.0397	0.0147	0.0391	0.0397
S_2	0	0	0.019	0.019	0.019	0	0	0	0.019	0	0
S_3	0	0	0	0		0	0	0	0	0	0

	this	best	scientist	perform	analysis
S_1	0	0	0	0	0
S_2	0.05247	0.05247	0	0	0
S_3	0	0	0.0797	0.0797	0.0797

Question 2

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

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

Compute Similarity b/w S_1, S_2, S_3 using cosine

$$\cos(S_1, S_2) = \frac{S_1 \cdot S_2}{\sqrt{S_1^2}\ \sqrt{S_2^2}}$$

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

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

$= 8$

$|S_1| = (1*1 + 2*2 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 0*0 + 0*0 + 0*0 + 0*0 + 0*0)\ ^\wedge\ 0.5$

$= (1 + 4 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1)\ ^\wedge\ 0.5$

$= (14)\ ^\wedge\ 0.5 = 7$

$|S_2| = (1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 1*1 + 0*0 + 0*0 + 1*1 + 0*0 + 0*0 + 1*1 + 1*1 + 0*0 + 0*0 + 0*0)\ ^\wedge\ 0.5$

$= (1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1)\ ^\wedge\ 0.5$

$= (9)\ ^\wedge\ 0.5 = 4.5$

$$\cos(S_1, S_2) = \frac{8}{4.5 \times 7} = \frac{8}{31.5} = 0.25 \text{ Ans}$$

S_1 [1 2 1 1 1 1 1 1 1 0 0 0 0 0]

S_3 [2 0 0 0 1 0 0 0 0 0 0 0 1 1 1]

$$\cos(S_1, S_3) = \frac{S_1 \cdot S_3}{\sqrt{S_1^2}\sqrt{S_3^2}}$$

$S_1 \cdot S_3 = 2 + 1 + 0$

$= 3$

$\sqrt{S_1^2} = 1 + 4 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1 + 1$

$= 14 \wedge 0.5$

$= 7$

$\sqrt{S_3^2} = 4 + 1 + 1 + 1 + 1$

$= 8 \wedge 0.5$

$= 4$

$$\cos(S_1, S_3) = \frac{3}{7 \times 4} = \frac{3}{28} = 0.107 \text{ Ans}$$

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

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

$$\cos(S_2, S_3) = \frac{S_2 \cdot S_3}{\sqrt{S_2^2}\ \sqrt{S_3^2}}$$

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

$\sqrt{S_2^2} = (1+1+1+1+1+1+1+1+1)\ \hat{}\ 0.5$

$= 9\ \hat{}\ 0.5$

$= 4.5$

$\sqrt{S_3^2} = (4+1+1+1+1)\ \hat{}\ 0.5$

$= 8\ \hat{}\ 0.5$

$= 4$

$$\cos(S_2, S_3) = \frac{3}{4 \times 4.5} = \frac{3}{18} = 0.167$$

Using Manhattan

Manhattan formula $|P_3 - P_2| = \{|P_3 - P_2|$

$S_1, S_3 = |1-2| + |2-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |1-0| + |0-0| + |0-0| + |0-1| + |0-1| + |0-1|$

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

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

$= 14$

$S_1, S_2 = |1-1| + |2-1| + |1-1| + |1-1| + |1-1| + |1-1| + |1-0| + |1-0| + |1-1| + |1-0| +$
$|1-0| + |0-1| + |0-1| + |0-0| + |0-0| + |0-0|$

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

$= 7$

$S_2, S_3 = |1-2| + |1-0| + |1-0| + |1-0| + |1-0| + |1-1| + |0-0| + |0-0| + |1-0|$
$+ |0-0| + |0-0| + |1-0| + |1-0| + |0-1| + |0-1| + |0-1|$

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

$= 11$

Using Euclidean Distances

formula $P_3 - P_2 = \sqrt{\sum (P_3 \cdot P_2)^2}$

$S_1, S_2 = [(1-1)^2 + (2-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2 +$
$(1-0)^2 + (1-0)^2 + (0-1)^2 + (0-1)^2 + (0-0)^2 + (0-0)^2 + (0-0)^2]^{\wedge} 0.5$

$= [(1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (-1)^2 + (-1)^2]^{\wedge} 0.5$

$= 7^{\wedge} 0.5$

$= 3.5$

$S_2, S_3 = [(1-2)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2 + (0-0)^2$
$+ (0-0)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 + (1-0)^2 + (1-0)^2 + (0-1)^2$
$+ (0-1)^2 + (0-1)^2]^{\wedge} 0.5$

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

$= 11^{\wedge} 0.5$

$= 5.5$

$S_1, S_3 = (1-2)^2 + (2-0)^2 + (1-0)^2 + (1-0)^2 + (1-0)^2 + (1-1)^2 + (1-0)^2 + (1-0$
$(1-0)^2 + (1-0)^2 + (1-0)^2 + (0-0)^2 + (0-0)^2 + (0-1)^2 + (0-1)^2 + (0-1)^2$

$= (-1)^2 + (2)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (-1)^2 + (-1)^2 + (-$
$= 12 + 4$
$= 16^{\wedge} 0.5$
$= 8$