COMSATS University Islamabad, Lahore Campus

Course Title:	Introduction to Data Science			Course Code:	CSC461	Credit Hours:	3(3,0)
Resource Person:	Dr. Muhammad Sharjeel			Programme Name:	BSSE		
Semester:	5th	Batch:	FA21	Section:	C	Max Marks:	10

HAMNA ASHRAF
SP20-BSE-047

Assignment 4

Due Date: 11-12-2023

Submission: Upload the assignment solution (PDF file and Python code, preferably iPython notebook) to your GitHub account (private repository).

Important instructions: Please write the following information at the start of your ipython file.
# Date
# CSC461 – Assignment4 – NLP
# Your Full Name
# You Complete Registration Number #
A brief description of the task

Important Instruction:
Solve the following questions manually as well as implement the solution using Python. Submit both.

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”

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

Hamna Ashraf
SP20-BSE-047
Section-A

Assignment: 4

Q:1 Compute BoW, TF, IDF & TF-IDF.

Ans:

Vocabulary (Unique Terms):

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

Bag of words (BoW):

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

Bag of words.

Term	S1	S2	S3
Scientists	0	0	1
perform	0	0	1
Analysis	0	0	1

Vector S1 : [1.2 1 1 1 1 1 1 1 1 1 0 0 0 0 0]

Vector S2 : [1 1 1 1 1 1 0 0 1 0 0 1 1 0 0 0]

Vector S3 : [2 0 0 0 0 1 0 0 0 0 0 0 0 1 1 1]

TF (Term Frequency):

Term	S1	S2	S3
tf(data)	1/12	1/9	2/6
tf(science)	2/12	1/9	0
tf(is)	1/12	1/9	0
tf(one)	1/12	1/9	0
tf(of)	1/12	1/9	0
tf(the)	1/12	1/9	1/6
tf(most)	1/12	0	0
tf(important)	1/12	0	0
tf(courses)	1/12	1/9	0
tf(in)	1/12	0	0
tf(computer)	1/12	0	0
tf(this)	0	1/9	0
tf(best)	0	1/9	0
tf(scientists)	0	0	1/6
tf(perform)	0	0	1/6
tf(analysis)	0	0	1/6

Inverse Document Frequency (IDF)

idf (data) = $\log(3/3) = 0$

idf(science) = $\log(3/2) = 0.18$

idf(is) = $\log(3/2) = 0.18$

idf(one) = $\log(3/2) = 0.18$

idf(of) = $\log(3/2) = 0.18$

idf ~~of~~ (the) = $\log(3/3) = 0$

idf(most) = $\log(3/1) = 0.48$

idf(important) = $\log(3/1) = 0.48$

idf(courses) = $\log(3/2) = 0.18$

idf(in) = $\log(3/1) = 0.48$

idf(computer) = $\log(3/1) = 0.48$

idf(this) = $\log(3/1) =$ 0.48

idf(best) = $\log(3/1) =$ 0.48

idf(scientists) = $\log(3/1) = 0.48$

idf(perfom) = $\log(3/1) =$ 0.48

idf(analysis) = $\log(3/1) =$ 0.48

TF-IDF :-

Term	tf × idf (S1)	tf × id.f (S2)	tf×idf(S3)
data	$\frac{1}{12} \times 0 = 0.$	$\frac{1}{9} \times 0 = 0$	$\frac{2}{6} \times 0 = 0$
science	$\frac{2}{12} \times 0.18 = 0.03$	$\frac{1}{9} \times 0.18 = 0.02$	0
is	$\frac{1}{12} \times 0.18 = 0.015$	$\frac{1}{9} \times 0.18 = 0.02$	0
one	$\frac{1}{12} \times 0.18 = 0.015$	$\frac{1}{9} \times 0.18 = 0.02$	0
of	$\frac{1}{12} \times 0.18 = 0.015$	$\frac{1}{9} \times 0.18 = 0.02$	0
the	$\frac{1}{12} \times 0 = 0$	$\frac{1}{9} \times 0 = 0$	$\frac{1}{6} \times 0 = 0$
most	$\frac{1}{12} \times 0.48 = 0.04$	0	0
important	$\frac{1}{12} \times 0.48 = 0.04$	0	0
courses	$\frac{1}{12} \times 0.18 = 0.04$	$\frac{1}{9} \times 0.18 = 0.02$	0
in	$\frac{1}{12} \times 0.48 = 0.04$	0	0
computer	$\frac{1}{12} \times 0.48 = 0.04$	0	0
this	0 *	$\frac{1}{9} \times 0.48 = 0.053$	0
best	0 *	$\frac{1}{9} \times 0.48 = 0.053$	0
scientists	0 *	0	$\frac{1}{6} \times 0.48 = 0.08$
perform	0 *	0	$\frac{1}{6} \times 0.48 = 0.08$
analysis	0 *	0	$\frac{1}{6} \times 0.48 = 0.08$

Hamna Ashraf
SP20-BSE-047

Q:2

Cosine :

Bag of words :-

$$\cos\theta = \frac{\bar{S}_1 \cdot \bar{S}_2 \cdot \bar{S}_3}{|S_1| |S_2| |S_3|}$$

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

$$S_1 \cdot S_2 = 1\cdot1 + 2\cdot1 + 1\cdot1 + 1\cdot1 + 1\cdot1 + 1\cdot1 + 1\cdot0 + 1\cdot0$$

$$+ 1\cdot1 + 1\cdot0 + 1\cdot0 + 1\cdot1 + 1\cdot0 + 1\cdot0 + 0\cdot1 + 0\cdot1 +$$

$$0\cdot0 + 0\cdot0 + 0\cdot0$$

$$S_1 \cdot S_2 = 9$$

$$|S_1| = (1+4+1+1+1+1+1+1+1+1+1+0+0+0+0+0)^{\wedge 0.5}$$

$$= 14^{\wedge 0.5} = 3.7417$$

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

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

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

$$\cos(S_1, S_2) = 0.80.17$$

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

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

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

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

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

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

$S_1 . S_3 = 1+2+1+1+1+1+1+0 = 8$

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

$|S_1| = 3.7417$

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

$\cos(S_1, S_3) = 0.3\,27$