Name _____ Roll No _____ Section _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	Information Retrieval	Course Code:	CS4051
Program:	BS(Data Science)	Semester:	Spring 2023
Duration:	25 Minutes	Total Marks:	10
Paper Date:	22 Feb 2023	Weight	0
Section:	BDS-6A	Page(s):	2
Exam:	Quiz 2	Roll No:	

Question 1

a) What is size of vocabulary (number of unique words) if total words in a collection are 100,000 Use Heap's law with k = 10 and β = 0.5 [2 Marks]

$V = k(N)^{\beta}$
$V = 10(100{,}000)^{0.5}$

$V = 3160$

(b) What is size of vocabulary if 500 new documents are added to the above collection? Assume each document has 1000 words on average. [2 Marks]

$N = 100{,}000 + 500*1000 = 600{,}000$

$V = k(N)^{\beta}$
$V = 10(600{,}000)^{0.5}$

$V = 7740$

Name _____ Roll No _____ Section _____

Question 2

Represent following 4 documents as vectors. Use TF.IDF weights. [3 Marks]

Document 1: Omar drove the car.
Document 2: Nadia and Omar ate oranges and snacks.
Document 3: Nadia ate oranges not snacks.
Document 4: Nadia and Omar ate snacks.

	Omar	drove	the	car	Nadia	and	ate	oranges	snacks	not
<D1>	1	1	1	1	0	0	0	0	0	0
<D2>	1	0	0	0	1	1	1	1	1	0
<D3>	0	0	0	0	1	0	0	1	1	1
<D4>	1	0	0	0	1	1	1	0	1	0

(a) Calculate cosine similarity between vectors of document 2 and document 3. [3 Marks]

Cosine sim (D2,D3) = D2.D3 / (|D2|*|D3|)

D2.D3 = 1*0 + 0*0 + 0*0 + 0*0 + 1*1 + 1*0 + 1*0 + 1*1 + 1*1 + 0*1 = 1+1+1 = 3

$|D2| = \text{sqrt}((1)^2 + (0)^2 + (0)^2 + (0)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (1)^2 + (0)^2) = \text{sqrt}(6) = 2.45$

$|D3| = \text{sqrt}((0)^2 + (0)^2 + (0)^2 + (0)^2 + (1)^2 + (0)^2 + (0)^2 + (1)^2 + (1)^2 + (1)^2) = \text{sqrt}(4) = 2$

Cosine sim (D2,D3) = D2.D3 / (|D2|*|D3|) = 3 / (2.45*2) = 0.61