Carnegie Mellon University
Department of Computer Science
15-826 Multimedia Databases and Data Mining
C. Faloutsos, Fall 2025

Homework 3

Due: **pdf, on canvas, at 2:00pm, on 10/31/2025**

VERY IMPORTANT:

- Upload **e-copy** of your answers, on canvas.

Reminders:

- *Plagiarism*: Homework is to be completed *individually*.
- *Typeset* your answers. Illegible handwriting may get zero points.
- *Late homeworks*: Follow the published policy

For your information:

- Explanations are *optional*, and will only be used to for partial credit, if the main answer is off.
- Graded out of **100** points; **6** questions total
- Rough time estimate: *1-3 hours (≈ 10-30 minutes per question)*

Revision : 2025/10/03 00:42

Question	Points	Score
Text - compression	15	
Text - String Editing	15	
Text - Signature files	15	
Fractals - Correlation integrals	30	
Text and Zipf	10	
Fractals - Estimations	15	
Total:	100	

Question 1: Text - compression . . . . . [15 points]

Consider the self-delimiting ("Elias") codes from class. As you recall, '2' is encoded as '0 10'; '3' is encoded as '0 11'.

(a) Consider the code for '7'

i. **[2 points]** How many bits does it have?

i. _____

ii. **[3 points]** Give the code for '7'

ii. _____

(b) **[4 points]** Give the code for '8'

(b) _____

(c) **[3 points]** Which is the smallest integer, that will require $\geq$10 bits in its code? (give it as a power of 2 plus/minus a small integer; e.g., a possibly correct answer, could be: $2^{10} - 1$)

(c) _____

(d) **[3 points]** How many integers require *exactly* 10 bits in their code?

(d) _____

Question 2: Text - String Editing ..... [15 points]

Consider the strings 'ANNA' and 'ANTENA', with insertion cost = 1, deletion cost = 1, and substitution cost = 0.5.

	ϕ	A	N	N	A
ϕ	0	1	2	3	4
A	1				
N	2				
T	3				
E	4				
N	5				
A	6				

Table 1: String editing distance.

(a) **[12 points]** Fill in the matrix in Table 1, with the best distances for matching the prefix at the top with the prefix at the left. The first row and column are filled in already, with ϕ meaning the empty string.

(b) **[3 points]** What is the editing distance of the two given strings?

(b) _____

Question 3: Text - Signature files ..... [15 points]

Consider a query word, whose signature is the string shown in Table 2 (line marked as 'query'), as well as the documents and their signatures, in the same Table. Assume that the signatures were created with superimposed coding.

(a) **[15 points]** For each document, indicate with an 'X' the result of the signature test

The meanings are as in the lectures, and repeated here:

- 'YES': the document definitely contains the query word
- 'NO': the document definitely does **not** contain the query word
- 'MAYBE': the document might, or might not, contain the query word

YES	MAYBE	NO	id	signature
N/A	N/A	N/A	query	11 10 00 00
			Document1	11 10 11 11
			Document2	10 00 11 11
			Document3	11 10 00 00
			Document4	11 11 11 11
			Document5	00 00 00 00

Table 2: Query and document signatures - mark one of YES/MAYBE/NO

Question 4: Fractals - Correlation integrals ..... [30 points]

Consider a 'mystery' dataset, with N points in a E-dimensional space, with the correlation integral of Figure 1: It starts flat at the left; then it has slope s; and then it finishes flat at the right.

Reminder: The vertical axis is the count of pairs within distance r or less, *including self-pairs and mirror-pairs*. We are told that

- the slope is s=2
- N= 1,000 points
- r_2=1.0 (the rightmost break-point = characteristic distance; diameter of the dataset)

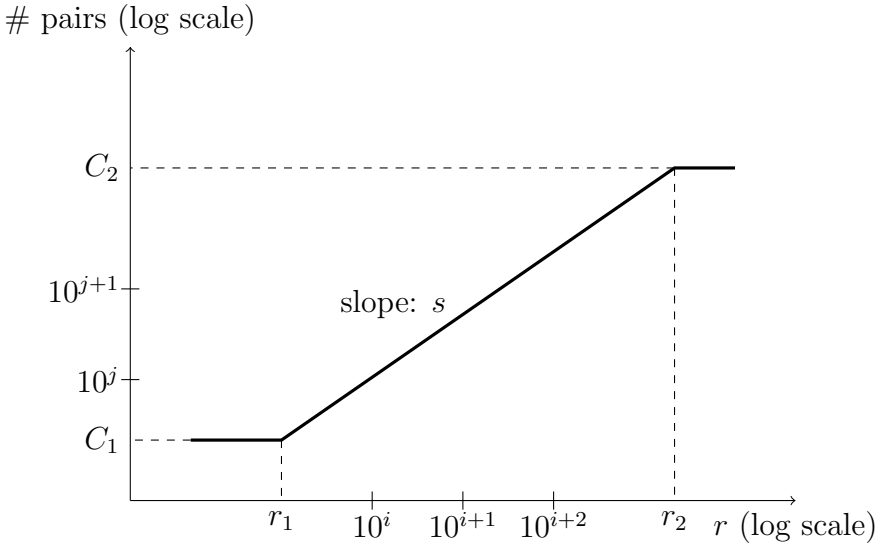

Figure 1: The correlation integral of the 'mystery' dataset.

(a) Mark with 'True' all the settings, that could give a correlation integral like the one in Figure 1:

i. **[2 points]** ☐ T ☐ F Points randomly (ie., uniformly) distributed in the unit square.

ii. **[2 points]** ☐ T ☐ F Random points on the unit-side Sierpinski triangle.

iii. **[2 points]** ☐ T ☐ F Random points on the surface of the unit sphere.

iv. **[2 points]** ☐ T ☐ F Random points inside the unit sphere.

v. **[2 points]** ☐ T ☐ F Random points on a diagonal of the unit cube.

(b) **[5 points]** What is your estimate for C_1?

(b) _____

(c) **[5 points]** What is your estimate for C_2?

(c) _____

(d) **[5 points]** We want the formula to estimate r_1 (the smallest distance between two points), as a function of N, r_2 and the slope s. Mark the correct one, or write down your own formula.

A. $r_1 = r_2 * N^{1/s}$
B. $r_1 = r_2/N^s$
C. $r_1 = r_2 * N^{-1/s}$
D. $r_1 = r_2/(s * \log(N))$
E. $r_1 = r_2^s/N$
F. none of the above - the correct formula is below

. . . . . .
. . . . . .
. . . . . .
. . . . . .

(e) **[5 points]** What is your estimate for r_1, the smallest distance between two points? (We did the arithmetic computation for your convenience.)

A. $1/N = 0.001$
B. $1/N^2 = 10^{-6}$
C. $1/N^{\sqrt{2}} = 5.71 * 10^{-5}$
D. $1/\log(N) = 0.144$
E. $1/(2 * \log(N)) = 0.0723$
F. $1/\sqrt{\log(N)}) = 0.380$
G. $1/\sqrt{N} = 0.0316$
H. $1/\sqrt[3]{N} = 0.1$
I. none of the above - the correct one is below, up to 3 significant digits.

. . . . . .
. . . . . .
. . . . . .
. . . . . .

Question 5: Text and Zipf ..... [10 points]

Consider a piece of text, which obeys the Zipf distribution, that is $frequency = C/rank$ or, more compactly $f_r = C/r$. Suppose that the vocabulary size is V=1,000.

(a) **[2 points]** What is your estimate for the constant C? *Hint:* The least frequent word appears once: $f_V = 1$.

(a) _____

(b) **[3 points]** What is your estimate for the occurrence frequency f_1 of the most common word?

(b) _____

(c) **[5 points]** What is your estimate for the total number of words in the document? Use the approximation $\sum_{i=1}^{n} 1/i \approx \ln(n)$ where ln() is the natural logarithm; give your answer numerically, with accuracy up to 3 significant digits,

.....

.....

.....

.....

Question 6: Fractals - Estimations . . . . . [15 points]

We are given a cloud of points in E=10 dimensions, consisting of $N = 10^6$ points.

We are told that the dataset is **self-similar**. The number of pairs $P(r)$ within distance $\leq r$ is shown in Table 3. (Notice that we include self-pairs and mirror-pairs).

We want to estimate the result-sizes for spatial joins of different radii, and other statistics from the dataset.

r	$P(r)$
1	$10,000,000$
2	$40,000,000$
4	$160,000,000$

Table 3: Number of pairs $P(r)$ within a given distance r.

(a) **[1 point]** Assuming self-similarity, guess the result size of the spatial join with radius r=8. Give a *numerical* answer, up to the third significant digit.

(a) _____

(b) **[2 points]** What is your estimate for D_2, the correlation fractal dimension? Give the *numerical* answer, up to the third significant digit.

(b) _____

(c) **[2 points]** Let D_2 stand for the correlation fractal dimension of our dataset, and let $P(1) = 10,000,000$, as in Table 3. Mark the correct formula for the count of pairs $P(r)$, within distance r or less - or write-in your own:

A. $P(r) = P(1) * r$

B. $P(r) = P(1) * r^{-D_2}$

C. $P(r) = P(1) * r^{2D_2}$

D. $P(r) = P(1) * r^{1/D_2}$

E. $P(r) = P(1) * r^{D_2}$

F. None of the above - the correct is: $P(r) =$ . . . . . .

(d) **[5 points]** Assuming self-similarity, we want to guess the *diameter* r_{max} of the dataset, that is, the distance between the two furthest-apart points. Mark the correct formula for r_{max}, or provide your own:

A. $r_{max} = N/P(1)$
B. $r_{max} = (N/P(1))^{D_2}$
C. $r_{max} = (N/P(1))^{-D_2}$
D. $r_{max} = N^2/P(1)$
E. $r_{max} = (N^2/P(1))^{D_2}$
F. $r_{max} = (N^2/P(1))^{-D_2}$
G. $r_{max} = (N^2/P(1))^{1/D_2}$
H. None of the above - the correct is: $r_{max} = \ldots\ldots\ldots$

(e) **[5 points]** For your estimate of the diameter r_{max}, give the **numerical** answer, up to the third significant digit. Eg., if the answer is *1234.5678*, you may give *1230*.

(e) _____