BRAC UNIVERSITY
Department of Computer Science and Engineering
CSE431: Natural Language Processing
Examination: Midterm
No. of Questions: 2, Semester: Fall 2020
Full Marks: 25, No. of Pages: 1
Answer the following questions. Figures in the right margin indicate respective marks.

Student ID:
Name in UPPERCASE:
Email address for viva notification:

Question 1) Show the minimum edit distance between any two parts of your name (preferably first & last) using the alternative Levenshtein algorithm where
Insertion cost is 1
Deletion cost is 1
Substitution cost is 2
Show the two-dimensional table, all intermediate steps, and separately mention the distance.
[10 points]

==========

Imagine a mini corpus containing three sentences is

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>

Calculations for some of the bigram probabilities from this corpus are:
P(I|<s>) = C (...) / C(…) = 2 / 3 = 0.67
P(Sam|<s>) = C (...) / C(…) = 1 / 3 = 0.33
P(am|I) = … = … = 0.67
P(</s>|Sam) = … = … = 0.5
P(Sam|am) = … = … = 0.5
P(do|I) = ... = … = 0.33

Question 2) Similar to the mini corpus above about "**Sam**" and food preferences, please design your own mini corpus consisting of 3 short sentences about the "**GPT3 language model**", including the markers: <s> and </s> [7 points]

Question 3) Calculate any two bigram probabilities from your corpus in the format shown above. [8 points]
That is using $P(W_n \mid W_{n-1}) = C(W_{n-1} W_n) / C(W_{n-1})$ = result
Please show detailed calculations including all intermediate steps including completion of the dotted part above for your corpus.