COT 6417
Algorithms on Strings and Sequences
Fall 2020
Homework Assignment 3

Please complete this assignment individually (and not in a group). Please email me a single pdf document with the solutions (*only pdf is accepted*). This homework is due on ***Thursday, November 19, 2020***.

1. Let $F(n,m)$ denote the total number of alignments between two strings X and Y, of lengths n and m respectively. Provide a recurrence relation to compute $F(n,m)$. Use this recurrence relation to compute $F(3,4)$.

2. We wish to compute the edit distance between the strings X= *acagatta* and Y = *tagctta*. Assume a unit cost model, where the cost of a match is 0, and the cost of a substitution and a gap are both 1. Provide a recurrence relation to calculate the optimal edit distance between two strings (of lengths n and m) under the unit cost model, and use that recurrence to compute the edit distance between X and Y. Show the dynamic programming matrix, and also an optimal cost path in this matrix.

3. Let S be a string of length n generated from a constant size alphabet. We are interested in queries of the form $Q(i,j)$, where query $Q(i,j)$ returns the length of the longest common prefix of $S[i, n]$ and $S[j,n]$. Provide an algorithm that takes O(1) time to answer a query $Q(i,j)$ and that uses no more than O(n log n) bits in total. You are allowed O(n) time to pre-process the string S, and this one time pre-processing cost is not included in the time to answer a query.

4. Recall the definition of a (d_1, d_2, p_1, p_2)-sensitive function. Prove that the family of minhash functions is a $(d_1, d_2, 1-d_1, 1-d_2)$-sensitive family for any d_1 and d_2, where $0 \leq d_1 < d_2 \leq 1$.