**Natural Language Processing**

Homework and Programming Assignment 1

**Total Points: 100**

**Deadline: Sep 9, 2022**

1. [**Points 10**] Please write regular expressions for the following.
   1. All binary strings. Example binary strings, 1001, 1011, 1111, etc.
   2. The email address contains only letters, and @, \. Symbols (both lower and upper cases). Example:- [alice@gmail.com](mailto:alice@gmail.com), [bob@yahoo.com](mailto:bob@yahoo.com), etc.
   3. Valid integer numbers. Examples: 1, 12843, -89232, +1262, etc.
   4. Valid phone number that contains ten (10) digits. Consider valid phone number formats are given below.
      1. xxx-xxx-xxxx
      2. (xxx) xxx-xxxx

Examples: 453-126-4570

(453) 126-4560

1. [**Points 5**] Determine the number of tokens and vocabulary, and types from the below text. Please list them in your answer too.

**Text:** “The quick brown fox jumps over the lazy dog.”

1. [**points 5**] Write down all the steps of text normalization and give an example for each step.
2. [**points 15**] We know how to compute similarity distance between two given strings using the edit distance algorithm.
   1. [**Points 10**] Please write down the distance matrix for the following strings. Consider space “ “ as a single character.

Strings 1: **Spokesman confirms**

String 2: **Spokeswoman said**

* 1. [**Points 5**] List down all the operations you need to perform. Please show backtracing matrix to validate your answer for the above example strings.

1. [**Points 20**] Please formulate your language model for the following text. Show the details of your LM formulation.

**Text:** “The day was grey and bitter cold, and the dogs would not take the scent. The big black bitch had taken one sniff at the bear tracks, backed off, and skulked back to the pack with her tail between her legs.”

1. Unigram model [Points 10]
2. Bigram model [Points 10]
3. [**Points 15**] You are given a training set of 30 numbers that consists of 21 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity?
4. [**Points 15**] Write a program to implement a minimum edit distance algorithm that takes two strings of any length and returns the total number of distances required to convert one string to another. Your program should be flexible to take flexible costs for the insertion, deletion, and substitute operations.
5. [**Points 15**] Write a program to construct bi-gram language model and provide source code to generate text using your bi-gram language model. Note: you could use any text dataset to construct your bi-gram language model. For simplicity I’m providing a short text example here.

**Example text:** “The day was grey and bitter cold, and the dogs would not take the scent. The big black bitch had taken one sniff at the bear tracks, backed off, and skulked back to the pack with her tail between her legs.”

**Submission Instructions:**

**Important.**

**Python notebook:** If you are using python, I would recommend using jupyter notebook/lab. So that I can see the visible output of your source code. the cell outputs should be visible with a sample test example if applicable. In case of you are using python scripts instead of notebook, then create a demo video presentation of your source execution showing the intermediate states and submit this demo with your source.

**C/C++ user:** Who uses c/c++, should create a demo video presentation of your source execution showing the intermediate states and submit this demo with your source.

Please note that rephrasing or rename variables, functions names does not mean you are not copying/cheating. For this similar situation, you will be graded zero.

**You should not zip your submission. – you may submit two files – one for question answering and another for coding.**

**Both cases:** source should be executable without error. Without visible outputs, you may get zero for those corresponding questions. You can submit your text question answer in a separate PDF/Doc (if needed) or in the same pdf/doc.

**Late submission or Extension:** Late HomeWorks/assignment will not be accepted unless an extension is approved by me in advance. Requests for extensions must be made at least three days before the due date with valid reason. **3 points** will be deducted for each day after the submission deadline from your grade even if you are approved for extension. For details, please see the **Homework and Exam Policies** section of your syllabus for more details.

**Grading Policy/Rule:** Copying/cheating/plagiarism is strictly prohibited as mentioned in our introductory lectures and syllabus. This policy holds for each assignment/homework/exam. In case of copying/cheating/plagiarism etc. you will be graded zero for the assignment as well as ‘F’ for the subject. Note that the first incident of cheating will result in the student getting a final grade of ‘F’ for the course. The second incident, by CCSE rules, will result in a semester suspension from the College.