**Natural Language Processing**

Homework and Programming Assignment 1

**Total Points: 100**

**Deadline: Sep 9, 2023**

1. [**Points 10**] Please write regular expressions for the following.
   1. 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.
   2. 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 10**] Determine the number of tokens and vocabulary, and types from the below text. Please list them in your answer too.

**Text:** “I came in in the middle of this film so I had no idea about any credits or even its title till I looked it up here, where I see that it has received a mixed reception by your commentators. I'm on the positive side regarding this film but one thing really caught my attention as I watched: the beautiful and sensitive score written in a Coplandesque Americana style. My surprise was great when I discovered the score to have been written by none other than John Williams himself.”

1. [**points 10**] Write down all the steps of text normalization and give an example for each step.
2. [**points 30**] We know how to compute similarity distance between two given strings using the edit distance algorithm.
   1. [**Points 20**] Please write down the distance matrix for the following strings. Consider space “ “ as a single character.

Strings 1: **Spokesman confirm**

String 2: **Spokeswoman said**

* 1. [**Points 10**] List down all the operations you need to perform. Please show backtracing matrix to validate your answer for the above example strings.

1. [**Points 25**] Please formulate your char-language model for the following text. Consider each character as a single word to formulate your language model. Show the details of your LM formulation.

**Training Text:** “aaaa bbb aaa bbb ababab acacac cacacad ccca dcdcdccdddccc cbbcbccb acac bdbdbd dbdbdb dadaaddadadddaaa ddd ccc bbb cdcdcdcd ccddcd dcdcdcdc”

Testing text: aabcacddbcbbdaadda

1. Unigram language model [Points 15]
2. Compute perplexity of your model [Point 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?

**Submission Instructions:**

**Important.**

**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.