CS 4742 Natural Language Processing
Homework 1

1. Please write regular expressions for the following.
   Points [20]
   a. All binary strings. Example binary strings, 1001, 1011, 1111, etc.
   b. The email address contains only letters, and @, \. Symbols (both lower and upper cases). Example:- alice@gmail.com, bob@yahoo.com, etc.
   c. Valid integer numbers. Examples: 1, 12843, -89232, +1262, etc.
   d. Valid phone number that contains ten (10) digits. Consider valid phone number formats are given below.
      i. xxx-xxx-xxxx
      ii. (xxx) xxx-xxxx
      Examples: 453-126-4570
      (453) 126-4560

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

   Text: "The quick brown fox jumps over the lazy dog."

3. Write down all the steps of text normalization and give an example for each step. [points 5]

4. We know how to compute similarity distance between two given strings using the edit distance algorithm. [points 25]
   a. Please write down the distance matrix for the following strings. Consider space " " as a single character. [Points 15]

      Strings 1: **Spokesman confirms**
      String 2: **Spokeswoman said**

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

5. Please formulate your language model for the following text. Show the details of your LM formulation. [Points 25]

   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."

   a. Unigram model [Points 10]
   b. Bigram model [Points 15]

6. 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? [Points 20]