Natural Language Processing  
Homework and Programming Assignment 1

Parker Smith

9/6/2022

1. Please write regular expressions for the following:
   1. All binary strings. Example binary strings, 1001, 1011, 1111, etc.  
        
      [01]+
   2. The email address contains only letters, and @, \. Symbols (both lower and upper cases). Example:- alice@gmail.com, bob@yahoo.com, etc.  
        
      \w+@\w+.\w+
   3. Valid integer numbers. Examples: 1, 12843, -89232, +1262, etc.  
        
      [+-]?\d+
   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

(\(\d{3}\) |\d{3}-)\d{3}-\d{4}

1. Determine the number of tokens and vocabulary and types from the below text:  
   “The quick brown fox jumps over the lazy dog.”  
     
   Tokens: the, quick, brown, fox, jumps, over, the, lazy, dog (9 tokens)  
   Types: the, quick, brown, fox, jumps, over, lazy, dog (8 types)  
   Vocabulary: Set of all types: {the, quick, brown, fox, jumps, over, lazy, dog}  
   |Vocabulary| = 8
2. Write down all the steps of text normalization and give an example for each step.

* Case Folding: Remove all capitals from the input. Ex: This is MY sentence. -> this is my sentence.
* Term Equivalence: If two words mean the same but have different spelling/an “s” on the end, swap them all to the same base. BUT check if the word means something different first. Ex: seal, seals, SEAL -> seal, seal, SEAL (Keep SEAL because it is an acronym)
* Lemmatization: Reduce all types of words to their most basic form. Ex: This is my mother’s sandwich -> This be my mother sandwich
* Stemming: Separate words into their stems and affixes. (Most of the time just keeping the stems). Ex: In continuation of the theme, I will continue to produce sentences. -> In continu of the theme, I will continu to produc sentence.
* Sentence Segmentation: Determine where the beginning/end of a sentence is. (Periods are ambiguous). Ex: The words Dr., Mrs., and Ph.D all contain periods where the sentence does not stop. -> The words Dr., Mrs., and Ph.D all contain periods where the sentence does not stop EOS

1. We know how to compute similarly distance between two given strings using the edit distance algorithm.
   1. Please write down the distance matrix for the following strings:  
      String 1: Spokesman confirms  
      String 2: Spokeswoman said

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| s | 18 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| m | 17 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| r | 16 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| i | 15 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| f | 14 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| n | 13 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| o | 12 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| c | 11 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | 10 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| n | 9 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| a | 8 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| m | 7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| s | 6 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| e | 5 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| k | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| o | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| p | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| S | 1 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| # | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|  | # | S | p | o | k | e | s | w | o | m | a | n |  | s | a | i | d |

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

1. Please formulate your language model for the following text. Show the details of your LM formulation.  
   “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
   2. Bigram model
2. 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?