Python Programming –Final Term Exam

ESS 112 - Programming I

International Institute of Information Technology – Bangalore

Problem 5

Marks: 6

Problem Description:

We want develop a simple Natural Language Processing program. The input of this program is para which is string. A para is made of sentences. Sentences are separated from one another by the string '.' (full stop). Each sentence is a sequence of tokens (words) separated by spaces. For the sake of simplicity we will ignore the possibility of other separators such as ',' or ';'. The list of all unique tokens in a paragraph is called its vocabulary.
In order to simplify further processing each token in a vocabulary is represented as a vector. One simple encoding function for tokens in NLP is the one-hot encoding.

In this encoding scheme each token in the vocabulary is assigned a unique index or position in a vector of fixed length (equal to the size of the vocabulary). Each vector has a value of 1 in the position corresponding to the index of the token and 0 elsewhere.

For example consider a vocabulary of 6 tokens: {"cat" "dog" "bird" "fish" "The" "ate"}. The one-hot encoding for these tokens would be:

- "ate": [1 0 0 0 0 0]
- "bird": [0 1 0 0 0 0]
- "cat": [0 0 1 0 0 0]
- "dog": [0 0 0 1 0 0]
- "fish": [0 0 0 0 1 0]
- "the": [0 0 0 0 0 1]

Write the program that does the following:

1. Given a paragraph computes the vocabulary for the program
2. Given a vocabulary that is sorted alphabetically computes a unique one-hot encoding for every token in the vocabulary
3. Given a paragraph output the paragraph encoded with the one hot encoding based on the vocabulary of the paragraph

As an example the paragraph "the cat ate the bird. the dog ate the fish." should be encoded as below:
[[0 0 0 0 0 1] [0 0 1 0 0 0] [1 0 0 0 0 0] [0 0 0 0 0 1] [0 1 0 0 0 0] [0 0 0 0 0 1] [0 0 0 1 0 0] [1 0 0 0 0 0] [0 0 0 0 0 1] [0 0 0 0 1 0]]

Input format:

The program will take a single paragraph as an input terminated with a return character. Tokens will be seperated either by a space ' ' or a period '.'. The input will only contain characters for the following sets: a-z, A-Z,' ', '.'

Python Programming –Final Term Exam

ESS 112 - Programming I

International Institute of Information Technology – Bangalore

Output format:

The first line of the output will print the vocabulary of the paragraph. The null token ‘’ is not considered a valid token and should be discarded from the vocabulary.

The subsequent lines will print the input paragraph coded with one hot encoding. The one hot encoding of each word will be printed on a new line

Sample Input:

the cat ate the bird. the dog ate the fish.

Sample output:

'ate' 'bird' 'cat' 'dog' 'fish' 'the'

0 0 0 0 0 1

0 0 1 0 0 0

1 0 0 0 0 0

0 0 0 0 0 1

0 1 0 0 0 0

0 0 0 0 0 1

0 0 0 1 0 0

1 0 0 0 0 0

0 0 0 0 0 1

0 0 0 0 1 0