Ali Lashini

Iran University of Science and Tech.
www.iust.ac.ir

Assignment 4 Problems

NLP : Fall 1401 : Dr. Minaei
Due Sunday, Azar 20, 1401

Contents

Edit Distance And Spell Correction (30 points)

(a) (5 points)

State the types of spelling errors. Give a brief explanation and ways to recognize each of them.

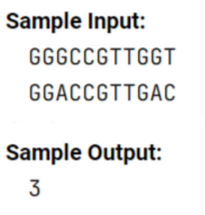
(b) (5 points)

What is a noisy channel and explain how it works.

(c) (20 points)

For this problem, we’re going to work with DNA string and genome replication
DNA string is a string of **nucleotides** which is one **{A, C, G, T}** letters.
Genome replication is one of the most important tasks carried out in the cell. Before a cell can divide, it must first replicate its genome so that each of the two daughter cells inherits its own copy.
Replication begins in a genomic region called **the replication origin (denoted ori)** and is carried out by molecular copy machines called **DNA polymerases** . Locating ori presents an important task not only for understanding how cells replicate but also for various biomedical problems. For example, some gene therapy methods.
Research has shown that certain proteins can only bind to DNA if a specific string of nucleotides is present, and if there are more occurrences of the string in a small subset of DNA, then it is more likely that binding will successfully occur and this subset is the ori of DNA
But in this problem, we don’t want to deal with the whole DNA string of a genome. We will focus on the relatively easy case of finding ori in bacterial genomes, most of which consist of a single circular chromosome. Research has shown that the region of the bacterial genome encoding ori is typically a few hundred nucleotides long. Our plan is to deal with a bacterium in which ori is known.
Suppose that we have a string (from new we use ”pattern” instead of a string) like this *”ACGT”* . In some cases, we don’t look for patterns with the exact abovementioned characters, sometimes a few mismatches could be okay like ”ATGT”. In this case, we considered some penalty for our problem, or in another word we accepted patterns with some mismatches. In the following questions, we’re going to use code to deal with such problems.

1) The first step is to implement hamming distance algorithm. Hamming distance function gets two strings with equal lengths and checks the corresponding character in each string to find mismatches. input and output of your function should be like this. (2 points)



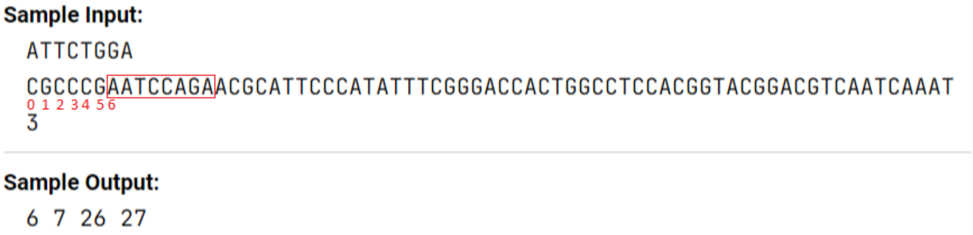
Sample Input:

```
GGGCCGTTGGT
GGACCGTTGAC
```

Sample Output:

```
3
```

2) In the next step, we give you a pattern, DNA string, and an integer as the maximum number of mismatches (provided in the assignment file as *”prob1_c2.txt”*) and your function has to return all starting positions where Pattern appears as a substring of DNA string with at most d mismatches. (5 points)
sample input and output of this problem is like this:



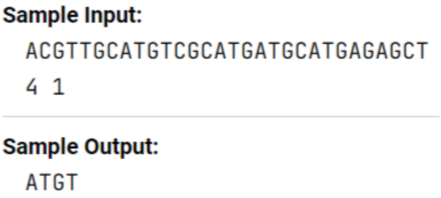
Sample Input:

```
ATTCTGGA
CGCCCGAATCCAGAACGCATTCCCATATTTCGGGACCACTGGCCTCCACGGTACGGACGTCAATCAAAT
0 1 2 34 5 6
3
```

Sample Output:

```
6 7 26 27
```

3) In the final step, we gave you a DNA string, the length of the pattern, and the maximum number of mismatches (provided in the assignment file as *"prob1_c3.txt"*). Your function should return a pattern with the most frequent occurrence in the DNA string. (13 points)
sample input and output of this problem is like this:



Sample Input:

```
ACGTTGCATGTCGCATGATGCATGAGAGCT
4 1
```

Sample Output:

```
ATGT
```

Generative and Discriminative models (20 points)

(a) (5 points)

What is the difference between a generative and discriminative model? Explain with a statistical view.

(b) (5 bonus points)

Compare generative and discriminative model performance based on the volume of a dataset and missing data.

(c) (10 points)

Consider the following features:
f1(c, d) : [c = PERSON and w-1 = "Mrs." and isCapitalized(w)]
f2(c, d) : [c = LOCATION and startsWith(w, "A") and startOfSentence(w-1)]
f3(c, d) : [c = VERB and endsWith(w, "ed")]

which:
w : current word
w-1 : previous word
isCapitalized : first letter of word has to be Capitalized
startsWith : start a word with certain string
endsWith : end a word with certain string
startOfSentence : that specific word has to be first word in the sentence

Now consider following sentence:
In Shiraz, Mrs. Sue viewed Pasargad last year.
For Shiraz, Sue, and Pasargard calculate one of the {LOCATION, PERSON, VERB} tags. (λ for f1, f2, f3 is 0.4, 0.5, 0.8.)

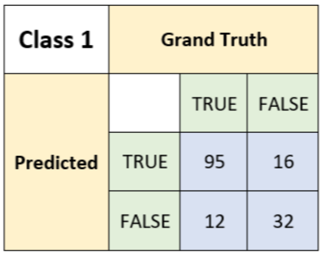
Text Classification (60 points)

(a) (5 points)

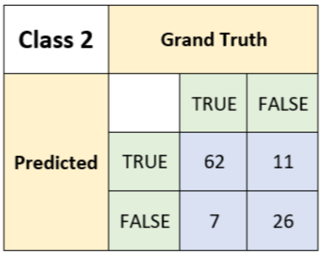
Explain the reason for using logarithm in Naïve Bayes model calculation, and how it affects learning.

(b) (5 points)

Consider the two following tables:
Compute micro and macro averaging precision, recall, and F1 score.



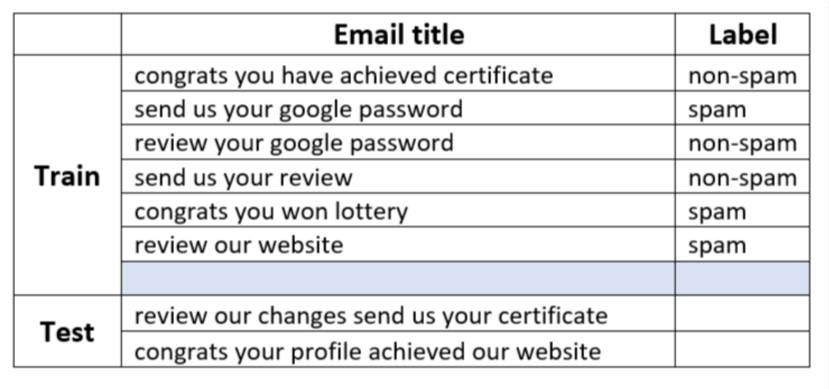
Class 1	Grand Truth		
Predicted		TRUE	FALSE
	TRUE	95	16
	FALSE	12	32



Class 2	Grand Truth		
Predicted		TRUE	FALSE
	TRUE	62	11
	FALSE	7	26

(c) (10 points)

Train a naïve Bayes Language model with add-1 smoothing on the following email title and corresponding spam/non-spam as the label, then test the model on the test set. You should also add your own sample in the empty rows (denoted by blue) of the train set.



	Email title	Label
Train	congrats you have achieved certificate	non-spam
	send us your google password	spam
	review your google password	non-spam
	send us your review	non-spam
	congrats you won lottery	spam
	review our website	spam
Test	review our changes send us your certificate	
	congrats your profile achieved our website	

(d) (35 points + 5 bonus)

In this part, we're going to train a Naïve Bayes Classifier for the task of sentiment analysis on the Hugging face emotion dataset (For more information check this link). Since this is a multi-class dataset and we want to train a binary classifier, we will use classes 0 and 1 (0:sadness, 1:joy). On the other side, You should merge the training set's data with the validation set. Please complete the notebook provided in your assignment folder (35 points).

Criterion:
You can't import any libraries in the notebook.
You have to write comments in your code that makes it fully apprehensible.
A code that meets the above criterion will complete this section's score.
Bonus: Your model should have accuracy above **90 percent** on the test set. (5 points)

Notes

Codes should be implemented in .ipynb format (notebooks)

- All Code cells should be executed before turning in the assignment (Make sure your outputs are there before you submit your assignment)
- Please explain the code and the results in the document or notebook
- If you have any questions, feel free to ask. You can ask your questions in the Telegram group.
- Please upload your assignments as a zipped folder with all necessary components. Upload your file in HW4-NLP-YourStudentID-YourName.zip format.