Amirali Molaei

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

Assignment 2 Problems

NLP: Fall 1400: Dr. Minaei
Due Friday, Azar 19th, 1400

Table of Contents

Spell Correction

Problem 1 (10 pts)

In a text, a misspelled word exists. we have three options to replace the misspelled word with. By using Levenshtein distance, suggest the word that is an optimal match from these options (Don't forget to draw the distance table for each option!).

Assume that the misspelled word in your text is in the row corresponding to the last digit of your student id from the following table.

Table 1) misspelled words

row	misspelled word	options		
1	Nureal	Natural	Neural	Neral
2	Overhet	Overrate	Overheat	overhead
3	terain	terrain	terran	train
4	netural	neutral	natural	neural
5	sieze	size	seize	seas
6	soldger	seldom	soldier	seller
7	Sleam	slim	sleep	sleeve
8	sirrow	swallow	sorrow	serow
9	supris	supper	surprise	separate

Problem 2 (10 pts)

A laboratory has provided DNA's sequences in a table. They ask you to find the best global alignment of two of these sequences:

First sequence: The row corresponding to your last student id digit

Second sequence: An arbitrary sequence from the table

Use Needleman & Wunsch algorithm to find the **optimal global** alignment by drawing the table and specifying **arrows** and the **traceback**.

Table 2) DNA sequences

DNA	Sequence
1	AGTATTTCCT
2	GGAATAATC
3	ACTGAT

4	CAAGACC
5	ACATCCAGA
6	TAATAAGC
7	GTGATTA
8	CTCAAACCAT
9	AAATGCTC

Problem 3 (Bonus-10 pts)

In this section, you're going to implement a spelling correction algorithm that takes a sentence as an input and outputs the best suggestions for each word in the sentence. Choose the first sentence of your paper abstract (the paper you've chosen for your project) and copy the sentence on this website to generate a misspelled sentence, Then use it as an input to your algorithm. The vocabulary provided in your assignment folder should be enough.

Generative and Discriminative models

Problem 1 (5 pts)

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

Problem 2 (5 pts)

Please explain the difference between Naïve Bayes and Maxnet classifiers.

Problem 3 (10 pts)

We want to train a model that converts a natural language text into logical programs. What kind of model (Generative or Discriminative) do you suggest? Explain why and describe the training process briefly.

Problem 4 (5 pts)

Assume we have table one for a Generative and table two for a Dicriminative algorithm. Complete these tables with the these data:
(1,3) , (1,1) , (1,1) , (2,3) , (2,1) , (2,2) , (0,3) , (0,3) , (0,2) , (0,1)

Table 3) Generative algorithm table

P(a,b)	b = 1	b = 2	b = 3
a = 0			
a = 1			
a = 2			

Table 4) Discriminative algorithm table

P(a | b)	b = 1	b = 2	b = 3
a = 0			
a = 1			
a = 2			

Text Classification

Problem 1 (5 pts)

What's the intuition behind the word "Naïve" in this classifier?

Problem 2 (10 pts)

Train a naïve Bayes Language model with add-1 smoothing on the following tweets and corresponding emotion 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 and test set.

Table 5) text samples from twitter

	Tweet	Emotion
Train set	I am feeling sad and hopeless today	sadness
	I feel very relaxed and fine today	Happy
	Today is my lucky day	Happy
	Today was boring as hell	sadness
	I can't do my job anymore	sadness
	I watched a fun movie	Happy

Test set	I feel relaxed and lucky after my job interview	
	This movie was boring and predictable with no fun	

Problem 3 (40 pts)

In this part, we're going to train a Naïve Bayes Classifier for the task of sentiment analysis on the IMDB movie reviews dataset. Please complete the notebook provided in your assignment folder (40 PTs).

Criterion:

You can't import any libraries in the notebook.

You have to write comments in your code that makes it fully apprehensible.

Your model should have accuracy above **80 percent** on the test set.

A code that meets the above criterion will result in a complete score for this section.

A Warm-Up for Deep Learning (Bonus-10 pts)

1. What kind of activation function is used for a multiclass classification problem? Why?
2. Why is CUDA architecture utilized in Deep Learning?
3. what cognitive characteristic of humans inspired the idea of transformers?
4. What made the Symbolic AI paradigm fall and the connectionist paradigm rise?
5. Explain the following items:
    a. Tensor
    b. Embedding
    c. Representation
    d. Optimizer
    e. Scheduler

Notes

- 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 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 *HW2_NLP_YourStudentID_YourName.zip* format.