CSE 4392 Special Topic: Natural Language Processing

Homework 3 - Spring 2025

Due Date: Feb 11, 2025, 11:59 p.m. Central Time

Problem 1 - 40%

You are given the below training set.

	Email ID	Content	Class
Training	1	Limited offer, apply now	Spam
	2	Your subscription offer expires soon	Spam
	3	Meeting scheduled for tomorrow	Ham
	4	Congratulations! You won a prize	Spam
	5	Can we reschedule the meeting?	Ham
Test	6	Your prize offer expires tomorrow	?

Question 1.1 - 30%

Compute the conditional probabilities matrix P with add-1 smoothing and the prior probabilities vector p.

Question 1.2 - 10%

Compute $P(c \mid d_6)$. You must write the original inference formula as covered in the slides then write it in terms of your defined parameter tensors.

Problem 2 - 60%

Question 2.1 - 30%

Define a class **NaiveBayesClassifier** that has a train method that computes the training parameters given an arbitrary list of sentences (with optional add-1 smoothing). Apply the class on the training data presented in the previous problem to ensure the results are correct.

Question 2.2 - 20%

Add to the class an inference method that given a sentences returns the most likely class. Test it on the previous problem and ensure it returns the same probability and the correct class.

Question 2.3 - 10%

Write an evaluate function that given a list of test sentences and a list of corresponding labels, returns the corresponding F1-Score by running inference on each sentence then comparing the result to the true labels. In comments, precisely explain why your computes the F1-Score correctly.

Your submission should include:

- Written answer to the first problem (ideally in LaTeX), but could be handwritten.
- A `.py` file for your answer to the second problem. Please note that the implemented class should not contain any information about the data (i.e., implement generically). When this file is run, it should print the test cases used in the two functions and the results (i.e., the first problem as test case).
- Include a screenshot of the output from running your `.py` file. This should include all the information mentioned in the last bullet.
- Convert the `.py` file to a PDF using a tool such as `https://www.i2pdf.com/source-code-to-pdf` and include it in your submission.
- Using libraries is okay as long as they do not do the probability computation end-to-end (ie, you can use libraries to construct the count matrices or equivalent datastructures to that will be used to later compute conditional and joint probabilities in inference and evaluation)
- You must use Numpy (or PyTorch) in the implementation. Inefficient loops will be penalized.