CSCI 3223: Tashfeen's Machine Learning II

Midterm Project

1. Disclaimer

"Insofern sich die Sätze der Mathematik auf die Wirklichkeit beziehen, sind sie nicht sicher, und insofern sie sicher sind, beziehen sie sich nicht auf die Wirklichkeit."

—Albert Einstein. 1921. *Geometrie und Erfahrung*. Julius Springer, Berlin.

The above quote is commonly translated in English to "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality." Usually I try to make my assignments as objective as possible. However, this one will have some room for creative flexibility to implement a real world solution. You are after all taking Machine Learning *à deux*. As such, if you find any ambiguity, confirm your choice of disambiguation with the professor. I do expect some communication from you as you decide on your approach.

2. Context

A while ago, a YouTuber (Luke Smith) uploaded a video demonstrating and complaining about the excessive ads and trackers on the internet. In particular, on cooking/recipe websites. Since I shared this experience with the author, I followed the two more videos they made in the series.

This resulted in an effort to start an open source (and possibly *libre*) website: https://based.cooking. The videos I refer to are linked in the "About this site" section.

The website is essentially generated by a static site generator Hugo. And contributions for new recipes are made through pull requests to their GitHub repository:
https://github.com/lukesmithxyz/based.cooking

The original website is no longer maintained but it has two popular forks,

- Fork 1: https://github.com/ronaldl29/public-domain-recipes
- Fork 2: https://github.com/Rushmore75/foss.cooking

The first fork seems to be updated more frequently, while the second one has its recipes normalised (which maybe desired for this project).

3. The Project

I want you to use the data from these websites and produce,

1) A clean dataset
2) A model that uses this dataset
3) A website that lets users interact with this model

The input of the model and the subsequent website should be a list of ingredients and the output should be a recipe for a food that uses these (and as much as possible, only these) ingredients.

The dataset you produce should be independent of your implementation, i. e., someone else should be able to get your data and solve a different problem.

Further details maybe filled in by your creativity.

4. Spirit of the Project

One possible way to solve the problem above is to fine tune a Large Language Model (LLM) with the recipe markdown files and then simply query it for new recipes. This is low hanging fruit and will earn you a "C" at best. This is not to say that you can't use an LLM at all, I am okay if it is a part of your solution's pipeline as long as it is not the whole pipeline. I want you to use all that you have learned in Machine Learning I and II so far and come up with as deterministic of a solution as possible. Do some *science!* Analyse your data, research different models, use different metrics, experiment with data transformation techniques and come out with an original solution.

Submission Instructions

1) A ten minute PDF presentation titled `presentation.pdf`. Here you will include all the plots, techniques, metrics and results and at the end demonstrate your website.
2) A `README.md` that shows how can someone run your project on their machine.
3) A zip of your cleaned root-project-directory. Remove any auxiliary files or environment directories, e. g., `.venv` or `.git`.