Project B0: Spam/Ham Classification — Design Document

Data 100/200: Principles and Techniques of Data Science
Fall/Spring 20XX

Background

In Project B, you and a partner will take on the roles of data scientists at a large email service. Your manager provides you both with a large collection of emails and assigns the job of building and evaluating a model that classifies each email as *spam* or *ham (not spam).*

Objective

This is the first part of this project. Here, you and your partner will plan your work and design a clear path for building a logistic regression classifier. Your design document should demonstrate understanding of the data by describing what is provided, suggesting data cleaning methods, and proposing what featuers to engineer.

Teamwork

In Project B, you will work in **pairs** (2 students). Both members are jointly responsible for each submission in the project, and will be graded equally all throughout.

Timeline

- Project B0 (Design Document) Released: `MM/DD`
- Pair Registration Due: `MM/DD`
- **Project B0 Due:** `MM/DD`
- Project B1 Released: `MM/DD`
- Project B1 Due: `MM/DD`
- Project B2 Released: `MM/DD`
- Project B2 Due: `MM/DD`
- Project B3 Released: `MM/DD`
- Project B3 Due: `MM/DD`

Design Document Format

Your design document should be a typed, 1-2–page document written in short sections outlined below that will help guide you in a more comprehensive EDA in the subsequent parts of the project. While your document does not need to be written in complete sentences, we strongly encourage you to provide descriptive, succinct bullet points in each section. Altogether, your design document should contain the following:

- **Asking a Question.** This section should be brief as this is essentially provided for you. However, still be concrete on what question we are trying to answer.
- **Describing the Data.** Describe what kind of data is provided and the format it is provided in.
- **Understanding the Data.**
  - **Data Processing.** Note down at least 1 potential data cleaning technique you may want to carry out. As guidance, you might want to consider the format in which the data is currently provided. If you were to construct a DataFrame, what columns would it have? What would its granularity be?
  - **Feature Engineering.** Describe at least 5 different features you might want to include in your model that can help with differentiating an email from being spam or ham. For each, explain why the feature might be effective (in 2-3 sentences) and provide at least 2 email examples that showcase the prevalence (or lack thereof) of the feature. In the examples you provide, please write down the file name along with the relevant text.

Submission

You will submit to Pensieve and add your partner to the submission. Do NOT individually submit to Pensieve!

Grading

Project Component	%
Asking a Question	10
Describing the Data	10
Understanding the Data (Data Processing)	20
Understanding the Data (Feature Engineering)	50
Clarity & Formatting	10