Chris Piech
CS109

Section #1
October 2nd, 2025

Section 1: Core Probability

1 Warm-Up with Music Preferences

In one prior offering of CS109, the distribution of students by year and their likelihood of liking the song *We Are the Champions* by Queen is shown in the table below. Here, "Graduate +" includes graduate students, SCPD students, coterminal students (everyone not in the other partitions). Let L_1 denote the event that a student likes the song.

Year	**% of Students**	$P(L_1 \mid \text{Year})$
Freshman	20%	0.23
Sophomore	25%	0.42
Junior	25%	0.39
Senior	10%	0.74
Graduate +	20%	0.89

What is the probability that a randomly chosen student likes *We Are the Champions* by Queen?

2 WebMD mini

In this problem, we will compute the probability that a person has a particular disease given that they present with the symptom of a fever. A person may have either *Influenza (Flu)*, *Streptococcal Pharyngitis (Strep)*, or *No Disease*. Assume these outcomes are mutually exclusive, so each person belongs to exactly one of the three groups: Flu, Strep, or No Disease.

In the general population, 8% of people have Influenza (Flu), 4.1% have Streptococcal Pharyngitis (Strep), and the remaining 87.9% have no disease. If a person has the flu, the probability they have a fever is 0.92. If they have strep, the probability they have a fever is 0.86. If they have no disease, the probability they have a fever is 0.01.

a. For each disease (Flu, Strep, No Disease), compute the probability that a person has that disease given that they present with a **fever**.

b. In part (a) $P(\text{Fever}|\text{Flu}) = 0.92$ was estimated from historical data of patients who were identified as having the flu. Another option for estimating the same probability would be to use a language model with the assumption:

   Assume $P(\text{Fever}|\text{Flu})$ is instead given by:

$$\texttt{prefix} = \text{“Summary of symptoms for patient with Flu: ”}$$
$$\texttt{phrase} = \text{“Has fever”}$$
$$\texttt{pr\_fever\_given\_flu} = \frac{\texttt{string\_pr(prefix + phrase)}}{\texttt{string\_pr(prefix)}}$$

   What are the advantages and disadvantages of the two methods of computing $P(\text{Observation}|\text{Flu})$?

3 Will your Friend Like This Song?

In CS109 we have 500 song recommendations! There are many randomized algorithms that can help us choose the top 16, but they rely on solving the following problem:

Imagine a student has already rated n songs as "like" or "dislike". For any new target song i that the student has not rated yet (out of the original 500) define the event L_i as the event that the student likes song i. Estimate $P(L_i \mid \text{student's previous } n \text{ ratings})$.

To get you started, we list out a few things that you can optionally use in your solution.

Datasets
Current Quarter: Assume it is currently a few weeks into the quarter, and you have a dataset with votes from other students in the class. Assume you have about 9 songs rated per student, and about 10 votes per song. Each student will have voted on a different randomly sampled set of 9 songs.

Historical Data: You also have a dataset from prior CS109 offerings with students and the songs that they like/dislike. Warning, the target song, and the songs they rated are likely not in the dataset.

Helper Functions
`string_pr(prompt)`
Returns the probability the LLM assigns to the *entire* input string.

`get_summary(student_song_ratings)`
Uses an LLM to produce a short natural-language description of the student's tastes.

`similarity(input_1, input_2)` Returns a number that represents how similar two inputs are in meaning (according to an LLM). For example, `similarity("rock music", "metal music")` would be a lot higher than `similarity("rock music", "broccoli")`.

This is not a problem with one correct solution. In fact, there are many possible approaches and it is still considered an open problem! The goal is to encourage you to be creative. Have fun.