Serverless Data Ingestion & Fuzzy Matching API

IMG**:** Me and coworker (Lynee) at our work.
Thanks for the image, Lynee!!

General Description of this Project

In the summer of 2025, I joined a new University job at the University of Wisconsin-Madison! The job entailed answering guest questions and redirecting contacts to the best possible resources on campus. Every employee was required to log the questions they were asked for the day in a shared google sheet. This project's purpose was to take in the extracted data and output an easily-readable pie labelled pie chart and list of the most frequently asked question. This served to help our team determine what resources every employee must be trained to know, and maximize the efficiency of our service!

Planning of the Project

Every good project starts with a good and clear planning. Here is what I hope to implement:

1. Python Development
    - Extracting the CSV File and load into Pandas Dataframe
    - Clean Data by removing extra "noise", fixing common misspelling, lowercasing, and working around different spacings.
    - Adjust and work with Fuzzy-Matching algorithm to ensure each data point is classificed under one roof. I.E "I&f", "Lost and Found for hat" turns into "Lost and Found (2)"
    - Report Generation to show graph and readable figures of the data
2. Workflow Packaging on Nextflow (RE-DESIGN -> Dockerize Flask API?)
    - Define Nextflow processes based on the python script
    - Create Docker/Singularity Image on cloud computing.
3. Deployment
    - Showcase and showboat! Show off the project to staff and explain its usage.
    - Allow access to staff. This will thus make my project into a real tool that can be used by my workplace
    - Dashboard integration. Have nextflow output to a readable visualization tool, such as google sheets. This allows an easy interpretation of the data.

Skills Used for the Project

Each skill is separated by columns and rows to specifically tell what I did with each.

Python & Pandas

Python was used for the backend of this program which was to process the csv file, cleaning and organizing data, to output usable and accurate results. The Pandas library was used to store and clean the inputted csv file.

Fuzzy-Matching Algorithm

Fuzzy-matching is a python algorithm that is commonly used for process and projects like this. It allows the data to be grouped together even if too different for my python code to notice.

Dockerize & Flask API

The flask API will be the bridge that helps communicate between the HTMP/Javascript and the python code. In order to avoid issues for users, I dockerized the Flask API to allow all processes to run smoothly.

HTML/CSS/Javascript

Used HTML, CSS, and Javascript to produce the website, allow for the drag-and-drop of files, and display the data. This makes it easy for the user and makes the project overall more usable.

Development Log

Title of The Log 12/1/25

This project is currently underway. Developments will be made here to showcase the development process