Project Documentation

BITEBUDDY

Bhakti Ukey
Harsh Shah
Jared Videlefsky

Table of Contents

INTRODUCTION 2

IDEATION 3

DATA: ESSENTIAL THREADS TO TRAINING LLM 4

APPROACHES EXPLORED 4

EXPLORATORY DATA ANALYSIS 7

PREPROCESSING 7

DATA CLEANING 8

LARGE LANGUAGE MODELS 9

POST PROCESSING FOR LLM 11

SENTENCE TRANSFORMATION MODEL & K-MEANS CLUSTERIN 11

TESTING REVIEW AGGREGATION 12

PROMPT ENGINEERING 13

RECOMMENDATION SCORE 14

DIETARY RESTRICTIONS. 14

TECH-STACK. 15

INTRODUCTION

In the dynamic realm of culinary choices, the question of "What to order?" at a restaurant is a universal puzzle that has perplexed everyone at some point in their gastronomic journey. The desire to savor a delectable meal while ensuring optimal value for money often leaves us at the mercy of restaurant staff recommendations, frequently accompanied by persuasive upselling tactics. What if there were a more democratic and efficient way to navigate the menu, drawing on the collective wisdom of past diners?

Introducing BiteBuddy, a revolutionary solution designed to transform the dining experience into a seamless and informed adventure. In a world where dining out is a celebration of flavors, BiteBuddy emerges as the ultimate companion for those seeking not just a meal but a curated culinary delight.

IDEATION

At the core of this innovation lies the "What to Order?" app, a groundbreaking tool poised to redefine how we approach restaurant menus. Functioning as an interactive guide, this app empowers users to swiftly access personalized food recommendations with just a few taps. The concept is simple yet powerful: users input the restaurant they find themselves in, and within moments, the app leverages the wealth of information embedded in Google reviews.

Here's how it works: BiteBuddy scours Google reviews for the specified restaurant, meticulously sifting through the myriad of experiences shared by previous patrons. The app intelligently identifies reviews that offer insights into specific meals, gathering valuable data about these culinary experiences. Harnessing the collective knowledge of the dining community, BiteBuddy then compiles this information to deliver tailored suggestions for what to order.

In essence, BiteBuddy is not just an app; it's a culinary confidant, a digital ally that democratizes the dining decision-making process. Whether you're a seasoned food enthusiast or someone embarking on a culinary exploration, BiteBuddy promises to be the compass that guides you to an unparalleled dining adventure based on the authentic experiences of those who have gone before you. Embrace the future of informed dining with BiteBuddy, where every bite is a discovery and every meal a memorable journey.

Data

Essential Threads in Training LLMs

In the intricate process of training Large Language Models (LLMs), the significance of high-quality data cannot be overstated. Accurate and diverse datasets form the bedrock upon which these models acquire language understanding and generate contextually relevant outputs. However, obtaining reliable data is a complex and often messy undertaking. The challenge lies not only in the vastness of the required information but also in navigating the nuances of language diversity, context, and ever-evolving linguistic trends. Our documentation explores the multifaceted journey of sourcing accurate data for LLM training. From leveraging curated datasets to wrangling unstructured information, we delve into the methodologies employed, such as web scraping, API integration, and manual curation. We further unravel the intricacies of data acquisition, an indispensable phase in sculpting the proficiency of LLMs.

Approaches Explored

1. **Google Local Data -USCD**
2. **API – Google Maps API, SerpAPI,**
3. **Open-Source Data Repository. - Common Crawl , Hugging Face**
4. **Synthetic Data using LLM – OpenAI , gretelAI, palm2**

Navigating an In-depth Exploration of Approaches

1. **Google Local Data**
   This Dataset contains review information on Google map (ratings, text, images, etc.), business metadata (address, geographical info, descriptions, category information, price, open hours, and MISC info), and links (relative businesses) up to Sep 2021 in the United States. We used this dataset (UCSD - Google Local Data) to train our LLM model. It contains:
   - Reviews - 666,324,103
   - Users - 113,643,107
   - Businesses: 4,963,111

2. **API**

In our quest to amass relevant data, we initially experimented with various APIs:

1. **Google Maps API**

   The logical choice for gathering restaurant-related information led us to the Google Maps API. The Google Maps API provides valuable details about businesses, including names, locations, and reviews. However, a notable drawback surfaced during our exploration: the API restriction, allowing retrieval of only 20 businesses with a maximum of 5 reviews per business. This limitation posed a challenge to our objective of acquiring a robust dataset for training our Language Models, prompting us to reevaluate our approach and consider alternative methods for comprehensive data collection.

2. **SerpAPI**

The limitations encountered with the Google API prompted us to investigate SerpAPI as an alternative solution. SerpAPI operates as a versatile search engine results page (SERP) data extraction tool, empowering us to extract detailed information from various sources efficiently. The transition to SerpAPI marked a strategic move in our pursuit of comprehensive and diverse data sets for refining our Language Models.

Encountering hurdles in the API integration, we navigated a challenge where the API call necessitated a data field known as "data_id," obtainable from the Google Maps URL for each business. To surmount this obstacle, we strategically employed the GMAP_ID sourced from Google Local Data, enabling a seamless resolution to this challenge.

3. **Common Crawl**

The Common Crawl Repository is a publicly accessible archive of web crawl data. It is a vast collection that encompasses snapshots of web pages gathered during periodic crawls of the internet. Common Crawl aims to democratize access to web data and promote open and free access to information on the internet. The repository contains petabytes of data, including text content, metadata, and information about the structure of websites.

Navigating uncharted terrain, we encountered challenges in grappling with various data formats, including warc, wet, segments, and indexes. Unzipping the .paths.gz files for each format added an additional layer of complexity. The data, hosted on S3, is conveniently accessible for download using HTTPS methods, facilitating the retrieval of a manageable subset for local exploration.

Upon extraction, a meticulous examination revealed that the restaurant review data wasn't available within the dataset. In tandem, we explored Hugging Face Datasets, investigating the possibility of uncovering restaurant-related information within the datasets hosted by this platform. This multi-faceted approach allowed us to comprehensively survey different sources, ensuring a thorough exploration of the available data landscape.

4. Synthetic Data

The surge in the use of synthetic data has become increasingly prevalent. Faced with the challenge of acquiring a substantial volume of data through API calls and other conventional sources, we embarked on an exploration into the realm of generating synthetic data using Large Language Models (LLMs). Our investigation encompassed several avenues, each offering unique insights into the process:

1. Vertex AI Palm 2:
   Initially promising, Vertex AI Palm 2 garnered attention. However, it posed constraints, requiring data in a specific JSON format with input and output as the sole parameters. Unfortunately, our dataset demanded additional parameters, leading to a misalignment with our requirements.

2. OpenAI
   Optimism centered around OpenAI, a potential solution for crafting the necessary dataset. Despite its promise, we encountered a significant hurdle in the form of API limits, restricting data retrieval to just five rows. Even after attempting to escalate the API limit, our efforts were thwarted by persistent challenges.

3. Gretel.ai:
   A standout alternative, Gretel.ai emerged as a powerful solution allowing us to work with datasets in various formats, such as JSON, JSONL, and CSV. Notably, Gretel.ai accommodated our need for substantial data volumes, offering options like 5000/10000 rows. The platform's flexibility and scalability proved instrumental in overcoming previous limitations.

While each avenue explored provided valuable insights, there remains room for improvement. Ongoing efforts are directed towards enhancing the capabilities of the platforms, ensuring more robust and flexible synthetic data generation. The landscape of synthetic data creation is evolving, and we are actively monitoring and adapting to forthcoming upgrades in these platforms for continued advancements in our data generation strategies.

Exploratory Data Analysis

Preprocessing

To streamline our dataset, several exclusion rules have been implemented:

1. **No Meal Items in Reviews**
   - Reviews lacking mentions of specific meal items are excluded from our analysis. This criterion ensures that our focus is on reviews providing detailed insights into dining experiences.

2. **Determined by LLM Response**
   - The Language Model's response plays a crucial role in filtering reviews. Only those reviews aligning with the model's criteria for relevance are retained, contributing to a more refined dataset.

3. **Less than 10 Reviews for a Restaurant**
   - Restaurants with fewer than 10 reviews are excluded from the dataset. This threshold ensures a minimum level of data density, enhancing the robustness of our analysis.

4. **At Least 5 Reviews with Text**
   - To ensure meaningful insights, restaurants must have a minimum of 5 reviews with associated text. This criterion prevents the inclusion of reviews with minimal or no textual content.

Data Cleaning

In addition to exclusion rules, a set of data cleaning procedures has been applied:

1. **Make the Review Lowercase**
   - Standardizing the text by converting all reviews to lowercase ensures uniformity in our dataset, minimizing variations and facilitating consistent analysis.

3. Remove Special Characters
   - Special characters have been systematically removed from reviews to eliminate potential distortions and enhance the accuracy of sentiment analysis and data processing.

3. **Exclude Reviews with No Review Text**
   - Reviews lacking substantive text content have been excluded from the dataset. This step focuses our analysis on reviews that contribute meaningful information to our understanding of dining experiences.

Large Language Models

Selecting the Right Language Model

In our quest for the optimal Large Language Model (LLM) , we conducted a thorough exploration of multiple prominent LLMs currently available. The search included renowned models such as GPT-3, GPT-4, PaLM 2, LaMDA, among others.

After meticulous consideration and evaluation, we strategically zeroed in on PaLM 2 as the backbone for Bard AI. PaLM 2 proved itself to be a highly capable Language Model, standing toe-to-toe with GPT-4 in terms of proficiency. The decision to choose PaLM 2 over others was further reinforced by Bard AI's status as the latest, most updated, and remarkably accessible option for integration. With a focus on cutting-edge technology, Bard AI not only aligns with our product vision but also offers users a seamlessly integrated and cost-effective solution.

Post Processing for LLM

In the post-processing phase, we implement a series of sophisticated techniques to refine and organize our dataset for enhanced analytical insights:

1. **Grouping Similar Meals Together**

- Employing advanced algorithms, we systematically group similar meals together. This ensures a coherent representation of culinary offerings, allowing for more nuanced analysis.

2. **Removal of Stop Words**

- To focus on the most meaningful content, non-essential words, known as stop words, are systematically removed from our dataset. This process enhances the relevance and clarity of our textual information.

3. **Utilizing NLTK Python Library**

- Leveraging the Natural Language Toolkit (NLTK) in Python, we employ a comprehensive suite of tools for text processing, enabling more precise and context-aware linguistic analysis.

4. **Stemming for Word Reduction**
   - The process of stemming is applied to reduce words to their base or root form. This not only streamlines the dataset but also enhances the efficiency of subsequent analyses.

5. **Tokenization for Numerical Representation**
   - Utilizing tokenization, we represent words numerically, paving the way for the application of clustering algorithms. This step facilitates a more structured and quantitative approach to understanding textual content.

6. **TF-IDF Vectorization or Cosine Similarity**
   - Incorporating advanced techniques such as TF-IDF vectorization and cosine similarity, we enhance the semantic understanding of our dataset. These methodologies contribute to a more nuanced representation of relationships and patterns within the data.

7. **LLM Application for Grouping:**
   - Employing a Large Language Model (LLM) on our meal name outputs, we harness advanced natural language processing capabilities to intelligently group related meal names. This step adds a layer of contextual sophistication to our dataset.

Exclusions:
To ensure data integrity and relevancy, the post-processing phase also involves exclusions based on specific criteria:

1. **Meal Name Results Threshold**
   - Meals with fewer than 2 results after the grouping process are excluded from the dataset. This criterion ensures that only meals with a sufficient presence in the dataset are considered, enhancing the robustness of subsequent analyses and interpretations.

Enhancing Meal Name Consistency: Leveraging Sentence Transformer Models and KMeans Clustering

In the pursuit of standardizing meal names, we employ a Sentence Transformer model to generate text embeddings for various meal descriptions. Our current experimentation involves the utilization of the MiniLM-L6-v2 architecture. The ensuing results are then organized into clusters using KMeans, facilitating the comparison of meal name counts before and after clustering.

The objective is to evaluate the impact of clustering on meal name consistency by comparing counts before and after the clustering process.

Key Steps in the Experiment:

1. **Sentence Transformer Model Integration**
   Employing the MiniLM-L6-v2 model to transform meal names into text embeddings for enhanced consistency.

2. **KMeans Clustering**
   Utilizing KMeans to group the generated embeddings, allowing for a more organized and coherent representation of meal variations.

3. **Cluster Labeling**
   Defining cluster labels by selecting three random names within each cluster. In cases where there are fewer than three results, all available results become the cluster label.

By systematically implementing these steps, our aim is to streamline meal names and foster greater coherence in the representation of diverse meals.

Testing Review Aggregation with Language Models

LLM Testing Approach: Combining Restaurant Reviews into a Single Prompt

Pros:
- Speed: Accelerated processing of restaurant reviews.

Cons:
- Sentiment Loss: Aggregation may lead to the loss of nuanced sentiments associated with individual meal reviews.

Pricing (GPT 3.5):
- Cost: $0.25 for a combined text of 100 reviews.
- Variable Factor: Pricing is contingent on the length of the reviews.

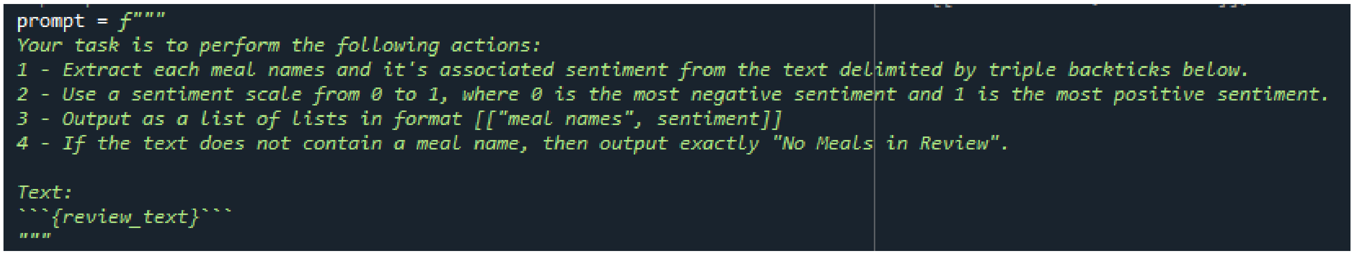
PaLM2 Performance:
- Records Processed: 8,500
- Processing Time: 215 minutes
- Free Tier: Allows 90 requests per minute.

GPT 3.5 Turbo:
- Cost: $5/month
- Free Tier:Available, providing limited access.

In the pursuit of efficient review processing, the LLM testing involves consolidating all reviews for a restaurant into a single prompt. This approach offers a faster turnaround, albeit at the expense of losing the nuanced sentiments associated with individual meal reviews. The pricing structure varies, with GPT 3.5 charging $0.25 for combining 100 reviews, while PaLM2 boasts the processing of 8,500 records in 215 minutes under its free tier. GPT 3.5 Turbo presents an alternative with a $5/month cost and a limited free tier. These insights are crucial for optimizing the selection of language models based on specific use cases and priorities.

Continuous Testing model : Prompt Engineering

Final Prompt:



~~~
prompt = f"""
Your task is to perform the following actions:
1 - Extract each meal names and it's associated sentiment from the text delimited by triple backticks below.
2 - Use a sentiment scale from 0 to 1, where 0 is the most negative sentiment and 1 is the most positive sentiment.
3 - Output as a list of lists in format [["meal names", sentiment]]
4 - If the text does not contain a meal name, then output exactly "No Meals in Review".

Text:
```{review_text}```
"""
~~~

Final Prompt Configuration:

- Temperature: 0
- Processing Speed:
  - Limitation: Owing to API constraints, processing is capped at 90 requests per minute.

Performance Testing Results:

- Requests Processed:461
- Time Taken:700 seconds (~11.5 minutes)

In our ongoing efforts to refine and assess the model's capabilities, the final prompt is configured with a temperature setting of 0. The processing speed is subject to API limitations, restricting us to handling a maximum of 90 requests per minute.

Our recent testing session involved the processing of 461 requests, with the total duration amounting to approximately 11.5 minutes. These metrics serve as valuable indicators of the model's efficiency and provide insights into its real-world responsiveness under specified conditions. As we continue to iterate and enhance, this continual testing approach ensures that our model consistently meets performance expectations.

Recommendation Score

In the development of our recommendation system, we meticulously crafted a score based on several key components, each assigned varying weights to reflect their significance. To ensure the robustness of our score, we implemented a clustering mechanism for meal names, effectively mitigating spelling variations and enhancing accuracy.

The components integrated into our recommendation score encompass the total number of reviews, the average rating derived from these reviews, and the sentiment associated with mentions of each meal.

Furthermore, our system incorporates real-time user feedback obtained through the Bitebuddy App, adding an additional layer of dynamic input for continuous improvement. This comprehensive approach allows us to provide users with nuanced and reliable meal recommendations, enriching their overall experience.

Dietary Restrictions

In order to enhance the end-user experience, we have implemented a user-friendly feature that empowers individuals to tailor their meal recommendations according to specific dietary preferences and restrictions. Users can seamlessly navigate through a set of pre-set dietary restriction questions, allowing them to discern whether a selected meal aligns with their dietary needs. These questions cover a spectrum of considerations, such as determining:-

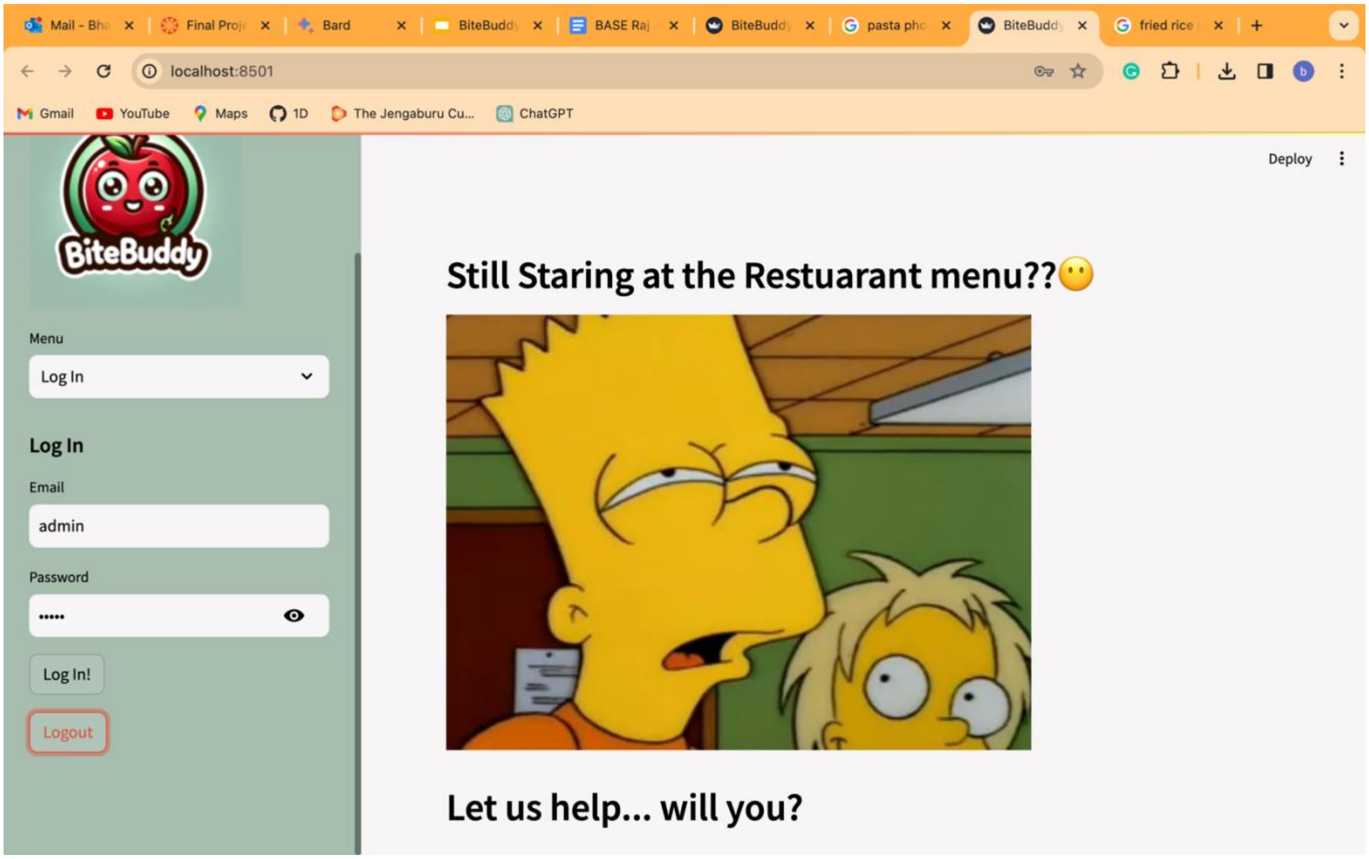
Questions:

- Is {meal_name} vegetarian or vegan?
- Is {meal_name} halal/ kosher?
- Does {meal_name} contain {restriction}?

The integration of the Language Model (LLM) further refines this experience by dynamically generating responses based on the context provided, specifically the name of the selected meal. This ensures that the responses are highly relevant and contextually accurate, fostering a more personalized and informed decision-making process for users. By combining a user-friendly interface with advanced language processing capabilities, we aim to offer a comprehensive and tailored solution that accommodates diverse dietary preferences and restrictions, ultimately enriching the user's interaction with our platform.

Tech- Stack

- Snowflake Database
- Python
- Streamlit
- Dbt



Menu

Log In

Log In

Email

admin

Password

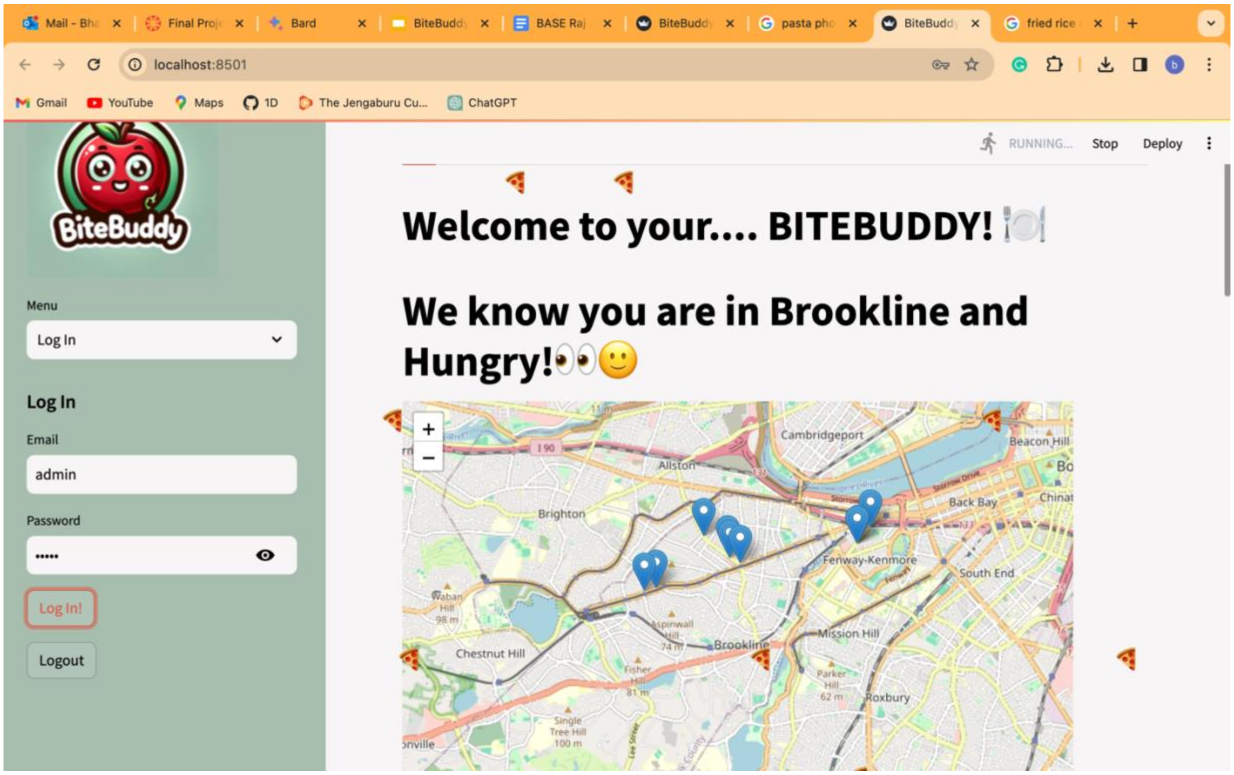
•••••

Log In!

Logout

Still Staring at the Restuarant menu??😶

Let us help... will you?



Menu

Log In

Log In

Email

admin

Password

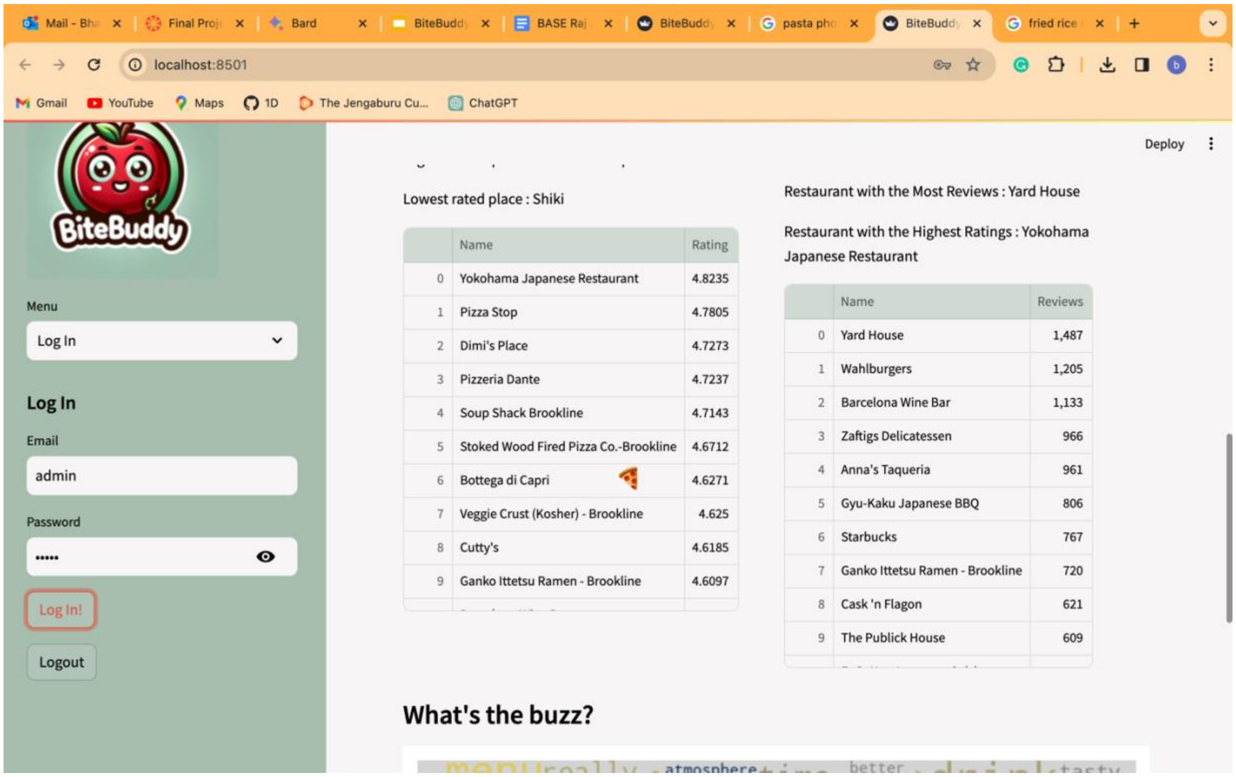
•••••

Log In!

Logout

Welcome to your.... BITEBUDDY! 🍽️

We know you are in Brookline and Hungry!👀🙂



Menu

Log In

Log In

Email

admin

Password

•••••

Log In!

Logout

Lowest rated place : Shiki

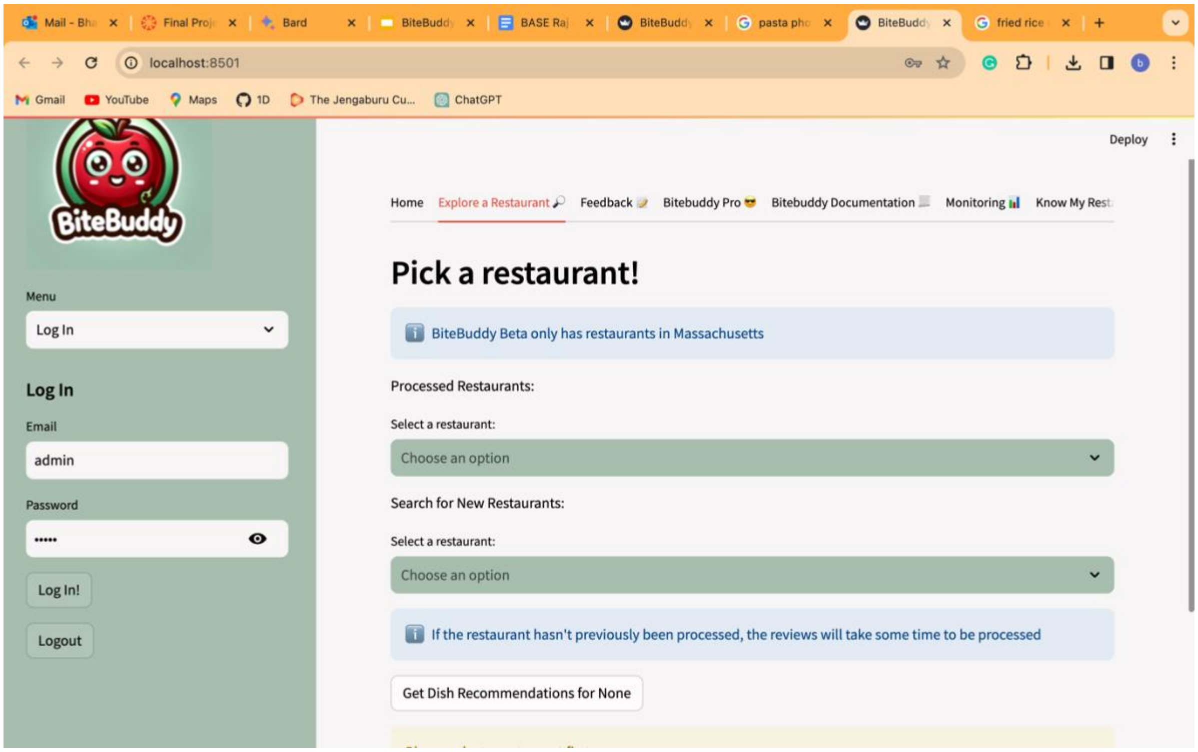
	Name	Rating
0	Yokohama Japanese Restaurant	4.8235
1	Pizza Stop	4.7805
2	Dimi's Place	4.7273
3	Pizzeria Dante	4.7237
4	Soup Shack Brookline	4.7143
5	Stoked Wood Fired Pizza Co.-Brookline	4.6712
6	Bottega di Capri	4.6271
7	Veggie Crust (Kosher) - Brookline	4.625
8	Cutty's	4.6185
9	Ganko Ittetsu Ramen - Brookline	4.6097

Restaurant with the Most Reviews : Yard House

Restaurant with the Highest Ratings : Yokohama Japanese Restaurant

	Name	Reviews
0	Yard House	1,487
1	Wahlburgers	1,205
2	Barcelona Wine Bar	1,133
3	Zaftigs Delicatessen	966
4	Anna's Taqueria	961
5	Gyu-Kaku Japanese BBQ	806
6	Starbucks	767
7	Ganko Ittetsu Ramen - Brookline	720
8	Cask 'n Flagon	621
9	The Publick House	609

What's the buzz?



Menu: Log In

Log In

Email: admin

Password: •••••

Pick a restaurant!

BiteBuddy Beta only has restaurants in Massachusetts

Processed Restaurants:

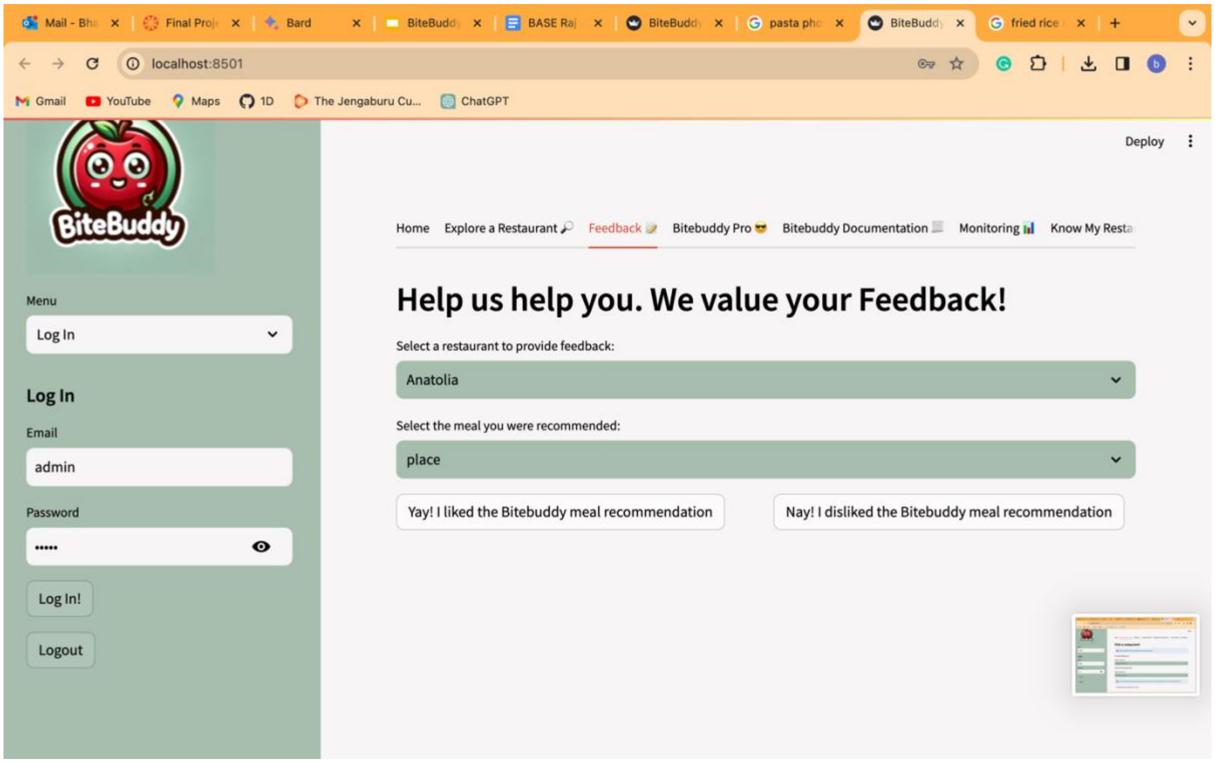
Select a restaurant: Choose an option

Search for New Restaurants:

Select a restaurant: Choose an option

If the restaurant hasn't previously been processed, the reviews will take some time to be processed

Get Dish Recommendations for None



Menu: Log In

Log In

Email: admin

Password: •••••

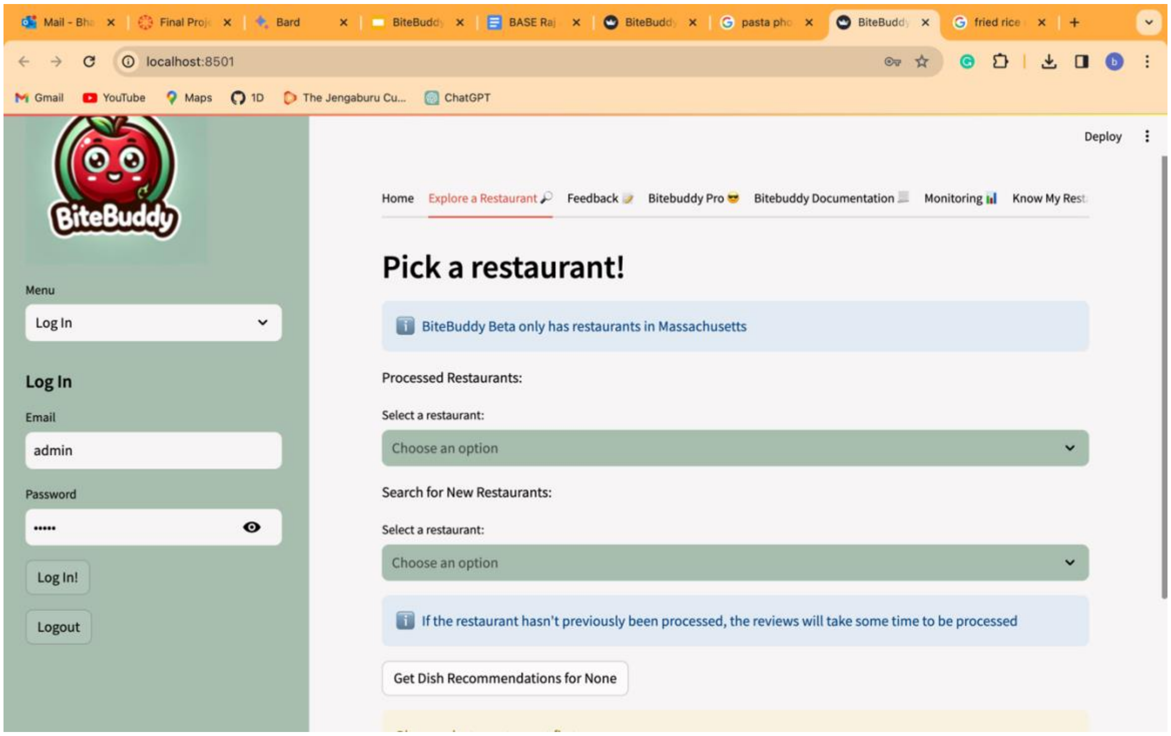
Help us help you. We value your Feedback!

Select a restaurant to provide feedback: Anatolia

Select the meal you were recommended: place

Yay! I liked the Bitebuddy meal recommendation

Nay! I disliked the Bitebuddy meal recommendation



Menu

Log In

Log In

Email

admin

Password

•••••

Pick a restaurant!

ℹ BiteBuddy Beta only has restaurants in Massachusetts

Processed Restaurants:

Select a restaurant:

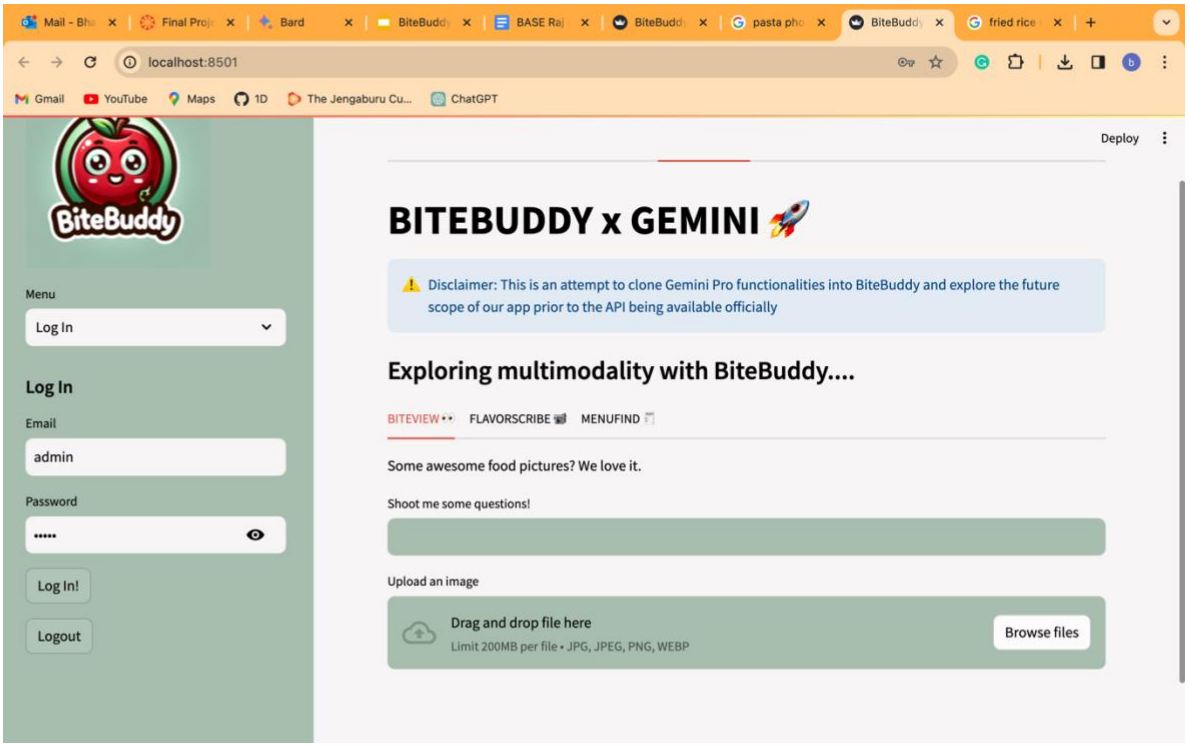
Choose an option

Search for New Restaurants:

Select a restaurant:

Choose an option

ℹ If the restaurant hasn't previously been processed, the reviews will take some time to be processed



Menu

Log In

Log In

Email

admin

Password

•••••

BITEBUDDY x GEMINI 🚀

⚠ Disclaimer: This is an attempt to clone Gemini Pro functionalities into BiteBuddy and explore the future scope of our app prior to the API being available officially

Exploring multimodality with BiteBuddy....

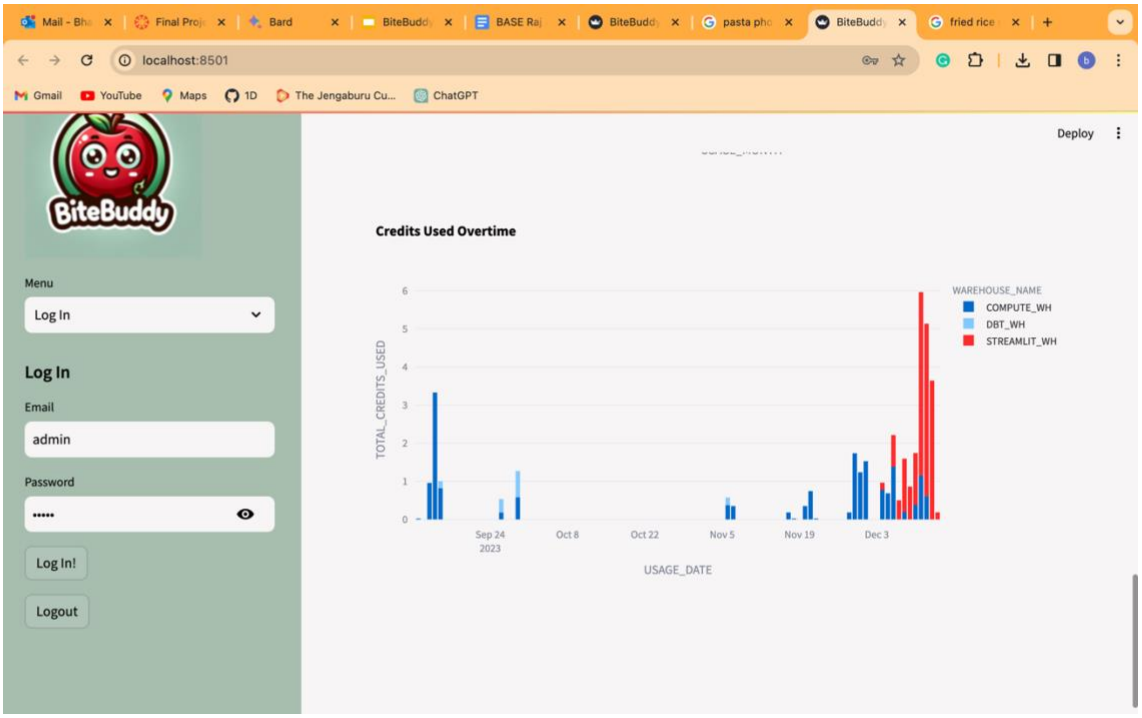
Some awesome food pictures? We love it.

Shoot me some questions!

Upload an image

Drag and drop file here

Limit 200MB per file • JPG, JPEG, PNG, WEBP



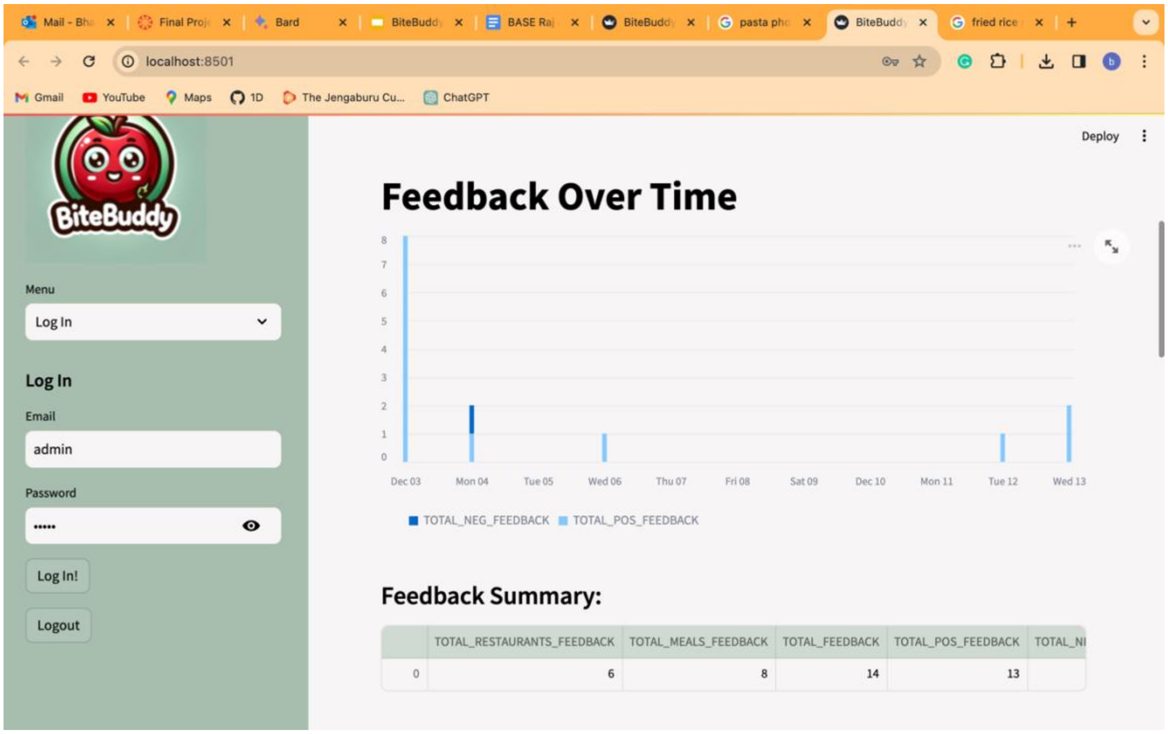
Menu: Log In

Log In

Email: admin

Password: •••••

Credits Used Overtime



Menu: Log In

Log In

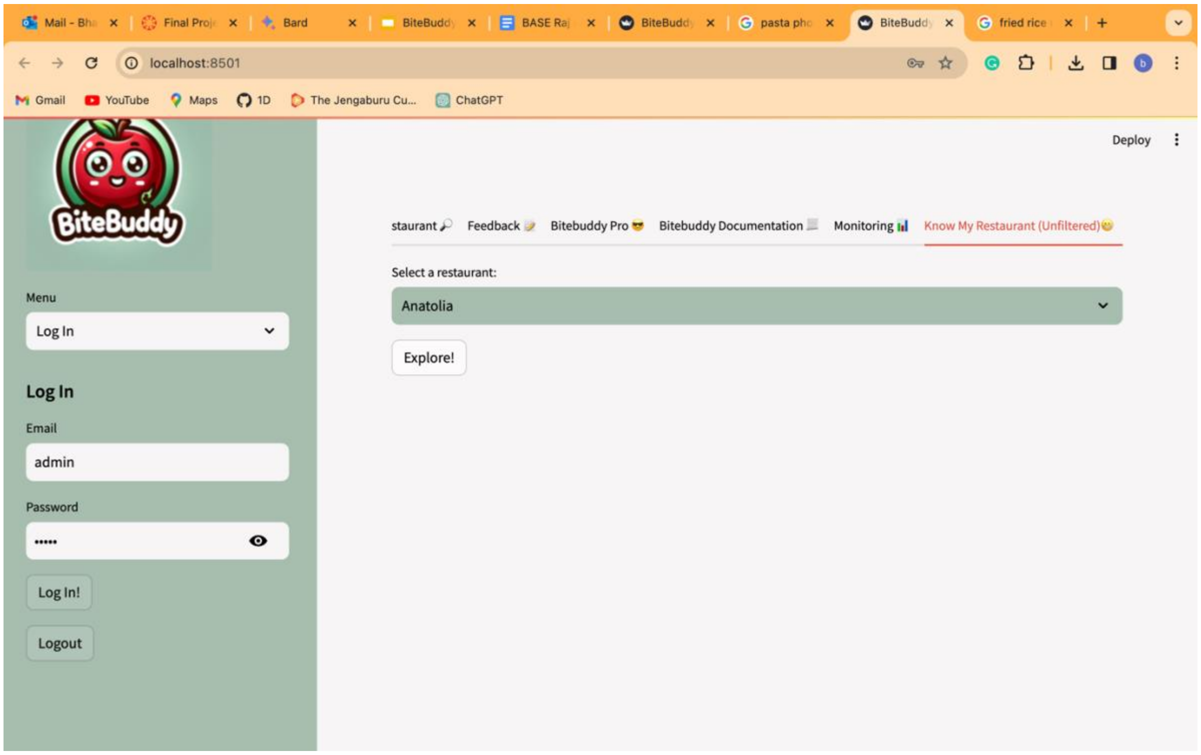
Email: admin

Password: •••••

Feedback Over Time

Feedback Summary:

	TOTAL_RESTAURANTS_FEEDBACK	TOTAL_MEALS_FEEDBACK	TOTAL_FEEDBACK	TOTAL_POS_FEEDBACK	TOTAL_NI
0	6	8	14	13	



Menu: Log In

Log In

Email: admin

Password: •••••

Select a restaurant: Anatolia