RESTAURANT HEALTH VIOLATIONS

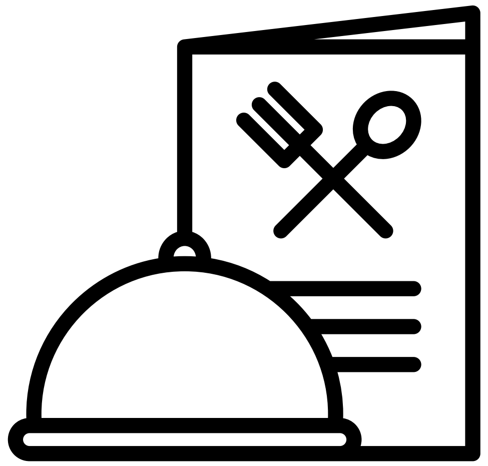
```python
# P2: Prior Work, Data Overview, and Project Plan

import pandas as pd

data = {'Names':
    ['Carolina Aldana Yabur',
    'Shlok Nandkishor Goud',
    'Zahid Rahman']}

df = pd.DataFrame(data)
df
```

```
+---+-----------------------+
|   |                 Names |
+---+-----------------------+
| 0 |   Carolina Aldana Yabur |
| 1 |   Shlok Nandkishor Goud |
| 2 |          Zahid Rahman |
+---+-----------------------+
```



CONTENTS

PROBLEM STATEMENT

Foodborne illnesses impact millions of people every year, and one major way to prevent them is through restaurant health inspections.

However, the official inspection reports themselves can be lengthy, technical, and scattered across different sources. This makes it hard for diners, restaurant owners, and even health officials to see which issues come up most often or which places repeatedly fail to follow safety rules.

Our project aims to take all that raw inspection information and turn it into concise summaries and easy-to-spot trends. Ultimately, we want to help everyone—from customers to health departments—quickly understand where and when common violations happen, improving transparency and safety in the dining world.

CDC estimates 48 million people get sick, 128,000 are hospitalized, and 3,000 die from foodborne diseases each year in the United States.

PAPER OVERVIEWS

Zahid

- Predictive Analytics Using Text Classification for Restaurant Inspections (Wang et al., 2017)
- Automatic Text Summarization and Keyword Extraction using Text Rank Algorithm (Dumne et al., 2020)

Carolina

- Hindsight Analysis of the Chicago Food Inspection Forecasting Model (Kannan et al. 2019)
- A for Effort? Using the Crowd to Identify Moral Hazard in New York City Restaurant Hygiene Inspections (Mejia et al., 2019)
- Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews (Kang et al., 2013)

Shlok

- Identification of critical factors for assessing the quality of restaurants using data mining approaches (Mahmood & Khan, 2019)
- Supplementing Public Health Inspection via Social Media (Schomberg et al., 2016)

Paper # 1: Predictive Analytics Using Text Classification for Restaurant Inspections (Wang et al., 2017)

NLP WORK IN RESTAURANT DOMAIN

ZAHID'S 1ST PAPER

Brief Study Summary	<u>**First study proposing predictive analytics to detect foodborne illnesses from Yelp reviews.**</u> **Approach:** • Used Yelp Academic Dataset due to API limitations. • Classified reviews as indicating foodborne illness (1) or not (0), addressing imbalanced data and computational complexity (large N-gram feature sets). • Balanced the dataset by selecting 15,213 reviews, filtering for foodborne illness keywords before training. • Applied text-mining (TF-IDF, LIWC sentiment scoring) and GLMs for feature selection. • Trained Naïve Bayes, SVM, Random Forest, and RNN, with SVM & RNN performing best.
Connection to Our Project, Gaps, and Differences	• **Similar NLP Methods** – Both use text-mining (TF-IDF, LIWC) and classification models to extract key insights. • **Different Focus** – They predict foodborne illness from reviews, while we summarize official health violations. • **Feature Selection** – They use keywords and sentiment, while we apply key phrase extraction and summarization. • **Integration Potential** – Their category-based approach could help us group violation types effectively

Paper #2: Automatic Text Summarization and Keyword Extraction using Text Rank Algorithm

Summarization using new algoritihm

(Dumne et al., 2020)

Zahid's 2nd Paper

Brief Study Summary	• **Goal**: Focuses on extractive text summarization using the TextRank algorithm, illustrated by summarizing lengthy technical documents to quickly capture essential information. • **Method**: Employs a graph-based approach with cosine similarity to rank sentences, ensuring the most representative content is extracted. • **Modules**: Allows for URL, and web scraping summarization, accommodating diverse input sources.
Connection to Our Project, Gaps, and Differences	• **Focus:** Chosen for dedicated study on text summarization, a topic not deeply covered in class. • **Transferrable:** Directly applies to extracting health code violations from scraped web pages. • **Enhance:** Strengthens the project's web scraping and key phrase extraction by tackling summarization. • **Gap:** Fills a gap in summarization methods, distinct from key phrase extraction already covered in class.

PAPER # 3: HINDSIGHT ANALYSIS OF THE CHICAGO FOOD INSPECTION FORECASTING MODEL

Objective	Analyze the Chicago Department of Public Health (CDPH) machine learning model that prioritizes inspections based on predicted risk of critical food violations.
Methodology	🔹 **Data:** Routine inspections: 17,075 training (Sep 2011 - Apr 2014) and 1,637 testing (Sept 2014 - Oct 2014). 🔹 **Model:** Logistic Regression. 🔹 **Target Variable:** Binary label (whether at least one critical violation was found). 🔹 **Predictor Variables:** Past serious/critical violations, time since last inspection, business age, alcohol and tobacco licenses, daily high temperature, burglary rate, sanitation complaints, garbage cart requests, and inspector cluster group (6 categories)
Findings CAROLINA 1ST PAPER	✅ Model Performance: • Reduces time to find critical violations by 7.4 days on average. • Evaluated using three metrics: 1️⃣ Average time reduction for detecting violations. 2️⃣ Standard deviation of time reduction. 3️⃣ Fraction of critical violations found early. ✅ Concerns & Limitations: ⚠ Sanitarian Bias: Inspection outcomes unfairly influenced depending on inspector. ⚠ Time Invariance Issue: A violation may not occur on a different inspection day. ⚠ Uncertain Model Impact: Post-2015 increases in violations could be due to external factors. ⚠ Flawed Hit Rate Metric: Inspectors may unconsciously adjust behavior when prioritizing high-risk locations. ⚠ Limited Predictors: Model lacks key food safety indicators (ingredients used, food storage, pest control history).

PAPER # 4: A FOR EFFORT? USING THE CROWD TO IDENTIFY MORAL HAZARD IN NEW YORK CITY RESTAURANT HYGIENE INSPECTIONS (MEJIA ET AL., 2019)

Objective	Investigate how online restaurant reviews can help detect hygiene-related issues in NYC restaurants (2010-2016).
Methodology	🔹 **Datasets Used:** • NYC Open Data Program (NYCOD): Contains inspection data (grades, violations, inspector details). • Yelp Reviews Dataset: 1.3M reviews from 24,625 restaurants, matched to NYCOD (95% match rate). 🔹 **Creating the SMASH Dictionary:** • Naïve Bayes Classifier identifies hygiene-related words in Yelp reviews. • MTurk crowd-sourced labeling + WordNet synonym expansion refine the word list. • Final dictionary includes: single words, two-word & three-word phrases (n-grams). 🔹 **Using SMASH to Identify Moral Hazard:** • Track daily SMASH word counts to monitor hygiene between 12-15 month inspection cycles. • Compare hygiene trends of two groups: "PAPA" restaurants (re-inspection required) ⚠️, and "AA" restaurants (passed on first attempt) ✅ • Apply longitudinal linear regression to model hygiene decline over 90 days post-inspection.
Findings **CAROLINA 2ST PAPER**	✅ Model Performance: • 📉 30% of NYC restaurants decline in hygiene within 90 days post-inspection. • 🍛 South Asian, Caribbean, and Pizza restaurants, along with low-cost eateries ($-$$), show higher regression. • 🏪 Franchise chains & cafés maintain stable hygiene. 🏙️ Brooklyn restaurants (especially low-cost pizza places) show the fastest hygiene decline within 30 days.

PAPER # 5: WHERE NOT TO EAT? IMPROVING PUBLIC POLICY BY PREDICTING HYGIENE INSPECTIONS USING ONLINE REVIEWS (KANG ET AL., 2013)

Objective

Analyze Yelp restaurant reviews in Seattle to predict hygiene inspection scores.

Methodology

🔹 **Data Collection:** Scraped Yelp restaurant reviews in Seattle from 2006–2013 and matched them with inspection records.

🔹 **Dataset:** 152K reviews across 1,756 restaurants, covering ~13K inspections.

🔹 **Feature Analysis:**

- Sentiment of reviews (average rating, negative review count).
- Deceptiveness of reviews (bimodal distribution, fake review detection).
- Filtering methods to remove outliers and potentially deceptive reviews.

🔹 **Prediction Model:**

- Features based on:
- ⭐Customers' Opinion: Aggregated opinion (average review rating) and review content (unigrams, bigrams).
- 🍽️Restaurant's Metadata: Cuisine, location, inspection history, review count, non-positive review count.

🔹 **Model Used:** Support Vector Machines (SVM) & Support Vector Regression (SVR) with 10-fold cross-validation.

Findings

CAROLINA 3ST PAPER

✅ Model Performance:

- Restaurant Metadata (Location, Cuisine): 🌍🍴 Predicts hygiene with ~66% accuracy.
- Review Content (Unigrams, Bigrams): 📝💬 Achieves the highest accuracy (~82.68%).
- Inspection History: 📅🔍 Highly predictive (~72%), indicating past performance is a strong indicator of future hygiene.

PAPER # 6: IDENTIFICATION OF CRITICAL FACTORS FOR ASSESSING THE QUALITY OF RESTAURANTS USING DATA MINING APPROACHES (MAHMOOD & KHAN, 2019)

SHLOK

ML Work on Inspection Data	First study to use official NYC inspection data to classify restaurant quality using ML. **Approach**: • Used **New York City Department of Health Inspection** Dataset (~215K records). • Applied **feature selection techniques (mRMR, LVQ)** to find most important inspection features. • Used **SVM (linear & nonlinear), Naïve Bayes, and Random Forest** for classification. • Evaluated performance using **Accuracy, Sensitivity, Specificity, AUC, and Kappa Coefficient.**
Connection to Our Project, Gaps, and Differences	• **Relevant Data Type** – They used official inspection data, just like our Hillsborough County plan. • **Key Features** – Score and grade were most predictive; other metadata (e.g., zip, cuisine) were weak. • **Gap in Consistency** – Results showed inspection decisions can be inconsistent or biased. • **Opportunity** – We can improve upon their approach by introducing NLP methods (e.g., from Yelp reviews) and validating on Florida data.

PAPER # 7: SUPPLEMENTING PUBLIC HEALTH INSPECTION VIA SOCIAL MEDIA (SCHOMBERG ET AL., 2016)

SHLOK

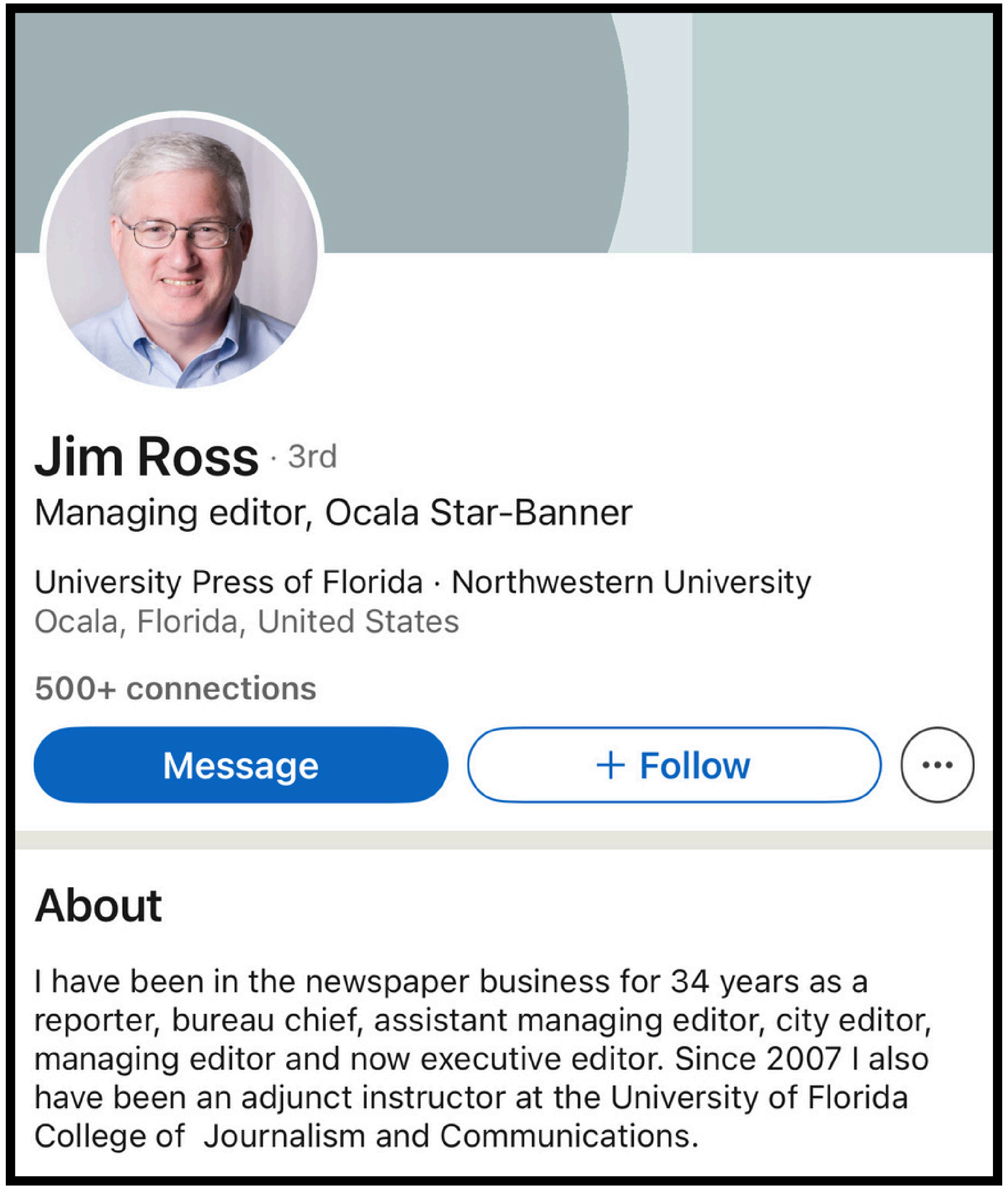
NLP for Public Health	Shows social media can supplement traditional inspections to detect food safety risks. **Approach**: • Collected Yelp reviews + Twitter posts with terms like **"vomiting," "food poisoning,"** etc. • Effectively combined **keyword-based filtering with manual curation and geospatial analysis**, making it a hybrid NLP + spatial analysis approach • **Mapped social signals to official inspection record**s to detect high-risk restaurants. • Created an **early warning system** to alert authorities of possible foodborne illness outbreaks.
Connection to Our Project, Gaps, and Differences	• **Shared Methodology** – They use NLP and real-time signals; we plan to incorporate public reviews too. • **Different Objective** – Their focus is early detection of outbreaks, ours is summarizing violations. • **Data Integration Potential** – Their geo-mapped approach could enrich our Florida inspection dataset. • **Inspiration** - Strong case for fusing official and public datasets to build more robust violation predictors.

Data Overview

- **Data Source:** Gannett → Oracle Star-Banner → Data Central → Restaurant Inspections

- **Data Restrictions:** No available API, so web scraping is necessary. However, according to https://data.ocala.com/robots.txt, only web scrapers for search engines (Google, Bing, OpenAI) are allowed and all custom user agents are disallowed.

- **What to do from here?**
  a. Mimic one of the allowed user agents. However, this is considered deceptive and unethical
  b. Contact the Site & Request Permission. This is slower but is the safer, transparent, and ethical method.



Data Overview

Request for Data Access for Academic Research – Ocala Restaurant Inspections

Zahid Rahman
To: jross@gannett.com
Cc: Shlok Nandkishor Goud; Carolina Aldana Yabur

Reply | Reply all | Forward
Tue 4/1/2025 8:33 AM

Dear Mr. Ross,

I hope this message finds you well. My name is Zahid Rahman, and I am a student in the ISM 6564 Text Analytics course at the University of South Florida. I am currently working on an academic project under the supervision of Dr. Anol Bhattacherjee (ABhatt@usf.edu), along with my teammates Carolina Aldana Yabur (caldanayabur@usf.edu) and Shlok Nandkishor Goud (shlokgoud@usf.edu).

The purpose of my project is to utilize data from the Ocala Restaurant Inspections page to automatically summarize health code violations. Specifically, the project will apply key phrase extraction and text summarization techniques to generate concise descriptions of common infractions. In addition, we will categorize these violations and identify trends over time to assess patterns within and across restaurants. Ultimately, this work may be extended to build a searchable dashboard or to highlight restaurants with routine violations.

To access the necessary data from your website in a manner that complies with your policies, I kindly request your assistance in one of the following ways:

1. **Whitelisting a Custom User Agent:**
   I intend to use an automated tool for data collection. To ensure transparency and compliance with your site's guidelines, the tool will include a custom user agent string.
   The user agent I plan to use is:
   "ZahidRahman_ISM6564_TextAnalytics/1.0 (+mailto:zrahman1@usf.edu)"
   Would it be possible to have this custom user agent whitelisted so that my requests are permitted?

2. **Providing API Access or a Structured Data Feed:**
   Alternatively, if you offer an API or any form of structured data feed (e.g., CSV or JSON) for the restaurant inspection data, I would greatly appreciate guidance or documentation on how to access it.

I want to emphasize that this project is solely for academic purposes. I will ensure proper attribution of your data in all my work and strictly adhere to any usage policies you have in place.

If any part of this request is outside your usual responsibilities or falls into more technical territory, I'd be very grateful if you could forward this to the appropriate member of your technical or IT team.

Thank you very much for considering my request. I look forward to your guidance on how best to proceed. Please feel free to contact me at zrahman1@usf.edu if you require any additional information.

Request for Data Access for Academic Research – Ocala Restaurant Inspections

GC GANNETT Corporate Communications<PR@gannett.com>
To: Zahid Rahman

Reply

You don't often get email from pr@gannett.com. Learn why this is important

Hi Zahid –

We will not be able to share access to the data you requested.

From: Zahid Rahman <zrahman1@usf.edu>
Sent: Tuesday, April 1, 2025 8:32 AM
To: Ross, Jim <jim.ross@starbanner.com>
Cc: Shlok Nandkishor Goud <shlokgoud@usf.edu>; Carolina Aldana Yabur <caldanayabur@usf.edu>
Subject: Request for Data Access for Academic Research – Ocala Restaurant Inspections

You don't often get email from zrahman1@usf.edu. Learn why this is important

...

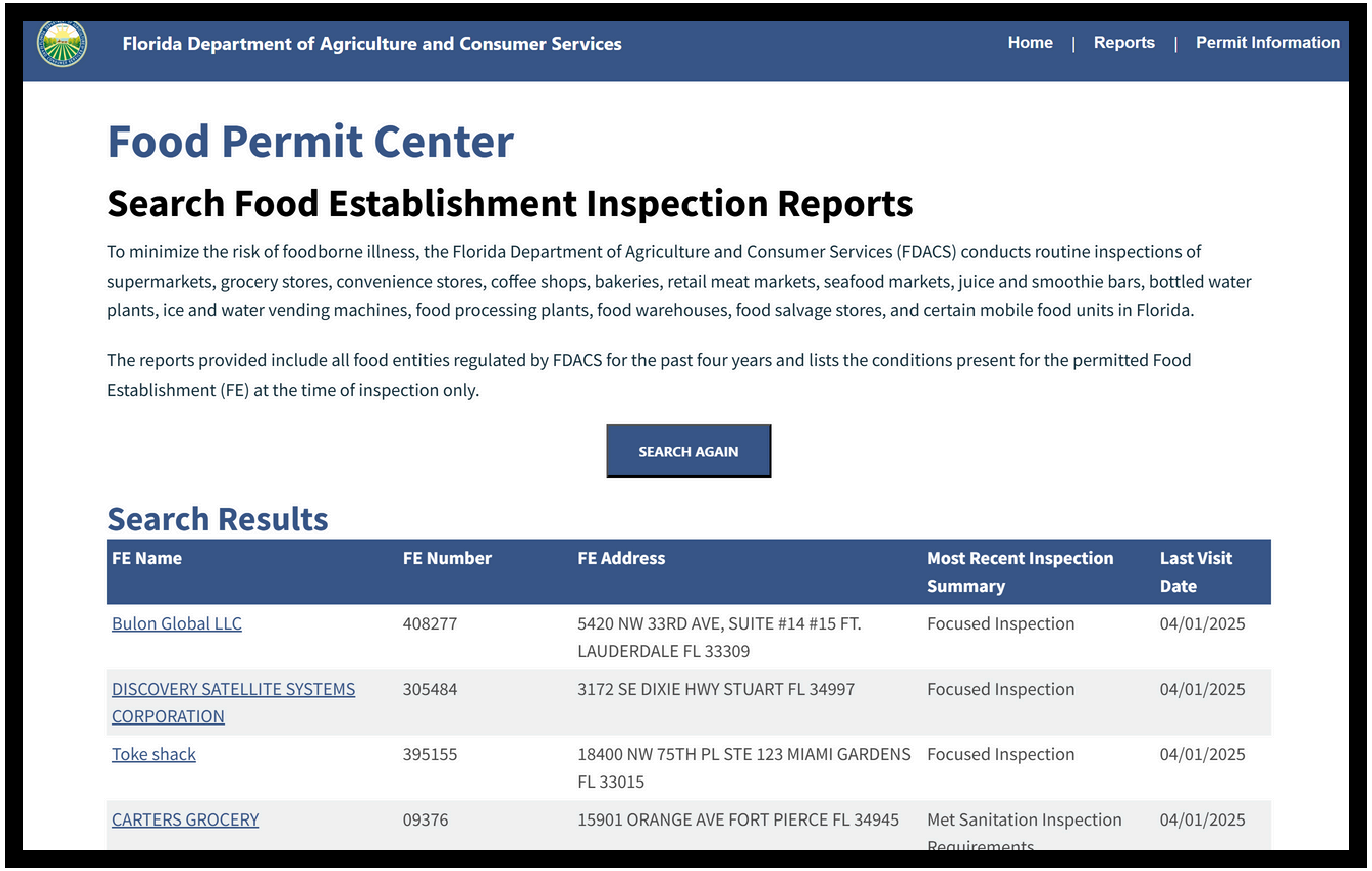
REQUEST FOR DATA ACCESS WAS DENIED FROM OWNER, SO NEED TO FIND DIFFERENT DATA SOURCE

LAST MINUTE DATASET ALTERNATIVE?

- State Government website, better primary source & is public information

- However, each entry is stored in a PDF file, which can add complexity for scraping

- More than 49,000 inspection records from present day dating back to 2021 of various food businesses across FL state.



Florida Department of Agriculture and Consumer Services

Home | Reports | Permit Information

Food Permit Center

Search Food Establishment Inspection Reports

To minimize the risk of foodborne illness, the Florida Department of Agriculture and Consumer Services (FDACS) conducts routine inspections of supermarkets, grocery stores, convenience stores, coffee shops, bakeries, retail meat markets, seafood markets, juice and smoothie bars, bottled water plants, ice and water vending machines, food processing plants, food warehouses, food salvage stores, and certain mobile food units in Florida.

The reports provided include all food entities regulated by FDACS for the past four years and lists the conditions present for the permitted Food Establishment (FE) at the time of inspection only.

SEARCH AGAIN

Search Results

FE Name	FE Number	FE Address	Most Recent Inspection Summary	Last Visit Date
Bulon Global LLC	408277	5420 NW 33RD AVE, SUITE #14 #15 FT. LAUDERDALE FL 33309	Focused Inspection	04/01/2025
DISCOVERY SATELLITE SYSTEMS CORPORATION	305484	3172 SE DIXIE HWY STUART FL 34997	Focused Inspection	04/01/2025
Toke shack	395155	18400 NW 75TH PL STE 123 MIAMI GARDENS FL 33015	Focused Inspection	04/01/2025
CARTERS GROCERY	09376	15901 ORANGE AVE FORT PIERCE FL 34945	Met Sanitation Inspection Requirements	04/01/2025

LAST MINUTE DATASET ALTERNATIVE?

- State Government website, better primary source & is public information

- However, each entry is stored in a PDF file, which can add complexity for scraping

- More than 49,000 inspection records from present day dating back to 2021 of various food businesses across FL state.

WILTON SIMPSON
COMMISSIONER

Florida Department of Agriculture and Consumer Services
Division of Food Safety

FOOD SAFETY INSPECTION REPORT

Chapter 500, Florida Statutes
(850) 245-5520

Visit #: 9999-7182-2847-39
Bureau of Food Inspection
Attention: Business Center
3125 Conner Boulevard, C-26
Tallahassee, FL 32399-1650

Name:	Dunkin Donuts # 8492	Establishment #:	61243
Owner:	Best Donut Inc	Date of Visit:	February 19, 2025
Type:	Retail Bakery w/FS	Inspected By:	TARIQUL ISLAM
Address:	9774 Glades RD Ste A11 Boca Raton, FL 33434-3993		

INSPECTION SUMMARY - Focused Inspection

A Focused Inspection is a visit focused on a specific aspect that will not result in an inspection summary.

NOTICE OF FEES

To review your account balance or to renew your permit, please visit our Food Permit Center at https://FoodPermit.FDACS.gov.

COMMENTS

This Focused Inspection is being conducted offsite to attach water and sewer bill.

A copy of this report has been provided to the person in charge of the food establishment and will be available online at https://foodpermit.fdacs.gov/Reports/SearchFoodEntity.aspx.

Tariqul Islam

TARIQUL ISLAM, SENIOR SANITATION AND SAFETY SPECIALIST

Name and Title of Whom This Report was Issued

WEB SCRAPING RESTAURANT DATA FROM THE FDACS WEBSITE

🔹 Automation

➤ Uses Selenium WebDriver to automate web browser interactions.

➤ Navigates to the FDACS food permit search page: https://foodpermit.fdacs.gov/Reports/SearchFoodEntity.aspx?mode=2

🔹 Data Extraction Process

➤ Selects a specific county (e.g., Hillsborough).

➤Triggers a restaurant search within the county.

➤ Parses the HTML table and extracts key details:

✅ Restaurant Name

✅ Address

✅ Inspection Report Links

🔹 Data Handling

➤ Stores extracted data in a structured dictionary.

➤ Converts the dictionary into a Pandas DataFrame.

➤ Saves the data as a CSV file: restaurants_info.csv

	A	B	C
1	Restaurant Name	Address	Inspections Link
2	MEMORIAL MARATHON	5701 MEMORIAL HWY TAMPA	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=332769
3	HOMEGOODS # 0576	18061 HIGHWOODS PRESER\	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=362763
4	MICHAEL'S # 2726	18081 HIGHWOODS PRESER\	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=282013
5	PERFORMANCE FOOD G	3140 GALLAGHER RD DOVER	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=279280
6	Horn of Plenty Produce &	802 W SAM ALLEN RD PLANT (	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=416150
7	Publix Super Market Inc.	2801 E. COUNTY LINE RD. LUT	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=408142
8	Publix Liquor Store #168	947 E BLOOMINGDALE AVE BI	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=382240
9	Dollar General Store #20	4860 S 78TH ST TAMPA FL 336	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=382257
10	7-ELEVEN #33019B - D-L	5102 POINTE OF TAMPA WAY	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=336184
11	WALGREENS # 5437	17511 BRUCE B DOWNS BLVD	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=265401
12	WALGREENS # 3145	1860 E FOWLER AVE TAMPA F	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=93331
13	EXTRA CARE PHARMACY	2001 E FLETCHER AVE TAMPA	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=314329
14	Save A Lot #30003	305 W HILLSBOROUGH AVE T	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=428948
15	PREMIER BEVERAGE CO	6031 MADISON AVE TAMPA FL	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=342037
16	Vapor Unlimited LLC	730 S DALE MABRY HWY TAMI	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=428940
17	TODD VIDEO INC	13417 N NEBRASKA AVE TAMI	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=328364
18	AUTOZONE # 104826	2560 E BEARSS AVE TAMPA FL	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=17900000
19	GULF COAST ICE (FLETC	102 W FLETCHER AVE TAMPA	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=350336
20	GULF COAST ICE	5919 W LINEBAUGH AVE TAM	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=355282
21	Smoke House Haven	5333 CAUSEWAY BLVD TAMP/	https://foodpermit.fdacs.gov/Visit/VisitList.aspx?id=428163

WORK TO BE DONE

1. **Scrape Data**
   a. **Determine how far back to scrape**
   b. **How to deal with PDFs**
   c. **Pagination, JavaScript rendering, token expiration, rate limiting, and nested data make large-scale scraping tricky.**
2. **Extracting Insights**
3. **Extending to live and searchable dashboard (hopefully)**

REFERENCES

- Dumne, R., Gavankar, N. L., Bokare, M. M., & Waghmare, V. N. (2024). Automatic Text Summarization using Text Rank Algorithm. 2024 3rd International Conference for Advancement in Technology (ICONAT), 1–6. https://doi.org/10.1109/ICONAT61936.2024.10775241
- Kang, J. S., Kuznetsova, P., Luca, M., & Choi, Y. (2013). Where not to eat? Improving public policy by predicting hygiene inspections using online reviews. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (pp. 1443-1448).
- Kannan, V., Shapiro, M. A., & Bilgic, M. (2019). Hindsight analysis of the Chicago food inspection forecasting model. arXiv.org.
- Mahmood, A., & Khan, H. U. (2019). Identification of critical factors for assessing the quality of restaurants using data mining approaches. Electronic Library, 37(6), 952–969. https://doi.org/10.1108/EL-12-2018-0241
- Mejia, J., Mankad, S., & Gopal, A. (2019). A for effort? Using the crowd to identify moral hazard in New York City restaurant hygiene inspections. Information Systems Research, 30(4), 1363–1386. https://doi.org/10.1287/isre.2019.0866
- Schomberg, J. P., Haimson, O. L., Hayes, G. R., & Anton-Culver, H. (2016). Supplementing public health inspection via social media. PLOS One, 11(3), e0152117–e0152117. https://doi.org/10.1371/journal.pone.0152117
- Wang, Z., Balasubramani, B. & Cruz, I (2017). Predictive analytics using text classification for restaurant inspections. Proceedings of the 3rd ACM SIGSPATIAL Workshop on Smart Cities and Urban Analytics. https://doi.org/10.1145/3152178.3152192