Parking and Camera Violations IR System

Group members: Boyuan Zhou, Congxiao Wang, Ziyiyang Wang

Introduction

This research is dedicated to the development of an IR system specifically designed to analyze and process comprehensive data on New York's parking tickets in order to direct drivers to the parking spaces where they are least likely to be in violation of the rules. The impact of solving this problem may reduce traffic congestion and enhance urban mobility. Concretely, we would like to locate the most appropriate parking locations within a well-defined geographic area based on the location information provided by queries.

IR Task Definition

When considering minimizing the risk of parking ticket violations, the IR task we plan to tackle will be defined as: analyzing historical and real-time updated parking ticket data from different spots in New York City to suggest the nearest parking spot with the least probability of violation and being ticketed to the driver. However, this IR system is not a real-time navigation system. It is more like a kind of query-answer format presentation. For example, a driver enters "112th Street and 2nd Avenue" and their personal driving information. The system will analyze parking ticket data around that location and combine it with the user's personal information to identify and recommend a series of possible parking locations with a relatively low probability of violation, in this case, the output might be “115th in front of Cherry Valley Marketplace.”

Regarding input information, this IR system will accept the following possible text inputs. Firstly, text-based location information can be submitted, as queries contain detailed location information. Considering multiple crossings and zones, it should support multiple text string inputs, such as "112th Street, 2nd Avenue", in order to precisely target the search area. The coordinates provided by the navigation system can also be entered into the IR system as a text string. For example, the input of the driver's personal information, such as plate expiration information, vehicle length, etc., to be used for further optimization of parking space recommendations. The system will output textual location information, which is recommended by the system as the parking location near the input location with the lowest probability of violation based on the analysis of past parking ticket data. As well as statistical information on past tickets for the recommended parking locations, such as the number of tickets at the location in the past year, common types of violations, and so on.

Data

We are harnessing a dataset detailing traffic and parking violations. This dataset is sourced from Data.gov, a platform promoting government transparency by offering open-access data. The Data.gov platform ensures versatile access to the data by offering it in multiple downloadable formats. We can obtain the dataset in either CSV or JSON formats.

The dataset is undeniably vast, boasting a total of 11,809,233 records. Each entry in this dataset represents a distinct violation. It provides comprehensive insights, such as the vehicle's registration details, the specifics of the violation, the location of the infraction, details about the issuing authority, and various vehicle attributes. Given the sheer volume, combined with some potential inaccuracies in time-related information, there's an impending need for rigorous data cleaning.

When utilizing a query tool to identify violation records based on specific parameters, such as vehicle details, violation type, location, or a particular time frame, it's imperative to gauge the relevance of each record to ensure the user's query is addressed accurately. In this scenario, determining the relevance score becomes paramount, and it can be derived from several pivotal factors:

- Match of the Query: The degree to which a record aligns with the user's input is crucial. For instance, if a user is specifically looking for "speed violations in New York in 2020," those records that precisely match this criterion will inherently possess a higher relevance score.
- Severity of Violations: The nature of the violation also plays a vital role in determining relevance. More grievous violations, given their significant implications, might be deemed of higher relevance compared to minor infractions. This means that certain violations, owing to their severity, might be prioritized and assigned a loftier relevance score.

Using these relevance scores, we can use traditional information retrieval evaluation metrics such as accuracy, recall, and F1 scores to evaluate the performance of our retrieval systems.

Related Work

Lightweight Deep Neural Network Approach for Parking Violation Detection: Utilizing an IP camera and processing images on a Raspberry Pi, they identified vehicles parked illegally. To aid enforcers, they introduced the "Enforcer App" for smartphones, which visually displays live violations. In tests, the system showcased a 98.7% precision rate.

Automated Traffic Rule Violation Detection with E-Challan Generation for Smart Societies: The research introduces an automated system for detecting traffic violations using Automated Number Plate Recognition (ANPR). The system captures vehicle images at intersections, and through image processing and neural networks, it identifies violations like signal breaches and illegal parking. After detecting a violation, the vehicle's license plate is recognized and an electronic challan (e-challan) is automatically sent to the vehicle owner's email. This system aims to enhance road safety and ensure adherence to traffic rules.

Full-text information retrieval: Further analysis and clarification: In the 1980s, the STAIRS study evaluated the efficacy of a full-text search and retrieval system, drawing criticism from pioneer Gerard Salton. The article underscores the challenges of full-text retrieval, especially in large databases, highlighting issues like linguistic variation, the unpredictability of keyword presence in relevant documents, and the subsequent "output overload."

Approaches to passage retrieval in full-text information systems: The paper investigates methods for retrieving specific portions of extensive texts, aiming to provide concentrated information and facilitate user overload. Using the SMART vector processing technology on an encyclopedia database, the research found that prioritizing subtext units like paragraphs over full texts enhanced recall.

Exploratory Data Analysis on Traffic Violations Issued in NYC: This is an exploratory data analysis (EDA) project focusing on parking violations in New York City. The project utilizes open government data and delves into the dataset's characteristics, challenges, and methodologies. The goal is to derive insights and patterns from the parking violation data.

The first two mentioned studies use technology to detect traffic violations. One focuses on a specific method for parking issues using IP cameras and Raspberry Pi, while the other has a broader approach with ANPR for various violations. Historically, research like the STAIRS study has focused on the challenges of full-text information retrieval, while other works used SMART vector processing for specific text passage retrieval. Separate studies, such as the EDA on NYC parking violations, utilized data analysis to understand violation patterns without advanced retrieval methods. Unlike these studies, our project aims to use sophisticated information retrieval specifically to assist users in checking penalty tickets, filling a unique niche. Our system is different;

it's designed to help users prevent violations and avoid fines, making it potentially more user-friendly and appealing.

Evaluation and Results

Evaluation Metrics:

- Precision at k (P@k):
    Definition: Measures the accuracy of the top-k retrieved parking violations.
    Importance: Users often search for recent or specific violations; hence it's crucial the most relevant results appear at the top.
- Recall at k (R@k):
    Definition: Gauges how many relevant violations appear in the top-k results.
    Importance: Ensures no critical violation or camera record is missed, especially if a user is verifying their compliance or understanding penalties.
- Mean Average Precision (MAP):
    Definition: Aggregates precision scores for all possible parking violation queries.
    Importance: An overall measure of the system's efficacy across multiple search scenarios, from plate numbers to precincts.
- Normalized Discounted Cumulative Gain (nDCG):
    Definition: Evaluates the ranking quality of the violations.
    Importance: Given that some violations might have graver consequences (higher fines, penalties, etc.), ensuring they rank higher can be pivotal for users.
- Click-Through Rate (CTR):
    Definition: Tracks which violation results users most engage with.
    Importance: Indirect feedback on the quality and relevance of results, especially for violations with associated summons images.
- User Satisfaction Surveys:
    Definition: Specific questions related to the ease of finding violations, clarity of summons images, and accuracy of violation details.
    Importance: Direct feedback to refine the system further based on user preferences and pain points.

Baseline Comparisons:

- Random Performance:
    Approach: Randomly display a parking violation or camera record.

Purpose: A foundational benchmark ensuring any refined search surpasses mere chance.

- Simple BM25:

  Approach: Use BM25 for common searches like plate numbers, precincts, or violation types.

  Purpose: Provides a base-level expectation of how a standard IR approach handles this specific dataset.
- Heuristic Baseline:

  Approach: Prioritize recent violations or those with higher fines.

  Purpose: Reflects a basic user inclination towards understanding recent infractions or more consequential ones.

Implementation and Testing:

- Test Set Creation:

  Design test queries mimicking real-world searches, such as "Violations for Plate XYZ in the last month" or "Camera violations in Precinct ABC."
- Quantitative Analysis:

  Compare results for queries like "Most common violation in County DEF" or "Average fines in State GHI" against known statistics.
- Qualitative Analysis:

  Engage users who've recently received violations to validate the system's accuracy and user-friendliness. Check the clarity and accessibility of summons images.
- Continuous Feedback Loop:

  Given the dataset's evolving nature (with new violations added weekly), regularly collect feedback to ensure the system remains updated and accurate.

Work Plan

What We've Done So Far:

Initial Problem Definition and Understanding

Recognized the need for an IR system specific to parking and camera violations.

Got acquainted with the dataset and its attributes.

Work Plan for the 9 Weeks Ahead:

Week 1: Data Preprocessing and Initial System Design

Clean the dataset, address missing values, and standardize formats.
Outline the system's basic design and features.

Week 2-3: Building the Search Functionality

Implement keyword-based search for columns like Plate, State, Violation, etc.
Set up a simple BM25-based baseline system.

Week 4: User Interface Design and Development

Draft a user-friendly UI/UX design.
Start frontend development using the chosen technology (e.g., React).

Week 5: Advanced Search, Analytics, and Backend Development

Incorporate date ranges, financial ranges, and filters.
Initiate backend integration for analytics and advanced search features.

Week 6: Visual Analytics and Summons Image Retrieval Integration

Integrate geospatial visualizations based on precincts and counties.
Embed functionalities for summons image retrieval and display.

Week 7: System Testing and Baseline Comparisons

Execute predefined queries on the system and gauge against baselines.
Address and rectify any identified bugs or issues.

Week 8: User Feedback Collection and System Refinements

Conduct user testing sessions for feedback.
Iteratively refine the system based on gathered insights.

Week 9: Finalization, Documentation, and Presentation Preparation

Complete any lingering development tasks and finalize the system.
Craft documentation detailing system usage and potential enhancements.
Prepare a comprehensive presentation showcasing the system's journey, capabilities, and value.

Relevance

Author links open overlay panelDavid C. Blair, 1985, A., Inc., A. D. L., Berring, R., Blair, D. C., Cleverdon, C. W., Dabney, D. P., Flores, F., Lancaster, F. W., Langendoen, D., Malone, T., & Maron, M. E. (2002, July 19). *Full-text information retrieval: Further analysis and clarification*. Information Processing & Management. https://www.sciencedirect.com/science/article/pii/0306457390901028

City of New York - Organizations - Catalog, catalog.data.gov/organization/city-of-new-york. https://catalog.data.gov/organization/city-of-new-york?q=parking+violation&sort=score+desc%2C+metadata_modified+desc Accessed 11 Oct. 2023.

Martin, J. D. (2021, July 18). *NYC parking 2017 final - notebook by Jonathan David Martin (jonathanmartinnyc)*. Jovian. https://jovian.com/jonathanmartinnyc/nyc-parking-2017-final

Thombre, S. S., Wazalwar, S. S., Tarekar, R., Ramgirwar, S., Zulkanthiwar, S., & Dubey, S. (1970, January 1). *Automated traffic rule violation detection with E-challan generation for smart societies*. SpringerLink. https://link.springer.com/chapter/10.1007/978-981-19-1689-2_5

University, C.-K. N. M., Ng, C.-K., University, M., University, S.-N. C. M., Cheong, S.-N., University, Y.-L. F. M., Foo, Y.-L., & Metrics, O. M. A. (2018, December 1). *Lightweight deep neural network approach for parking violation detection: Proceedings of the 2018 VII international conference on network, communication and computing*. ACM Other conferences. https://dl.acm.org/doi/abs/10.1145/3301326.3301387

Univ., G. S. C., Salton, G., Univ., C., Allan, J., Profile, V., Buckley, C., & Metrics, O. M. A. (1993, July 1). *Approaches to passage retrieval in full text information systems: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM Conferences. https://dl.acm.org/doi/abs/10.1145/160688.160693