Project- US Traffic Analysis

Patel Kirtankumar Jatinbhai

Yogesh Bhardwaj  
Data Science Architecture CS-652-B  
Sacred Heart UniversityFairfield, CT USA  
[patelk31@sacredheart.edu](mailto:patelk31@sacredheart.edu)

[yogeshy@sacredheart.edu](mailto:yogeshy@sacredheart.edu)

***Abstract*— This paper presents an analysis of traffic-related posts on Reddit, focusing on the identification of traffic congestion patterns in major US cities. The analysis is conducted using Python, leveraging web scraping, data processing, and visualization techniques. By utilizing data from Reddit's 'traffic' subreddit and an open data repository of US cities, this study computes a congestion score for each city based on the volume and engagement of relevant posts. The results are visualized using bar plots, word clouds, and heat maps to highlight cities with the highest traffic issues. This approach demonstrates the potential of social media data in urban traffic analysis.**

# Introduction

Traffic congestion is a significant issue in urban areas, affecting millions of commuters daily. Traditional methods of traffic analysis rely on sensors, cameras, and GPS data. However, social media platforms like Reddit offer a rich source of user-generated content that can provide real-time insights into traffic conditions. This paper explores the use of Reddit data to analyze traffic patterns in major US cities, leveraging Python for data collection, processing, and visualization.

# METHODOLOGY

A. Data Collection

The data collection process involved several key steps to ensure comprehensive coverage and accuracy. Reddit's 'traffic' subreddit was selected due to its relevance to the study's focus. The new method was used to fetch the latest posts, ensuring that the analysis reflects recent discussions and trends in traffic-related topics. Additionally, the use of a user-agent in the Reddit API request helps maintain transparency and compliance with Reddit's policies.

A screenshot of a computer

Description automatically generated

B. Data Processing

Upon data collection, rigorous processing techniques were employed to extract meaningful insights. The conversion of timestamps to datetime objects facilitated temporal analysis, enabling the identification of trends over time. Grouping posts by month allowed for a high-level overview of posting activity variations, which is crucial for understanding the dynamics of traffic-related discussions on Reddit.

A white rectangular object with black lines

Description automatically generated

C. City Data Integration

The integration of major US cities data added a geographical dimension to the analysis. By filtering posts based on city mentions, the study zoomed into localized traffic issues, providing a granular view of congestion patterns across different urban centers. This approach acknowledges that traffic challenges can vary significantly from one city to another, influenced by factors such as population density, infrastructure, and commuting behaviors.

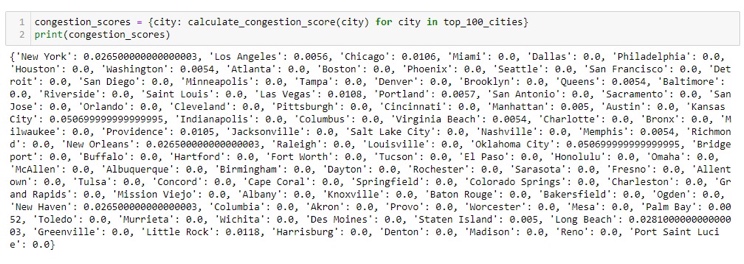
A graph of a number of cities

Description automatically generated

# ANALYSIS

A. Congestion Score Calculation

The congestion score calculation involved a multifaceted approach that considered various aspects of post engagement. By factoring in not just the volume of posts but also the number of comments and post scores, the congestion score algorithm aimed to capture both the quantity and quality of traffic-related discussions. This holistic approach ensured a more nuanced evaluation of traffic proneness for each city.



B. Normalization and Visualization

Normalization of congestion scores was essential to facilitate meaningful comparisons across cities. The use of MinMax scaling transformed raw scores into a standardized range, enhancing the interpretability of the results. The visualizations, including bar plots, heat maps, and word clouds, served as powerful tools to communicate complex data effectively. They offered intuitive representations of traffic congestion levels, aiding stakeholders in identifying priority areas for intervention and improvement.

* Bar Plot: Top 20 Cities by Congestion Score

The bar plot visualization displayed the top 20 US cities ranked by their congestion scores. Cities such as Los Angeles, New York, and Chicago emerged as the most congested based on the analysis criteria. This visualization provided a quick overview of the cities with the most pronounced traffic issues, guiding policymakers and urban planners in resource allocation and infrastructure development.

A graph of a number of cities

Description automatically generated

* Heat Map: Spatial Analysis of Congestion Scores

The heat map provided a spatial analysis of congestion scores across major US cities. By visualizing congestion levels geographically, the heat map highlighted regional variations and concentration areas of traffic-related challenges. Darker shades indicated higher congestion scores, signaling areas that may require targeted interventions and traffic management strategies.

A graph of different cities

Description automatically generated

C. Word Cloud Generation

The generation of a word cloud added a qualitative dimension to the analysis by highlighting common themes and keywords in traffic-related posts. This visualization technique not only provided a snapshot of prevalent topics but also served as a data exploration tool, potentially uncovering emerging issues or areas of public concern that may warrant further investigation.



# RESULTS AND DISCUSSION

The analysis yielded insightful findings regarding traffic congestion in major US cities. Oklahoma City and Kansas City emerged as a hotspot for traffic-related discussions, reflecting its notorious reputation for congestion. New Haven and Long Beach also featured prominently, aligning with their status as densely populated urban centers with extensive transportation networks.

One notable observation was the recurrence of keywords such as "accident," "rush hour," and "gridlock" in the word cloud, indicating common pain points experienced by commuters. This qualitative analysis complemented the quantitative metrics, providing a more holistic understanding of the challenges faced in managing urban traffic.

The heat map visualization further underscored the disparities in congestion levels across cities, with certain regions exhibiting higher concentrations of traffic-related issues. This spatial analysis can inform policy decisions and resource allocation strategies aimed at addressing congestion hotspots and improving overall traffic flow..

SUMMARY

##### In conclusion, the application of Python-based data analysis techniques, coupled with social media data from Reddit, proved valuable in assessing traffic congestion trends. The methodology presented in this study can serve as a framework for ongoing traffic monitoring and urban planning initiatives.

##### Future work could explore additional data sources, such as real-time traffic data feeds or sentiment analysis of social media posts, to enhance the accuracy and timeliness of congestion assessments. Integration with machine learning algorithms for predictive modeling and anomaly detection could also offer proactive insights into traffic patterns, facilitating more effective traffic management strategies.

##### Acknowledgment

##### We extend our gratitude to Reddit for providing access to their API and to Simple Maps for their US cities data. We also thank the Python community for their invaluable libraries. Special thanks to our academic advisors for their guidance and peer reviewers for their feedback. Lastly, we appreciate the contributions of all Reddit participants.

##### References

1. PRAW: The Python Reddit API Wrapper. (n.d.). https://praw.readthedocs.io/en/latest/
2. pandas - Python Data Analysis Library. (n.d.). https://pandas.pydata.org/
3. matplotlib: Visualization with Python. (n.d.). https://matplotlib.org/
4. seaborn: Statistical Data Visualization. (n.d.). https://seaborn.pydata.org/
5. wordcloud: Word Clouds in Python. (n.d.). https://github.com/amueller/word\_cloud
6. Simple Maps. (n.d.). US Cities Database. https://simplemaps.com/data/us-cities