Datasheet for Dataset: Augmented Police Incident Reports

**Dataset Information**

**Dataset Name:**

Augmented Police Incident Reports

**Description:**

This dataset contains augmented data extracted from public police department PDF reports. The augmentation includes additional derived attributes like the day of the week, time of day, weather conditions, location rank, side of town, incident rank, nature of the incident, and EMS status.

**Version:**

1.0

**Creation Date:**

April 6, 2024

**Last Updated:**

April 6, 2024

**Data Source:**

Public police department PDF reports.

**Data Augmentation Details:**

The data augmentation process includes extracting URLs from PDF files, deriving additional attributes such as day of the week, time of day, weather conditions (using WMO CODE), location and incident frequency ranking, side of town (using geospatial analysis), nature of the incident, and EMS status. The weather data is fetched using the Open-Meteo API based on the incident's time and location.

**Dataset Creators**

**Creators:**

Kushagra Sikka

**Contact Information:**

kushagrasikka@ufl.edu

+1 352-740-6029

**Data and File Overview**

**Data Format:**

The dataset is stored in CSV format.

**File Structure:**

* Day of the Week: Integer (1-7, where 1 corresponds to Sunday)
* Time of Day: Integer (0-23 hours)
* Weather: Integer (WMO weather code)
* Location Rank: Integer (Frequency ranking of incident locations)
* Side of Town: String (N, S, E, W, NW, NE, SW, SE)
* Incident Rank: Integer (Frequency ranking of incident types)
* Nature: String (Description of the incident)
* EMSSTAT: Boolean Integer (1 for presence of EMS, 0 otherwise)

**Number of Records:**

Can be however many based on the input csv file

**Data Fields Description:**

* **Day of the Week:** Encoded as integers representing each day of the week.
* **Time of Day:** Hour of the day the incident was reported.
* **Weather:** Weather conditions at the time and location of the incident, represented by WMO weather codes.
* **Location Rank:** Ranking of the frequency of incidents at the location.
* **Side of Town:** Categorical representation of the incident location's approximate orientation from the town center.
* **Incident Rank:** Ranking of the frequency of the nature of incidents.
* **Nature:** Text description of the nature of the incident.
* **EMSSTAT:** Indicator of EMS involvement in the incident.

**Data Collection Method**

The data is collected from publicly available police department PDF reports. URLs to these reports are stored in a CSV file. A Python script automates the download of these PDFs, extracts incident data, and performs data augmentation.

**Data Preprocessing**

The preprocessing steps include:

* PDF text extraction.
* Data cleaning to remove irrelevant information.
* Geospatial analysis to determine the side of town.
* Fetching weather conditions using the Open-Meteo API.
* Ranking of locations and incident natures based on their frequency.

**Data Augmentation**

Augmentation involves adding derived information to the dataset to enhance its usefulness for analysis. This includes adding weather conditions, location ranks, and side of town information based on geospatial analysis.

A screenshot of a computer

Description automatically generated

**References**

* Open-Meteo API Documentation: [Open-Meteo Docs](https://open-meteo.com/en/docs)
* Google Cloud geocoding API