DSCI 551 Proposal Report -- Los Angeles traffic collision and weather

• Project Members

Xinrui Ying: mscs + data science; Software Engineering; Python, SQL
Mingliao Xu: Applied Data Science Program. Statistics Background; Familiar Program Skills: R, SAS, Java, C, Python
Houqingchen Zhu: Applied Data Science Program. Computer Science Background; Python, Java, Android development

• Project Description

In this project, we would like to analyze Los Angeles city weather conditions and traffic collision incidents and try to find if there's a relationship between these two data. This project's goal is to build an interactive User Interface in order to present certain weather conditions given dates and to obtain traffic collision data with further provided location information. Basically, users should be able to use keywords like dates, zip code or street address name to search for weather conditions and happened traffic collisions so that they could be forewarned.

• Data sets to be used

1. Los Angeles traffic collision data retrieved from Kaggle:
https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data?select=traffic-collision-data-from-2010-to-present.csv

Our traffic collision datasets contain the information about the location and area name, and have features such as date reported, reporting district, date occurred, time occurred and crime description, etc.
This dataset retrieved from Kaggle is 174 Megabytes, which reflects 10 years of Los Angeles traffic collision incidents.

2. Weather conditions retrieved from Kaggle:
https://www.kaggle.com/selfishgene/historical-hourly-weather-data?select=weather_description.csv

This dataset is 21 Megabytes, which includes the weather description for Los Angeles from 2012-2017.

• Data problems to be addressed

Data Storage: Store in mySQL and Firebase after processing our datasets
Data Cleaning: Clean the second dataset so that only LA city weather data remains
Data Transformation: Transform Weather Dataset into json format.
Data Integration: Integrate weather data and traffic collision data according to date and location
Data Aggregation: Summarize the weather data and traffic data so that users can be forewarned to drive carefully at certain times and weather conditions

Parallel Data Processing: Use Spark to perform data cleaning and data aggregation

• Databases usage

We will first use python to pre-process our dataset. We will perform data cleaning, aggregation and integration and then store the second dataset in JSON format. Then we will store the processed datasets into mySQL and Firebase respectively. We will use Spark to do data, combining timelines. Finally, We will do data visualization through a webpage.

• Milestones and timelines

Time	Task	Detailed Task
Week 5	Data preprocessing	Preprocess both weather and traffic collision data and conduct data cleaning.
Week 7	Modeling and Analysis	Build data models for both datasets and find relationships between them.
Week 10	Web development and debugging	Build a website which enables users to view the data and searches for historical traffic collisions.
Week 12	Advance functions implementation	Provide active and vivid data visualization that can change with time slider.
Week 13	Summarize	Summarize the project and prepare the final report.