- AI & Robotics –
Visite São Paulo sem o Primeiro Comando da Capital

Assignment

Tim Dupont

<tim.dupont@pxl.be>

Sam Van Rijn

<sam.vanrijn@pxl.be>

May 2019

Project description

This is the assignment of the Group Project for the AI & Robotics course. You will, as a group of approximately five, be processing two large datasets for this assignment. One data set contains weather information of São Paulo. The other bulk of information stores the Crime information of the same city.

Datasets

	Weather	**Crime**
Daterange	2000 - 2016	2007 - 2016
# Records	~10 million	~15 million
Size (MB)	~2000MB	~8000MB
Link	https://www.kaggle.com/PROPPG-PPG/hourly-weather-surface-brazil-southeast-region	https://www.kaggle.com/inquisitivecrow/crime-data-in-brazil

Data set 1: Weather

Context

It's covers hourly weather data from 122 weathers stations of southeast region (Brazil). The southeast include the states of Rio de Janeiro, São Paulo, Minas Gerais e Espirito Santo.

Dataset Source: INMET (National Meteorological Institute - Brazil).

Equipment: Vaisala Automatic Weather Station AWS310

Category: Weather

Data set 2: Crime

Context

Brazil has a very powerful Freedom of Information law which allows any citizen to request any data from the government which is not restricted, and where these restrictions are well defined exceptions. But still, having the right to request the information does not mean it is easy to get it. Bureaucracy and ignorance of the law gets in the way many times. In order to encourage the government to put their databases in order and to inspire people to have the courage to ask the government for information, we made a massive request of information, for all the complete dataset of crime data available for the last 10 years, in the biggest city of South America.

Content

This dataset contains structured data about all crime occurrences that have been acted upon by the PM, the main police force in Sao Paulo. The dataset is not consistent in its completeness, as some of the towns comprising the Greater Sao Paulo were slow in collecting full data. It also does not contain the actual historic of each crime report, as that would violate privacy.

Acknowledgements

We would like to acknowledge the prompt assistance from the SSP (Secretaria de Seguranca Publica), for providing the data with minimal resistance.

Inspiration

Primarily we would like to see a visualisation of this data, so that the people can have an idea of how crime has evolved in their city, which crimes are more prevalent in which areas, etc. In addition, any model which can predict at what times and where the police is most needed would be helpful, as this can then be sent to the SSP to help them in planning.

Important to remember here is that we're dealing with large datasets. It is your job to figure out how to deal with this large dataset appropriately. We strongly recommend using Google Colab for this assignment.

Tools

Google Colab

Google Colab gives you access to Google's Cloud Machine Learning platform, including their powerful GPUs and TPUs. Colab allows you up to 12h of CPU, GPU or TPU computation time **for free** every day.

To get up and running in Google Colab, surf to https://colab.research.google.com and sign in with a Google/GMail Account.

Follow the steps in the following guides:

- Importing libraries:
  - https://colab.research.google.com/notebooks/snippets/importing_libraries.ipynb
  - `!pip install fastai==0.7.0`
  - `!pip install kaggle`

- Setting up the Kaggle API:
  - https://medium.com/robotics-club-sastra/working-with-kaggle-dataset-on-google-colab-d3d4a77ead62
  - https://medium.com/@move37timm/using-kaggle-api-for-google-colaboratory-d18645f93648
  - https://github.com/Kaggle/kaggle-api

If you get an error trying to use the Kaggle API, you might have to execute the following commands:

```
!mkdir ~/.kaggle
!cp /content/.kaggle/kaggle.json ~/.kaggle/kaggle.json
```

Issue description and solution: https://github.com/Kaggle/kaggle-api/issues/97

Other tools

For the rest pandas, numpy, scikit-learn and/or fastai are recommended to use for this assignment.

Tasks and Questions

Data processing

The first task is to try to get relevant insights out of the datasets

I.E. for the weather data set:

- Temperature data
  - Monthly/yearly average
  - Min temperature, max temperature
- Precipitation data
  - Monthly/yearly average
  - Min precipitation, max precipitation
- Draught periods

        - Summers
        - Years
    - Rain periods
    - When is it optimal to visit this city?
    - ...

Crime dataset:

- What regions of the city are worst for violent crimes?
- What regions of the city are worst for nonviolent crimes?
- Has there been an increase or decrease in violent crime throughout the years?
- Has there been an increase or decrease in nonviolent crime throughout the years?
- Which populations are most likely to be victims?
- Which populations are most likely to be perpetrators?
- ...

You don't have to limit yourself to these questions, these are just a few suggestions.

Data visualization

Generate meaningful charts and visualizations for these insights.
Find inspiration at:
http://www.bom.gov.au/climate/change/index.shtml#tabs=Tracker&tracker=timeseries
https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Crime_statistics

Prediction and exploration

1. Generate predictions for the coming year using techniques from the course (Random Forests, Neural Networks, Clustering)
    a. about temperature and precipitation
    b. About crimes happening in different parts of the city
2. Optimize and compare your models
3. Find correlations in the data (i.e. links between weather and crime?)

Website and recommendation

- Use Flask to upload your model (like in the spring break assignment)
- Display the visualizations (i.e. heatmaps and charts) about crime in the city
- Display the visualizations (i.e. temperature and precipitation graphs) in the city

Use the data acquired in the previous steps to make a recommendation to a tourist, wishing to visit the colorful city of São Paulo, Brazil.

- When is the best time to visit the city?
    - When the temperature is not extreme
    - Lowest amount of precipitation
    - ...
- What are the parts of the city to avoid as a tourist?

The tourist should supply some basic information like Age, Gender and Race.

Keep the website simple. It doesn’t need to be ready for deployment in a business context, but it will have to display your data analysis in a clear and comprehensible manner.

Good luck!!!

(and see you on a sunny beach in São Paulo sometime in the future)