CS2006: Python 2 Report

Data Analysis

Analysing Crawdad's Taxicrabs

120010055, 120010090 & 120010815

December 2, 2013

Contents

1 Summary of Functionality

1.1 Requirements Met

We believe our program meets all of the basic requirements specified for this practical. Our program can successfully load, parse and interpret the *Cabspotting* dataset.

The program can calculate five basic statistics on the dataset. I.e. the total number of unique cabs, the number of cabs over time (per hour), the total distance travelled by each cab, the average speed of each cab and the efficiency of each cab. The efficiency of a cab is the percentage of time that it had a passenger.

The program can parse subsets of the dataset, all of the above statistitcs can be calculated on these subsets. Our program can parse subsets in two different ways. Either a start and end time in ISO 8601 format can be provided, or list of cab ids to indicate particular users.

The program can take command line arguments to indicate the name of the dataset to be processed or the name of the starting file that contains the list of ids in the dataset.

Our program can parse the metadata file about a particular dataset and display it to the user. There is also an option to download the metadata file.

We have explored and answered a question, this is addressed in the attached file.

Our code works on the machines in the Mac Lab.

Our code is PEP 8 compliant

1.2 Extensions Attempted

We attempted to do some social networking analysis of the dataset. Our program can display the degree centrality of the network.

2 Problems Encountered

2.1 Storing the Dataset

Our first attempt at storing the dataset was to load the entire dataset into RAM. We quickly discovered that was not an ideal solution as our program would slow down and eventually crash when attempting to load the entire cabmobility dataset.

Our solution was to calculate the statistics we need as we loaded the dataset then only store those. Because of this our program can easily handle even larger datasets than the one provided with only a linear increase in loading time.

2.2 Taking a Subset

Because we calculate all our statistics as we load the data, taking a subset of the data requires us to reload the entire dataset. This can be tedious for large datasets and so is probably not the best solution to this requirement. But without massively restructuring our code it is impossible to avoid.

2.3 Social Networking Analysis

The first problem with completing the social networking analysis was that we need a network to analyse while we are given a list of cabs and their routes through the city. We decided to analyse the network of locations that passengers travel to. This means that only the locations where passengers leave a cab are in our graph.

The second problem was that our network of locations is perhaps not suitable for social networking analysis. To calculate the closeness centrality of a network for example you must know which nodes pass through each other. But in our network passing through a node to get to another node is meaningless. I think that we misunderstood what was meant by ”the social network of the nodes in a dataset”

2.4 GUI

We had to make a decision on how to display the graphs produced for the networking analysis extension. Originally we used google maps to plot points. But between having to generate kml files, uploading these to dropbox and then displaying them on google maps to check if they were correct this became very tedious very quickly. I decided to copy an image of San Franciso from google maps, then use this as a background image for a Tkinter interface. Then use Tkinter to draw the graph onto this image. One issue with this is that the image is not interactive, so points outside of the image are not drawn. I used an image that would cover the majority of the locations, but some key locations like the airport are not shown. To fit the airport in I would have had to zoomed out a lot and the image would not have been readable.

2.5 Creating charts

Since no one of us has worked with data visualisation in Python previously, we had to learn how to do it on the go. We chose to use the library ReportLab because it seemed quite easy to understand it. However, as we worked on the charts, we discovered that it was extremely difficult to find any documentation or tutorials, or code examples. Because of that, we spent quite a lot of time trying to figure everything out, and composed our part of the program which creates charts piece by piece.

3 Summary of provenance

3.1 What we wrote

With the exception of our chart creation code all of the code is our own work. We used the build in module Tkinter to create our graphical output.

3.2 What we modified

Since we never worked with charts in Python, it was useful to find some code examples in all sorts of external sources and use some parts of it in our program. The code was never blindly taken from the source, but it was studied sthough and understood thoroughly.

3.3 What we sourced from elsewhere

We used the reportlab module to generate charts to answer our question. The license for this module is included with the source.

4 Testing

- We tested that our program would not crash if an invalid command was entered into the console.
- We tested that the statistics, graphs and charts produced alligned with the datasets. To do this we run our code on a small subset of the dataset so that we could more accurately predict the correct results.

5 My Contribution to the Project

5.1 120010055

I wrote the code that loads and calculates basic statistics on the dataset. I also wrote the code that visualizes graphs of the data.

5.2 120010090

Wherever we used branches or created multiple heads in the repository - either deliberatey or accidentally - I was responsible for most of the merging and conflict resolution. I also worked on the XML Metadata, including the extension of allowing downloading of the dataset from the metadata. I did large portions of the commentry for the code, as this is a necessary feature for debugging and testing the program - something which I also worked on in the later stages of the project.

5.3 120010815

I wrote the code that finds answers on different questions based on the data that is stored. I also worked with the ReportLab library and created some of the charts that show the results of the data analysis. I also commented part of the code and wrote the small report in IMRAD style which shows how we used the data for the analysis.