Identifying Crime Hotspots using Twitter

By Redmond Todd-Bennett
Student Number: c1124856

CM3203- One-Semester Individual Project
Module Credits: 40
Supervisor: Dr Peter Burnap
Moderator: Dr Xianfang Sun

School of Computer Science and Informatics
Cardiff University
May 2015

Abstract

Data is being created all the time leading to the development of big data sets such as social media platforms like Twitter. By mining big data sets it is possible to extract useful patterns and trends. The main focus of this project is to analyse Twitter and identify references to crime or insecurity in the plain text of tweets then be able to visualise hotspots of crime onto a map by leveraging the location metadata attached to these tweets.

A software environment has been developed to plot locations onto a map to provide a geospatial distribution of crime by crime type. A further map to plot instances of actual crime sourced from the Police Application Program Interface (API) has been implemented to provide a real visual comparison. To understand whether there is value in analysing social media data in relation to crime, a Pearson chi-squared statistical test has been carried out looking for correlation between the number of crime references on Twitter and actual crimes for the spatial area covering the Metropolitan area of London. The investigation concludes that there is no correlation between the two data sets based on modelling the crime-related references which have been identified from Twitter against the number of actual crimes. The web-based software developed for this project was created using HTML, CSS and JavaScript programming languages incorporating the JavaScript-based Google Maps API as a mapping tool.

Acknowledgements

I would like to acknowledge the valuable support and mentoring from my project supervisor, Dr Peter Burnap throughout the course of this project. The provision of Twitter data and other useful resources from Peter has been vital to the development of this project. A further acknowledgement must go to my family for supporting me during the course of my Computer Science degree programme at Cardiff University.

Contents

Table of Figures

1 Introduction

Social media is a growing presence in society and business. Facebook, Twitter and Google+ are all examples of social media platforms on the internet. Widespread use is now possible due to the 'Internet of Things'. The internet of things refers to the "ever-growing network of physical objects [which have] internet connectivity and the communication that occurs between these objects" [1]. This means more people can connect to social media at any time and in any place using the plethora of connected devices available. The high volume of interaction with social media has created 'Big Data' with each social media platform needing to be able to store all of the data its users create. Big Data is an across-the-board term to describe "a massive volume of…data that is so large it is difficult to process using traditional database and software techniques" [2]. This project will be analysing a big data set in the form of social media platform, Twitter to see what value can be discovered in this area of focus, crime.

This project will analyse the natural language of tweets for specific crime-related terms or phrases. Then by leveraging the location metadata which is linked with a tweet, a geospatial distribution of these crime references will be plotted to show where there are particular hotspots of crime. In order to do this a web-based software environment will be required using Google Maps API to plot locations of such tweets. Actual crimes will be plotted onto a separate map by sourcing data from the Police API which archives monthly crime by location reports. To see whether there is a correlation between actual reported crime and those references of crime from Twitter some kind of statistical test will be conducted.

The key aims and scope for this project are:

- Identify specific crime-related terms or hashtags in Twitter data
- Develop system to visualise the distribution of these Twitter criminal hotspots by crime type onto a map using Google Maps API
- Visualisation of particular hotspots of criminal activity based on successful implementation of the above aims
- Plot actual reported crimes (sourced from Police API) onto a similar map in the software environment
- Undertake a statistical analysis to see if there is a correlation between actual reported crimes and the crimes identified from Twitter data

The key outcomes of this project:

- Source code which evaluates Twitter data set for keywords related to crime for each crime type by pattern matching
- Web-based Software which shows comparison between actual crime and references to crime on Twitter as a visualisation for each crime type
- Google Maps API implemented as primary mapping tool within the software environment
- Complete an appropriate statistical test and analyse results to check for any correlation between the two data sets for each of the crime types

2 Background

2.1 Related Work

Value is very important when it comes to 'Big Data' so it important to understand what we want to gain from analysing a big data set. Social media platform Twitter is a perfect example of a big data set where the volume of data is constantly increasing but potentially useful and interesting patterns can be extracted from the data which it stores.

This project will specifically focus on references to crime on Twitter. Much work has already been carried out in this area particularly to predict crime happening. An investigation by Xiaofeng Wang et al into the automatic prediction of crime using events from Twitter posts [3] examines the semantics and natural language of tweets to predict future hit-and-run crimes. This is an important study because it can help us to understand what specific keywords and phrases relating to, in this case hit-and-run crimes are regularly used on Twitter. This can support predicting future hit-and-run crimes and perhaps give us a wider perspective of potential crime-related references which users are articulating on Twitter for a larger set of crime types.

Further to this, Nick Malleson and Martin Andresen [4] have carried out a study to see the impact of social media in crime rate calculations. They have tried to investigate using two local spatial statistical methods, Getis-Ord GI* and the Geographical Analysis Machine to see how crime hotspots change when looking purely at the volume of social media data. This work looks particularly at how crime prevention methods focus on residential population as opposed to the 'mobile' population and how modelling using crowd sourced data from Twitter to represent the population 'on the move' can cause shifts in criminal hotspots.

A particularly relevant and ongoing piece of similar work is the COSMOS project which refers to the 'Collaborative Online Social Media Observatory' of which Cardiff University is a co-researcher. As part of this research a COSMOS software environment has been created which "reduces the technical and methodological barriers to accessing and analysing social media" [5]. The platform will harvest data from social media and analyse natural language of a social media post or tweet while adapting to a particular user's focus i.e. in the case of this project looking for crime references on social media. The platform also captures the metadata which is linked with the main data allowing the potential geolocation of tweets to be plotted amongst other things.

Some research on crime prediction distinct from social media includes spatio-temporal analysis carried out on crime scenes in a German City to discover hotspots of crime. Using a GIS-based application, the study visualises these hotspots by "[integrating the findings] into a geo-virtual environment" [6]. This paper focusses particularly on the visualisation side in contrast to the other related work discussed in this section which is primarily analysing the natural language and semantics of tweets to understand how particular topics are being referenced on social media.

The next step my project will take is to develop a method to identifying specific crime-related references from Twitter in conjunction with being able to visualise the locations of these references in a tangible fashion. The existing solutions in this area concentrate more on the semantics of the language relating to crime whereas this project is more concerned with creating a method which can evaluate the crime-related data into a usable format. Then using the Google Maps API the software will plot a geospatial distribution so we can see where particular hotspots of crime are based on the specific terms identified in the set of Twitter data. Finally, by comparing with actual reported crime we can see if there is some kind of relationship or correlation with crime-related references on Twitter.

2.2 Potential Uses and Stakeholder Identification

The Police are potential stakeholders to this project. It may be helpful for them to see visualisation of crime hotspots to see whether Twitter or indeed any other social media platform could be a useful tool to exploit. Seeing particular crime hotspots may support them to allocate their resources more effectively or predict future crime. The statistical analysis which will be carried out to see if actual reported crime correlates with that of crime references on Twitter could be potentially beneficial to see if there would be value for them in analysing social media.

COSMOS is an ongoing research project which may also benefit from the outcome of this project. In particular, the visualisation software environment delivered as part of this project's fulfilment is a potential bolt-on to COSMOS so the data being harvested can be visualised and plotted to see the actual location or a heat-map of data points where appropriate. This could be applied in a number of different areas not just to the main focus of this project, crime.

2.3 Project Constraints

Twitter as a big data set can be particularly challenging to analyse due to the huge volume of tweets per day. According to Twitter themselves "500 million Tweets are sent per day" [7] which shows the sheer volume of tweets which could be analysed.

Of course, this project is only across one University semester and therefore time is a major constraint. This means that data has been sourced from the Cardiff School of Computer Science and Informatics. To further filter the data set in general the focus will be on the London Metropolitan area as a spatial case study although initially work will begin with tweets with a London location which are within a certain timeframe i.e. one month's worth of tweets. For testing purposes it is more feasible to work with a smaller data such as this one so that during implementation the processing time is much quicker. It would be inefficient to work with a data set which takes a long time to evaluate for crime-related tweets in addition to further processing time for plotting the geospatial distribution using Google Maps API.

As the project continues a larger data set will be evaluated for specific crime-related references because the data will be for a number of months. As the volume of data increases there will be a lengthier time required to process it and I must be mindful of this ensuring this is taken into account when working in line with my initial project plan.

The data set being used will be static data so a dynamic, current visualisation of crime references on Twitter will not be provided. Taking this project further it would be useful to have a live feed of Twitter data which could be analysed for crime-related references and placed into a visualisation to see live hotspots of crime.

3 Specification and Design

3.1 Software Specification Overview

The aim for the software which will support my efforts to solve the problem which this project poses is to be functional and easily operated. With the additional requirement of needing a mapping tool to plot a geospatial distribution of crime related references from Twitter I settled on a web-based system because I would need the use of the Google Maps API and its heat map visualisation layer. A website will not only return a tangible visual output to help see how much progress I am making on the software solution but is also very flexible with many web programming languages available to use. Furthermore there are many free, open-source libraries which can easily be leveraged to support the development of this system.

The data I have sourced from the Cardiff School of Computer Science and Informatics contains tweets across a particular spatial area (London) and time period (January to April 2014). I will use Java code to search for tweets with specific crime references. These terms will need to be defined by a user in the script itself. Returned would be a standard Comma Separated Values (CSV) file with timestamp, Latitude, Longitude, London Borough and then the plain text of the tweet itself. These values are always returned in same column order meaning the code will not have to be very flexible when referencing the CSV files. The Latitude and Longitude values, in particular, will be required to plot a geospatial distribution onto a map. Additionally, the Police API will be used to pull data concerning actual reported crimes which will also be visualised on a map to see a comparison between that and crime references on Twitter. This will aid me when trying to perform the statistical analysis to test if there is any correlation between the two in a month-by-month format for the timeframe which the data covers. If it is possible I would like to split the data by crime type so we can view hotspots of particular crime categories in addition to crime all-up. A common crime type framework to follow will be required in order to provide a relational link between the crime referenced on Twitter and actual crime.

3.2 Crime References on Twitter

3.2.1 Twitter Data Flow

The below diagram details how the data sourced from Twitter will flow in the proposed system.

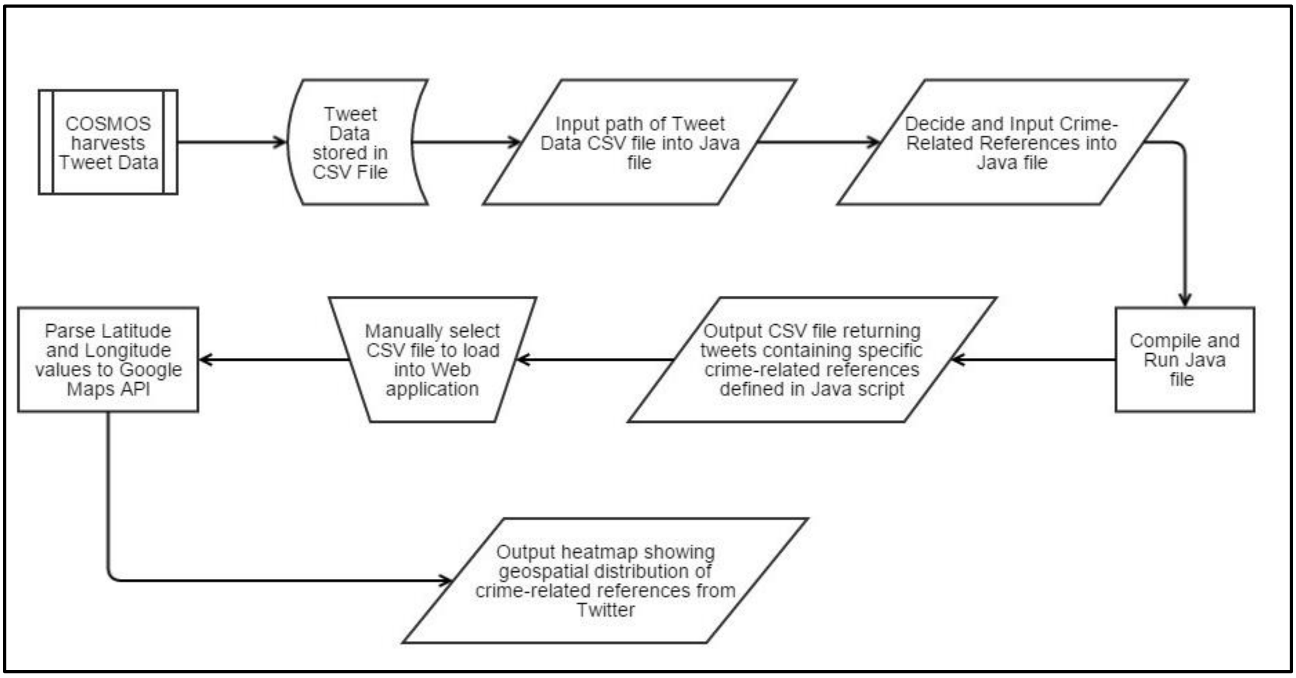
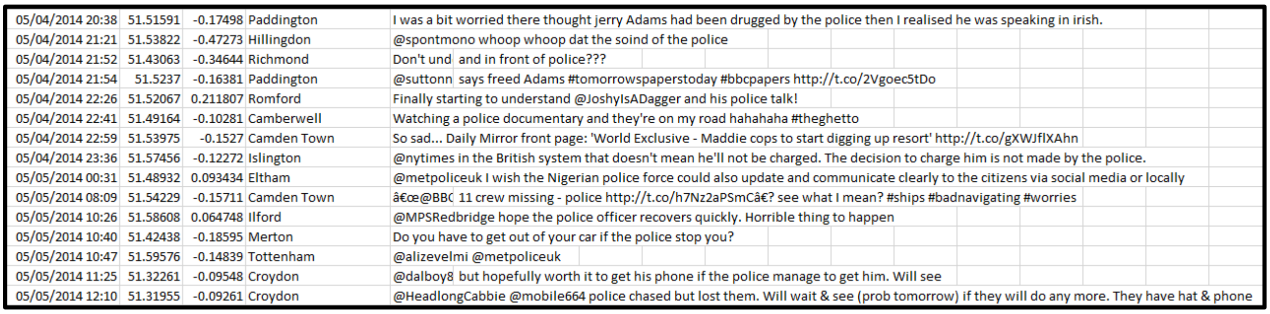

COSMOS harvests Tweet Data
Tweet Data stored in CSV File
Input path of Tweet Data CSV file into Java file
Decide and Input Crime-Related References into Java file
Compile and Run Java file
Output CSV file returning tweets containing specific crime-related references defined in Java script
Manually select CSV file to load into Web application
Parse Latitude and Longitude values to Google Maps API
Output heatmap showing geospatial distribution of crime-related references from Twitter

Figure 1: Twitter Data Flow Diagram

3.2.2 Twitter Data Set Analysis

In order to generate a usable data set a script will be developed in Java which is able to take an input of the raw dataset, search for specific crime terms within the plain text of tweets (which a user can define themselves in the code) and eventually return a CSV file similar to below.



05/04/2014 20:38	51.51591	-0.17498	Paddington	I was a bit worried there thought jerry Adams had been drugged by the police then I realised he was speaking in irish.	
05/04/2014 21:21	51.53822	-0.47273	Hillingdon	@spontmono whoop whoop dat the soind of the police	
05/04/2014 21:52	51.43063	-0.34644	Richmond	Don't und	and in front of police???
05/04/2014 21:54	51.5237	-0.16381	Paddington	@suttonn	says freed Adams #tomorrowspaperstoday #bbcpapers http://t.co/2Vgoec5tDo
05/04/2014 22:26	51.52067	0.211807	Romford	Finally starting to understand @JoshyIsADagger and his police talk!	
05/04/2014 22:41	51.49164	-0.10281	Camberwell	Watching a police documentary and they're on my road hahahaha #theghetto	
05/04/2014 22:59	51.53975	-0.1527	Camden Town	So sad... Daily Mirror front page: 'World Exclusive - Maddie cops to start digging up resort' http://t.co/gXWJflXAhn	
05/04/2014 23:36	51.57456	-0.12272	Islington	@nytimes in the British system that doesn't mean he'll not be charged. The decision to charge him is not made by the police.	
05/05/2014 00:31	51.48932	0.093434	Eltham	@metpoliceuk I wish the Nigerian police force could also update and communicate clearly to the citizens via social media or locally	
05/05/2014 08:09	51.54229	-0.15711	Camden Town	â€œ@BBC	11 crew missing - police http://t.co/h7Nz2aPSmCâ€? see what I mean? #ships #badnavigating #worries
05/05/2014 10:26	51.58608	0.064748	Ilford	@MPSRedbridge hope the police officer recovers quickly. Horrible thing to happen	
05/05/2014 10:40	51.42438	-0.18595	Merton	Do you have to get out of your car if the police stop you?	
05/05/2014 10:47	51.59576	-0.14839	Tottenham	@alizevelmi @metpoliceuk	
05/05/2014 11:25	51.32261	-0.09548	Croydon	@dalboy8	but hopefully worth it to get his phone if the police manage to get him. Will see
05/05/2014 12:10	51.31955	-0.09261	Croydon	@HeadlongCabbie @mobile664 police chased but lost them. Will wait & see (prob tomorrow) if they will do any more. They have hat & phone	

Figure 2: Sample Data Set for Crime-Related Tweets in London

Columns from left to right: Timestamp (Date and 24 hour Time), Latitude, Longitude, London Borough (for the purpose of London Tweets Only), Tweet Plain Text

3.3 Crime Types and References

3.3.1 Home Office Crime Types

When looking for references to crime in the data from Twitter the Home Office crime type framework is to be used. This is an official framework which will add more reliability to the investigation and help me to produce a visualisation of hotspots for different crime types which addresses one of the aims in my project plan.

The crime type categories which will be used based on this framework are:

- **Violent Crime** –The victim of this crime type would be "intentionally stabbed, punched, kicked, pushed…or threatened with violence whether there is an injury or not" [8]. Example offences include the likes of Homicide, all kinds of Assault and Weapon crime
- **Acquisitive Crime** – This is a reference to any "household and personal crime where items are stolen" [8]. Burglary and Vehicle theft are covered under this crime type
- **Vandalism and Criminal Damage** – With vandalism being malicious damage, criminal damage is where any person "without lawful excuse destroys or damages any property belonging to another" [8]. Graffiti and scratching a Car is considered an act of criminal damage but this category covers crimes even as extreme as Arson
- **Fraud and Forgery** – The definition of Fraud according to the Fraud Act 2006 is somebody "dishonestly making a false representation to obtain property or money for themselves or another" [8]. Stealing somebody's bank details from credit card fraud is a prime example of what is being described by this crime type category
- **Racially and Religiously Aggravated Offences** –This category is based on religious or racially motivated attacks. For example an offence of criminal damage could be racially motivated and therefore would be placed in the numbers for this category
- **Drug Offences** – Any crime which involves the possession, dealing or taking of illegal drugs

For each of these crime types key crime-related words and phrases will be identified which are relevant to the particular crime type. In turn, the Java code which is planned to be implemented will be run to create a CSV file for each crime type in the same style and layout as described in the Data Set Analysis section. Finally to show an all-up view of crime hotspots work will be done to output one file which shows a combination of all the crime types together. To do this I will combine all of the crime-related references for each crime type into one array and then search for those references in the Twitter data.

3.3.2 Selecting Key Crime-Related Words and Phrases

When using the Java script to produce a CSV file for each crime type I will manually choose which crime-related keywords and phrases for the script to search for in the data set. It is very difficult to cover every possible term so using the Home Office Framework description for each crime type I will pick out the key terms and build up a list so for each crime type there will be a separate array of related words and phrases.

There is no 'correct' approach to this task and therefore I see this as the best way to link actual crime and crime references on Twitter. The problem with this particular approach is despite the clear logic behind using the Home Office framework to link what the British Government use for recording actual crime to search for these type of references on Twitter it is not guaranteed to produce accurate results. Users of social media tend to use different, less formal ways to describe their point and therefore inaccurate results could be returned. For example, a user could tweet the word ‘smashed’. Although the word could very much be used as a crime-related tweet in the context of ‘smashed window’, it could also be used to informally describe somebody ‘smashed’ which according to Urban Dictionary refers to being “heavily intoxicated” [9] or to describe comfortably winning a sports match i.e. “we smashed them”.

The table below shows the crime-related references which were settled upon having analysed the Home Office Crime Types document:

Home Office Crime Type	**Crime-Related References/Terms**
Violent Crime	"violence", "stabbed", "stabbing", "knife attack", "punched", "assault", "robbery", "homicide", "wounding", "domestic violence", "mugging", "murder", "murdered", "killed", "manslaughter", "infanticide", "sexual assault", "rape", "drink" + "driving", "drug" + "driving", "GBH", "possession" + "weapons", "harassment", "firearm", "stalking"
Acquisitive Crime	"theft", "burglary", "vehicle" + "stolen", "nicked", "robbery", "breaking" + "entering", "bike" + "stolen"
Criminal Damage	“vandalism", "vandalised", "vandal", "arson", "graffiti", "deliberate damage", "malicious", "set fire", "criminal damage", "destroyed", "damaged", "keying" + "car", "reckless"
Fraud and Forgery	"fraud", "forgery", "identity theft", "bank card" + "stolen", "false accounting", "bankruptcy", "credit card", "debit card", "bank details" + "stolen", "fraudulent"
Racial/Religiously Aggravated Crime	"racial", "racial" + "abuse", "racist", "hate crime", "religion", "hatred", "religious" + "aggravated"
Drug Offences	"drugs", "drug" + "possession", "intent to supply", "cannabis", "class a", "class b", "class c", "drug trafficking", "heroin", "cocaine", "LSD", "amphetamine", "ketamine"
All Crime (combination of all references for each crime type)	"violence", "stabbed", "stabbing", "knife attack", "punched", "assault", "robbery", "homicide", "wounding", "domestic violence", "mugging", "murder", "murdered", "killed", "manslaughter", "infanticide", "sexual assault", "rape", "drink" + "driving", "drug" + "driving", "GBH", "possession" + "weapons", "harassment", "firearm", "stalking", "theft", "burglary", "vehicle" + "stolen", "nicked", "robbery", "breaking" + "entering", "bike" + "stolen", "vandalism", "vandalised", "vandal", "arson", "graffiti", "deliberate damage", "malicious", "set fire", "criminal damage", "destroyed", "damaged", "keying" + "car", "reckless", "fraud", "forgery", "identity theft", "bank card" + "stolen", "false accounting", "bankruptcy", "credit card", "debit card", "bank details" + "stolen", "fraudulent", "racial", "racial" + "abuse", "racist", "hate crime", "religion", "hatred", "religious" + "aggravated", "drugs", "drug"+ "possession", "intent to supply", "cannabis", "class a", "class b", "class c", "drug trafficking", "heroin", "cocaine", "LSD", "amphetamine", "ketamine"

Table 1: Crime References to search for in Twitter data set (By Crime type)

3.4 Actual Reported Crime - Police API

In order to gather data for actual crimes the project will make use of the Police API which has much archived data. The ‘Crimes at a Location’ data source will provide me with the required data to plot a geospatial distribution using Google Maps API. With my focus very much on the London area, the data for the Metropolitan Police should give me sufficient data to provide a useful comparison against the Twitter crime.

3.4.1 Actual Crime Data Flow

The below diagram details how the data sourced from the Police API will flow in proposed software environment. The data is stored in a slightly different format to the Twitter data set hence there is differences in the data flow and how the data is proposed to be handled. This is not ideal but would mean implementing another evaluation tool to breakdown the monthly CSV files which is not feasible in the time available. The main distinction is that data is already provided in a usable format for actual crime in contrast to the Twitter data which needs to be evaluated first because it initially comes in a raw form.

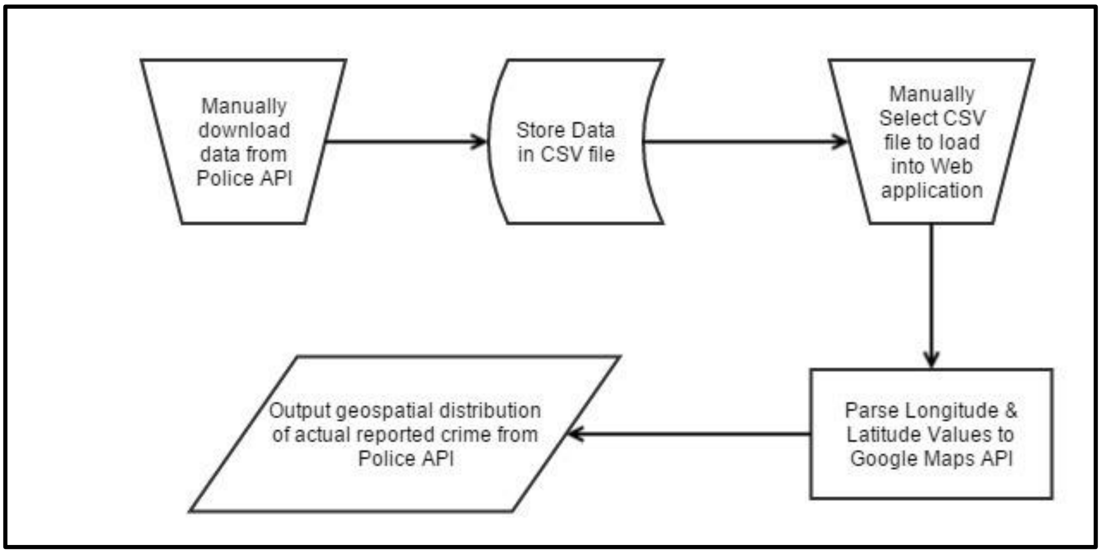
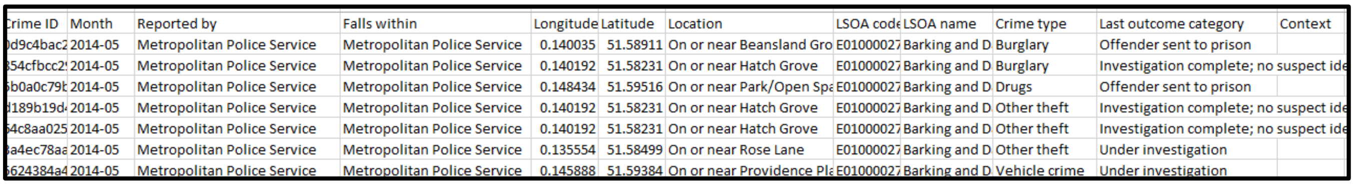

Manually download data from Police API
Store Data in CSV file
Manually Select CSV file to load into Web application
Parse Longitude & Latitude Values to Google Maps API
Output geospatial distribution of actual reported crime from Police API

Figure 3: Data Flow Diagram for Actual Reported Crime (from Police API)

3.4.2 Actual Crime Data Set Analysis

In order to make a comparison with the crime which has been referenced on Twitter the Police API provides a trusted data source with the required parameters to create a geospatial distribution in similar fashion to the one which will be created for crime-related references on Twitter. Each data set comes in a monthly format and will fall within the Metropolitan Police area which covers the same region (London) which this project is focussing on.



Crime ID	Month	Reported by	Falls within	Longitude	Latitude	Location	LSOA code	LSOA name	Crime type	Last outcome category	Context
0d9c4bac2	2014-05	Metropolitan Police Service	Metropolitan Police Service	0.140035	51.58911	On or near Beansland Gro	E01000027	Barking and D	Burglary	Offender sent to prison	
354cfbcc2	2014-05	Metropolitan Police Service	Metropolitan Police Service	0.140192	51.58231	On or near Hatch Grove	E01000027	Barking and D	Burglary	Investigation complete; no suspect ide	
6b0a0c79b	2014-05	Metropolitan Police Service	Metropolitan Police Service	0.148434	51.59516	On or near Park/Open Spa	E01000027	Barking and D	Drugs	Offender sent to prison	
d189b19d	2014-05	Metropolitan Police Service	Metropolitan Police Service	0.140192	51.58231	On or near Hatch Grove	E01000027	Barking and D	Other theft	Investigation complete; no suspect ide	
54c8aa025	2014-05	Metropolitan Police Service	Metropolitan Police Service	0.140192	51.58231	On or near Hatch Grove	E01000027	Barking and D	Other theft	Investigation complete; no suspect ide	
3a4ec78aa	2014-05	Metropolitan Police Service	Metropolitan Police Service	0.135554	51.58499	On or near Rose Lane	E01000027	Barking and D	Other theft	Under investigation	
5624384a4	2014-05	Metropolitan Police Service	Metropolitan Police Service	0.145888	51.59384	On or near Providence Pla	E01000027	Barking and D	Vehicle crime	Under investigation	

Figure 4: Sample Data Set from Police API (Actual Recorded Crimes)

Columns from Left to Right(NB: fields in **Bold** text are relevant): Crime ID, **Month**, Reported By (Police Constabulary), Falls within (Police Constabulary), **Longitude, Latitude**, Location, LSOA code, LSOA name, **Crime Type**, Last outcome category, Context

3.5 Statistical Analysis Design

To test for any correlation between actual crime and the references to crime on Twitter a statistical analysis is to be completed. The statistical method which will be used complimenting the categorical data sets which have been sourced is a Pearson Chi-Squared test. By building a simple matrix of actual crimes versus crimes referenced on Twitter a good starting point for the test will be formed.

The following formula will help with this calculation:

$$\boxed{\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}}$$

Figure 5: Pearson Chi-Squared formula [10]

*Mathematical Notation-

X^2 = Chi-Squared Statistic
O = Observed Frequency (Actual Crime)
E = Expected Frequency (Crime References from Twitter)

Actual crime is what has been observed and in order for correlation to occur we expect it to be at a similar level to the crime references on Twitter which justifies the orientation of the 'Observed' and 'Expected' variables in this formula.

Ideally, a month-by-month comparison would give a much more accurate comparison and a truer representation of the actual relationship between the two data sets. The test will focus on how all-up crime correlates against each other for January to April 2014 in addition to measuring the relationship for each crime type.

The test itself is purely focussed on a comparison of the number of data points between the two sets of data and does not take into account the distribution of the crime hotspots which will be plotted. Due to the expected high volume of data points returned, a spatial comparison will take too long to complete and it is difficult to tell whether we would gain any more understanding of the relationship from completing such a test. This means that the Pearson test is the most appropriate choice of statistical analysis under the constraints of this project.

3.6 Google Maps API Heat-Map Layer Customisation

In order to draw further conclusions on the location (comprising of a longitude and latitude value) of crime hotspots in the project's proposed software environment some customisation tools will be used from the Google Maps API heat-map layer.

The heat-map layer provides a client-side data visualisation which is ideal for this project. Although a large amount of data points can result in reduced performance, the layer is well supported by all major browsers and can easily be customised. The other option would be to use the Fusion Table layer but this is still in Beta testing with Google Developers and therefore less reliable. In addition, there is no support for customisation and with a number of different heat-maps to display for each crime type the heat-map layer is much more suitable for my proposed software.

3.6.1 Max. Intensity

The 'Max Intensity' property can be used to set a fixed maximum number of points for the highest concentration. In Figure 6 below, dark red denotes the highest concentration, light green the lowest concentration and transparent indicates no longitude and latitude points. By setting maximum intensity to a particular value such as 50 would indicate that if 50 locations were to be plotted then the heat-map layer would render the dark red colour which represents the highest concentration of points.

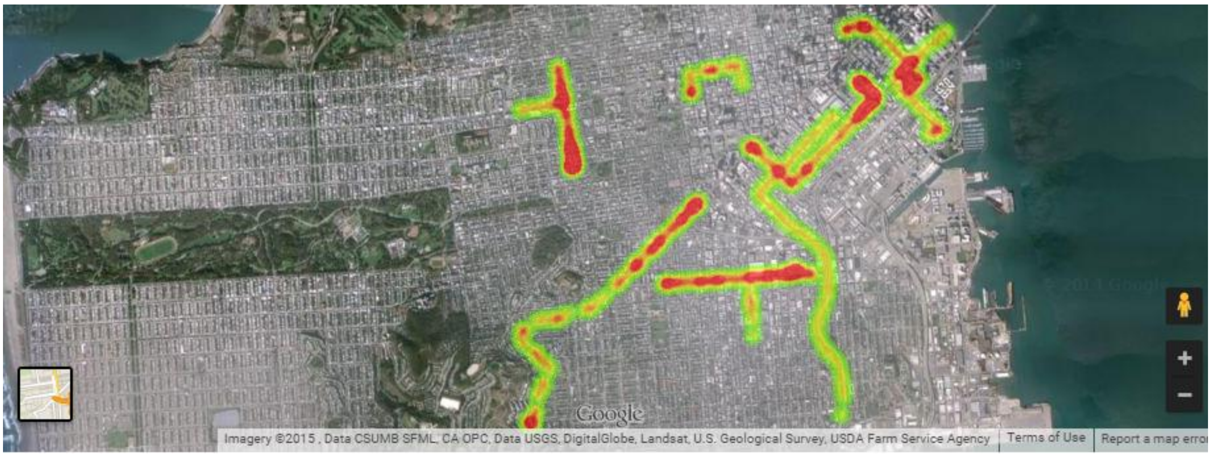
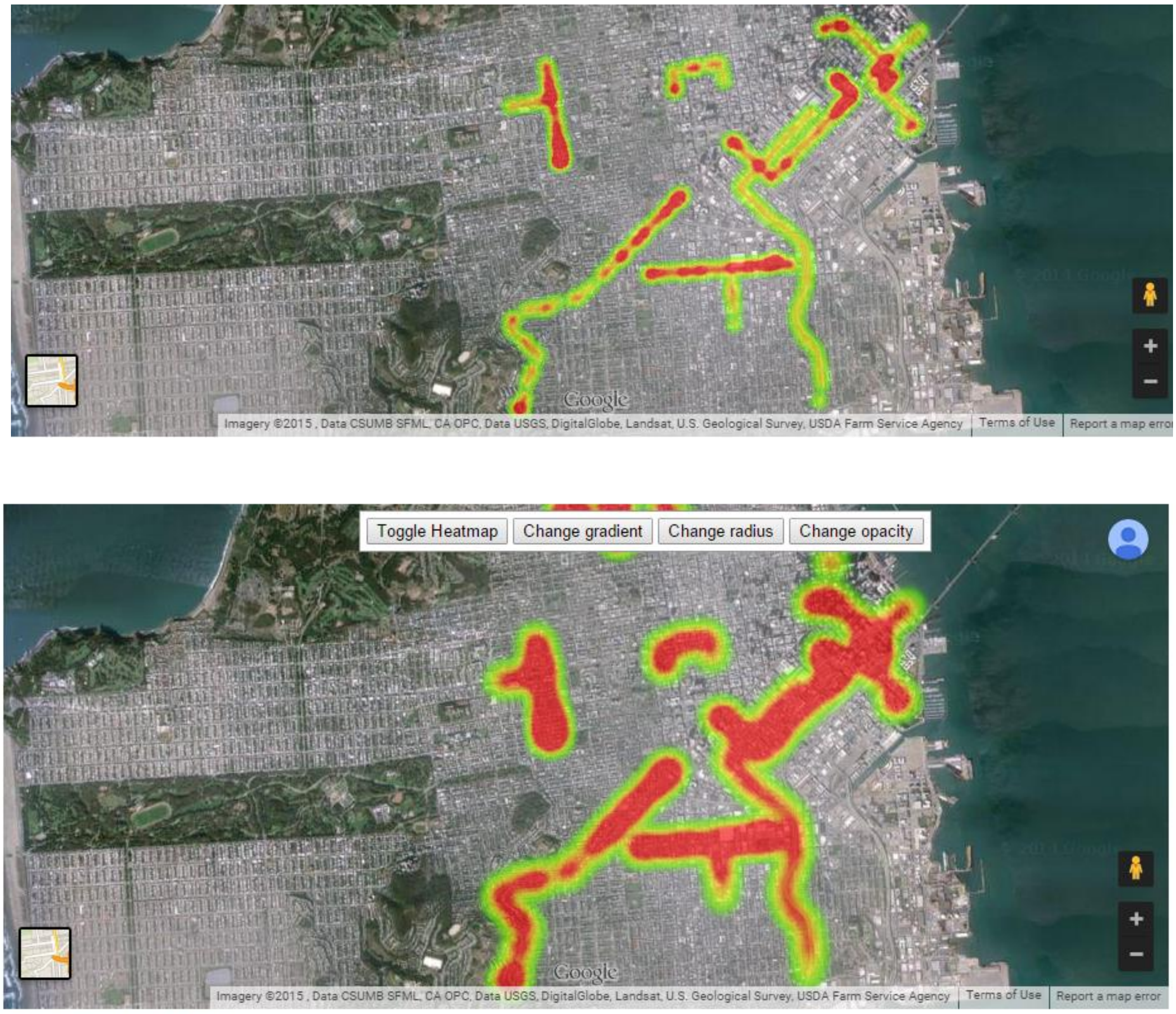

Google
Imagery ©2015 , Data CSUMB SFML, CA OPC, Data USGS, DigitalGlobe, Landsat, U.S. Geological Survey, USDA Farm Service Agency
Terms of Use
Report a map error
+
−

Figure 6: Example Heat-map Layer of San Francisco, USA [11]

This will be useful for this project, particularly during implementation and testing of the software environment. During the development stage when working with a smaller data set there will be less data points and consequently a sparsely populated heat map from which to conclude particular criminal hotspots from will be rendered. By using the 'Max. Intensity' property a more useful visualisation will be produced which will help me to understand whether the data is being displayed in an effective manner. Eventually the heat-maps for both the actual crime data and Twitter references to crime will need a fixed maximum intensity to be defined to ensure a fair visual comparison between the two sets of data.

3.6.2 Change Radius

I plan to use the change radius functionality from the heat-map layer of the Google Maps API. This property means you can set the “radius of influence for each data point in pixels” [12]. In Figure 7, there is an example from the Google Maps API Documentation where you can see a comparison between the default radius and when the radius of each data point has changed to a larger value to have more influence. The bottom image is much clearer to see where there is a greater concentration of data points and although does not give the true reflection of the exact location of each data point it shows a useful visualisation. With the sourced Twitter data, changing the radius will support the visualisation of crime hotspots by making them much clearer to see across a larger spatial area.


Google
Imagery ©2015 , Data CSUMB SFML, CA OPC, Data USGS, DigitalGlobe, Landsat, U.S. Geological Survey, USDA Farm Service Agency
Terms of Use
Report a map error
Toggle Heatmap
Change gradient
Change radius
Change opacity
Google
Imagery ©2015 , Data CSUMB SFML, CA OPC, Data USGS, DigitalGlobe, Landsat, U.S. Geological Survey, USDA Farm Service Agency
Terms of Use
Report a map error

View Heatmap Layer example

Figure 7: A Heat-map Layer Map of San Francisco, USA with a default radius value (top). The same Image with a larger radius (bottom) [11]

4 Implementation

4.1 Technology Acknowledgements

During the development of the software environment which is vital to achieving the aims of the project set in the initial planning phase, a number of technologies have been used which I feel should be acknowledged.

4.1.1 Dropbox

I have used Dropbox as my primary backup for this project. The tool is a free, file storage host which can be accessed anywhere on multiple devices through the Cloud. With the ability to download a desktop client it makes a perfect backup which can easily be accessed on my development machine. This tool has stored all versions of my source code and documentation.

4.1.2 Google Maps API

Google Maps API forms the backbone of this piece of software. In order to meet the aims of this project a geospatial distribution of crime locations was imperative. Some kind of mapping and visualisation tool was therefore required and Google Maps API was the obvious choice as it is available in a web-based format which I decided to develop with. The API makes for an interactive and easy to use mapping tool which can take longitude and latitude values as parameters to plot data points. This was extremely useful considering my data sets use longitude and latitude as their location references.

The heat-map layer provided further suitability for use in this project with a need to visualise hotspots of crime. The flexibility of the layer adds further value with the ability to alter the colour gradients (useful when splitting by crime type) and process thousands of data points was a key factor in my choice to use this API and its heat map layer.

4.1.3 jQuery

jQuery is a free, open source JavaScript library which in the context of this project has been useful to "select DOM [(*Document Object Model)*] elements…[and] handle events" [13]. Particularly when handling the input CSV files from both the Twitter data and Police API in addition to a lightweight method of loading the Google Maps API.

4.1.3.1 jQuery UI

With a highly populated user interface (UI) in my system I needed an interactive UI which could help to show the visualised data. The jQuery UI is a "set of user interface interactions...built on top of the jQuery JavaScript library" [14] and was an excellent choice because it was compatible with the jQuery and JavaScript programming languages which were used to predominantly develop this software environment. The ‘draggable’ menu property was particularly useful because it enabled me to alter the position of the menu of the screen to view a crime hotspot of interest in more detail on the map.

4.1.4 PapaParse

PapaParse is a JavaScript library which acts as a “powerful, in-browser CSV parser” [15]. With both my data sources being in similar CSV format a lightweight parser which was able to process files with thousands of rows was required. PapaParse provides a fast implementation of parsing local CSV files in a browser environment and therefore was a good choice to use as the method to handle files. Its JavaScript implementation also meant it fit seamlessly into my source code which was in the same programming language.

4.2 Changes to Design

During the implementation of the software environment changes to the design were required to help meet the aims of the project and produce a more effective code implementation. Although the design represents a conceptual model of how this software will be built it is important there is room for change and flexibility which will help to improve the final prototype.

4.2.1 Change of Crime Type Framework

In the specification and design section I stated that I would use the Home Office Crime Type framework to provide a reliable, structured approach to producing separate visualisations for each crime type. On reflection having further analysed the data for actual crime sourced from the Police API, I discovered that it would be more appropriate to use the crime types which were already outlined in this data. Instead of predicting which crime type from the Police API data fits into which crime type from the Home Office framework, a more accurate link between the crimes referenced on Twitter and actual crime data sets can be provided. In summary, using this method it is simply mapping one entity against another as opposed to using a third entity to map the two against. This reduces much ambiguity in the way the visualisations for each crime type are created.

The crime types which have now been followed during the development of this system are as follows (sourced from the Police API [16]):

- **Anti-Social Behaviour -** A particularly broad term but describes “day-to-day incidents of nuisance and disorder that affects people's lives” [17]. This includes crimes such as littering, excessive noise and being rowdy from alcohol or drug use
- **Bike Theft -** This covers the illegal removal or theft of a non-motorised pedal bicycle. Examples methods to steal a bike includes the use of ‘sucker poles’ which are effectively bike racks erected by criminals which can easily be dismantled if somebody locks their bicycle against it

- **Burglary-** A category referring to “illegal entry into a building for the purposes of committing an offence” [18] which could be theft for example
- **Criminal Damage -** Criminal damage is where any person “without lawful excuse destroys or damages any property belonging to another” [8]. Graffiti and scratching a car is considered an act of criminal damage but this category covers crimes even as extreme as Arson. (NB: This category remains similar to the one outlined in the Home Office crime type framework)
- **Drug Offences -** Any crime which involves the possession, dealing or taking of illegal drugs
- **Possession of Weapons – “**carrying an offensive weapon, or a knife, or a bladed/pointed article is a serious offence” [19]. The possession or provision of such weapons is against the law and can constitute to conviction within this crime category.
- **Public Order –** Public Order references the “use of violence and/or intimidation by individuals or groups” [20] including crimes such as rioting.
- **Robbery -** Robbery differs from Burglary by the fact it is “the taking of money or goods in the [immediate] possession of another…by force or intimidation” [21] . There are different categories of Robbery including aggravated, armed and highway.
- **Shoplifting -** This crime occurs when “someone steals merchandise offered for sale in a retail store” [22]. Often the criminal will conceal the items they intend to shoplift in a coat or bag to reduce the chance of them being caught.
- **Theft from the person -** This crime category covers “items [which] are stolen from someone without the threat or use of physical force” [23]. Example of theft from the person crime are pickpocketing and snatching possessions.
- **Vehicle Crime -** This crime type normally involves the theft of or from a vehicle or damage to a vehicle.
- **Violent Crime and sexual offences -** The victim of this crime type would be “intentionally stabbed, punched, kicked, pushed…or threatened with violence whether there is an injury or not” [8]. This category also includes sexual offences such as sexual assault or rape.

4.2.2 Changes to Crime-Related Twitter References

Having previously based my approach to splitting by crime type on the Home Office crime type framework the plan has now altered to rely upon the Police API crime categories as discussed in section 4.2.1 above. Consequently by changing the crime type framework to follow, an overhaul of the crime-related words and phrases to search for within the Twitter data set was necessary.

The method previously used to find crime-related words and phrases for each crime type using the Home Office framework involved analysing each crime type category and extracting key words and phrases. The Police API does not give a description for each crime category and therefore the approach to this problem had to be adapted. The method now being used is to search for articles on the Web for each crime type, picking out crime-related words and phrases in similar fashion to the extraction performed on the Home Office framework. Additionally Urban Dictionary has been used to identify possible slang words or phrases which could be used to reference each crime type.

In relation to crime references in the Twitter data, each crime type has its own array of crime-related words which gets searched using Java code outputting a separate CSV file in the same style as in section 3.2.2 which outlines the structure of an output Twitter data CSV file.

Following the change in approach these are the crime-related words and phrases which have been deduced for each crime type:

Police API Crime Category	**Crime-Related References/Terms**
Anti-Social Behaviour	"anti-social behaviour", "ASBO", "distress", "arson", "begging", "urinating in public", "drunk", "spitting", "littering", "intimidation", "fare evasion", "smoking illegally"
Bike Theft	"bike theft", "bike nicked", "bike stolen", "bicycle nicked", "bicycle stole", "bike gone", "stole bike", "bike lock", "stole cycle", "sucker pole", "bike thief", "bike"
Burglary	"burglary", "breaking and entering", "housebreaking", "thief", "trespass property", "steal", "home invasion", "nick property", "burgle", "burglaries", "stole"
Criminal Damage	"vandalism", "vandalised", "vandal", "arson", "graffiti", "deliberate damage", "malicious", "set fire", "criminal damage", "destroyed", "damaged", "keying car", "reckless"
Drug Offences	“drugs", "drug possession", "intent to supply", "cannabis", "class a", "class b", "class c", "drug trafficking", "heroin", "cocaine", "LSD", "amphetamine", "ketamine"
Possession of Weapons	"weapon", "firearm", "possession weapon", "bomb", "artillery", "gun", "knife"
Public Order	"public order", "riot", "lashing out", "fight", "abusive", "violent disorder", "affray", "harassment", "protest", "trespass", "racial", "racist", "religious hatred"
Robbery	"robbery", "threat", "armed robbery", "highway robbery", "aggravated robbery", "steal", "blagging", "stick-up", "violent theft","mugging","car jacking"
Shoplifting	"shoplift", "shoplifting", "boosting", "five finger discount", "shoplifter", "stealing", "nicked from shop", "take without paying", "stolen goods", "stole", "thief", "lifting"
Theft from the person	"theft", "nicked", "stolen", "mugged", "pickpocket", "snatched", "theft from the person", "thieving", "stealing", "stole", "filching", "taking property", "handbag taken"
Vehicle Crime	"vehicle crime", "stolen car", "car nicked", "vehicle trafficking", "car theft", "car broken", "getaway vehicle", "getaway car", "car thief", "car crime", "car damaged", "stole motorbike", "taking driving away", "twockers", "twocking","grand theft auto", "carjacking", "taken without owner's consent", "hotwire", "hotwiring", "car alarm"

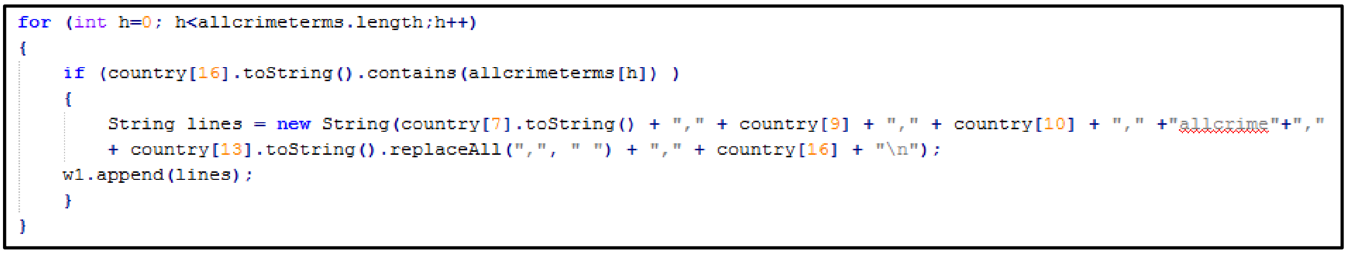
Violent Crime	"violence", "stabbed", "stabbing", "knife attack", "punched", "assault", "robbery", "homicide", "wounding", "domestic violence", "mugging", "murder", "murdered", "killed", "manslaughter", "infanticide", "sexual assault", "rape", "drink driving", "drug driving", "GBH", "harassment", "stalking"
All Crime (combination of all references for each crime type)	violence", "stabbed", "stabbing", "knife attack", "punched", "assault", "robbery", "homicide", "wounding", "domestic violence", "mugging", "murder", "murdered", "killed", "manslaughter", "infanticide", "sexual assault", "rape", "drink driving", "drug driving", "GBH", "harassment", "stalking", "weapon", "firearm", "possession weapon", "bomb", "artillery", "gun", "knife", "vandalism", "vandalised", "vandal", "arson", "graffiti", "deliberate damage", "malicious", "set fire", "criminal damage", "destroyed", "damaged", "keying car", "reckless", "public order", "riot", "lashing out", "fight", "abusive", "violent disorder", "affray", "harassment", "protest", "trespass", "racial", "racist", "religious hatred", "anti-social behaviour", "ASBO", "distress", "arson", "begging", "urinating in public", "drunk", "spitting", "littering", "intimidation", "fare evasion", "smoking illegally", "drugs", "drug possession", "intent to supply", "cannabis", "class a", "class b", "class c", "drug trafficking", "heroin", "cocaine", "LSD", "amphetamine", "ketamine", "robbery", "threat", "armed robbery", "highway robbery", "aggravated robbery", "steal", "blagging", "stick-up", "violent theft", "mugging", "car jacking", "burglary", "breaking and entering", "housebreaking", "thief", "trespass property", "steal", "home invasion", "nick property", "burgle", "burglaries", "stole", "theft", "nicked", "stolen", "mugged", "pickpocket", "snatched", "theft from the person", "thieving", "stealing", "stole", "filching", "taking property", "handbag taken", "bike theft", "bike nicked", "bike stolen", "bicycle nicked", "bicycle stolen", "bike gone", "stolen bike", "bike lock", "stole cycle", "sucker pole", "bike thief", "bike", "shoplift", "shoplifting", "boosting", "five finger discount", "shoplifter", "stealing", "nicked from shop", "take without paying", "stolen goods", "stole", "thief", "lifting", "vehicle crime", "stolen car", "car nicked", "vehicle trafficking", "car theft", "car broken", "getaway vehicle", "getaway car", "car thief", "car crime", "car damaged", "stole motorbike", "taking driving away", "twockers", "twocking", "grand theft auto", "carjacking", "taken without owner's consent", "hotwire", "hotwiring", "car alarm"

Table 2: Crime-Related References/Terms (using Police API Crime Categories)

4.3 Source Code Implementation

4.3.1 Searching Twitter data with Java

In order to search for crime-related terms some Java code has been adapted which takes the input of a raw data file such as a month's worth of Twitter data and refines the relevant data into a manageable CSV file format. In effect this is the pattern-matching element of the project and this tool is separate from the web-based software which has been created.



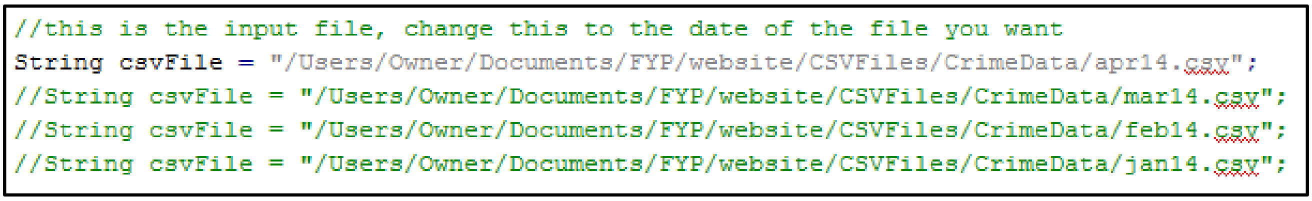
```
for (int h=0; h<allcrimeterms.length;h++)
{
    if (country[16].toString().contains(allcrimeterms[h]) )
    {
        String lines = new String(country[7].toString() + "," + country[9] + "," + country[10] + "," +"allcrime"+","
        + country[13].toString().replaceAll(",", " ") + "," + country[16] + "\n");
    w1.append(lines);
    }
}
```

Figure 8: Example Java Code- searching for 'All Crime' References in the Twitter Data

The code searches the data for the array of crime terms which in the case of the above Figure 8 is the combination of crime-related terms for all crime types defined as 'all crime'. If the plain text of a tweet in the raw data contains a word which is in the particular crime-related array then a new line will be created in the resulting CSV file. Each line of the data which is retrieved will be in the order which is specified in the code so in column order exactly the same as the example Twitter data set outlined in the Design section. By adding a further manual flag into the data this can be used to distinguish between crime types i.e. 'allcrime'. Although it means that duplicate data is created for each line in a CSV file it provides a crime type which can be leveraged to create separate heat-maps and give a meaningful identity to each CSV file. For example, without this crime type flag each output CSV file would look identical in structure. Only the content of the file (plain text of tweets) would give us a clue which crime type the file is representing but a high attention to detail would be required decipher this.

By appending after each created line of data, an eventual line-by-line CSV file is created (where >1 lines of data are found) and output to a filename of choice referenced in this Java code. In order to make it easy to work with, the files have been named in line with the crime type array which is being searched for so that they are easy to load into the tool. This code is repeated for each crime type defined in the Police API so that we have a full output of separate CSV files.

In order to alter the month of Twitter data we wish to work on, it is simply a case of altering the path of the input file to point to a different month's data (by removing the comments in Figure 9) and re-compile the Java file.



```
//this is the input file, change this to the date of the file you want
String csvFile = "/Users/Owner/Documents/FYP/website/CSVFiles/CrimeData/apr14.csv";
//String csvFile = "/Users/Owner/Documents/FYP/website/CSVFiles/CrimeData/mar14.csv";
//String csvFile = "/Users/Owner/Documents/FYP/website/CSVFiles/CrimeData/feb14.csv";
//String csvFile = "/Users/Owner/Documents/FYP/website/CSVFiles/CrimeData/jan14.csv";
```

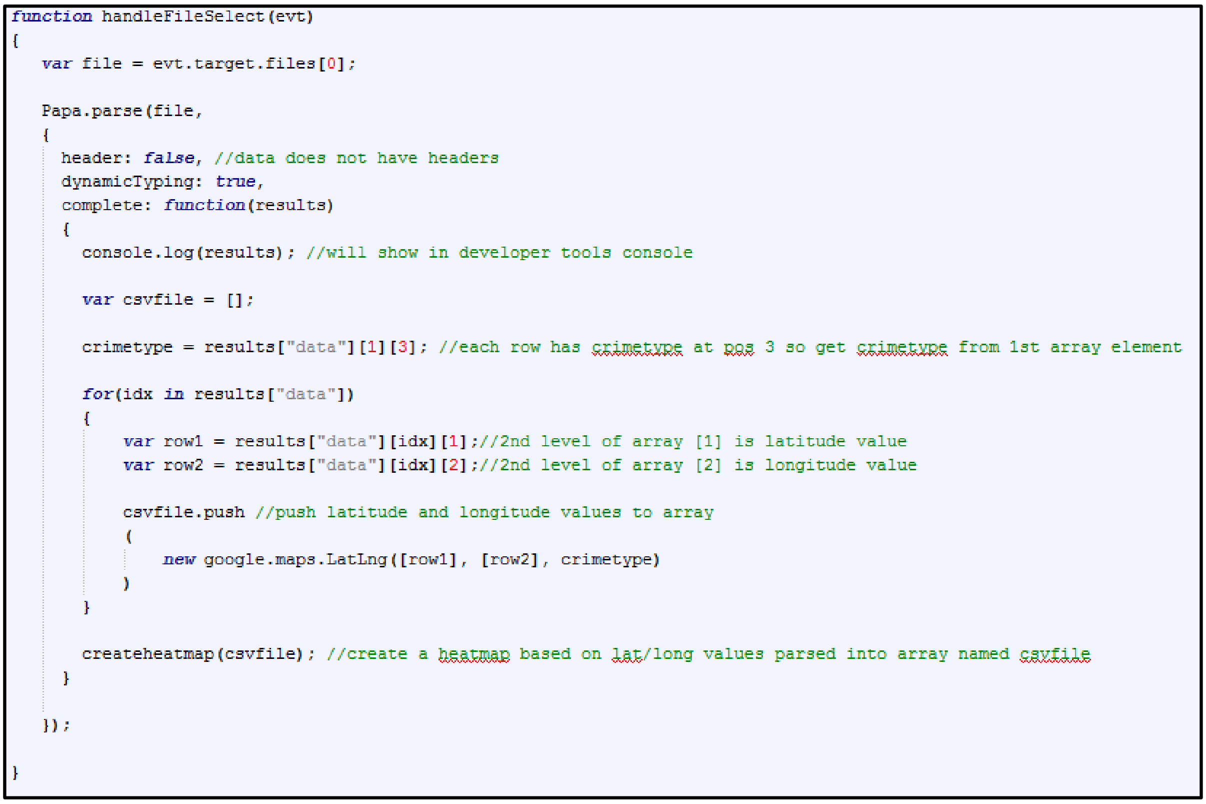
Figure 9: Input file is for April 2014 tweets. Change input to evaluate another month's data

4.3.2 Parsing CSV Files to Web Browser

Despite the difference in the compilation of the CSV files for Actual crime and the Twitter crime references the same JavaScript parsing library, PapaParse [24] has been used to parse the data into the web browser but in slightly different ways. With the goal to see a month-by-month comparison, for each month there are many Twitter CSV files to parse while the actual crime CSV files are consolidated into one file with all the data. While this implementation makes the code slightly less flexible it helps it become fit for purpose.

4.3.2.1 Parsing Twitter Data

In order to parse the CSV files for each crime type which were created using Java code on the raw Twitter data, a jQuery function was used to handle the event of the user choosing the initial file in the interface of the web-based software. When the event of selecting a file occurs then that particular file is parsed into the below JavaScript function (in Figure 10).



```
function handleFileSelect(evt)
{
   var file = evt.target.files[0];

   Papa.parse(file,
   {
     header: false, //data does not have headers
     dynamicTyping: true,
     complete: function(results)
     {
       console.log(results); //will show in developer tools console

       var csvfile = [];

       crimetype = results["data"][1][3]; //each row has crimetype at pos 3 so get crimetype from 1st array element

       for(idx in results["data"])
       {
           var row1 = results["data"][idx][1];//2nd level of array [1] is latitude value
           var row2 = results["data"][idx][2];//2nd level of array [2] is longitude value

           csvfile.push //push latitude and longitude values to array
           (
               new google.maps.LatLng([row1], [row2], crimetype)
           )
       }

       createheatmap(csvfile); //create a heatmap based on lat/long values parsed into array named csvfile
     }

   });

}
```

Figure 10: Parsing the Twitter data from a CSV file to browser

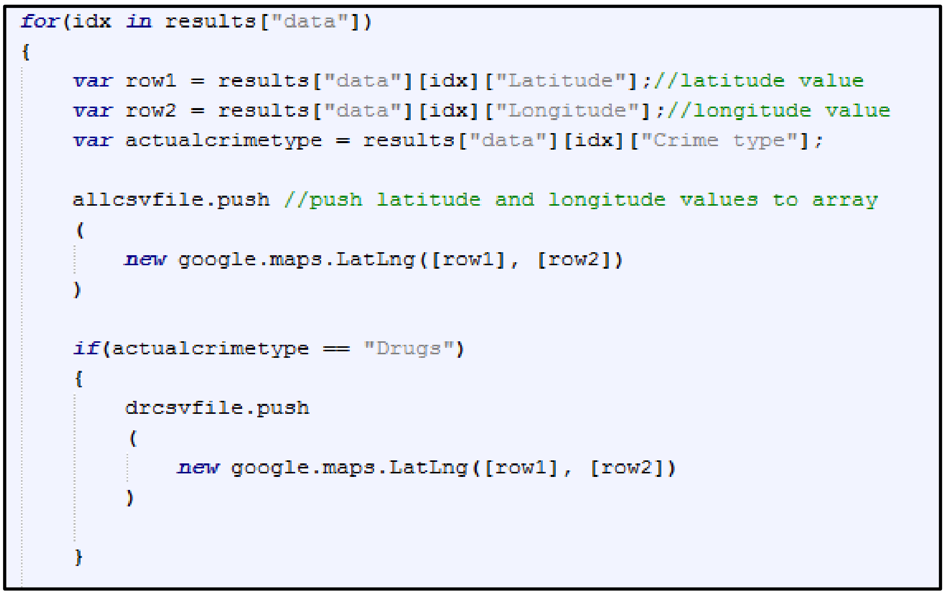
The file itself is then evaluated using the PapaParse JavaScript library. There are a number of configuration settings which you can set to help the library parse the CSV file more effectively. In this case the header has been set to false because the files do not have a row of headers defining what each column is, meaning that it will not eliminate the first line of data. The dynamicTyping property when set to true will make any numeric data resort to either integer or more likely in this case float type which is particularly useful with the eventual need to push latitude and longitude numeric values to an array in order to plot onto a map.

Once the library has completed parsing the file to the web browser the complete call-back then executes to do ‘something’ to the file. In our results, each array in the results relates to one complete row of the CSV file data. Having already described that each Twitter CSV file has a manual flag to describe its crime type, this is where it becomes particularly vital to implementation. In this case we specify the variable ‘crimetype’ which can be found in a certain row (row 3) in the arrays created from the results of parsing the file. We will use this to support building the heat-map for each crime type using the Google Maps API.

By defining an array called ‘csvfile’ we now have a data structure to store our longitude, latitude and crimetype in to parse to Google Maps API. So for each index or array in our results we push a pair of longitude and latitude coordinates and the crime type into ‘csvfile’. Finally we set ‘csvfile’ as a parameter to the ‘createheatmap’ function which is called to starting rendering the heat-maps.

The processing to produce a heat-map visualisation is completed after this whole function has been run. Therefore it is possible to load in each and every crime type CSV file using the same HandleFileSelect function.

4.3.2.2 Parsing Actual Crime Data



```
for(idx in results["data"])
{
    var row1 = results["data"][idx]["Latitude"];//latitude value
    var row2 = results["data"][idx]["Longitude"];//longitude value
    var actualcrimetype = results["data"][idx]["Crime type"];

    allcsvfile.push //push latitude and longitude values to array
    (
        new google.maps.LatLng([row1], [row2])
    )

    if(actualcrimetype == "Drugs")
    {
        drcsvfile.push
        (
            new google.maps.LatLng([row1], [row2])
        )

    }
```

Figure 11: Parsing Police API CSV data to browser

The actual crime CSV files downloaded from the Police API are structured slightly differently and additionally come in the form of one file per month as opposed to many files split by crime type as per the Twitter CSV files.

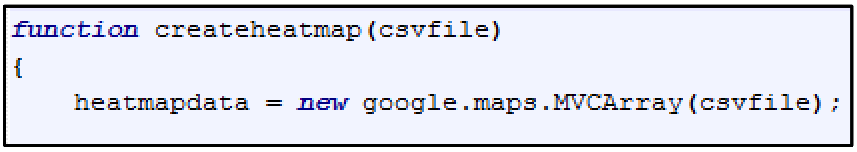
This means that parsing the file to the browser to access the data we need to plot a geospatial distribution has to be done in an alternative method to that of the Twitter CSV files. With the Police API data having column headers, when accessing each row of data that is parsed the longitude and latitude columns can be specified by column name. This is a feature implemented by the parsing library, PapaParse which makes it easier to access and manipulate the data. In addition the variable 'actualcrimetype' has been set which refers to the different crime types being using. As these are specified in each row of data in one large file, we must do the filtering of the data pre-rendering of the heat-map as opposed to the Twitter crime which has one file per crime type and therefore we can filter the data at the heat-map rendering stage.

As a consequence it means we have to define a separate array to store each crime type's longitude and latitude values in and create a number of 'if' statements for each crime type as displayed in Figure 11 above. To store the data for all crime it is a simple case of not creating an 'if' statement and then all rows of data are pushed to the array instead. A further repercussion of filtering the data pre-rendering is that in order to create a heat-map for each crime type a separate function for each crime type needs to be created and called. However, the code may not be 'nice' to look at with many 'if' statements and variables for each crime type but it is fit for purpose to parse the data into the browser in the way which it is needed.

4.3.3 Google Maps Heat-map Layer Implementation

An integral piece of work in this project was implementing the Google Maps API to visualise criminal hotspots. The API's heat-map layer due to its ability to customise appearance and display many data points (despite some possibility of performance reduction) was the tool I chose to use.

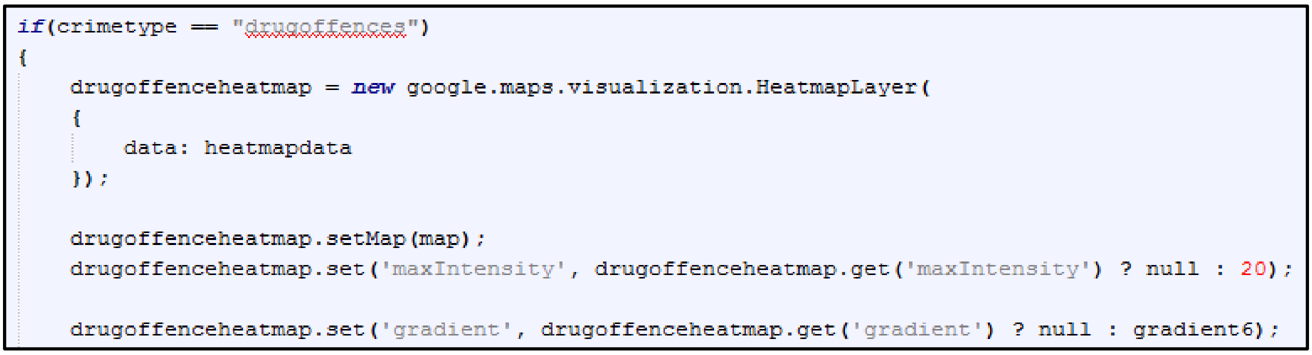
Figure 12 and 13 below shows the source code used to create a drug offences crime type heat-map for Twitter Crime. The main function 'createheatmap' takes the parameter 'csvfile' which contains the relevant data for a particular crime category in the form of an array. In the example below, the Drug Offences crime type CSV will have already been loaded in by the user using the interface, the longitude and latitude data has been parsed by the browser and stored in an array. The 'createheatmap' function is then called at the end of the parsing function. We then begin to implement the Google Maps API features in the code.



```
function createheatmap(csvfile)
{
    heatmapdata = new google.maps.MVCArray(csvfile);
```

Figure 12: createheatmap function which renders heat-maps for Twitter Crime

The variable 'heatmapdata' stores and creates an MVCArray which places the data we have already put into an array into a usable format for the API to produce a heat-map visualisation.

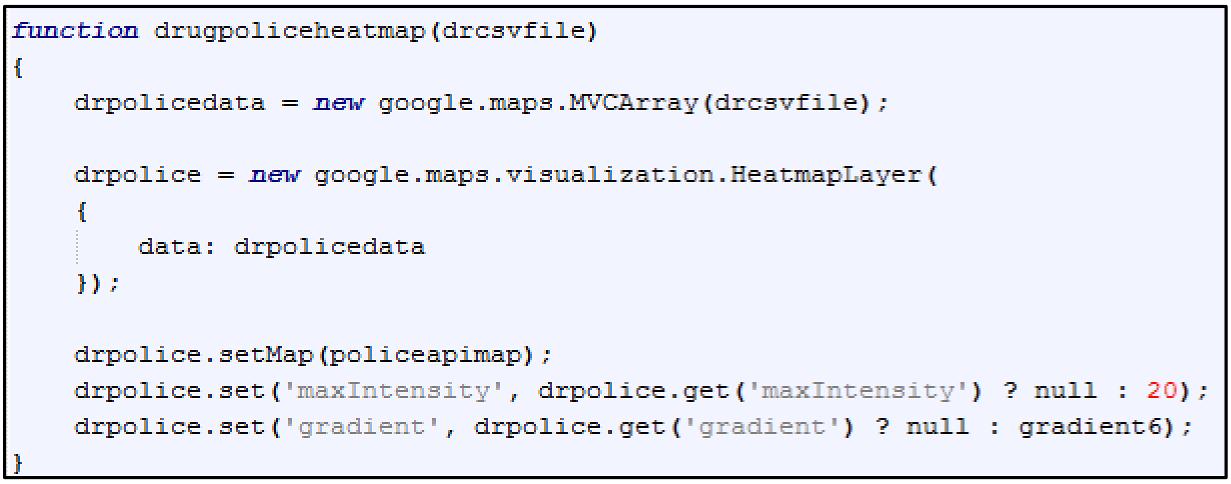


```
if(crimetype == "drugoffences")
{
    drugoffenceheatmap = new google.maps.visualization.HeatmapLayer(
    {
        data: heatmapdata
    });

    drugoffenceheatmap.setMap(map);
    drugoffenceheatmap.set('maxIntensity', drugoffenceheatmap.get('maxIntensity') ? null : 20);

    drugoffenceheatmap.set('gradient', drugoffenceheatmap.get('gradient') ? null : gradient6);
```

Figure 13: Creating and Customising the Drug Offences Crime Type Heat-map for Twitter Crime

The code begins with an 'if' statement because as discussed in section 4.3.2.1 the filtering of the data by crime type occurs during the heat-map rendering stage. A new heat-map layer visualisation is defined and stored using a meaningful variable name i.e. 'drugoffenceheatmap' and we parse the new visualisation the longitude and latitude values which are stored in the MVCArray object to be displayed.

The Max Intensity customisation feature is set to 20. I decided on 20 data points in the same location to be a reasonable figure to generate the darkest colour in a heat-map. This is set at the same level for the corresponding crime type in the actual crime data. The gradients have been developed using a CSS Gradient Generator *[25]* with a different RGBA (Red Green Blue Alpha colour model) colour gradient for each crime type so they are easily distinguished.



```
function drugpoliceheatmap(drcsvfile)
{
    drpolicedata = new google.maps.MVCArray(drcsvfile);

    drpolice = new google.maps.visualization.HeatmapLayer(
    {
        data: drpolicedata
    });

    drpolice.setMap(policeapimap);
    drpolice.set('maxIntensity', drpolice.get('maxIntensity') ? null : 20);
    drpolice.set('gradient', drpolice.get('gradient') ? null : gradient6);
}
```

Figure 14: Actual Crime Drug Offences Heat-map Implementation

Figure 14 shows the code to create the drug offences heat-map for the actual crime map. The main difference between the Police API data showing actual crime is that each crime type has its own function as opposed to one large function to develop each heat-map. This is due to the filtering of the data occurring during the parsing stage. The same heat-map layer customisation features also apply to ensure a fair visualisation comparison between the two data sets.

4.3.4 Software User Interface

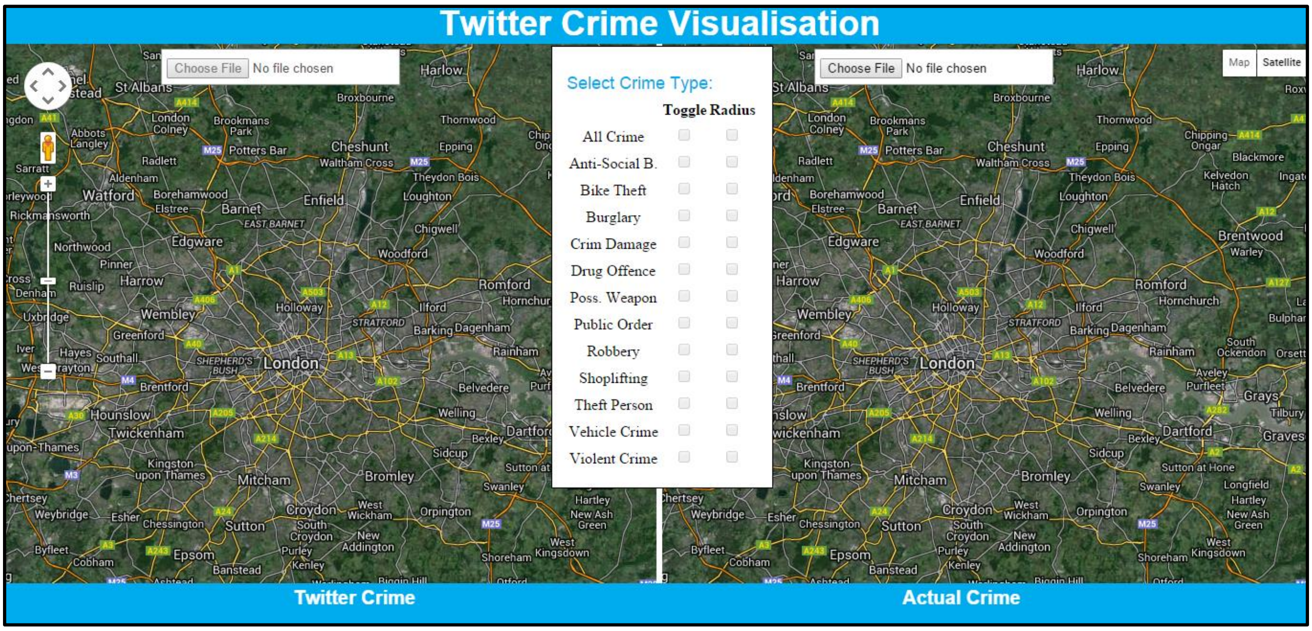


Figure 15: Software User Interface which appears on load in a web browser

The user interface which has been developed makes the software easy to use and interactive for the user. Figure 15 shows the interface which will appear once you load the main file in a web browser. A side-by-side map gives the user an 'easy on the eye' comparison of actual and Twitter crime. The interface prompts the user to load in an Actual Crime file for one month first and then will enable the button to load in each crime type file for Twitter crime. On load of each Twitter Crime file the checkbox menu becomes usable so the user can toggle the heat-maps on and off in addition to changing the radius of influence to show a clearer version of each heat-map. You can load in multiple files to the Twitter Crime map but they must be selected one at a time. It is not possible to pre-load the files or specify the filenames in the code because it causes a security issue which JavaScript does not support. For example, if an incorrect filename was specified a file which contained sensitive information could potentially be loaded instead causing data privacy issues.

jQuery UI helped to develop a 'draggable' main menu so it is possible to move it around the interface should it interfere with any area of either heat-map. It also provides the user with an interactive, check-box based main menu interface to view the particular heat-maps which they are interested in.

4.4 Implementation Issues

4.4.1 Parsing Challenges

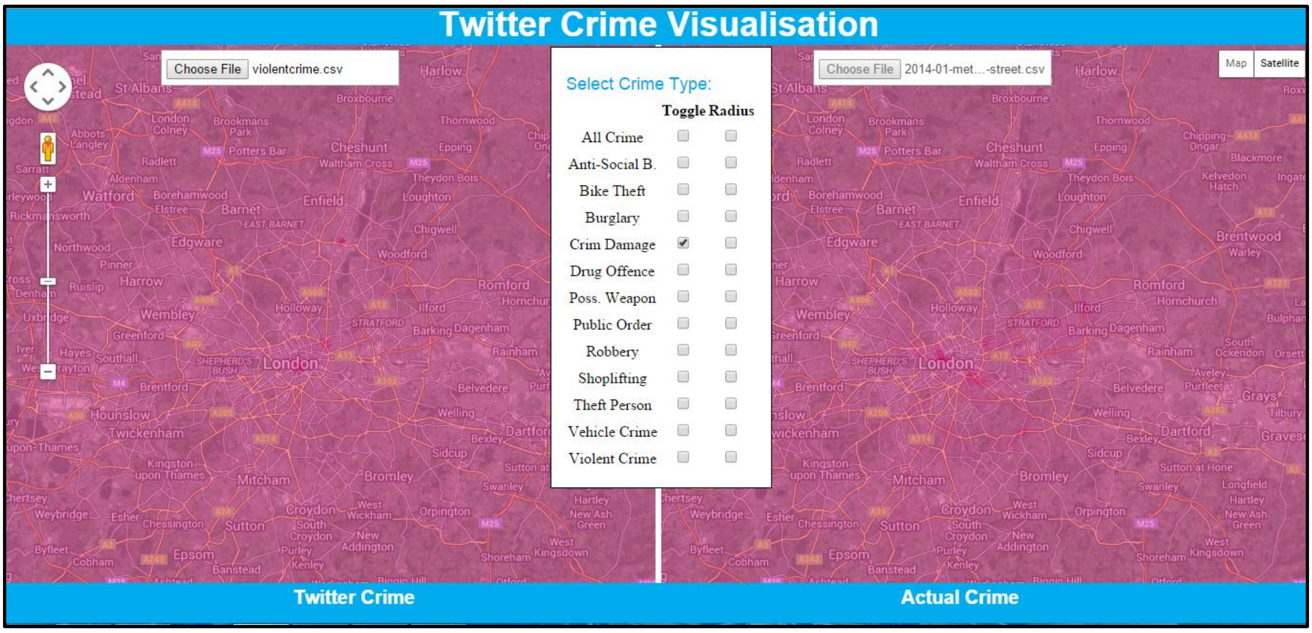
The parsing library PapaParse encourages large files to be handled using a web worker API. A worker thread runs scripts in the background and "can perform tasks without interfering with the user interface" [26] which would be useful with the need to load large CSV files in this application. In the code used to parse a CSV file to the web browser the configuration was set to use a worker thread which in theory should have meant improved performance because the large files would load in the background processes. After further investigation I discovered there is a potential bug which causes the JavaScript console to throw errors which could affect the running of the software. The PapaParse library is still looking for a solution to this particular bug. In order to solve this issue I removed the worker thread configuration in the parsing code which can have an impact on performance but does not throw errors.

4.4.2 Month-By-Month Visualisations

A stretch goal in this project was to create a slider which would allow the heat-maps to be viewed month by month for each crime type. Although I was able to produce a non-functional slider using the jQuery UI, it proved to be an over-ambitious task due to the way the files were structured in singular months by the Police API for actual crime. Combining files would have been required and then rendering further heat-maps would have resulted in an untidy implementation and unrealistic deliverables within the one-semester time constraint.

The solution I came up with to tackle this issue was to make use of the functionality which allows you to choose the file to load in manually. For both the actual crime and Twitter a file structure was set up so each month has a file with all of the relevant data contained inside it. Therefore the user would select Jan 2014 data in the actual crime folder then select all of the crime type files in Jan 2014 folder for Twitter crime. Although this is a manual operation it does provide a workable solution to this issue. Meaningful file and folder names are very important in this solution so it is clear to user the files which they are loading in.

4.4.3 Crime Type Colour Gradient Choice



	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	☑	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 16: Example Criminal Damage Heat-Map without default starting colour 'rgba (0,255,255,0)'

A difficulty encountered during implementation was creating enough RGBA colour gradients to make each of the 12 crime types unique. For the 'all crime' heat-maps I was able to persevere with the green to yellow to red gradient which is the default for the Google Maps API because it shows the all up comparison. The unique colours were required in order to be able to overlay more than one crime type heat-map so the user could see the interlinking of certain crimes in specific areas of London.

To rectify this using a CSS gradient generator, I had to resort to using multiple shades of certain colours. For example, using a light blue to dark blue and a light turquoise to a slightly darker version of turquoise. Some colours such as yellow do not have the flexibility to be able to use multiple shades because it is not as easy on the eye or the different shades are not as distinguishable.

In addition I discovered each colour gradient had to have a default starting value of 'rgba(0,255,255,0)' which gives a near transparent colour. Without this being specified for each crime type then the whole map is covered in a light shade of the colour relating to that crime type such as in Figure 16 above.

5 Results

The software environment produced has reached a successful outcome to help meet the aims of this project. The web-based implementation provides an interesting month-by-month visual comparison of the time period January to April 2014 for actual crime and crime references on Twitter by crime type. The numbers behind the visual representation form a solid foundation for the Pearson chi-squared statistical test which will look for correlation between the two data sets.

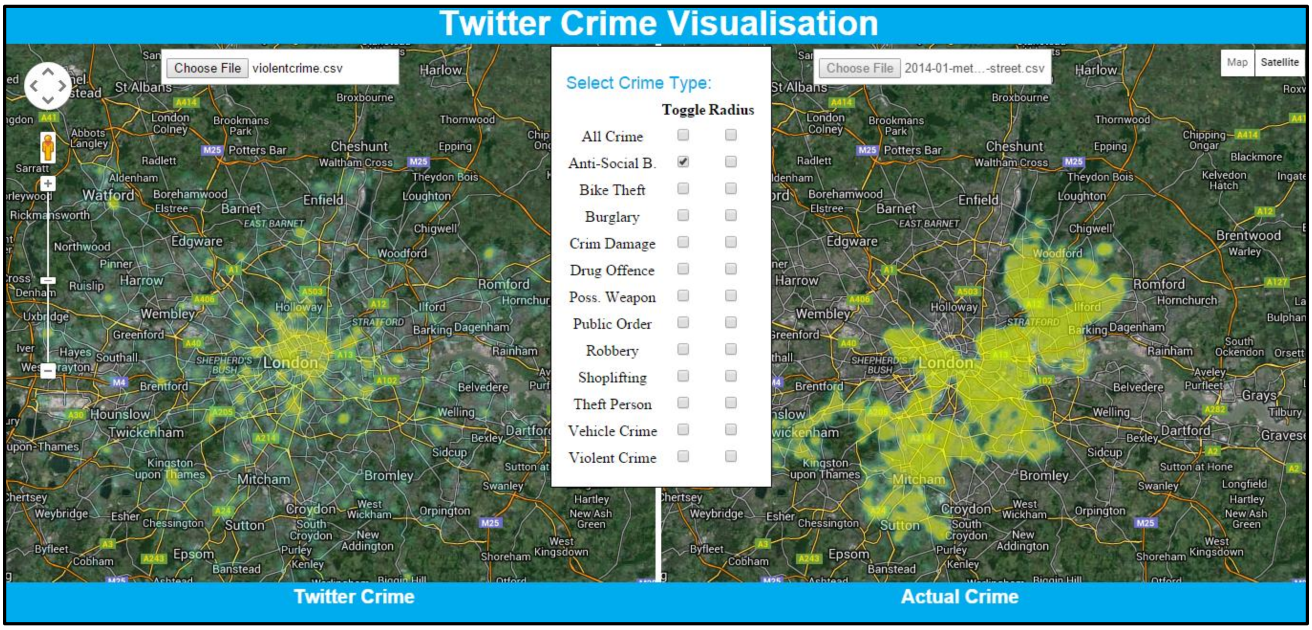
5.1 Testing Visualisations by Crime Type

The system which has been developed is able to show a heat-map visualisation comparison of Twitter crime against actual crime for each crime type specified in the Police API. In this sub-section only January 2014 crime hotspots will be analysed. Further example heat-maps for each crime type relating to February, March and April 2014 data can be located in Appendix A.

The main purpose of this sub-section is to test and prove that the system is able to render heat-map examples for each crime type although a concise analysis will be given of the resulting crime hotspots based on the heat-maps produced for January 2014.

In the visualisations below in accordance with the labelling on the bottom information bar of the interface, Crime references on Twitter is shown on the map on the left and actual crime on the right-hand side.

5.1.1 Anti-Social Behaviour



violentcrime.csv

2014-01-met...-street.csv

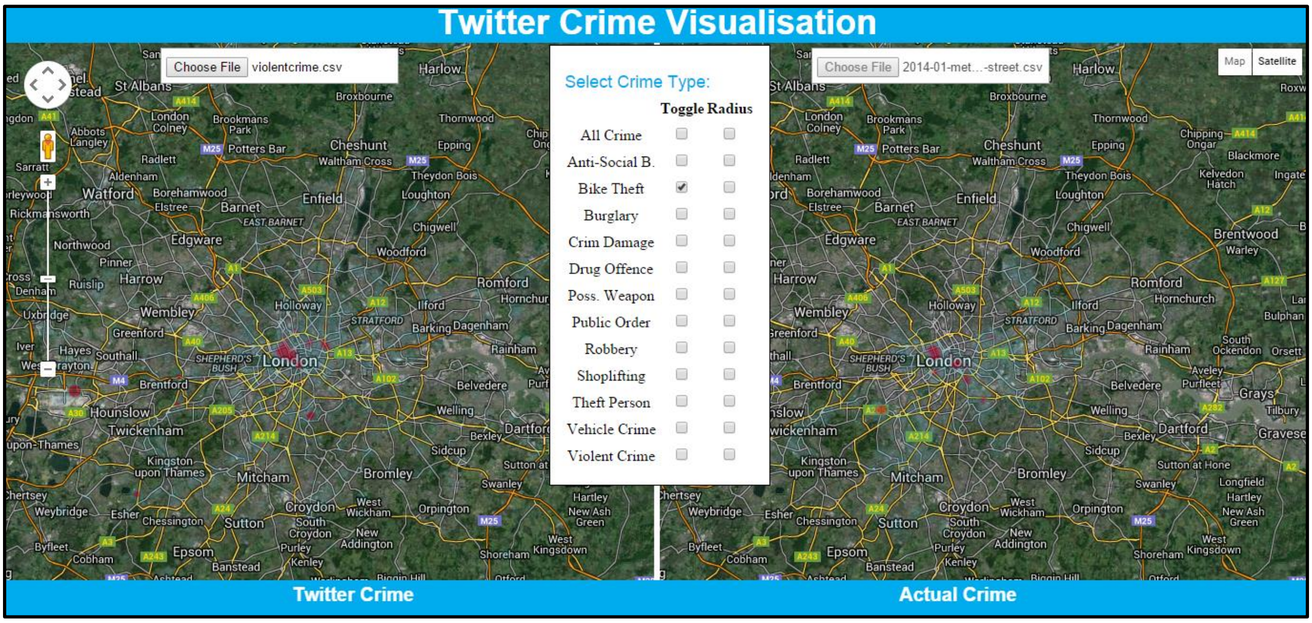
Select Crime Type:

	Toggle	Radius
All Crime	□	□
Anti-Social B.	☑	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 17: Anti-Social Behaviour Heat-maps for January 2014

Anti-Social Behaviour is one the larger crime types in regards to actual crime. The visualisation shows that the actual crime heat-map is much more densely populated in comparison with the Twitter crime heat-map. Actual crime shows high density of crime across a large spatial area whereas Twitter crime is particularly concentrated on central London. It is possible that this is due to there being more people in the central London area therefore anti-social behaviour is more likely to happen and be witnessed by people to tweet in reference to it.

5.1.2 Bicycle Theft



Choose File violentcrime.csv

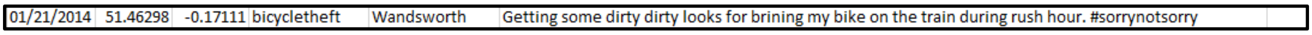
Choose File 2014-01-met…-street.csv

Select Crime Type:

	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	☑	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 18: Bicycle Theft Heat-maps for January 2014

As a very limited crime type it was fairly difficult to choose references to bicycle theft. The references to this crime could have been inflated due to the presence of the word "bike" standalone in the search array designed to create the CSV file for this particular category. Therefore a number of references could be inaccurate. Figure 19 shows an example tweet which is extracted from the raw data of the Bicycle Theft Twitter crime CSV file for January 2014. The tweet shows how the word 'bike' has been picked up in a non-crime context which is inflating the number of locations used to plot the heat-map.

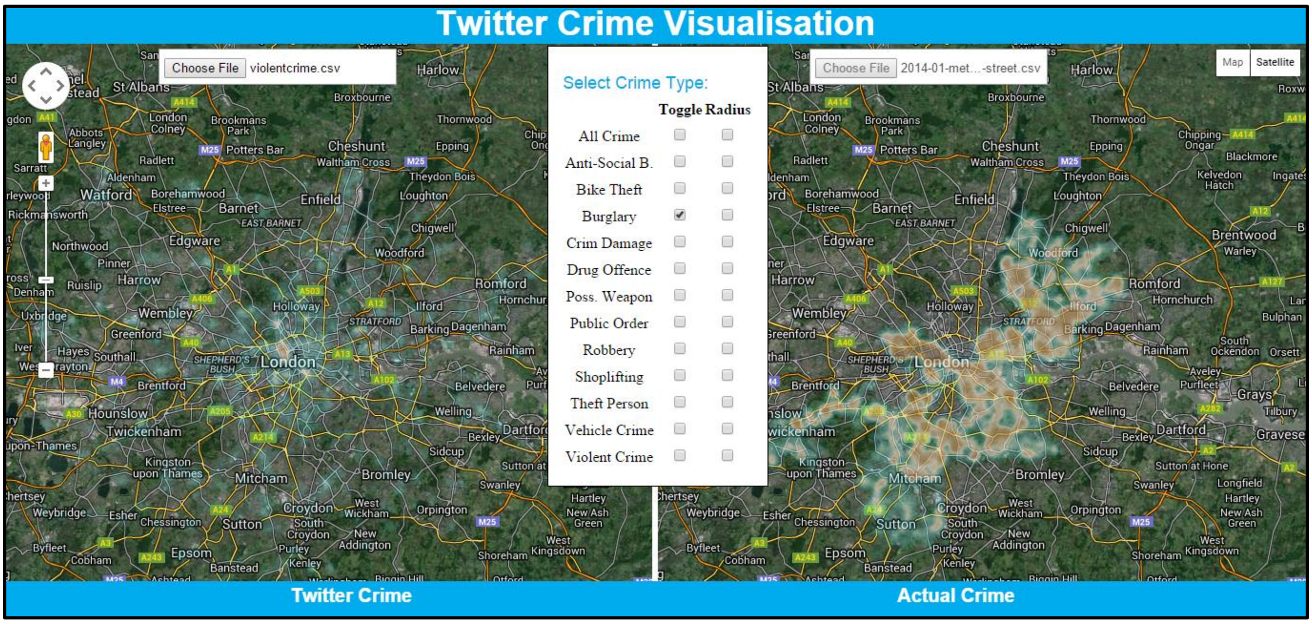


01/21/2014	51.46298	-0.17111	bicycletheft	Wandsworth	Getting some dirty dirty looks for brining my bike on the train during rush hour. #sorrynotsorry

Figure 19: Example plain text of a tweet where the word 'bike' is used in a non-crime context

According to results published by the Office of National Statistics during the same time period which this project visualises data for, "In London the number of people cycling to work [has] more than doubled in 10 years" [27]. Coupled with the inception of the congestion charge many commuters are choosing to use their bicycle to get to work in the central part of the city. Therefore small pockets of bicycle theft hotspots around the central London area are to be expected because more presence of pedal bicycles means an increase in the theft of them is likely. Hence the heat-maps for both data sets show a high concentration of colour around the central London zone.

5.1.3 Burglary


Twitter Crime Visualisation
Choose File violentcrime.csv
Choose File 2014-01-met…-street.csv
Map
Satellite
Twitter Crime
Actual Crime

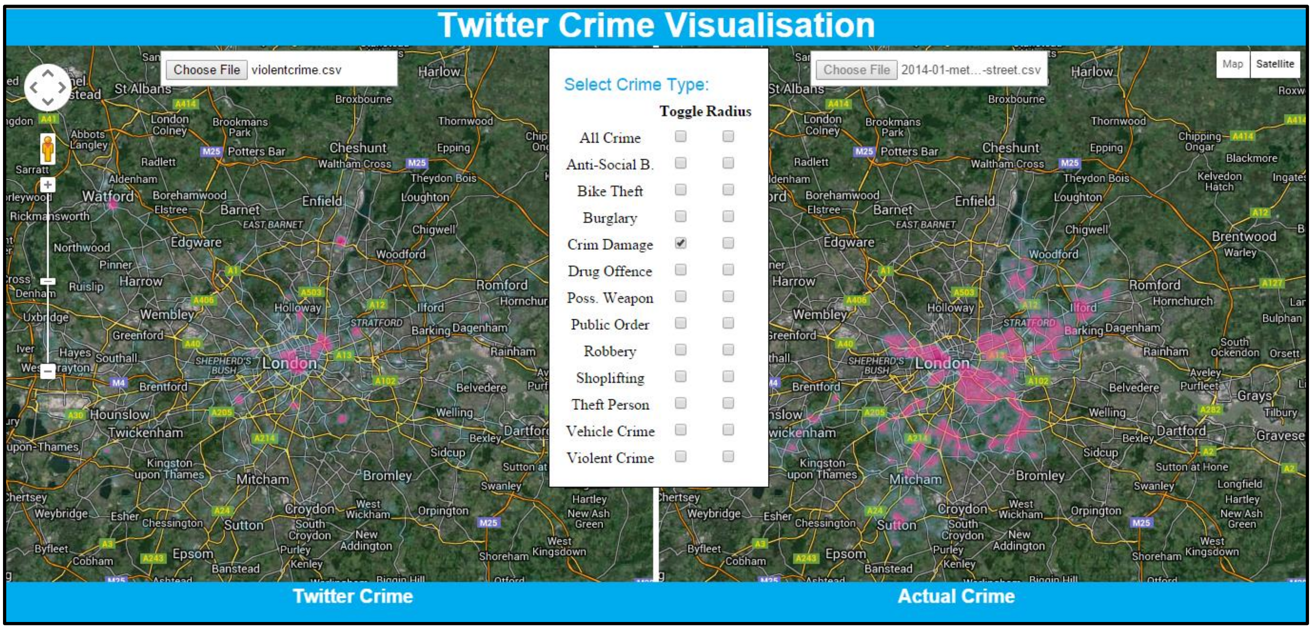
Select Crime Type:

		Toggle Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	☑	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 20: Burglary Heat-maps for January 2014

Burglary does not seem to be something which is widely tweeted in reference to judging by the heat-map rendered from Twitter. The actual crime heat-map has high concentrations of the brown colour gradient across much of London meaning Burglary is a problem for the Metropolitan Police. A potential reason for the lack of references on Twitter could be that burglary relates to illegally entering a private dwelling which is likely to be a house or business premises. The victims in this situation are unlikely to tweet about being burgled so not to broadcast the insecurity of their property. Further to this it is even less likely the criminal would tweet about burglary so to avoid being caught by the police for committing the crime.

5.1.4 Criminal Damage


Twitter Crime Visualisation

Choose File violentcrime.csv

Choose File 2014-01-met…-street.csv

Select Crime Type:

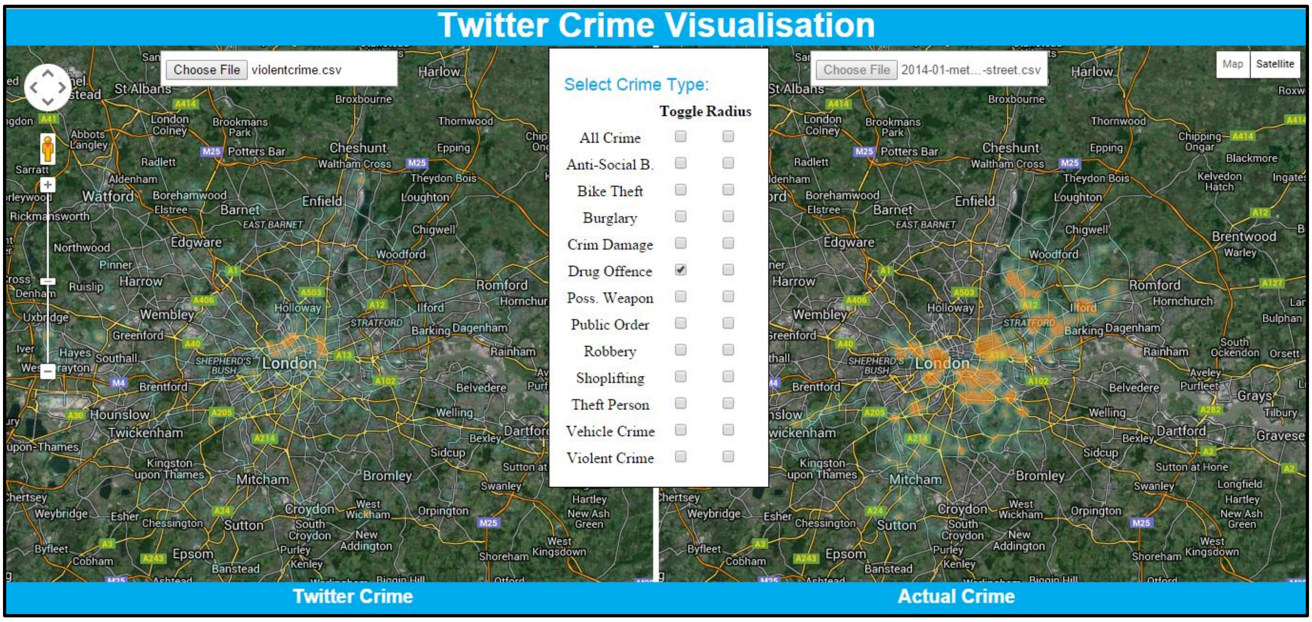
Toggle Radius

All Crime □ □
Anti-Social B. □ □
Bike Theft □ □
Burglary □ □
Crim Damage ☑ □
Drug Offence □ □
Poss. Weapon □ □
Public Order □ □
Robbery □ □
Shoplifting □ □
Theft Person □ □
Vehicle Crime □ □
Violent Crime □ □

Figure 21: Criminal Damage Heat-maps for January 2014

Criminal Damage is rife right across London for January 2014 according to the actual crime heat-map. However I would have expected a more densely populated heat-map for Twitter crime especially with this crime type covering crimes such as vandalism & graffiti which are present in all areas of the UK. Central London seems to be a prime hotspot again for this crime type with the 'mobile' population likely to tweet about such crimes which are often easily seen by the naked eye.

5.1.5 Drug Offences



Twitter Crime Visualisation

Choose File violentcrime.csv

Choose File 2014-01-met…-street.csv

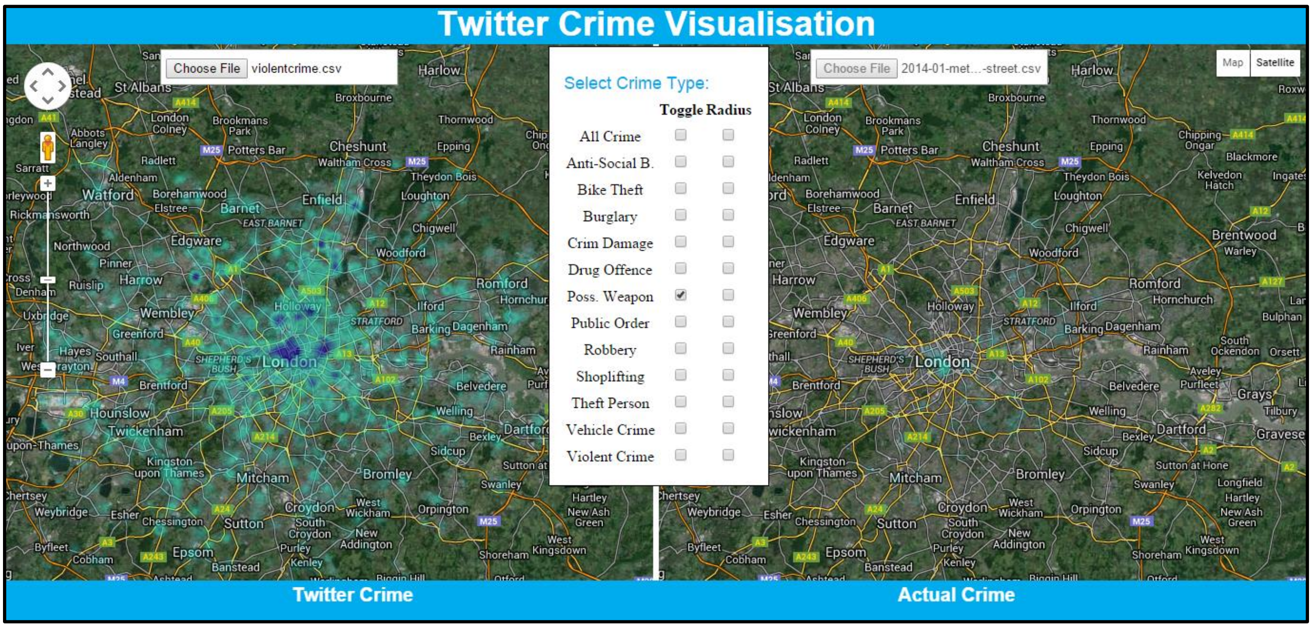
Select Crime Type:

	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	☑	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 22: Drug Offences Heat-maps for January 2014

Higher amounts of references to drugs on Twitter are almost entirely in Central London although actual crime would suggest it is a London-wide problem. More inhabitants, increased disposable income available to buy drugs and the nightclub scene are factors which could have led to the increase in Twitter references around the central area of London.

5.1.6 Possession of Weapons



Twitter Crime Visualisation

Choose File violentcrime.csv

Choose File 2014-01-met...-street.csv

Select Crime Type:

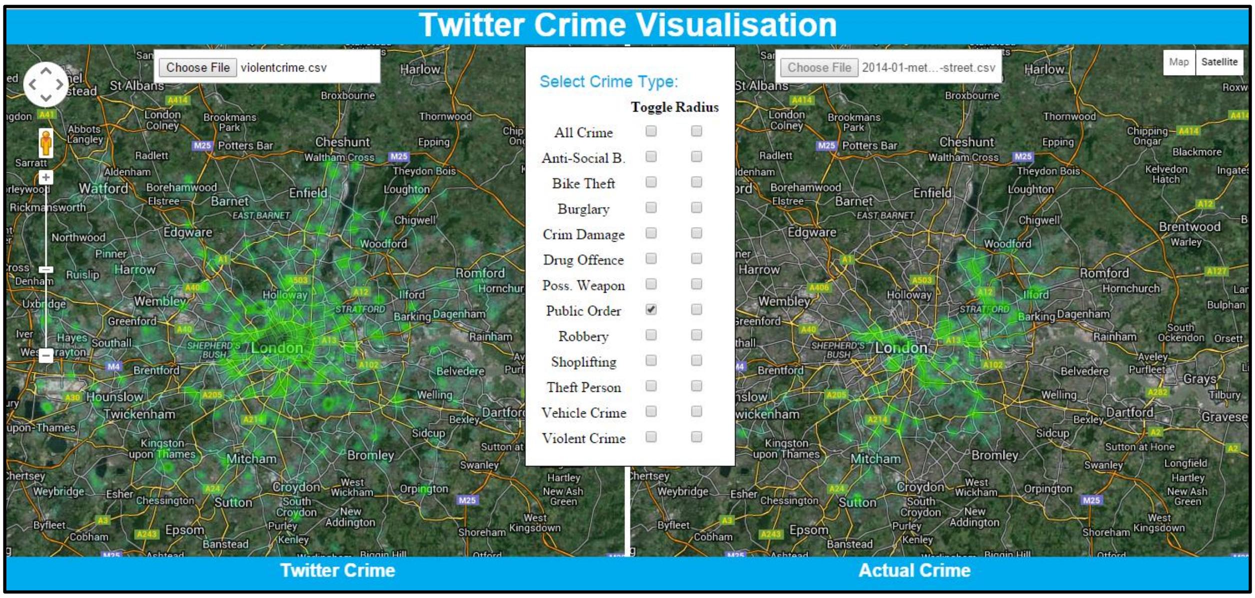
	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	☑	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 23: Possession of Weapons Heat-maps for January 2014

The dense heat-map rendered from the January 2014 Twitter CSV file for Possession of Weapons could be the results of increased coverage of weapons in every day communications. In video games, TV programmes and in the media you see regular references to weapons which means people are more likely to tweet about them but not in the context of committing a crime.

The lack of actual crime can be explained that often a weapon can be used in a crime but the main crime committed seems to be what the criminal is convicted for as opposed to possession of weapons. For example, an Armed Robbery will involve weapons but it is the Robbery offence which is likely to be priority for prosecutors.

5.1.7 Public Order

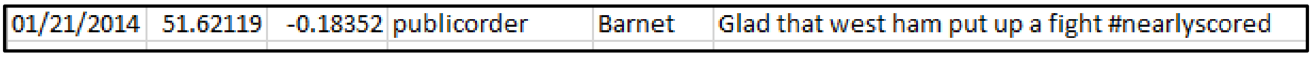

Twitter Crime Visualisation
Choose File violentcrime.csv
Choose File 2014-01-met...-street.csv
Map
Satellite
Twitter Crime
Actual Crime

Select Crime Type:

	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	☑	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 24: Public Order Heat-maps for January 2014

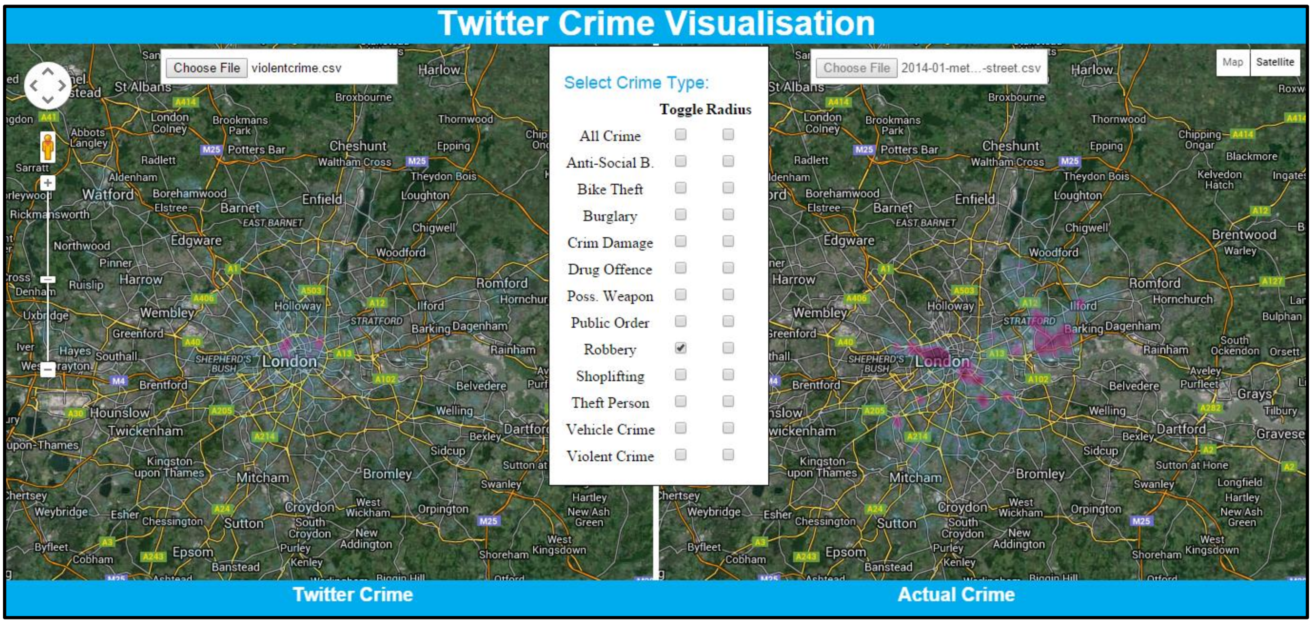
Public Order offences can often occur at large protests or riots which tend to get heavy coverage in the media. This would lead to an increase in tweeting about such events and the crimes committed during them because it provides many talking points. Central London is where the UK Government are based and therefore a likely place for individuals and groups to protest hence the intense green colour in that area on both maps. Twitter references relating to public order are also likely to occur relating to sporting events such as football matches which take place across London. The presence of well-known clubs Chelsea, Tottenham, West Ham and Arsenal amongst others leads to an increase in Twitter public order references with intense rivalry between supporters. Although football has always had links with hooliganism there are examples where the crime-related terms I have chosen are used in a different context. In Figure 25 below ‘fight’ has been used but in a football term to describe West Ham’s battling qualities as opposed to a violent fight which would likely carry a public order offence.



01/21/2014	51.62119	-0.18352	publicorder	Barnet	Glad that west ham put up a fight #nearlyscored

Figure 25: Plain text of a Tweet which uses 'fight' in a different context

5.1.8 Robbery



Select Crime Type:

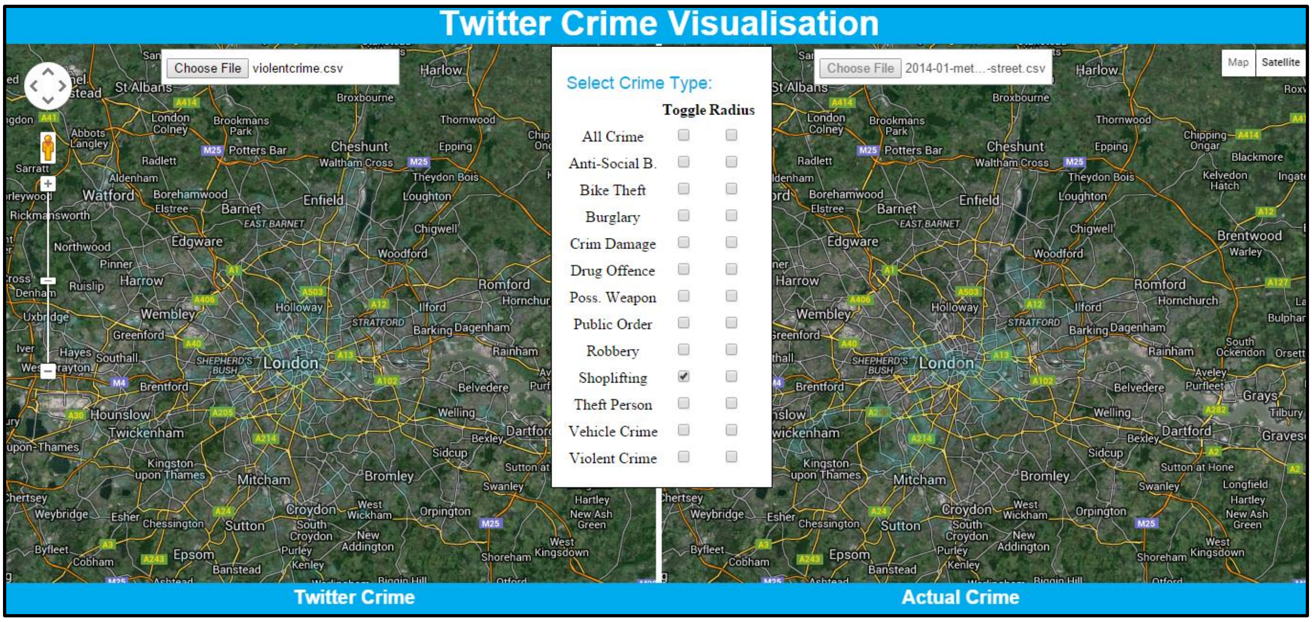
		Toggle Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	☑	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 26: Robbery Heat-maps for January 2014

The actual crime and Twitter crime heat-map for Robbery shows a particular hotspot in Mayfair. In addition to the areas' already extremely affluent reputation, Camilla Turner writes in The Telegraph that "private equity firms, boutique investment banks and hedge funds are descending on London's most exclusive district, with their presence more than doubling in the space of a year" [28]. Expensive jewellers located in the district are also likely targets making Mayfair a prime area for robberies to take place.

The Twitter crime heat-map is much less packed in comparison with the actual crime map. There is only a subtle difference between Robbery and Burglary which could account for the lack of references on Twitter.

5.1.9 Shoplifting



Twitter Crime Visualisation

Choose File violentcrime.csv

Choose File 2014-01-met...-street.csv

Select Crime Type:

	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	☑	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Twitter Crime

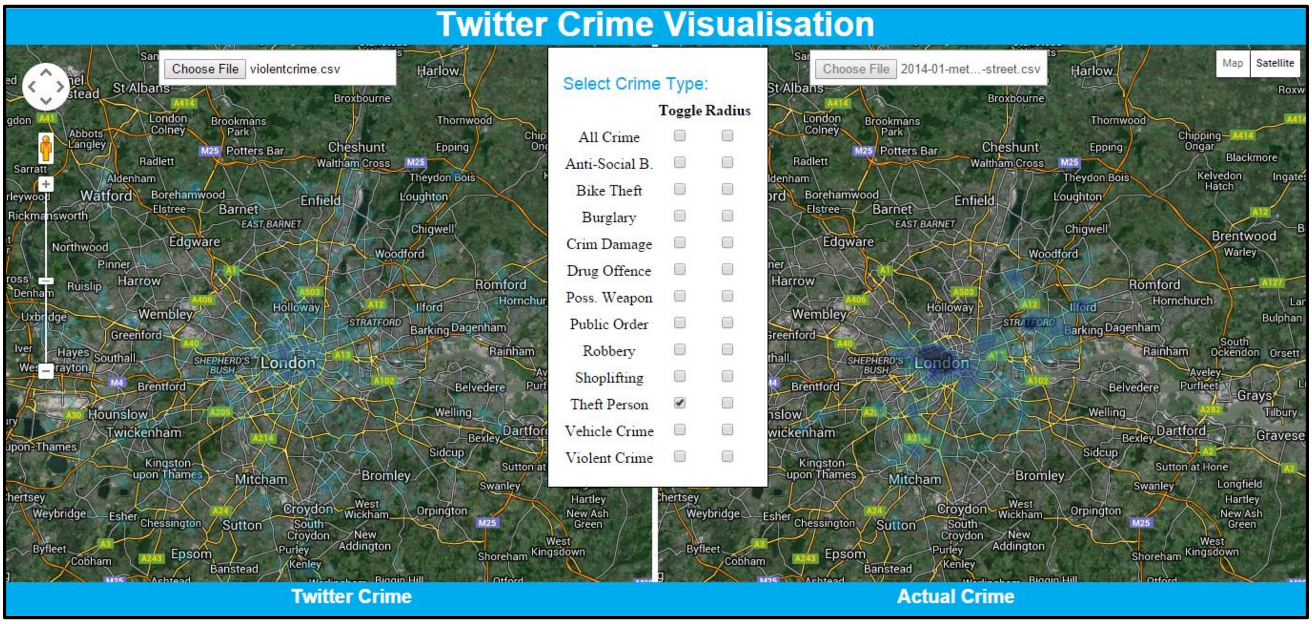
Actual Crime

Figure 27: Shoplifting Heat-maps for January 2014

The Shoplifting heat-maps for both data sets are primarily situated in Central London. Increased footfall compared with other areas of London results in busier shops and therefore shoplifting becomes easier. Consequently an increase in security for retail businesses is required meaning shoplifters are more likely to be caught in the act.

The lack of data points in both heat-maps suggest shoplifting is not a major problem or if it is the retailers tend to catch the shoplifter without involving the police. The lack of actual shoplifting offences suggests people are less prone to tweet in reference to it too.

5.1.10 Theft from the Person



Select Crime Type:

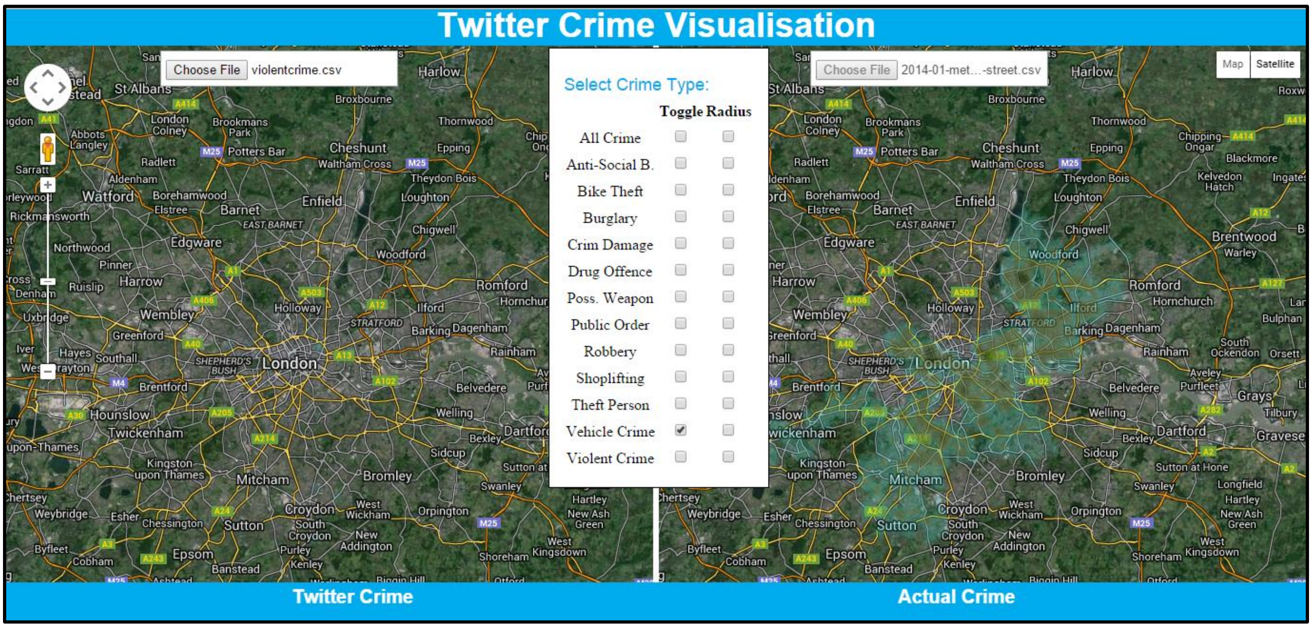
	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	☑	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 28: Theft from the Person Heat-maps for January 2014

Actual Theft from the person crimes are heavily based in Central London. High numbers of people on the street carrying valuable items in areas such as Westminster make it much easier for Pickpockets to commit crime.

There are no intense hotspots derived from the Twitter data for theft from the person references. However there are slighter darker shades of blue around a similar area to the obvious hotspot on the actual crime map which is around the Westminster area.

5.1.11 Vehicle Crime


Twitter Crime Visualisation
Choose File violentcrime.csv
Choose File 2014-01-met…-street.csv
Map
Satellite
London
Twitter Crime
Actual Crime

Select Crime Type:

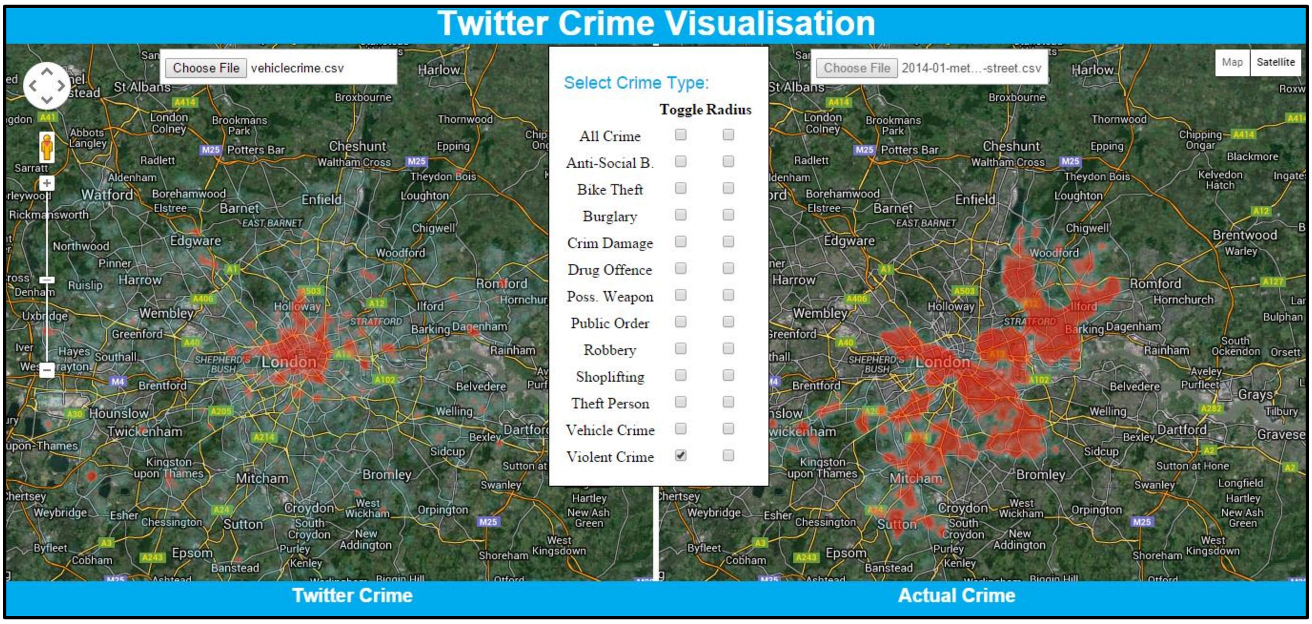
	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	☑	□
Violent Crime	□	□

Figure 29: Vehicle Crime Heat-maps for January 2014

Vehicle Crime was a particularly difficult crime type to identify crime-related keywords for. Despite choosing many references with different meanings of the same words and phrases there was very few references to plot onto the map.

The actual crime heat-map depicts that vehicle crime has pockets of hotspots across London. This shows that the references used to search the Twitter data are not effective or people are not tweeting in relation to vehicle crime.

5.1.12 Violent Crime and Sexual Offences


Twitter Crime Visualisation

Choose File vehiclecrime.csv

Choose File 2014-01-met...-street.csv

Select Crime Type:

		Toggle Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	☑	□

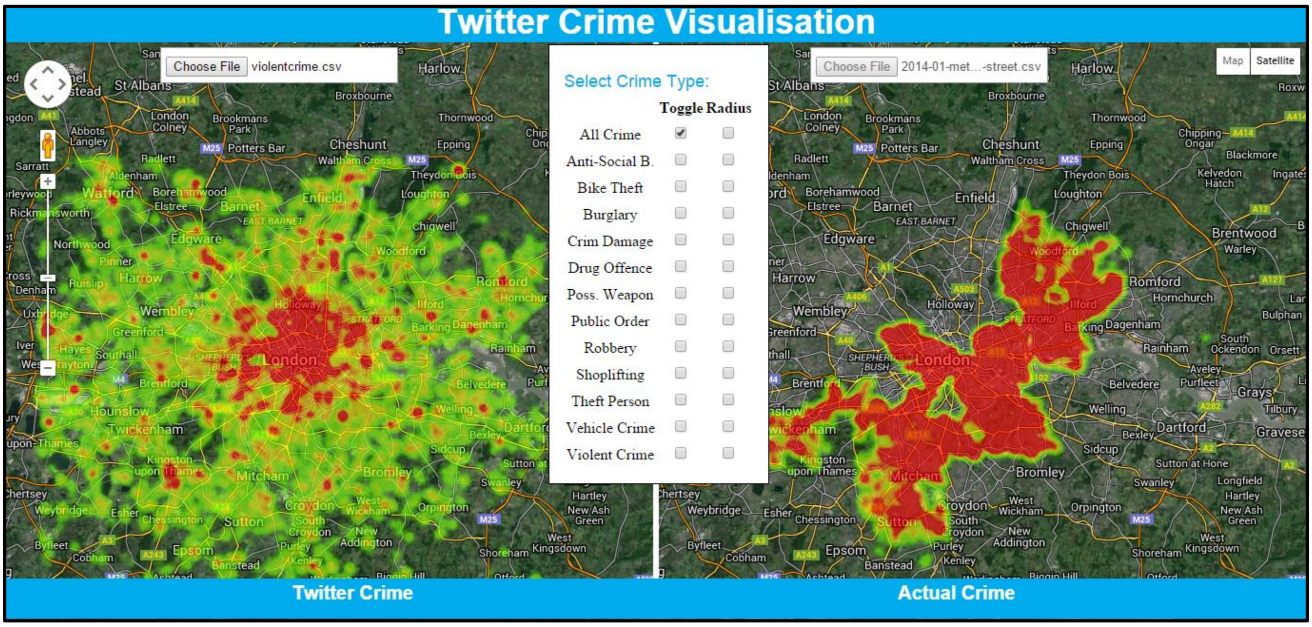
Figure 30: Violent Crime Heat-maps for January 2014

From the Twitter crime heat-map we can see many references to violent crime and sexual offences were picked up. The hotspots created from these references seem to be concentrated around the central area of London in addition to some territory occupied in North London. Violent Crime is generally quite shocking and provokes reaction from humans therefore people may feel compelled to post about it on social media.

Actual crime is also interesting in this case because there are large hotspots not only in Central London but in the North East of the City. This is a fairly large category in terms of what crimes are committed to fall within it therefore a widespread heat-map such as the one rendered is expected.

It is possible that some data could be missing from this crime type because the police or victims in question do not wish for crimes of this nature to be categorised and therefore they fall into the ‘other crime’ type. ‘Other crime’ is only visualised as part of the ‘all crime’ heat-map for actual crime.

5.1.13 All Crime



Twitter Crime Visualisation

Choose File violentcrime.csv

Choose File 2014-01-met...-street.csv

Select Crime Type:

		Toggle Radius
All Crime	☑	☐
Anti-Social B.	☐	☐
Bike Theft	☐	☐
Burglary	☐	☐
Crim Damage	☐	☐
Drug Offence	☐	☐
Poss. Weapon	☐	☐
Public Order	☐	☐
Robbery	☐	☐
Shoplifting	☐	☐
Theft Person	☐	☐
Vehicle Crime	☐	☐
Violent Crime	☐	☐

Figure 31: Heat-maps for All Crime - January 2014

When all the references used to create search arrays for each crime type in Twitter data are combined into one large array the ‘All crime’ heat-maps are the final outcome. From this it can be seen that actual crime is very dense London-wide denoted by the intense red colour in contrast to Twitter crime which is heavily referenced primarily in Central London. It is worth noting that the actual crime data also plots data for ‘other crime’ and ‘other theft’ which have not been included as part of the crime type framework because this would mean result in guesswork as to which crime types these fall into.

A further conclusion which can be drawn from this visualisation is the subtle difference in the area covered by the two data sets. The actual crime data does not cover as wider spatial area as the Twitter data so the comparison between them will not be as accurate. By conducting further tests it can help to understand whether this could be a problem with parsing the Police API data.

5.2 Data Set Spatial Area Discrepancy Testing

The ‘All crime’ maps show the full monthly data sets compared against each other side by side. While viewing this visualisation it is easy to see the discrepancy between the two areas which the data sets cover. When visualised the actual crime data covers a lot smaller spatial area than the area covered by the Twitter data sourced from the Cardiff School of Computer Science and Informatics. However, the below Figure 32 shows that the London Metropolitan Police area covers almost exactly the same spatial area as the Twitter data. It is extremely unlikely that there would be no actual crimes committed in the additional spatial area occupied by the Twitter crime as opposed to the actual crime heat-maps.

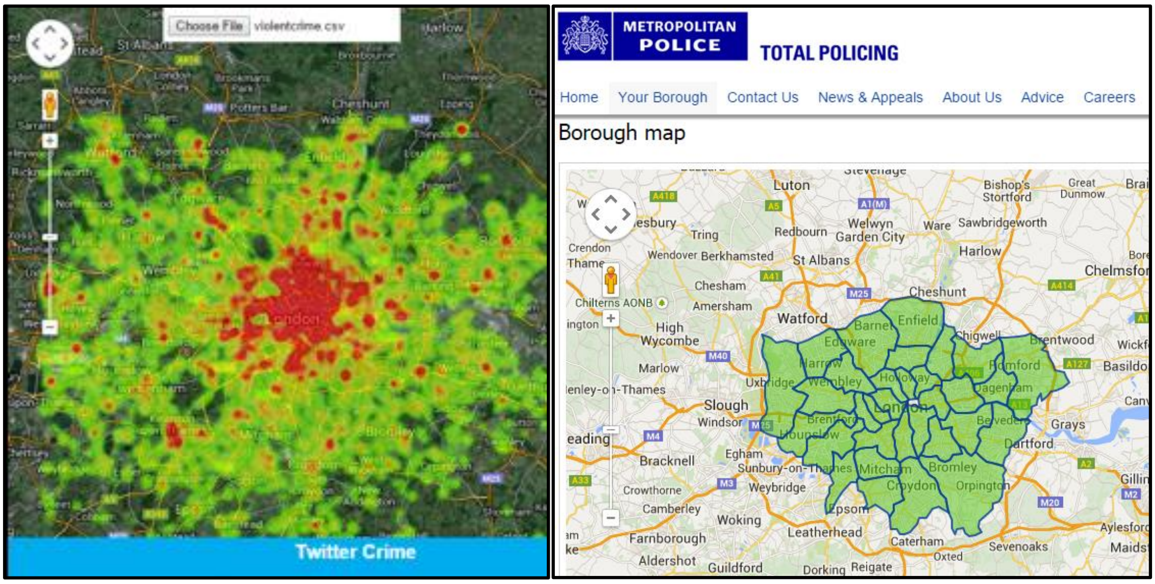


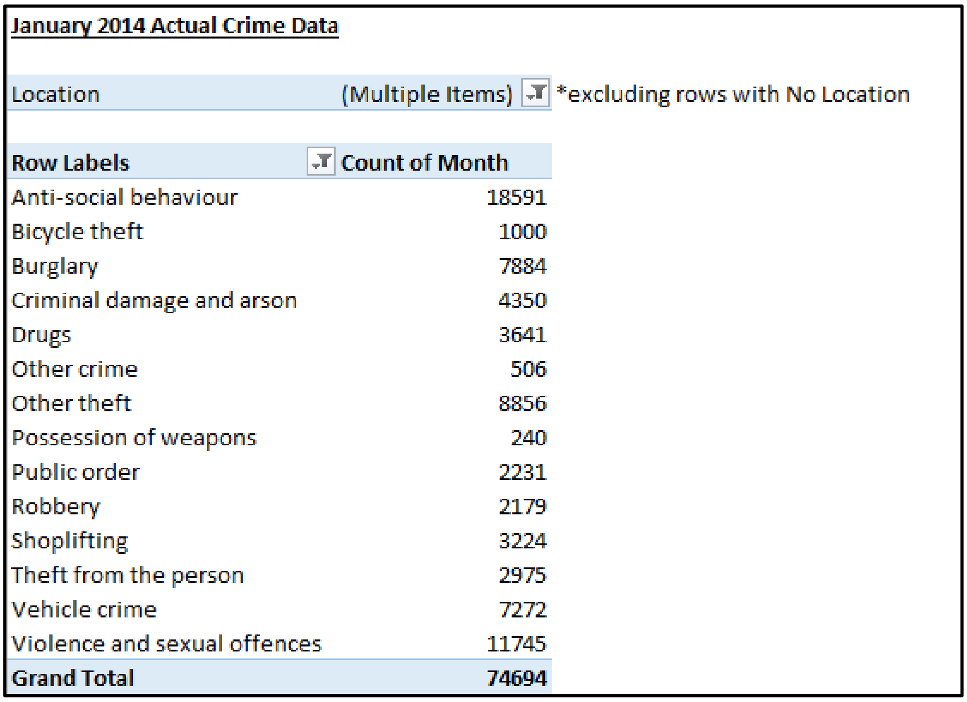
Figure 32: (Left) All Crime Map Twitter Crime for Jan 2014, (Right) London Metropolitan Police Coverage Map [29]

To test to see whether this is due to a problem parsing the data a series of tests will be conducted to compare the numbers in the monthly, actual crime data files against those parsed into the developer console in a web browser from the same files. The test will use January and February 2014 data in order to measure across a greater sample and prove that if there is a problem that it is not month specific. In addition, to confirm the Twitter data files are being parsed correctly a similar test will be performed.

5.2.1 Conducting Data Parsing Tests

<u>Actual Crime Parsing Tests</u>

- January 2014

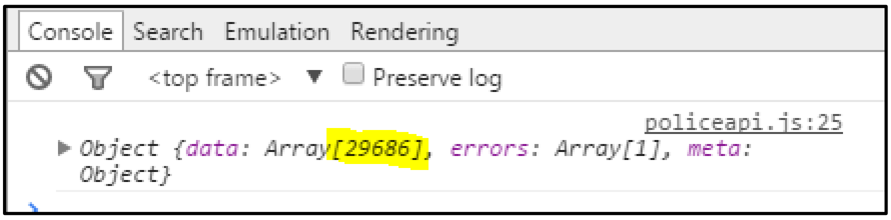


January 2014 Actual Crime Data

Location (Multiple Items) *excluding rows with No Location

Row Labels	Count of Month
Anti-social behaviour	18591
Bicycle theft	1000
Burglary	7884
Criminal damage and arson	4350
Drugs	3641
Other crime	506
Other theft	8856
Possession of weapons	240
Public order	2231
Robbery	2179
Shoplifting	3224
Theft from the person	2975
Vehicle crime	7272
Violence and sexual offences	11745
Grand Total	**74694**

Figure 33: January 2014 Actual Crime CSV Data by Crime Type



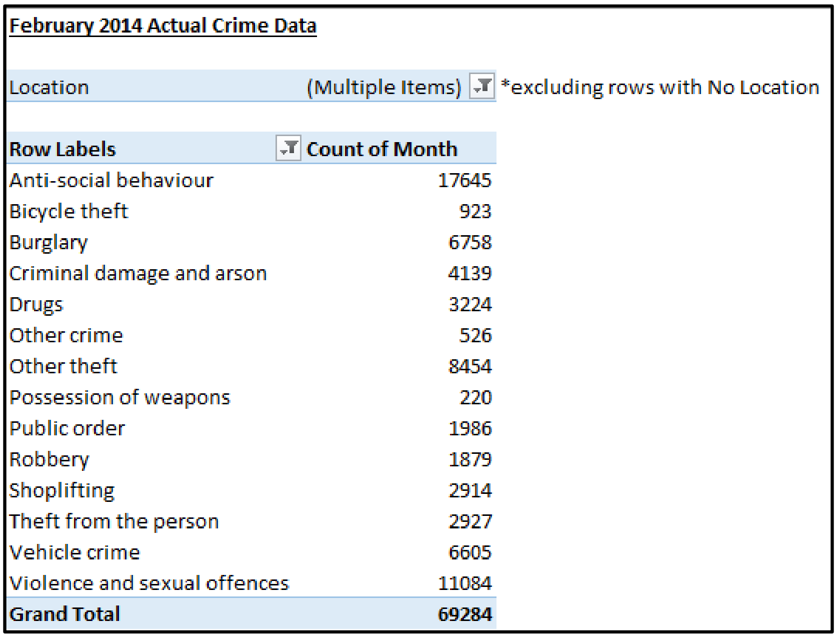
```
                                             policeapi.js:25
▶ Object {data: Array[29686], errors: Array[1], meta:
  Object}
```

Figure 34: January 2014 Actual Crime - Rows of CSV parsed to Browser (All Crime)

Crime Type	Data Points in CSV File	Data Points parsed to be plotted on Heat-Map	Difference (No of Data Points in CSV File – Parse Data Points)
Anti-Social Behaviour	18591	7337	11254
Bicycle Theft	1000	444	556
Burglary	7884	2914	4900
Criminal Damage and arson	4350	1602	2748
Drug Offences	3641	1457	2184
Possession of Weapons	240	98	142
Public Order	2231	884	1347
Robbery	2179	1123	1056
Shoplifting	3224	1136	2088
Theft from the Person	2975	1407	1568
Vehicle Crime	7272	2860	4412
Violent Crime and Sexual offences	11745	4556	6919
All Crime (inc. Other Crime and Other Theft)	74694	29686	45008

Table 3: January 2014 Actual Crime- Number of Data Points in CSV file vs Data Points parsed to browser

- February 2014

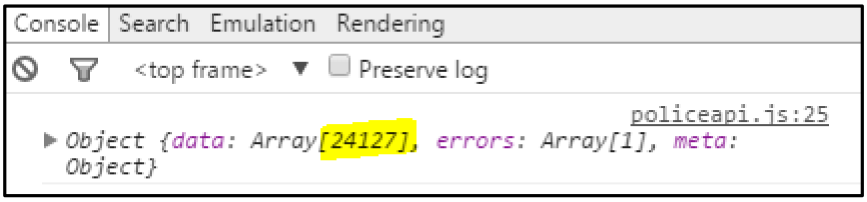


February 2014 Actual Crime Data

Location	(Multiple Items)	*excluding rows with No Location

Row Labels	Count of Month
Anti-social behaviour	17645
Bicycle theft	923
Burglary	6758
Criminal damage and arson	4139
Drugs	3224
Other crime	526
Other theft	8454
Possession of weapons	220
Public order	1986
Robbery	1879
Shoplifting	2914
Theft from the person	2927
Vehicle crime	6605
Violence and sexual offences	11084
Grand Total	**69284**

Figure 35: February 2014 Actual Crime CSV Data by Crime Type



```
                                              policeapi.js:25
▶ Object {data: Array[24127], errors: Array[1], meta:
  Object}
```

Figure 36: February 2014 Actual Crime - Rows of CSV parsed to Browser (All Crime)

Crime Type	Data Points in CSV File	Data Points parsed to be plotted on Heat-Map	Difference (No of Data Points in CSV File – Parse Data Points)
Anti-Social Behaviour	17465	6036	11429
Bicycle Theft	923	383	540
Burglary	6758	2095	4663
Criminal Damage and arson	4139	1310	2829
Drug Offences	3224	1122	2102
Possession of Weapons	220	78	142
Public Order	1986	632	1354
Robbery	1879	845	1034
Shoplifting	2914	958	1956
Theft from the Person	2927	1244	1683
Vehicle Crime	6605	2277	4328
Violent Crime and Sexual offences	11084	3740	8064
All Crime (inc. Other Crime and Other Theft)	69824	24127	45697

Table 4: February 2014 Actual Crime- Number of Data Points in CSV file vs Data Points parsed to browser

Twitter Crime Parsing Tests

To ensure the CSV files for Twitter Crime are also being parsed to the browser correctly, the same test has been conducted on each of crime type files for the same time period.

- January 2014

Crime Type	Number of Data Points in CSV file	Number of Data Points parsed by the browser	Difference (to number of data points in CSV)
Anti-Social Behaviour	2274	2274	0
Bike Theft	712	712	0
Burglary	854	854	0
Criminal Damage	784	784	0
Drug Offences	831	831	0
Possession of Weapons	3304	3304	0
Public Order	2625	2625	0
Robbery	829	829	0
Shoplifting	652	652	0
Theft from the Person	952	952	0
Vehicle Crime	15	15	0
Violent Crime	2569	2569	0
All Crime	16401	16401	0

Table 5: Twitter Crime References January 2014- Number of Data Points in File vs Number of Data points parsed into web browser

- February 2014

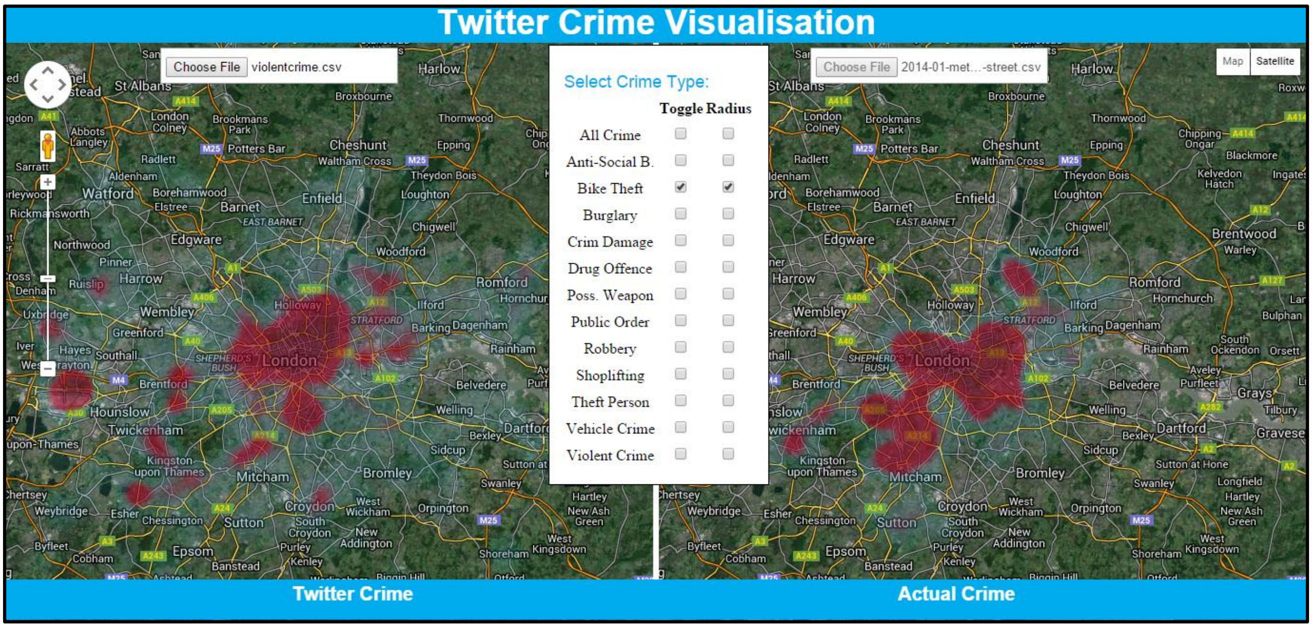
Crime Type	Number of Data Points in CSV file	Number of Data Points parsed by the browser	Difference (to number of data points in CSV)
Anti-Social Behaviour	6098	6098	0
Bike Theft	2187	2187	0
Burglary	1832	1832	0
Criminal Damage	2147	2147	0
Drug Offences	1897	1897	0
Possession of Weapons	7905	7905	0
Public Order	6249	6249	0
Robbery	1709	1709	0
Shoplifting	1522	1522	0
Theft from the Person	1878	1878	0
Vehicle Crime	55	55	0
Violent Crime	6067	6067	0
All Crime	39546	39546	0

Table 6: Twitter Crime References February 2014- Number of Data Points in File vs Number of Data points parsed into web browser

The above tests conclude that there is an issue parsing the actual crime CSV data into the browser to be plotted onto a heat-map. While many rows are parsed in correctly and plotted, there are over 60% of rows missing per month. The Twitter Crime CSV files for each crime type parse into the browser correctly but are much smaller so there is a performance issue with the actual crime files here. The likely cause of the issue with the actual crime data which is downloaded from the Police API is that some rows of data may not be structured correctly to be parsed by the library, PapaParse.

If I were to review the way I approached handling the parsing of actual crime data I would conduct earlier testing during implementation to check the numbers were being correctly parsed. A solution in the long-term of the project would be to use an alternative CSV parser or produce separate, smaller CSV files for each crime type in similar fashion to the way the crime references on Twitter are evaluated.

5.3 Customisation - Change Radius


Twitter Crime Visualisation
Choose File violentcrime.csv
Choose File 2014-01-met…-street.csv
Map
Satellite
London
Twitter Crime
Actual Crime

Select Crime Type:

	Toggle	Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	☑	☑
Burglary	□	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	□	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 37: Bicycle Theft Heat-Maps with Changed Radius of Influence (January 2014)

Using the customisation available in the Google Maps API, a feature discussed in the project's design section has been implemented. For each crime type visualisation the user is able to change the radius of influence relating to each data point.

By toggling the 'Radius' menu checkbox to change the radius of the heat-map for a particular crime type it helps the user to more clearly visualise hotspots of criminal activity. In Figure 37 above for example, we can more easily identify the similarity and differences of hotpots for January 2014 bicycle theft in the London area by changing the radius.

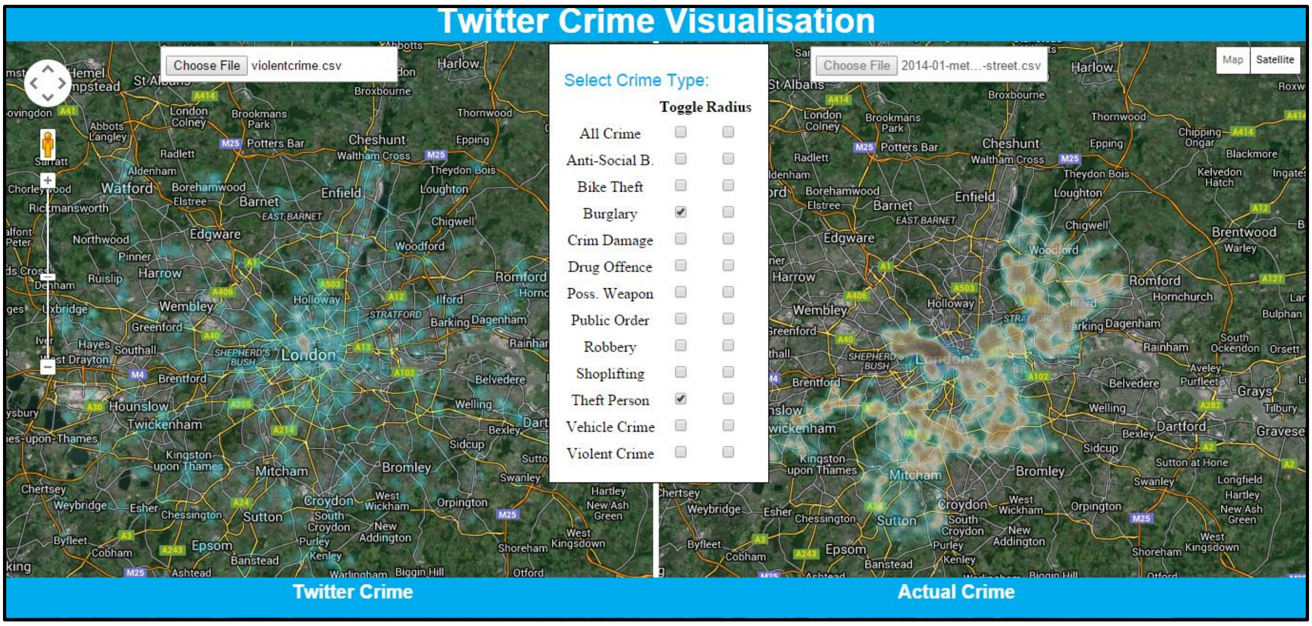
This is a usability feature helping to make the interface more personalised to the user because they are able to view the visualisations in a way which may be more appropriate to them if required. This tool's main purpose is to display data so therefore by implementing this it offers a greater range of options to visualise the data.

5.4 Software Constraints

Despite being able to implement a fully working prototype the system also has its constraints. In the short time frame available for this project it was important to identify realistic deliverables. Therefore it is difficult to implement other potentially useful features meaning the system will have some constraints.

5.4.1 Multiple Heat-Map Overlaying

When loading in the CSV files for both data sets the heat-map layers for each crime type overlay each other. This means that sometimes you are unable to see one heat-map below another because the colours do not mix effectively. Further viewing difficulty occurs when three or more crime type heat-maps of different colours are mixed. A method to avoid this occurring would be to put a limit on the number of heat-maps you can overlay at the same time. An effective limit to suggest based on my observations would be only being able to select two crime type heat-maps at any one time.



Twitter Crime Visualisation

Choose File violentcrime.csv

Choose File 2014-01-met…-street.csv

Select Crime Type:

		Toggle Radius
All Crime	□	□
Anti-Social B.	□	□
Bike Theft	□	□
Burglary	☑	□
Crim Damage	□	□
Drug Offence	□	□
Poss. Weapon	□	□
Public Order	□	□
Robbery	□	□
Shoplifting	□	□
Theft Person	☑	□
Vehicle Crime	□	□
Violent Crime	□	□

Figure 38: Blue gradient Possession of Weapons overlaid by brown gradient Burglary heat-map (January 2014 data)

5.4.2 Providing Actuals

The system does not give the actual number of data points behind the heat-maps meaning that it is down to human interpretation to see where the criminal hotspots are and which set of data has more locations plotted. The tool itself is not designed to be anything more than a data visualisation aid but this is something which may enhance its value and help users to understand the data.

5.4.3 Usability Constraints

The system although easy to use can be frustrating for the user. The difference in how actual crime and Twitter crime CSV files are loaded into the system adds complexity to the system usability. When loading the actual crime files it is simply one file loaded in and all the heat-maps for each crime type load as layers on top of each other. Then the user has to load in each and every crime type file separately because of the way the monthly Twitter data is evaluated into a number of different CSV files.

At this point both maps would now have thirteen heat-map layers (twelve crime types and the all crime heat-map) with the next step being to deselect each checkbox on the main menu to display two blank maps. Once the system is at the state of two blank maps only then can the user really begin to start viewing the visualisations they desire to view. The process to get the system into this state although fairly trivial is lengthy. In order to make the system more user friendly a solution to improve the loading in and handling of the files including being able to load in multiple files for Twitter crime would be useful.

5.5 Pearson's chi-square Statistical Testing

The method used to test for correlation between crime references on Twitter is a simple Pearson chi-squared test. The actual crime numbers were sourced by pivoting on the raw data files in Microsoft Excel. The reason for this is because the inaccurate parsing of actual crime data has resulted in missing data points in the visualisation which will not be missing in the raw data. Crimes with 'No location' are excluded from the data because these would not be plotted due to no longitude and latitude values being given.

For all of the following tests we are looking to see if this hypothesis is true:
Hypothesis 0: Actual Crime correlates with Twitter Crime

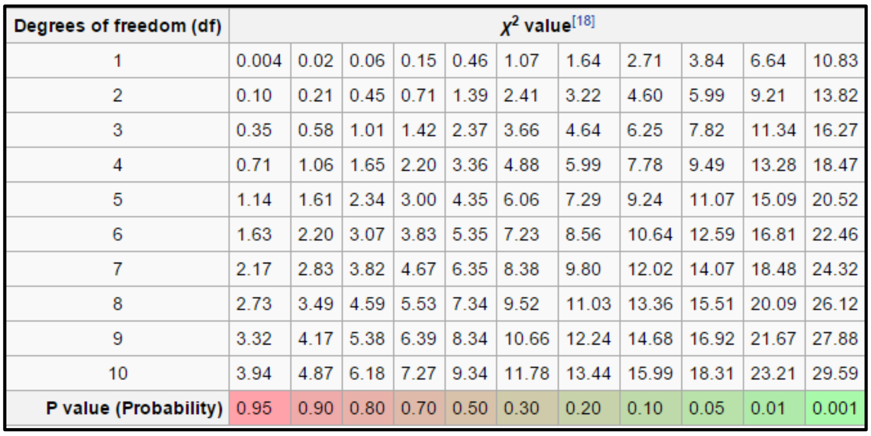
Additional information required for these calculations:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} :$$

*Mathematical Notation-

X^2 = Chi-Squared Statistic
O = Observed Frequency (Actual Crime)
E = Expected Frequency (Crime References from Twitter)
n = number of rows (number of Months)

Degrees of Freedom (Number of Months – 1) = **3**
Level of Significance = **0.05**



Degrees of freedom (df)	χ^2 value[18]										
1	0.004	0.02	0.06	0.15	0.46	1.07	1.64	2.71	3.84	6.64	10.83
2	0.10	0.21	0.45	0.71	1.39	2.41	3.22	4.60	5.99	9.21	13.82
3	0.35	0.58	1.01	1.42	2.37	3.66	4.64	6.25	7.82	11.34	16.27
4	0.71	1.06	1.65	2.20	3.36	4.88	5.99	7.78	9.49	13.28	18.47
5	1.14	1.61	2.34	3.00	4.35	6.06	7.29	9.24	11.07	15.09	20.52
6	1.63	2.20	3.07	3.83	5.35	7.23	8.56	10.64	12.59	16.81	22.46
7	2.17	2.83	3.82	4.67	6.35	8.38	9.80	12.02	14.07	18.48	24.32
8	2.73	3.49	4.59	5.53	7.34	9.52	11.03	13.36	15.51	20.09	26.12
9	3.32	4.17	5.38	6.39	8.34	10.66	12.24	14.68	16.92	21.67	27.88
10	3.94	4.87	6.18	7.27	9.34	11.78	13.44	15.99	18.31	23.21	29.59
P value (Probability)	0.95	0.90	0.80	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001

Figure 39: Table of X^2 Value [30]

I have decided that actual crime will be the observed frequency because this is what has actually happened. The expected frequency is the crime-related references from Twitter because that is our main focus in this project and what we expect to be in line with actual crime if there is correlation between the two data sets.

<u>All Crime</u>

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
January	16401	74694	**91095**
February	39456	69284	**108740**
March	30112	79200	**109312**
April	38959	74660	**113619**
Total	**124928**	**297838**	**422766**

$$= \frac{(74694-16401)^2}{16401} + \frac{(69284-39456)^2}{39456} + \frac{(79200-30112)^2}{30112} + \frac{(74660-38959)^2}{38959}$$

Sum = 342474.16
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

<u>Anti-Social Behaviour</u>

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	2274	18591	**20865**
Feb	6098	17645	**23743**
Mar	4634	20992	**25626**
Apr	5582	20647	**26229**
Total	**18588**	**77875**	**96463**

$$= \frac{(18591-2274)^2}{2274} + \frac{(17645-6098)^2}{6098} + \frac{(20992-4634)^2}{4634} + \frac{(20647-5582)^2}{5582}$$

Sum = 236672.61
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Bike Theft

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	712	1000	**1712**
Feb	2187	923	**3110**
Mar	1688	1318	**3006**
Apr	2453	1439	**3892**
Total	**7040**	**4680**	**11720**

$$= \frac{(1000-712)^2}{712} + \frac{(923-2187)^2}{2187} + \frac{(1318-1688)^2}{1688} + \frac{(1439-2453)^2}{2453}$$

Sum = 1347.29
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Burglary

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	854	7884	**8738**
Feb	1832	6758	**8590**
Mar	1434	6722	**8156**
Apr	1874	5916	**7790**
Total	**5994**	**27280**	**33274**

$$= \frac{(7884-854)^2}{854} + \frac{(6758-1832)^2}{1832} + \frac{(6722-1434)^2}{1434} + \frac{(5916-1874)^2}{1874}$$

Sum = 99333.34
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Criminal Damage

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	784	4350	**5134**
Feb	2147	4139	**6286**
Mar	1608	4720	**6328**
Apr	2049	4561	**6610**
Total	**6588**	**17770**	**24358**

$$= \frac{(4350-784)^2}{784} + \frac{(4139-2147)^2}{2147} + \frac{(4720-1608)^2}{1608} + \frac{(4561-2049)^2}{2049}$$

Sum = 27170.38
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

<u>Drug Offences</u>

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	831	3641	**4472**
Feb	1897	3224	**5121**
Mar	1340	3957	**5297**
Apr	1748	3522	**5270**
Total	**5816**	**14344**	**20160**

$$= \frac{(3641-831)^2}{831} + \frac{(3224-1897)^2}{1897} + \frac{(3957-1340)^2}{1340} + \frac{(3522-1748)^2}{1748}$$

Sum = 17341.53
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

<u>Possession of Weapons</u>

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	3304	240	**3544**
Feb	7905	220	**8125**
Mar	6049	315	**6364**
Apr	7726	278	**8004**
Total	**24984**	**1053**	**26037**

$$= \frac{(240-3304)^2}{3304} + \frac{(220-7905)^2}{7905} + \frac{(315-6049)^2}{6049} + \frac{(278-7726)^2}{7726}$$

Sum = 22927.95
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Public Order

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	2625	2231	**4856**
Feb	6249	1986	**8235**
Mar	4852	2495	**7347**
Apr	6075	2450	**8525**
Total	**19801**	**9162**	**28963**

$$= \frac{(2231-2625)^2}{2625} + \frac{(1986-6249)^2}{6249} + \frac{(2495-4852)^2}{4852} + \frac{(2450-6075)^2}{6075}$$

Sum = 7092.54
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Robbery

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	829	2179	**3008**
Feb	1709	1879	**3588**
Mar	1362	1882	**3244**
Apr	1786	1715	**3501**
Total	**5686**	**7655**	**13341**

$$= \frac{(2179-829)^2}{829} + \frac{(1879-1709)^2}{1709} + \frac{(1882-1362)^2}{1362} + \frac{(1715-1786)^2}{1786}$$

Sum = 2469.16
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Shoplifting

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	652	3224	**3876**
Feb	1522	2914	**4436**
Mar	1168	3372	**4540**
Apr	1473	3308	**4781**
Total	**4815**	**12818**	**17633**

$$= \frac{(3224-652)^2}{652} + \frac{(2914-1522)^2}{1522} + \frac{(3372-1168)^2}{1168} + \frac{(3308-1473)^2}{1473}$$

Sum = 17863.97
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Theft from the Person

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	952	2975	**3927**
Feb	1878	2927	**4805**
Mar	1523	2987	**4510**
Apr	1839	2417	**4256**
Total	**6192**	**11306**	**17498**

$$= \frac{(2975-952)^2}{952} + \frac{(2927-1878)^2}{1878} + \frac{(2987-1523)^2}{1523} + \frac{(2417-1839)^2}{1839}$$

Sum = 6602.71
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Vehicle Crime

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	15	7272	**7287**
Feb	55	6605	**6660**
Mar	33	7133	**7166**
Apr	27	6800	**6827**
Total	**130**	**27810**	**27940**

$$= \frac{(7272-15)^2}{15} + \frac{(6605-55)^2}{55} + \frac{(7133-33)^2}{33} + \frac{(6800-27)^2}{27}$$

Sum = 7517577.4
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

Violent Crime and Sexual Offences

Month (2014)	Expected (Crime Refs on Twitter)	Observed (Actual Crime)	Total
Jan	2569	11745	**14314**
Feb	6067	11084	**17151**
Mar	4421	13061	**17482**
Apr	6327	12205	**18532**
Total	**19384**	**48095**	**67479**

$$= \frac{(11475-2569)^2}{2569} + \frac{(11084-6067)^2}{6067} + \frac{(13061-4421)^2}{4421} + \frac{(12205-6327)^2}{6327}$$

Sum = 57369.41
Chi Squared Value (based on Figure 39 table) = 7.82

The chi squared value is much less than the value of our sum. Therefore Reject Hypothesis 0 so no correlation

5.6 Statistical Testing Results

Overall, it is clear that based on this statistical analysis in section 5.5 that for each crime type and all up there is no correlation between numbers of references identified from Twitter against actual crime in the London area. This is due to the large differences between the observed (actual crime) and expected (Twitter references to crime) numbers meaning there is no fit between the two variables. To take this analysis further there may be value in testing at London borough level or using spatial patterning. The Pearson chi-squared test does not take into account how the hotspots of crime are distributed within a spatial area.

It is worth noting that by altering the crime-related keywords in the crime type arrays in the Java file it is expected that different results will be yielded. However, it is unlikely that a major change in the number of references from Twitter will occur based on the alteration of some key terms. If I were to evaluate the data used in this test, it would have been useful to create multiple versions of each crime type for all months but searching using different sets of crime-related words to see the differences in results. In turn, this test could be applied to put the observed (actual crime) against the expected (Twitter crime) a number of times for each month. This would add more reliability to the investigation because the test would be on a larger sample size.

6 Evaluation

6.1 Results versus Project Aims

The software and source code which I have managed to implement has helped to meet deliverables outlined in the initial plan. Below are the aims and the outcome or result of how they have been progressed against during this project.

- Aim: Identify specific crime-related terms or hashtags in Twitter data

I have sourced Twitter data from the Cardiff School Computer Science & Informatics. The data comes in separate monthly, raw data CSV files for the time period January to April 2014 and is confined to London area locations.

I have adapted Java source code to create a file which takes the input of a raw Twitter dataset such as a month's worth of tweets for London. A number of arrays are used to store all the crime-related words for each crime type. The method I used to collect these crime-related references was to search for related articles pertaining to each category of crime. I extracted key words or phrases including slang which were documented as ways to refer to the particular crime type or examples of crimes which are committed within these categories.

Using pattern-matching the Java code searches through the plain text of each tweet looking for the presence of each keyword. If a keyword from an array is found then the line of data is stored. Once all iterations across the whole file for each crime-related reference are completed then a CSV file is output. This process happens for each crime type within one run of the Java file for one month's worth of Twitter data. The source code makes it possible to change the input and output file locations. Therefore separate monthly outputs of each crime type for the time period can be produced in CSV format. These files are in a manageable form to parse into a web browser to use in our visualisation software environment.

- Aim: Develop system to visualise the distribution of these Twitter criminal hotspots onto a map using Google Maps API by crime type

A web-based, software environment has been implemented with the ability to read in longitude and latitude location values from CSV files. Each crime type has its own CSV file which can be parsed via the web browser to the API in order to visualise its data contents. I used a JavaScript CSV parsing library, PapaParse to read the data into the web browser. The Google Maps API provides the functionality to implement heat-maps which take the longitude and latitude values to plot onto a map. The overlaying heat-map layer provides the visualisation component of the system to display the data which I have sourced and compiled in regards to crime on Twitter.

When a particular crime type CSV file is loaded in to the system, a checkbox becomes enabled on the main menu so you can toggle the heat-map on and off. This helps to provide interactivity for the user and allows them to customise the view which they want to see displayed on the map.

- Aim: Plot actual reported crimes (sourced from Police API) onto a similar map in the software environment

I downloaded the publicly available monthly CSV files from the Police API to obtain actual crime data for the January to April 2014 time period. The files were for the Metropolitan Police area of London which represents the most similar area of London to where my Twitter data holds location for. I used the same parsing method in a slightly different way to get the data required to be visualised on a separate actual crime Google Map.

Two Google Maps side-by-side have been implemented into the user interface with the left map showing a heat-map for crime references on Twitter and the right, the actual crime based on data from the Police API. Although the actual numbers are not specified for either of the heat-maps, the system provides a useful visual comparison of the data.

The system prompts the loading of the actual crime file first and then disables that ability once a file with a month’s worth of data has been loaded in. The user is then able to load in the different crime type CSV files for references to crime on Twitter onto the other Google Map (Left side of the interface). There is one main menu in the centre of the interface on which each crime type can be toggled on and off as well as changing the radius of influence of the data points. The checkboxes for each crime type relate to both maps so therefore if you toggle the criminal damage heat-map on for example then both the Twitter heat-map and actual crime heat-map will be displayed on their corresponding Google map.

After conducting rigorous testing, I discovered the parsing library I have used does not parse in all the rows of the data for actual crime from the Police API CSV files (although crime type files for Twitter crime are parsing correctly). If I were to do this again I would undertake more testing during the system implementation to allow for time to change the parsing library or manipulate the CSV data using a different method. The fact the data is created by another source unlike the Twitter CSV data means the structuring is slightly different and is the expected cause of this issue.

- Aim: Visualisation of particular hotspots of criminal activity based on successful implementation of the above aims

The implemented heat-maps are down to human interpretation but make it easy to identify hotspots of crime. The easy to comprehend lighter colour for less data points and darker for more points is an effective method to display location data such as within the CSV files I have been handling.

An additional usability feature is the change radius functionality for each crime type which when toggled on delivers an easy way to view hotspots of crime more clearly. By changing the radius each data point carries a larger 'influence area' and therefore the heat-map expands across a larger area but becomes much simpler to examine.

- Aim: Undertake a statistical analysis to see if there is a correlation between actual reported crimes and the crimes identified from Twitter data

Based on the analysis I have carried out in the Pearson chi-squared test there is no correlation between crime references on Twitter and actual crime in London. This is the case at overall level and at each crime type level for the time period January to April 2014 which I have data for.

There may be a number of different reasons for this outcome. For example, the key terms which I have chosen to search for may not be in reference to an actual crime or some crimes may not be released to the public at the time they happen meaning they would not be tweeted about. The main reason for the lack of correlation is because overall the number of actual crimes (which are the observed variable within the Pearson formula) are in general much larger than the number of crime-related references we picked up from Twitter.

7 Future Work

This section describes the additional features or work which could be carried out to take this project further. During this project these are ideas which were considered but the time to implement them was not available with the main focus on working towards the key deliverables.

7.1 Improved Software Usability & Data Parsing

The software environment needs some refinement in order to make it more effective as a visual aid to identify crime hotspots. While the Twitter data is broken down into a number of CSV files which parse correctly, the JavaScript parsing library PapaParse does not effectively parse the data to the web browser for actual crime data which is downloaded from the Police API. Being able to create separate, smaller CSV files for each crime type would be an option to improve performance of the current parsing library or alternatively looking at a different CSV parser which may be able to handle larger data sets.

With regards to usability, a smoother process to load the files into the tool would be useful. The current functionality allows only one file at a time to be loaded in. In usability terms this is done effectively for actual crime with one month's file rendering all the heat-maps for each crime type at the same time. However, each crime type for the Twitter CSV files must be selected one by one. A uniform way to load in for both maps would help to make the system easier to use. The solution here could be that both data sets are split into separate CSV files by crime type then multiple selections can be chosen to load in to the tool or all the data is in one file for each data set so all the heat-maps load at once.

7.2 User Choice of Crime-related Keywords

A feature which would be useful to implement to the system prototype would be to create an area of the website where a user is able to submit their own crime-related words to search for in the Twitter data. A benefit of this is it would not only provide extra flexibility to the system but will make it more personalised to the user.

The area of the website could be set up as a form-like interface which has each crime type with its corresponding array of related references below it. The user would be able to replace these references with their own and then choose which month they would like to produce results for. The back-end script would the run to evaluate the Twitter data into separate crime type files based on the user's choice of words. The user could download these CSV files to load into the tool as you would currently. This is a stretch feature which could help to integrate the Java back-end with the front-end visualisation software. At the moment they are two separate entities which run independently so linking them together would make a more complete system.

7.3 Dynamic Data

All the data currently used for both actual crime and references to crime from Twitter is static. The current data set is not up to date either because I have only been able to source full Twitter data sets relating to the London area for January to April 2014.

In the case of Twitter crime the data could be pulled directly as it happens using the Streaming Twitter API. This API's Public Stream provides a potential useful resource because it "streams the public data flowing through Twitter. Suitable for following specific users or topics, and data mining" [31]. This would be fit for purpose for this dynamic data concept because the tweets being evaluated for crime-related keywords would be live. If the script which identifies crime-related references from Twitter data was run in an automated fashion then it would be able to create a more up-to-date set of CSV file based on recent tweets which could be made available on perhaps a daily basis. Evaluating on a daily basis would also mean a visualisation which could show the changes from one day to the next lending further value to this idea. Unfortunately due to security constraints with the JavaScript programming language, specifying which file should be parsed in the code is not possible although with further research and resources then a workaround may be achievable. Therefore initially, the user would have to select the file they wish to load into the tool hence a daily basis would mean there is not a constant change in the latest file available (in the event the file was live as the tweets came in).

This concept may not be as feasible for actual crime because crimes are not logged immediately as they happen neither is the data likely to be released for privacy reasons. The latest data available seems to be at least a month behind so for example in March 2015 the data for February 2015 can be downloaded therefore a live or recent implementation is unlikely to be possible. However, having a separate up-to-date Twitter crime visualisation is viable and something which could be achieved in further iterations of this project.

7.4 Month-by-Month Implementation

Although a manual workaround was found to implement the monthly comparison between the two data sets, it would be advantageous to programme the system to be able to load in a full time period of data (for example in this project, the whole set of sample data from January to April 2014). Then with the full data set available in the background, some functionality e.g. a slider which would make it easier to view each month's data without requiring multiple data loads like in the current implementation. On alteration of the slider position to a different month, the current view would clear and the new month's heat-maps would be displayed. The jQuery UI provides a slider implementation which may be useful as a starting point for this piece of future work.

7.5 Spatial Patterning Analysis

Having conducted a Pearson chi squared test to understand if there is any correlation between actual crime and references to crime on Twitter, it is possible to take this analysis further. The statistical test which was carried out only looks purely at the overall numbers against each other, it does not consider each spatial area or the distribution or distances between data points across that spatial area.

Spatial patterning statistical methods do exist which could help to deduce a more comprehensive analysis of the relationship between crime references from Twitter against actual crime. A spatial autocorrelation method is required which looks for an association between the two 'signals' which are ultimately measuring the same thing such as the two comparison data sets in this project.

We could use London Boroughs as a way to split up our overall spatial area which is the Metropolitan area of London. A separate matrix storing the monthly counts of Twitter references and actual crime for each London Borough would provide a suitable foundation for most of the useful tests which could be carried out.

Using these matrices, the bivariate Moran's I test can help us to understand if there is spatial autocorrelation between the data sets. By using this test, we will get a better idea of whether the locations of the data points are correlated even if the number of points may be different as we have already identified from the Pearson test.

R is a useful statistical programming environment which is open source and free to download. It is possible that this tool could be used to model the data and run the relevant Moran I test to give us an eventual probability value. Using this value we can decide whether to accept or reject the null hypothesis in a similar fashion to in the Pearson chi-squared test which has already been completed in this project.

However, there is a problem if you were to conduct this analysis on the current data sets. The London boroughs in the Actual Crime data under the header 'LSOA Name' have the London Borough and then an Area Code meaning there is multiple entries for the same borough. For example in the current data there could be Barking and Dagenham 001A and Barking and Dagenham 001B. This would need to be mapped to the borough Barking and Dagenham in order to be usable for this test. In order to convert the data into a usable form to do further analysis a comprehensive mapping table to link all area codes to the same borough names returned in the Twitter data set would be required. Our data may not have all the potential area codes because crime may not have been committed there so sourcing the full set of area codes and borough names to create this linking table would be a useful starting point.

8 Conclusion

To conclude, this project has reached a successful outcome against the aims which were set out in the initial plan.

One of the initial aims was to be able to evaluate Twitter data identifying keywords and phrases in relation to crime. The source code which has been developed to do this searches the Twitter data looking for keywords stored in separate crime type specific arrays and outputs CSV files for each crime category.

A web-based software environment which displays a geospatial distribution of crime hotspots onto a map by crime type in the form of heat-maps also now exists. The CSV files produced for each crime type from evaluating the Twitter data can be loaded into it to view the heat-maps in addition to loading in the actual crime CSV files. This has required me to implement the Google Maps API and using this I have managed to incorporate a side-by-side comparison of the crime-related references on Twitter against actual crime hotspots. Although the parsing of the actual crime data is not entirely accurate, the tool shows the main crime hotspots for both data sets in a monthly format for the time period January to April 2014.

The visualisation tool in amalgamation with the Twitter data evaluation source code is a useful piece of software which with additional flexibility could be used to visualise any type of subject not just crime. But in this project, it provides an appropriate visualisation of criminal hotspots which meets the milestones in the project plan.

Based on the Pearson chi-squared statistical analysis carried out, it can be said there is no correlation between the numbers of crime-related references returned from Twitter versus the number of actual crimes. This is useful to learn because it provides understanding of the relationship between the social media tool and reported crimes to see whether there is value analysing Twitter as a big data set for this subject. However, taking this analysis further could yield an interesting result by looking for spatial autocorrelation between the two data sets which is basically looking to find if the plotted locations of the crime hotspots have a relationship.

9 Reflection on Learning

This project has posed a particularly challenging problem which has helped to develop both my technical and project management skills.

The development of a strong project plan was particularly important in order to provide focus and deliverable goals for the task. A weekly list of sub-goals has helped me to structure the delivery of a successful project within the one semester time constraint afforded to us. I feel that my year in industry in a business has supported me to develop strong time management skills already but initially working with my supervisor/stakeholder to construct a reasonable set of deliverables helped to set expectations with what could be achieved during the time available. If I were to improve my plan then I would look to break down and be more specific with the goals for the period week five to nine. This would have helped me to be more productive and keep me on task during this window of time. As a consequence of this, I fell behind in my objectives because I did not have a clear deadline for which the deliverables needed to be ready for.

I put myself out of my comfort zone by using a programming language which I had only a basic understanding of pre-project. Having decided JavaScript was the most appropriate programming language to use to develop the proposed web-based application it was very important that I grasped a more comprehensive view of the underlying standards behind it. In order to do this I worked on a number of tutorials which are widely available online. I feel this was extremely beneficial not only to help expand my knowledge of best practices but to obtain skills which will be desirable when applying for IT or web development roles in the future.

A further skill I have learnt is being able to implement APIs into my work. A key part of delivering this project was to incorporate the Google Maps API and an excellent decision which I made in the planning phase of the project was to experiment with this before starting the actual deliverable. By doing this I was able to learn how to use the API effectively in addition to understanding which were the most appropriate ways of visualising the data which I had sourced.

If I were to change something in this project it would be the way I conducted the testing. I feel that if I were to complete this project again the system testing would be much more extensive and on a more regular rhythm within my time plan. This would eradicate situations such as the inaccurate parsing of the actual crime data which would ultimately improve the tool I have developed. I will be able to draw on this experience in future projects knowing that quality unit testing is vital to the implementation of any project with a software deliverable.

Overall, I feel this project has been an excellent experience in managing a technical project. I have in particular enhanced my technical skills which was an area I identified as a personal development goal. With future employability in the more technical side of the technology sector in mind, to action this goal I decided that working on a project with a technical focus and a system as a deliverable rather than a pure research activity would be most beneficial to me.

Table of Abbreviations/Glossary

Term	Description
API	Application Program Interface – A set of tools for building software with already implemented components to specify how graphical user interfaces can be programmed
COSMOS	Collaborative Online Social Media Observatory – An Economic and Social Research Council (ESRC) strategic “Big Data” investment that brings together social, computer, political, health, statistical and mathematical scientists to study the methodological, theoretical, empirical and technical dimensions of social media data in social and policy contexts
CSS	Cascading Style Sheets – Style sheet language in markup form which is used to describe formatting of a document
CSV	Comma Separated Values – Stores tabular data in its plain text form. Fields in CSV files are separated by a comma or a tab.
HTML	Hypertext Markup Language – The standard markup language used to develop web pages
jQuery	jQuery JavaScript Library – cross-platform, lightweight JavaScript library
MVC	Model View Controller – Software Architecture Pattern which helps to implement user interfaces
RGBA	Red Green Blue Alpha – Red Green Blue Colour Model with extra information
UI	User Interface – Everything that a human can see and interact with on a system

Appendices

Appendix A – Crime Visualisations (February 2014 to April 2014 by Crime Type)

References

[1] F. Stroud, “Webopedia: IoT-Internet Of Things,” Quinstreet Inc, 20 February 2015. [Online]. Available: http://www.webopedia.com/TERM/I/internet_of_things.html. [Accessed 20 February 2015].

[2] V. Beal, “Big Data,” 02 February 2015. [Online]. Available: http://www.webopedia.com/TERM/B/big_data.html. [Accessed 04 February 2015].

[3] X. Wang, M. S. Gerber and D. E. Brown, “Automatic Crime Prediction Using Events Extracted from Twitter Posts,” *Social Computing, Behavioral - Cultural Modeling and Prediction,* vol. 7227, pp. 231-238, 2012.

[4] N. Malleson and M. Andresen, “The impact of using social media data in crime rate calculations,” 10 April 2014. [Online]. Available: http://nickmalleson.co.uk/wp-content/uploads/2014/05/CaGIS-AAM.pdf. [Accessed 6 February 2015].

[5] Collaborative Online Social Media Observatory, “What is COSMOS?,” 10 February 2015. [Online]. Available: http://www.cs.cf.ac.uk/cosmos/. [Accessed 10 February 2015].

[6] M. Wolff and H. Asche, “Geovisualization Approaches for Spatio-temporal Crime Scene Analysis – Towards 4D Crime Mapping,” *Computational Forensics,* vol. 5718, pp. 78-89, 2009.

[7] Twitter, Inc, “About: Company,” Twitter, 12 February 2015. [Online]. Available: https://about.twitter.com/company. [Accessed 12 February 2015].

[8] Home Office, “User Guide to Home Office Crime Statistics,” 01 October 2011. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/116226/user-guide-crime-statistics.pdf. [Accessed 06 March 2015].

[9] Urban Dictionary, “Smashed,” 16 March 2015. [Online]. Available: http://www.urbandictionary.com/define.php?term=smashed. [Accessed 16 March 2015].

[10] Wikipedia, “Pearson's chi-squared test,” 15 April 2015. [Online]. Available: http://en.wikipedia.org/wiki/Pearson%27s_chi-squared_test. [Accessed 16 April 2015].

[11] Google Developers, “Google Maps Javascript API v3 - Heatmap Layer,” Google, 04 March 2015. [Online]. Available: https://developers.google.com/maps/documentation/javascript/heatmaplayer#customize_a_heatmap_layer. [Accessed 08 March 2015].

[12] Google Developers, “Heatmap Layer Options Object Specification,” Google, 04 March 2015. [Online]. Available: https://developers.google.com/maps/documentation/javascript/reference#HeatmapLayerOptions. [Accessed 08 March 2015].

[13] Wikipedia, “jQuery,” Wikimedia Foundation, Inc, 29 March 2015. [Online]. Available: http://en.wikipedia.org/wiki/JQuery. [Accessed 30 March 2015].

[14] The jQuery Foundation, “jQuery UI 1.11 API Documentation,” 01 March 2015. [Online]. Available: http://api.jqueryui.com/. [Accessed 30 March 2015].

[15] M. Holt, “PapaParse,” 30 March 2015. [Online]. Available: http://papaparse.com/. [Accessed 30 March 2015].

[16] Gov.uk, “Police API Crime Categories,” 2015. [Online]. Available: http://data.police.uk/api/crime-categories?. [Accessed 01 April 2015].

[17] South Kesteven District Council, “What is Anti-Social Behaviour,” 21 October 2014. [Online]. Available: http://www.southkesteven.gov.uk/index.aspx?articleid=2137. [Accessed 01 April 2015].

[18] Wikipedia, “Burglary,” Wikimedia Foundation, Inc, 24 March 2015. [Online]. Available: http://en.wikipedia.org/wiki/Burglary. [Accessed 01 April 2015].

[19] Crown Prosecution Service, “Offensive Weapons, Knives, Bladed and Pointed Articles,” 2015. [Online]. Available: http://www.cps.gov.uk/legal/l_to_o/offensive_weapons_knives_bladed_and_pointed_articles/#a01. [Accessed 01 April 2015].

[20] Crown Prosecution Service, “Public Order Offences incorporating the Charging Standard,” 2015. [Online]. Available: http://www.cps.gov.uk/legal/p_to_r/public_order_offences/#Introduction. [Accessed 01 April 2015].

[21] Farlex, Inc, “Robbery,” 2015. [Online]. Available: http://legal-dictionary.thefreedictionary.com/robbery. [Accessed 01 April 2015].

[22] C. E. McGoey, “Shoplifting Facts,” 2014. [Online]. Available: http://www.crimedoctor.com/shoplifting-facts.htm. [Accessed 01 April 2015].

[23] Home Office, “Theft from the Person,” July 2013. [Online]. Available: https://crimestoppers-uk.org/keeping-safe/personal-safety/theft-from-the-person/. [Accessed 01 April 2015].

[24] M. Holt, *Papa Parse,* 4.1 ed., 2015.

[25] Freepik, “CSSmatic,” 2013. [Online]. Available: http://www.cssmatic.com/gradient-generator. [Accessed 04 March 2015].

[26] C. Mills, “Using Web Workers,” 27 March 2015. [Online]. Available: https://developer.mozilla.org/en-US/docs/Web/API/Web_Workers_API/Using_web_workers. [Accessed 08 April 2015].

[27] Bike Hub, “Census confirms London cycle commuting boom,” 26 March 2014. [Online]. Available: http://www.bikehub.co.uk/news/bike-to-work/census-confirms-london-cycle-commuting-boom/. [Accessed 13 April 2015].

[28] C. Turner, “Finance and investment firms descend on Mayfair,” 04 June 2015. [Online]. Available: http://www.telegraph.co.uk/finance/economics/10875566/Finance-and-investment-firms-descend-on-Mayfair.html. [Accessed 13 April 2015].

[29] London Metropolitan Police, “Borough Map,” 2015. [Online]. Available: http://content.met.police.uk/Page/YourBorough. [Accessed 15 April 2015].

[30] Wikipedia, “Chi-squared distrbution,” 17 March 2015. [Online]. Available: http://en.wikipedia.org/wiki/Chi-squared_distribution. [Accessed 04 April 2015].

[31] Twitter, Inc, “The Streaming APIs Overview,” 2015. [Online]. Available: https://dev.twitter.com/streaming/overview. [Accessed 17 April 2015].