**Pushing Privacy: Exploring Twitter Usage Across the Residences of Melbourne**

Gustavo Michael Carrion Lopez - 667597

Department of Computing and Information Systems

The University of Melbourne

Supervisors: Professor Richard O. Sinnott

1. **Description**

Gathering information about what is being tweeted is used by many researchers to understand the population and what they think/feel on particular topics. However it is often the case that Twitter use can reveal the location/identity of individuals directly. This project will explore this topic directly and develop automated systems that explore the use of Twitter across the residences/homes of Melbourne. It will support data analytics and sentiment analysis and identify the happiest/saddest household; the happiest/saddest street as well as identify the household that uses most social media. This work will not divulge information on the individuals themselves, but instead show how such scenarios are now directly possible. It will propose steps that might be taken to obfuscate the exact location of tweeters to protect their privacy. A key part of this work will be automating the process for harvesting data for each household – rather than for example a single large bounding box for Melbourne. The geolocations of all the residences in Melbourne will be identified/use and their latitude/longitude identified using data from the Australian Urban Research Infrastructure Network (AURIN) and other sources. These latitude/longitudes will be the centre-point of centroids that allow for targeted Twitter harvesting, i.e. tweets from that household (in that centroid). The system should allow an arbitrary address to be given and the Twitter data (if any?) captured from that address.

Although the idea seems pretty simple in reality the big data generated by Twitter presents challenges that must be addressed regarding is known as the four V’s:

* Volume: The volume of twitter data generated by Melbourne’s citizens is of the order of gigabytes of data daily, given that the study will be gathering data for more than a month it is expected to have enough resources to store and process this data.
* Velocity: In order to reflect the current opinion of tweeter user’s data must be gathered periodically, by querying twitter API automatically and keep feeding the database.
* Variety: The data can come in many forms for this particular study, only English text tweets, username, userIds, geolocation(if available). Will be processed discarding the remaining metadata.
* Veracity: Accurate tweets information must be enforced by not processing duplicated tweets, or non-human generated information (bots), tweets outside Melbourne, among other parameters.

1. **Approaches**

In order to fulfill the hardware infrastructure needs of the project the NECTAR research cloud will be used, concerning to software the project will aim to use open source tools like couchDB (Backend), ElasticSearch/MapReduce (Data processing), HTML5 + JS frameworks (Front end). Among other technologies that will be implemented depending on the project needs.

The project will be divided in four main stages dependent one from another as following:

* Infrastructure setup: In this stage the servers that will support all the implementation process will be deployed within NECTAR research cloud platform, in this stages things like harvesting processing power, storage capacity, data analysis processing unit, front end infrastructure, backup servers will be determined and created.
* Data gathering: The following stage consists in gathering the data and preprocessing it to ensure will overcome the four V’s challenges, things like software development to preprocess data, script setup to gather the tweets and residences information periodically, data storage will be implemented in this phase.
* Processing, analysis and display of facts: This phase will aim to find interesting facts about the data to query it efficiently, prior to the design of a web interface that shows the findings in an accurate manner.
* Report: This point consist in documenting the process and findings of the research scenario, specifying technical details as well as the most important social, political, economic facts found after the analysis.

However this is a preliminary approach which could be modified during the development of the project.

1. **Outcomes:**

* A report specifying the methods and the findings obtained during the study.
* A web portal that will have the data ready to be queried, processed and displayed in an adequate manner.
* A 24 / 7 functional harvester that gathers tweets, preprocess them to finally feed the data repository periodically, as well as all the configured servers mentioned above.
* Software source code used on each of the stages.