# "Establishing the location of a vernacular name using social media mining techniques" - Initial Project Plan

1109949
Billy Hickman
CM2303 - One Semester Project (40 Credits)
Supervisor: Chris Jones

February 2, 2015

## 1. Project Description

Qualitative vernacular names for places is a problem encountered on a day-to-day basis; the definitions are vague and may or may not correspond to an administrative gazetteer (Twaroch *et al.*, 2009). Through my own experience of using social media I have noticed the considerable use of vernacular names; for instance, someone may use the name 'Pompey' to describe Portsmouth, UK. There has been considerable research into detecting the location of tweets and detecting vernacular names, demonstrated in Nand *et al.* (2014). Additionally, 'IBM Research' in Tokyo has looked at establishing a users location based on names and locations mentioned in tweets before and after their current tweet (Ikawa *et al.*, 2013).

My proposal is to reverse engineer this process so you start with a vernacular name, or one that the user knows to be vernacular, and search social media outlets such as Twitter or Flickr for mentions of this name. Currently, around 5% of Twitter posts (based on my experience) are geo-tagged; some reports suggest this could be as high as 20% (We, 2013). Most social media sources provide location data through an API in some form. My general assumption is that clusters will be formed around the locations in which these vernacular names are mentioned most frequently; this will form a relationship between a vernacular name and a spatial region. There are problems with how many tweets can be mined, as well as how well the clusters will be formed; I envisage my system will fail to establish clear clusters when some vernacular names are used. There also has to be appropriate ways to remove outliers that are likely to be encountered. I will provide a platform for a user to enter vernacular names and a way of displaying these results on a map, through methods such as a heat map. I imagine these names will have to be queued in order to gather enough data and perform clustering.

The use case I imagine comes in two forms. Firstly, I see it as being particularly useful for emergency services as a way to collect information on vernacular names. They could use this platform as a way to compare and verify their own knowledge on vernacular names, which will be useful in an emergency scenario. Secondly, I see it as a tool that could be used for a member of the general public to compare his or her own knowledge about vernacular names against a general consensus.

Comparing my system results with data such as that collected from the website 'YourPlaceNames.com' (Twaroch, 2010), is where I feel the interesting conclusions can be drawn. 'YourPlaceNames.com' is a website that was created by researchers at Cardiff University to collect data about the everyday names people use to describe places; this website provides facilities to assign names to locations through various methods. It will be interesting to see how the results differ between manual collection and automated social media collection. Hopefully, it will lead to some interesting research conclusions and help to prove or disprove my assumption. Most of the previous work located has attempted to use lexical analysis to find locations mentioned in tweets or to establish a users location. My project aims to reverse this and use Twitter to establish the location of a known (to a user) vernacular name. Analogy can be found with an article by Jones *et al.* (2008) entitled "Modelling vague places with knowledge from the web" . This journal discusses the modelling of vague place names by harnessing web knowledge; this is done by associating vague place name mentions with administrative regions that are mentioned in the same document. This is similar to my assumption as it both harnesses web data and starts with vague names (which could be vernacular).

I believe this project provides me with an opportunity to explore a real world and complex problem, whilst also producing a tangible product. Finally, it will allow me to experiment with different programming concepts, tools and algorithms. I envisage, at this early stage, developing it as a web application with a front end UI (User Interface) and a back end API (Application Programming Interface) *(Fig 1)*.

## 1.1 Summarised Description

- My approach is to establish the location of colloquial/vernacular names by mining data from Twitter or other social media outlets.
- Users will enter what they believe to be known vernacular names. The system may not be able to cluster all vernacular names that are entered.
- Output will demonstrate relationship between vernacular names and spatial regions.
- Combined with other data sources comparisons can be drawn.

- Has a potential use case for emergency services.

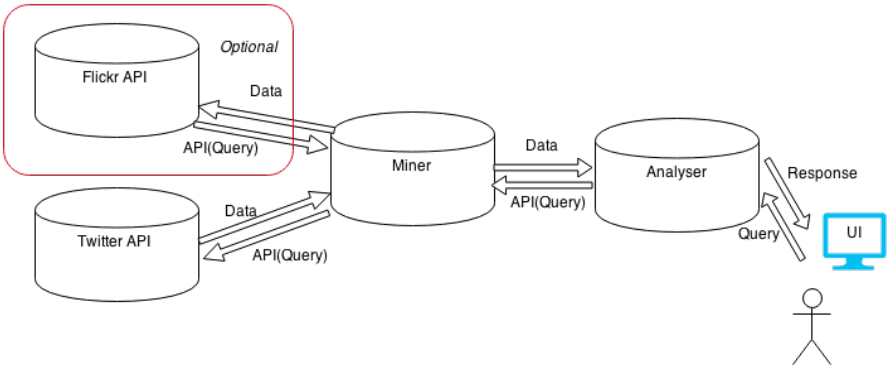



*Figure 1. This image shows an abstract view of intended system architecture.*

# 2. Project Aims and Objectives

## 2.1 Aims of the finished product

I have split these aims into a 'must have', 'should have', 'could have' type format in order to assign a priority/importance.

- "Must Have" - These have to be implemented for the project to be a success.

  1. A way of retrieving geo-tagged tweets that contain the colloquial name.
  2. A method of identifying clusters should they exist.
     - This involves the removal of outliers should they be considered not to form part of a valid cluster.
  3. A way of queuing and mining tweets in an asynchronous fashion.

- "Should Have" - These are secondary in priority but will assist in the project being considered a success.

  1. A user interface to enter a query (a vernacular name) and viewing results.
     - This means a method to insert queries (vernacular names).
     - A way of viewing these data sets in a visual form such as a heat map.
  2. A web service (API) that performs these queries.

- "Could Have" - These are additional objectives which are not required for the project to be a success but will add value.

1. A method of overlaying the results found on 'YourPlaceNames.com' onto the map in order to allow for comparison.
2. Data streams from other social media outlets e.g. Facebook or Flickr.
3. Provide a variety of visual representations (not just heat maps) such as bounding polygons etc.
4. A web app that allows for the user to log in and then view previously entered queries and the results of these.

### 2.2 Aims of the research

1. A system to investigate where vernacular names in social media posts, do in fact cluster around a particular geographical region. This involves proving or disproving my original theory about the clustering of these vernacular names.
2. Able to draw conclusions by comparing manually entered data such as that from 'YourPlaceNames.com' (Twaroch, 2010).
3. Draw conclusions about whether a system such as this could be used in a real world environment, for example with the emergency services.

## 3. Ethics

Reviewing the ethical guidelines on the Cardiff University website (Spasic, 2014) has allowed me to understand whether ethical approval is required. The social media data that will be mined is publically available through the various API's and is not private at any point. Also, the project does not involve any interviews or observations, data will be collected automatically through data mining. Finally, no other school is involved in the progress of this project. So at this initial stage, I see no reason to seek ethical approval; however this is subject to change and will be constantly reviewed.

## 4. Work Plan

I have split the work into four main sections, research, design, prototyping/development and the final report. I have included a Gantt Chart in Appendix 1 to show how I intend the project to progress. Note: This schedule is subject to inevitable change as the project continues.

### 4.1 Sections and Deliverables

- **Research:** This is the stage in which I will begin to understand the problem and related area. It will involve locating related material and establishing the most appropriate tools

for this type of problem.

  - Finding all related work - this will involve locating papers, news articles and software tools. Deliverable: A list of related papers and work that has been completed. Due Date: 29/01/2015
  - Researching the best tools to use. Deliverable: A list of tools and software dependencies. Due Date: 03/02/2015
  - Establishing a list of risks the project may encounter. Deliverable: A list of risks and their associated weighting. Due Date: 06/02/2015

- **Design:** This is the stage which will involve designing the system. It will involve extensive use of modelling tools such as UML, as well as the designing of the user interface. This stage must be completed in order for the prototyping and development stages to start.

  - Designing the structure of the program. Deliverables: UML Diagrams Due Date: 10/02/2015
  - Designing the database schema. Deliverables: Database file with the schema included. Due Date: 11/02/2015
  - Designing the UI. Deliverables: A file containing the mock UI. Due Date: 13/02/2015
  - Designing unit tests. Deliverables: Unit tests ready to run against the code.Due Date: 17/02/2015

- **Prototyping and Development** This will be an extensive cyclical stage of prototyping. I aim to produce at least two prototypes (this is subject to change) and these will form two of the deliverables. It will also involve completing extensive technical documentation and testing; all of which will be available alongside the final report.

  - Producing an initial prototype (stage 1). Key Deliverable: A prototype that demonstrates basic functionality. Due Date: 06/03/2015
  - Producing a final prototype (stage 2). Key Deliverable: A prototype that demonstrates further functionality. This will be the second prototype before the final stage of development. Due Date: 31/03/2015
  - Producing documentation. Deliverable: Final Documentation - due at the same time as the final development cycle. Due Date: 03/04/2015
  - Testing. Deliverable: Testing report - due at the same time as the final development cycle. Due Date: 03/04/2015
  - Final Development - this involves making any changes to deliver a final software solution. This shall meet to all of the "must have" objectives. Deliverable: Final Product. Due Date: 03/04/2015

- **Final Report** This is the stage that will involve collating the results. It will involve merging some of the previous deliverables and presenting my findings.
  - Final Report - Drafting. Deliverable: A first draft of the report. Due Date: 15/04/2015
  - Final Report - Finalizing. Deliverable: A finished version of the report. Due Date: 05/05/2015

## 4.2 Week by week plan

This plan shows an abstract week by week plan that is subject to change. I have also scheduled the weeks in which review meetings will be held (specific dates will be organised at a later time).
**Week 1 - 3**

- Finding all related work.
- Researching the most appropriate tools to use.
- Assessing risks.

  By the end of this time frame I will have delivered: List of research material, List of best tools, A list of risks

**Week 3 - 4**

- Designing the structure of the program.
- Designing the database schema.
- Designing the UI.
- Designing unit tests.

  By the end of this time frame I will have delivered: UML Diagrams, Database File, Mock UI's, Unit Tests

**Review meeting to be held in week 4**

**Week 4 - 10**

- Prototyping stages
- Final Development
- Testing and documentation

- Starting final report

  By the end of this time frame I will have delivered: Final Program, Documentation, Testing Report

**Review meeting to be held in week 9**

**Week 10 - 15**

- Drafting final report.
- Finalizing report.

  By the end of this time frame I will have delivered: Final Report

# References


Ikawa, Y., Vukovic, M. and Rogstadius, J. 2013. Location-based insights from the social web. In: *WWW '13 Companion*. WWW, pp. 1013–1016.

Jones, C., Purves, R., Clough, P. and Joho, H. 2008. Modelling vague places with knowledge from the web. *International Journal of Geographical Information Science* 22(10), pp. 1045–1065.

Nand, P., Perera, R., Sreekumar, A. and Lingmin, H. 2014. A multi-strategy approach for location mining in tweets:. In: Ferraro, G. and Wan, S., eds., *Australasian Language Technology Association Workshop 2014*. Brisbane, Australia: ALTA, pp. 163–170.

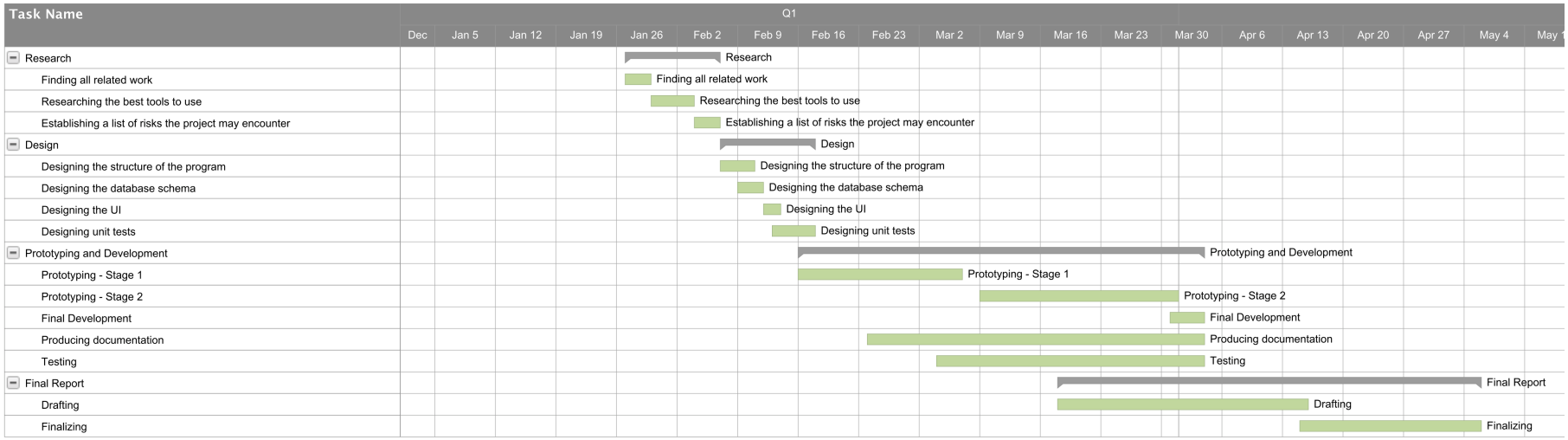
Spasic, I. 2014. *Research ethics - cardiff school of computer science and informatics*, [Online]. Available at: `http://users.cs.cf.ac.uk/I.Spasic/ethics/ [Accessed: 29/1/2015]`.

Twaroch, F. A. 2010. *Yourplacenames.com [accessed: 29/1/2015]*, [Online]. Available at: `http://www.yourplacenames.com`.

Twaroch, F. A., Purves, R. S. and Jones, C. B. 2009. Stability of qualitative spatial relations between vernacular regions mined from web data. In: *Workshop on Geographic Information on the Internet*. Toulouse, France.

We, S. 2013. *Twitter and privacy*, [Online]. Available at: `https://pressroom.usc.edu/twitter-and-privacy-nearly-one-in-five-tweets-divulge-user-location-through-geotagging-or-metadata/ [Accessed: 29/1/2015]`.

**Appendix 1**


Task Name
Q1
Dec
Jan 5
Jan 12
Jan 19
Jan 26
Feb 2
Feb 9
Feb 16
Feb 23
Mar 2
Mar 9
Mar 16
Mar 23
Mar 30
Apr 6
Apr 13
Apr 20
Apr 27
May 4
1 Research
2 Finding all related work
3 Researching the best tools to use
4 Establishing a list of risks the project may encounter
5 Design
6 Designing the structure of the program
7 Designing the database schema
8 Designing the UI
9 Designing unit tests
10 Prototyping and Development
11 Prototyping - Stage 1
12 Prototyping - Stage 2
13 Final Development
14 Producing documentation
15 Testing
16 Final Report
17 Drafting
18 Finalizing