INITIAL PLAN

Collecting Knowledge from Social Media

CM3203 One Semester Individual Project- 40 credits

Author: Zain Tahir
Supervisor: Prof Alun D Preece
Moderator: Dr Martin J Chorley

Project Description

Human involvement is particularly useful in sensing various processes in complex social and urban spaces [1]. The traditional approach of using embedded sensor networks is plagued with problems such as: gaps in spatiotemporal coverage, difficulty adapting to complex spaces and various others such as aesthetics etc. By using people who frequent/visit these places and by using their knowledge of the area, human-centric sensing makes it feasible to get information that otherwise is not possible.

This combined with natural language question-answering systems and contextually aware mobile applications allows for improved Sensemaking. Sensemaking is the process by which people give meaning to experience. Users equipped with mobile devices act as sensors (able to acquire information) [2].

Twitter is being used more and more for obtaining real time information on events as they unfold [3]. The tweets made by tweeters can be used to provide actionable information, which can be characterised by two distinct methods. The first method is location, tweets can be geo-tagged or specifically mention a place (e.g. Trevithick building, Cardiff, etc). The second method is by the topic. This can be achieved by characterising specific hashtags, such as “#JeSuisParis” used during/after the Paris attacks, or by the topic mentioned in the tweet itself (e.g. Paris Attack, etc).

This data can be collected with relative ease using Twitter's streaming API, which allows a developer to only retrieve tweets that meet a certain criteria - set terms or within a given geo-spatial region.

Once the data has been collected, it has to be processed in order to enhance the semantic value of the data so that it can be used to provide factual information. To achieve this several techniques can be used in conjunction, such as natural language processing and event detection. This data is then added to a knowledge base about that particular event.

The final stage is the implementation of some tool for the user to be able to query the knowledge base about a particular event. CENode is a JavaScript based “conversational agent designed to mediate interactions between human users and machine agents” [4]. It allows users to query the knowledge base by asking it questions in English, such as: “Who is Paul Heaney?”, and CENode would respond with:
“there is a journalist named `Paul Heaney' that
uses the twitter account `paulheaney67' and
works for the media organization `bbc'.”

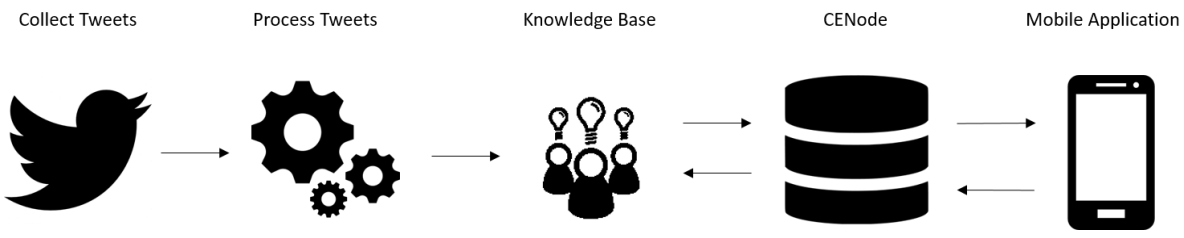

Figure 1. An overview of the proposed system

For this project, I will be attempting to develop a system that crowdsources up-to-date information about a major event such as an open day or a music event and allows users to query the knowledge base for up-to-date information.

Summarised Description

- Collect knowledge (Facts) about an event from twitter
- Use natural language processing to make sense of that information
- Feed that enriched into a knowledge base so that it allows attendees of that event to find relevant information via CENode

Project Aims and Objectives

Aims of the finished product

The finished product must have

- Working knowledge gathering
    - Collect data from twitter
    - Process this data to enrich it semantically
    - Store enriched in a knowledge base
- Working application to access this knowledge
    - Mobile Application
        - Develop application that uses the CENode API

It should:

- Be able to support for a single kind of event (e.g. festival or other big public event)

Aims of the research

The aim of the research is to answer the following question

- Do people tweeting about a particular event actually reflect what's going on in the real world?

Ethics

The twitter data that will be mined is publically available through the various API's and is not private. This project will not collect any personally-identifiable data from users who test the software. Also no other school is involved in the progress of this project. Hence myself and my supervisor see no reason to require ethical approval [5].

Work Plan

The duration of this project is approximately 15 weeks. I will be working a minimum of three hours every day, Monday - Friday. This gives me a minimum of 225 hours to work on this project. My approach to this project will be to documenting of the project report alongside with designing and implementing the proposed system solution, in order that design decisions are codified and documented for future reference. I will use the weekend to summarise my work for the week and plan for next week. I will be also holding weekly review meetings either in person or by email to discuss progress.

Research

I have to research and understanding the following topics before I begin development. This should take approximately two week maximum.

- Background research further into the domain

- Twitter API → capabilities of the API
- JavaScript
- How CENode works
- Bag of words approach to enrich semantics of collected data
- hardware requirements - server, how to query server

Design and Implementation

The following aspects have to be designed and implemented in order to meet the project's objectives and aims.

- Plan prototype test
    - Designed to test the functionality of the system
        - Does it collect relevant tweets?
        - Is the knowledge base providing the most up-to-date information?
        - Does the mobile application retrieve the correct information?
        - Can the user update the knowledge base?
    - No users involved
- Plan actual experiment
    - Trial of the system of the system under experimental conditions.
    - Use either pre-collected data or run it live
        - Pre collected → An event that has already occurred
        - An event on a large scale, such as a music festival (Glastonbury etc)
            - to be decided
- Back end functions
    - Collect relevant data from twitter
    - Process the collected data
    - Store in knowledge base
    - Develop functions to allow CENode to access the knowledge
- front end functions
    - Develop mobile application
        - Using CENode API

Final Report

- I will be documenting as I go along the project but I have dedicated the last two weeks to writing the report and proofreading for submission.

Week by week plan with a Gantt chart

- **Week 1 - 25th to 31st Jan**
    - Work on initial plan
    - Background reading
    - Deliverable: Initial Plan (23:00, 31/1/2016)
- **Week 2 - 1st to 7th Feb**
    - Research Twitter API
    - Research web frameworks and web languages
    - Research natural language processing
    -
- **Week 3 - 8th to 14th Feb**

    - Research how CENode works
    - Research hardware requirements
        - server requirements
        - how to query server
- **Week 4 - 15th to 21st Feb**
    - Begin Designing proposed system the interlinking of tools to meet the project aims
- **Week 5 - 22nd to 28th**
    - Develop backend functions
- **Week 6 - 29th to 6th of March**
    - Develop front end functions
    - Carry out system test on past data
- **Week 7 - 7th to 13th of March**
    - Plan the prototype test
- **Week 8 - 14th to 20th March**
    - Carry out prototype test
- **Week 9 - 21st to 27th March - Easter**
    - **1st Supervisor Review Meeting** to discuss project progress and any alterations required
- **Week 10- 28th to 3rd Arpil - Easter**
    - carry out the main experiment
    - analyse the data collected from the experiment
    - summarise the data collected and whether it meets projects aims & objectives
- **Week 11 - 4th to 10 April - Easter**
    - **2nd Supervisor Review Meeting** to discuss project progress and any alterations required
    - Work on final product solution
- **Week 12 - 11th to 17th April**
    - Deliver final product
- **Week 13 - 18th to 24 April**
    - writing report/documentation
- **Week 14 - 25th to 1st May**
    - writing report/documentation
- **Week 15 - 2nd to 5th May**
    - Proofreading & submission
    - Deliverable: Final Report (23:00, 6/5/2016)

Deliverables

- Initial plan: submit by 23:00, 31/1/2016
- Final report: submit by 23:00, 6/5/2016

References

[1]M. Srivastava, T. Abdelzaher and B. Szymanski, "Human-centric sensing", 2011.

[2]A. Preece, W. Webberley and D. Braines, "Conversational Sensemaking", 2015.

[3]A. Preece, W. Webberley and D. Braines, "Tasking the Tweeters: Obtaining Actionable Information from Human Sensors", 2015.

[4] Cenode.io, "CENode", 2016. [Online]. Available: http://cenode.io/. [Accessed: 31- Jan- 2016].

[5]I. Spasić, "Ethical Approval of Research: Procedures and Guidance", 2016. [Online]. Available: http://users.cs.cf.ac.uk/I.Spasic/ethics/COMSC%20Ethics%20Procedure%20&%20Policy.pdf. [Accessed: 31- Jan- 2016].