School of Computer Science and Informatics

Coursework Submission Cover Sheet

Please use Adobe Reader to complete this form. Other applications may cause incompatibility issues.

Student Number

Module Code

Submission date

Hours spent on this exercise

Special Provision

(Please place an x is the box above if you have provided appropriate evidence of need to the Disability & Dyslexia Service and have requested this adjustment).

Group Submission

For group submissions, *each member of the group must submit a copy of the coversheet.* Please include the student number of the group member tasked with submitting the assignment.

Student number of submitting group member

By submitting this cover sheet you are confirming that the submission has been checked, and that the submitted files are final and complete.

Declaration

By submitting this cover sheet you are accepting the terms of the following declaration.

I hereby declare that the attached submission (or my contribution to it in the case of group submissions) is all my own work, that it has not previously been submitted for assessment and that I have not knowingly allowed it to be copied by another student. I understand that deceiving or attempting to deceive examiners by passing off the work of another writer, as one's own is plagiarism. I also understand that plagiarising another's work or knowingly allowing another student to plagiarise from my work is against the University regulations and that doing so will result in loss of marks and possible disciplinary proceedings.

Initial Report

Recognising Place Names in Text Documents

Supervisor: Chris B Jones

Craig Harris
1-28-2016

Contents

Project Description and Overview

The primary objective of this project is to develop machine learning methods to correctly index documents with regards to geographic space, allowing place names to be recognised and geocoded. Due to the difficulty of distinguishing place names from other terms, such as names of people, objects and organisations, the machine learning methods will need to be capable of distinguishing the names of locations in an absolute form.

To attempt to overcome the ambiguous nature of place names in text documents, the machine learning process will have to utilize various pieces of evidence to aid in the correct categorisation and recognition of such names. Examples of the sort of evidence that will be employed in the machine learning process include: whether the name occurs in a gazetteer (a list of place names of the concerning region), if the name is preceded by spatial prepositions such as “near to” or “towards” and whether it is associated with place type terms, such as “town” or “river.”

The focus of this project will be aimed primarily in the areas of Wales and the United Kingdom as a whole. Therefore, the gazetteer used will be locally focused, such as the National Gazetteer of Wales and the Ordnance Survey OpenNames gazetteer product.

The use of spatial relationships will be key in determining if a name is a place or something else. This will be achieved by examining the qualitative and quantitative information that may precede or follow a place name within the particular document. For this particular project it will be more common to find qualitative spatial relationships mentioned within the documents, such as relative locations using proximal relations (“near”, “close”) and orientation based relations (“north”, “south”). Parsing this information will be key in developing the machine learning methods required to correctly identify absolute place names in their correct context.

It has been decided that rather than using traditional text documents to perform analysis on, the project will focus on recognising place names in a pre-made, database, of Twitter posts. This will mean that the documents being dealt with are smaller and somewhat easier to process. However, due to the nature of social media, it would also mean that there is more variation within the text documents. The use of Twitter documents will also introduce several interesting focus points within the project, including the analysis of place name trends, such as when an event happens in a particular area. Also, the way in which places are defined by individuals can change from person to person. If possible (i.e. if time permits) it will be an interesting analytical point of focus to determine specific boundaries that correspond to colloquial uses of place names. Thus some people may interpret an official name differently from the official administrative definition or they may use alternative names to the administrative ones.

Ethics

Due to the nature of data used within the project there is a possibility of ethical issues regarding human data. The data that will be used within the project has been precompiled outside of the scope of the project and, therefore, should fall well within the acceptable guidelines the university has outlined when using human data.

None of the data used within the project will be made publically available and therefore there should be no breach of ethics guidelines concerning the sharing of human data.

Project Aims and Objectives

The main aims and objectives of the project are highlighted below:

- **Primary:** Develop machine learning methods to correctly identify place names within their correct context in text documents (Twitter Posts).
- **Primary:** Correctly index and geocode locations discovered within text documents.

These aims and objectives are listed as primary as they are required to fulfil the project purpose. It should be noted however that this is a challenging task in the context of Tweets and it is not expected that all place names will be successfully recognised and geocoded.

The secondary aims and objectives I have decided on are highlighted below:

- **Secondary:** Develop the machine learning methods such that nicknames and local colloquialisms are understood and indexed correctly.
- **Secondary:** Develop a model for local boundaries determined by data, allowing for mapping of local town borders from the official administrative records in comparison to what the general public refer to as being within their town borders. There are areas within many towns that are actually outside of that towns borders but are commonly referred to by the public as being a part of the town they live.

These secondary aims and objectives are several things that would be deemed additional for the purpose of the project.

Work Plan

The basis of this projects work will follow a traditional waterfall model. However, where possible an iterative approach will be used allowing for project steps to be repeated to perfect them.

This work plan is set out in a week by week format from the 18th of January 2016 until the 5th May 2016 and includes all time outside of the academic year.

Week One

(18/01/2016): Initial Planning Research

Week one of the project will consist of initial research into machine learning and a preliminary overview of existing programs and technologies that can aid the project. The information will consist of documents provided by the supervisor, prior to the start of the project. Any other relevant materials and documents will be compiled and used as reference points throughout the project lifecycle.

Week Two

(25/01/2016): Preliminary analysis

Week two will involve the creation of an initial project plan. The plan is outlined within this report. This week will also allow the majority of the project milestones and deliverables to be mapped out, giving clear targets and goals for the forthcoming weeks.

(28/01/2016): Email Correspondence with Supervisor

This meeting will allow the review of a draft plan and report by the supervisor, allowing any adjustments to be made before concreting the project plan.

Deliverables (31/01/2016): Initial Report to be submitted.

Week Three / Week Four

(01/02/2016) - (14/02/2016): System / Software Requirements & System Analysis and Design

Weeks three and four will focus on the system requirements. During these weeks the initial system requirements will be developed. These weeks will help solidify what the projects scope will be and what will be required in regards to programming and planning to implement all primary and secondary features. This time will also be spend creating the initial system design in regards to the system requirements. This process will involve practice implementation of code snippets and information found in previous research steps.

(03/02/2016): Meeting with supervisor

(10/02/2016): Meeting with supervisor

Meetings with supervisor to review and discuss the initial system requirements and determine if the requirements cover all key areas.

Deliverables (14/02/2016): System Requirements with supporting report. Initial system design outline and review of coding practices to implement requirements.

Week Five / Week Six

(15/02/2016) - (28/02/2016): System Design and Library Selection

This period will involve the completion and finalisation of the system design process. This will allow a complete overview of the system and how it will function in regards to scalability and any issues that there may be during the implementation period. This will also be a suitable time to select any libraries that will be useful to use within the project.

(17/02/2016): Meeting with supervisor

(24/02/2016): Meeting with supervisor

Deliverables (28/02/2016): Completed system design report and complete list of libraries and software being used to implement the project.

Week Seven

(29/02/2016) - (07/03/2016): Initial Library Testing

Testing of libraries and software selected to use in the project.

(02/03/2016): Meeting with supervisor

(07/03/2016) Deliverables: List of libraries and supporting materials.

Week Eight / Week Nine / Week Ten / Week Eleven / Week Twelve

(08/03/2016) - (10/04/2016): Coding Implementation and Method Testing

During this period the coding will be implemented. This will involve creating a working prototype following thee system design put in place previously and implementing any libraries selected. This will also include the testing of the researched machine learning methods and implementation of methods that will be used to achieve a working prototype.

(09/03/2016): Meeting with supervisor

(16/03/2016): Meeting with supervisor

(23/03/2016): Meeting with supervisor

(30/03/2016): Meeting with supervisor

(06/04/2016): Meeting with supervisor

These meetings will provide support throughout the implementation process

(10/04/2016) Deliverables: Working prototype of system

Week Thirteen / Week Fourteen

(11/04/2016) - (17/04/2016): Testing and Evaluation

Weeks thirteen and fourteen will consist of critical system testing. This will allow for an overview of the systems capabilities and provide material for the evaluation process.

(13/04/2016): Meeting with supervisor

(17/04/2016) Deliverables: Test case reports and system evaluation report

Week Fifteen / Week Sixteen

(18/04/2016) - (05/05/2016): Final Report

The final weeks of the project will involve preparing and collating all notes taking to this point to allow for a concise and critical project report.

(20/04/2016): Meeting with supervisor

(27/04/2016): Meeting with supervisor

(03/05/2016): Meeting with supervisor

(05/05/2016) Deliverables: Final project report