Roman Bleier, DAH TCD
Supervisor: Seán Duffy

3rd Year Work Placement: Report For DAH

Internship period: 05.05.2014 to 25.07.2014

Host Organisation:
Trinity Centre for High Performance Computing (TCHPC)
Lloyd Institute, Trinity College Dublin
http://www.tchpc.tcd.ie/

Internship supervisor:
Dermot Frost, Manager TCHPC, and TCD Principal Investigator for the Digital Repository of Ireland

Background and Objectives

The Trinity Centre for High Performance Computing (TCHPC) is Ireland's premier High Performance Computing Centre with large scale supercomputing and visualisation facilities. The centre employs researchers and technical staff with expertise the areas of numerical modelling, risk analysis, web development, visualisation, systems administration, etc. In recent years the centre has also collaborated with humanities researchers and centres on a number of digital humanities projects. DH projects developed and hosted by TCHPC include the “Battle of Clontarf”[1] website, developed in collaboration with TCD medieval history researchers, digital editions of historical diaries and letters (“A Family at War: The Diary of Mary Martin”[2], “The 1916 Diary of Dorothy Price”[3], “Harry Clarke Correspondence Project”[4]) developed in collaboration with the Mphil in Digital Humanities. The biggest DH project hosted by TCHPC is the “Letters of 1916: Creating History” project[5].

The “Letters of 1916: Creating History” project is the first crowd-sourced humanities project in Ireland. The project was launched on Friday September 27th 2013 at Discover Research Night in TCD Longroom Hub and invites people all over the world to share letters written during and related to the Easter Rising of 1916. Images of letters can be uploaded, read online and

[1] http://dh.tcd.ie/clontarf/ (13.07.2014)
[2] http://dh.tcd.ie/martindiary/ (13.07.2014)
[3] http://dh.tcd.ie/pricediary/ (13.07.2014)
[4] Forthcoming
[5] http://dh.tcd.ie/letters1916/ (13.07.2014)

transcribed. This project focuses especially on private collections and the letters and voices of people that were less well known or even forgotten. The project was originally a TCD project, but with the appointment of Susan Schreibmann as Professor of Digital Humanities at NUIM the project moved to An Foras Feasa (DH centre at NUIM).

My work focused on data mining and topic modelling of the transcriptions collected by the "Letters of 1916: Creating History" project so far. The tasks were as follows: A program and scripts had to be developed to clean and prepare the texts for analysis, analysis software and libraries had to be tested, and finally visualisations of the collected data had to be created. The main research question to be answered was: Do computer generated topics differ from the topics assigned by humans?

I collaborated with experts in TCHPC and An Foras Feasa and was mainly based in TCD and attended occasional meetings in Maynooth. The internship's supervisor was Dermot Frost, manager of TCHPC. Dermot and his team showed great interest in questions of Digital Humanities, and thanks be to him for many good suggestions and helpful comments. The progress of the internship was reviewed in weekly meetings, and I was encouraged to keep an internship blog[6] for personal reference and to have a public documentation of the internship. The blog attracted a good number of followers during the internship.

Work Undertaken

The "Letters of 1916: Creating History" project is a crowd-sourcing project and follows a very similar approach as the famous UCL Transcribe Bentham[7]. Users can upload images of letters, add metadata, and transcribe these letters or letters uploaded by other people. The editor provided for transcribing allows minimal TEI/XML[8] markup and can be used to provide basic information about the original letters. A media-wiki backend allows to keep track of all edits that were made to a letter.

During the uploading of a letter image the uploader assigns the letter to one of 16 predefined categories: Easter Rising Ireland 1916, Art and literature, Business, Children, City and town life, Country life, Crime, Faith, Family life, Irish question, Last letters before death, Love letters, Official documents, Patronage, Politics, World War 1: 1914-1918. The research question I had

[6] http://1916letteranalysis.wordpress.com/ (12.07.2014)
[7] http://blogs.ucl.ac.uk/transcribe-bentham/ (12.07.2014)
[8] http://dh.tcd.ie/letters1916/contribute/instructions/ (12.07.2014); TEI: http://www.tei-c.org/index.xml (12.07.2014)

to address was to show differences or overlaps between computer generated topics and the topics assigned by humans.

The data was exported by Juliusz Filipowski from TCHPC to an Excel spreadsheet. The first task was to make sense of the data, sort it, merge multiple edits on the same letter or multiple pages of the same letter. For this purpose a program was developed using the programming language Python. Python is a powerful language when it comes to work processing and there are libraries available that helped with data cleaning and the creation of a word corpus. For this project the topic modelling library "gensim"[9] was used together with other word processing libraries - NLTK[10], natural language toolkit, pyenchant for spell checking, and others. I had to write Python modules for importing, cleaning, analysing the data and exporting the results.

Besides Gensim I also used the topic modelling software Mallet. Mallet is a well-known tool for topic modelling and there is ample online documentation on how Mallet can be used to explore humanities data. Mallet is an out-of-the-box tool and needs only little setup.

The results from Gensim and Mallet had to be exported and visualised to be human readable and better suitable for interpretation. For this task I used the network visualisation software Gephi[11] which is often used to display topic models[12]. In the final stage of the internship the results were interpreted using "distant" and close "reading" approaches.

Critical Analysis

The research question addressed during the internship was of immediate concern for the team of the "Letters of 1916: Creating History" project. For them it was important to know if the categories they chose for the letters are meaningful and are reflected by the corpus itself. The research conducted by me shows that the topics used so far to categories letters might need re-thinking and revision.

Corpus preparation, topic modelling and visualisation with Gephi were the main skills gained from this internship. My everyday PhD research focuses around electronic texts, and even if the skills gained through the internship might not be directly applicable to my PhD research they will be very useful for my future career.

[9] http://radimrehurek.com/gensim/ (12.07.2014)
[10] http://www.nltk.org/ (12.07.2014)
[11] http://gephi.github.io/ (12.04.2014)
[12] http://tedunderwood.com/2012/11/11/visualizing-topic-models/ (12.07.2014); http://altbibl.io/dst4l/topic-modeling-and-gephi/ (12.07.2014)

My own research is concerned with digital scholarly editing using TEI. Even if I had previously done workshops on text mining and visualisation, topic modelling was new to to and it will be a powerful tool in my DH toolbox. Over the last decade topic modelling projects have been growing in the Digital Humanities and text mining of a corpus of letters (or any text) and extract computer generated topics from it is a cutting edge methodology. Also, the software that was used during the internship, Mallet and Gephi, is often used in DH research, and there is a high chance that I will use these tools in the future.

The internship was a chance to practise the development of complex and multi-modular scripts using Test Driven Development. Python is a programming language widely used in DH. Researchers use it for text and data mining, visualisations, web development, etc. The internship provided time to try and experiment with different programming principles and code libraries. Even limited expertise in Python will be a great advantage for any quantitative research I will conduct in the future.

Furthermore, I was encouraged to attend conferences, such as the “1st International Workshop on Computational History”, held on 27th June in the Royal Irish Academy. This was a great networking opportunity and an interesting event addressing various areas around “Big Data”.

Another fantastic experience was the work in an interdisciplinary team. On the one hand I worked together with people from TCHPC, on the other hand with early modern historian and literary scholars.

To conclude, the experience gained through the internship were highly practical and very beneficial for my future development and research. Text mining and topic modelling are key skills for dealing with huge amounts of text. The software used throughout the internship and the programming language Python are frequently used in DH projects and to know how to use them will be valuable for my future career.

Thanks to Dermot Frost, and Susan Schreibman for their continuous support during the internship, and to Alex O’Conner and Emma Clarke for their helpful suggestions related to Topic Modelling and Data Mining.