**Group meeting draft**

**06/02/2014**

Diana Silvia Teodorescu, dst1@aber.ac.uk

**About the project (the aim and the background)**

The project will be an analysis of the data collection provided from the historical newspapers found at the National Library. The newspapers are digitised and have a basic API that will be used to search, request and extract specific data.

The National Library holds the entire data and can provide information regarding the API and the other resources available. They might be interested in the end product but they won't be involved the project progress and don't have specific requirements regarding it execution except for their need to represent the data in an exciting way that would attract people to use it.

I am choosing to analyse specific articles related to crime in the past. This means having an algorithm to separate only the crime specific articles and then splitting them on type of crime. After this I will apply some statistic techniques to find different connections between age of victims, sex, place, year. I am thinking of extracting the data into another database and then representing it in a user friendly way on a website.

**Current reading list**

* Ricardo Baeza-Yates. Challenges in the interaction of information retrieval and natural language processing. In Computational Linguistics and Intelligent Text Processing, pages 445–456. 2004.
* Satanjeev Banerjee and Ted Pedersen. An adapted lesk algorithm for word sense disambiguation using WordNet computational linguistics and intelligent text processing. In Alexander Gelbukh, editor, Computational Linguistics and Intelligent Text Processing, volume 2276 of Lecture Notes in Computer Science, chapter 11, pages 117–171. Springer Berlin / Heidelberg, Berlin, Heidelberg, February 2002.
* G. Corso, A. Gulli, and F. Romani. Ranking a stream of news, 2005.
* Monika Henzinger, Bay W. Chang, Brian Milch, and Sergey Brin. Query-free news search. In Proceedings of the 12th international conference on World Wide Web, WWW ’03, pages 1–10, New York, NY, USA, 2003. ACM.
* G. Holmes, A. Donkin, and I. H. Witten. WEKA: a machine learning workbench. In Intelligent Information Systems,1994. Proceedings of the 1994 Second Australian and New Zealand Conference on, pages 357–361, August 2002.
* Karen S. Jones. A statistical interpretation of term specificity and its application in retrieval. In Journal of Documentation, volume 28, pages 11–21, 1972.
* Karen S. Jones. Towards better NLP system evaluation. In HLT ’94: Proceedings of the workshop on Human Language Technology, pages 102–107, Morristown, NJ, USA, 1994. Association for Computational Linguistics.
* Jure Leskovec, Lars Backstrom, and Jon Kleinberg. Meme-tracking and the dynamics of the news cycle. In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ’09, pages 497–506, New York, NY, USA, 2009. ACM.
* David D. Lewis and Karen S. Jones. Natural language processing for information retrieval. Communications of the ACM, 39(1):92–101, 1996.
* Zhiwei Li, Bin Wang, Mingjing Li, and Wei Y. Ma. A probabilistic model for retrospective news event detection. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR ’05, pages 106–113, New York, NY, USA, 2005. ACM.
* Levon Lloyd, Dimitrios Kechagias, and Steven Skiena. Lydia: A system for Large-Scale news analysis. In Mariano Consens and Gonzalo Navarro, editors, String Processing and Information Retrieval, volume 3772 of Lecture Notes in Computer Science, chapter 18, pages 161–166. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2005.
* Dragomir Radev, Jahna Otterbacher, Adam Winkel, and Sasha B. Goldensohn. NewsInEssence: summarizing online news topics. Commun. ACM, 48(10):95–98, October 2005.
* Yusuke Shinyama and Satoshi Sekine. Named entity discovery using comparable news articles. In Proceedings of the 20th international conference on Computational Linguistics, COLING ’04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.
* K. Sparck. Some points in a time. Computational Linguistics, 31(1):1–14, March 2005.
* Ian H. Witten and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann, 1 edition, October 1999.
* Ron Zacharski. A programmer’s guide to data mining.[**http://guidetodatamining.com/,**](http://guidetodatamining.com/,) 2012.
* Natural Process API: Ling Pipe
* Google places vs. Open Street map - gazetteer (list of places)

**Technical/compsci challenges**

- access of API only from computer science computers

- finding the best attributes to use the machine learning techniques

- applying machine learning and statistics in code - this will be most time consuming as I am new to it.

- thinking of what to hold the data in to so I am able to do the analysis fast.

- not all data is ladled correctly and the articles can have misspelled words

**Other (non-technical) challenges**

- Newspapers have welsh articles - this might be an issue as don't want to not take them into consideration but I know nothing about the language;

**Ideal finished result and intermediate stages.**

The ideal finish result would be a web application that presents the data in a user friendly way(graphs, pie charts) depending on user input. And depending on results maybe an interactive map.

Intermediate stages:

- algorithm for finding only crime related articles

- algorithm of splitting crimes into different types

- analysing and finding statistics depending on age, or sex or place, or year

- making a website where these will be presented to users with nice charts and graphs

- look into making an interactive map of the crimes on years and places; (if data is sufficient)