Analysis of the digital historical newspaper collection at the National Library of Wales

Report Name	Outline Project Specification
Author (User Id)	Diana Silvia Teodorescu (dst1)
Supervisor	Amanda Clare (afc)
Module	CS39440
Degree Scheme	G600 (Software Engineering)
Date	February 10, 2014
Revision	0.1
Status	Final Release

1. Project description

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 point of the project is to discover interesting themes and concepts (research studies, advertised services, positive/negative) within the articles and using data mining processes and statistics to present them in an interesting way to the user.

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.

In my project I am choosing to analyse specific articles related to crimes in the past by classifying them into type of crime and making connections between age, sex, location and year to come up with interesting statistics.

2. Proposed tasks

The ideal finish result would be a web application that presents the analysis in a user friendly way(graphs, pie charts) depending on user input.

The intermediate stages proposed are:

- Initial research about machine learning techniques and natural language processing and which programming languages are the best to use with large amount of data and text processing. During this stage there will be journals and books to be studied but also APIs and libraries such as: Apache OpenNLP library, LingPipe API, Stanford NLP, Weka;

- Describing the structural patterns [1] that constitute a crime article at first and then the ones that represent the different type of crimes. Once the variables that define a crime article are found the machine algorithm can be trained.

- Training and testing the machine learning algorithm that will result in finding the crime related articles from the entire data. Choosing the learning technique: supervised, unsupervised or reinforcement learning [2].

- Finding the algorithm for splitting the crime articles into categories: murder, suicide, property damage, bodily harm, rubbery and using machine learning to get the desired results.

- **Choosing a new data structure** to hold the resulted articles;

- Analysing and finding statistics related to age, location, sex, year. This can result into interesting discoveries about the amount of crime and weather it went up or down in a certain location or at a certain point in the past. This step requires research and a knowledge of coding techniques used in statistical studies.

- Making a website where the results will be extracted from a database and presented using nice charts and graphs. This will be done using JavaScript and HTML5.

3. Project deliverables

Abstract - A deliverable used in my first supervisor meeting. This material contains in the first paragraph my impressions and ideas regarding what the project is about. It was also a presentation of the aspects of computer science I would need to use and which parts of the project require more in-depth research.

Dissertation blog - This will be used to keep a constant review of how the project is moving forward. It will be an online blog that presents the achievements and struggles during both the research and coding part in an informal way.

Outline Project Specification - The material requires research regarding the project idea and also a well thought strategy for what will happen moving forward. At this point the steps written will be prone to change so only the basic general stages are known.

Feature list/Diagram - A short document presenting the list of features and the connection between them. The features list will be created by carefully examining the proposed tasks and splitting them into smaller ones. This will also be used as a check list to review my progress in time.

Algorithm Report - This will describe some of the proposed algorithms to be used for the analysis of the newspaper text and the variable that can be used to predict that the right set of articles is selected.

Mid Project Demonstration - I aim to have a working algorithm that can extract correctly the crime articles from the entire data and save the results into a data structure.

Final Report and project code - This will be the final deliverable and it will contain the entire process from start to finish. There will be detailed information regarding the starting requirements, the design, the testing strategies and the changes that were made along the way.

4. Initial annotated bibliography

[1] 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.

This is an introduction to machine learning that will help me understand it's basics and how to correcly represent an algorithm

[2] Andrew Ng of Stanford University. Machine learning lectures. Online Course iTunes U.

The courses are meant to be an introduction into machine learning

[3] 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.

The journal presents one of the machine learning suit built for JAVA that I intend using to apply the algorithms with.

[4] Karen S. Jones. A statistical interpretation of term specificity and its application in retrieval. In *Journal of Documentation*, volume 28, pages 11–21, 1972.

This journal will help me with finding the structural paterns and gathering the right data for making the final statistics.

[5] 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.

This will provide more knowledge regarding the assement strategies of training and test data in natural language processing.

[6] 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.

The article presents a system that analyses news articles. It describes the steps the software has in it's process.

[7] 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.

A journal that addresses the name entity tagger idea and the importance of using part of speach in language processing

[8] K. Sparck. Some points in a time. *Computational Linguistics*, 31(1):1–14, March 2005.

An article about the information and language processing focusing on statistical methods that would help with achieving the right variables for the text analysis.

[9] Ron Zacharski. A programmer's guide to data mining. http://guidetodatamining.com/, 2012.

A ongoing blog written as a book about data mining basics for beginners.