Contents

Chapter 1

Introduction

1.1 Project summary

1.2 Background

1.2.1 Project choice

these are my first steps [**?**]
To justify the choice of my project I need to describe my background. I have previously graduated a undergraduate and master's degree in Marketing. This has developed my passion for research and as part of my study scheme I had to prepare surveys, apply them and analyse the gathered data in statistic analysis software such as SPSS (pretty graphs and statistics). Later, during my industrial placement while working with clients I felt that the company can really benefit from a system that could analyse client feedback(language processing). When asked to evaluate and choose from tens of major project suggestions I was naturally drawn to the research ones.

The Analysis of the Historical Newspaper data was my first choice as it seemed to have more freedom regarding what direction I want to go on. There were several options: improving the API, improving the text using image processing, correcting and transforming the articles by removing unnecessary characters and checking spelling errors done by the scanning process or using the actual articles to obtain information about historical events.

The National Library of Wales has made available online the historical newspaper archive for research purposes. The entire collections has over .... newspapers from a period of over 100 years. The articles contain news and

events that extends over a variety of domains such as: science, events, sport, politics, crime, advertisement. These are not classified as the news are today in sections depending on theme and there is no evident connection between the type of article and the page it's found on.

During the initial stage I considered different options and ideas about how the data can be used. I knew my goal was to make it interesting not only for me to work on but also for a future user that might want to discover specific facts about events that occurred a long time ago.
I fiddled with ideas such as:
Advertisements in Wales history - analysis of all advertisements, what type products/services are more advertised? How it changed in time? What product services are mentioned in news articles;
Scientists in Wales - discover all the mentions of scientists, and work out information such as: are they in the news for their science or something else (crime, business, politics etc), what kind of science, were they living in Wales, what kind of language was used to describe them (positive or negative), did this reporting change over time, what words were most frequently used to describe them? Crime Stories - Separating only the crime specific articles and then developing an algorithm that can split them into different types of crime: theft, murder, fraud, bodily harm, no crime. After applying the model on all the articles, time allowing, I will apply some statistic and text processing techniques to find different connections between place, sex, year and then showing them to the user on a website. The decision was made towards this topic because it combines natural language processing and machine learning and it also has the advantage of having interesting results that many can be shocked but also intrigued by.

Sentiment Analysis and formal news data - This was an interesting idea that I have researched but after reading more of the articles I realised it is a very difficult task for a human to realise weather a news article is positive or negative (for example text about a specific gathering) and this would make it even more difficult for a computer. Sentiment analysis would be more appropriate for subjective text or one that reflects an opinion and not formal text such as a news article.
Text classification is important especially nowadays during the amount of Big Data available over the internet. On most news websites the articles are labelled and introduced into categories by humans but with text from scanned books or newspapers the amount of data that would need to be analysed is far too large for manual labelling. This is why machine learning techniques have been used and improved during the years.

The UCOTP project will use general classification to split articles topic into separate types of crimes. I have chosen a fixed number of categories: murder, robbery, fraud, no crime, assault. Articles can result in being part of multiple of multiple categories. e.g. A rubbery that has as result the loss of human life would be considered apart of both murder and robbery classes.

Machine learning is a process that happens in stages and requires a training data set that it can make connections about the commune variables for each class and learn to make future classifications.

The National Library of Wales is the only entity that holds the digitised newspapers dating from 1804 - 1919 and in contrast to how newspapers are structured and written now these don't have a known structure divided by topics or page numbers, we can also think of a language barrier between the past and the present for a computer to process. Because of this using a data set found online to create a model for the classification could end up with a increased margin of error so finding and manually labelling the crime articles to set-up a training set would be the first stage of the project.

1.2.2 Using Natural Language Processing

1.2.3 Similar systems in scientific papers

1.3 Objectives