**Introduction**

The National Library of Wales has made available online the historical newspaper archive for research purposes. The entire collections has over 650k newspapers from a period of over 100 years. Currently the project is being developed on a test server with only 65k articles.

The articles contain news and events that extends over a variety of domains such as: science, events, sport, politics, crime, advertisement. These are not always 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.

The main goal of the project is 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.

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. For this after the extraction of all the articles in a database there will be a manual labelling stage.

**The work so far**

XML from the National Library API is being parsed using DOM implementation in Java.

Getting articles that will be in the implementation by making a query to the API. The query will select only articles that contain the following words in the text: robbery, murder, crime, thief, arrested, fraud, assault, kidnapping, killer, victim;

Getting verbs from sentences works: combining Stanford POS tagger Library and a few .replace and .split and .contains methods.

Object is crated with the necessary data from the Articles and an array of these type objects.

Part of background document written (1100 words ).

Set-up Git and my data/repository is on Dropbox (nice, easy and safe);

**Next steps**

Creating the database structure and filling it with the extracted data;

 Developing the model: play with weka, extract data from database in right format, test the model.

Applying model on rest of data

Depending on time and results there will be a decision regarding which way to go with the project: website and user interface or improving the ML algorithm

**Issues:** duplicate articles, extracting data from api doesn't work when over 20000 articles, split articles;