- make own interactive map representing through different colours events/news(colour depending on domain or feeling) - you could zoom in to a specific area and see more news in details.

calculate their importance depending on how many stories on them (maybe connecting to other data online)

1. get all articles applicable to zones in Wales - extract area or address if possible

2. sorted by domains or moods

3. sorted by dates

4. check which events reoccur in multiple newspapers or articles - compare to external resources -> have a top most important news depending on that

5. colours show domain/mood

6. zooming in into area results in change of colours and other news

7. get be applied on domain: biology, science, crimes, advertisements, sports

difficulties:

- no exact address or area for some news

- places that existed 100 years ago don't exist now

Classifying the types of news stories: can we use machine learning to   
label the news stories as politics, local, sport, education, sci+tech,   
business and so on?

So my immediate questions would be:

**The map:** what kind of map would you use? Google maps? open street map?

Do you know how they work and how you can add information to them?

API are free and available for both.

Do you know what the copyright restrictions on using them might be?

How would you recognise "regions and addresses"?

Google places API

Do you have a list of all the towns in Wales?

What do they look like?

What software exists for finding these?  
  
**Mood:** can you give several examples of what a positive mood story looks   
like (and negative mood stories)?

Can you make a decision for every story or are there some that are neither (if so, how many are there, is   
it too many or is it fine? and what would you do with these?).

What software exists for finding mood/sentiment?

Alchemy API - AlchemyAPI provides easy-to-use facilities for extracting positive/negative sentiment from any textual content

uClassify - free web service where you can easily create your own text classifiers. You can also directly use classifiers that are shared by the community

**Domain:** what do you mean by this, can you be more specific?

sport, crime, science, biology? How did you choose this list?

What would the full list look like?

If you picked 10 stories at random for a particular day, could you label them all?

What software exists for classifying domain of news stories?

Advertisements analysis - Advertisements in Wales history: analysis of all advertisements, what type products more advertised? , how it changed in time? , are there any products mentioned in news articles? (positive or negative);

**That day this place:** android app; connect Google coordinates and date to find articles about a specific day an place on desired domains; events, crimes

**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?  
  
  
  
  
Classifying the types of news stories: can we use machine learning to label the news stories as politics, local, sport, education, sci+tech, business and so on?  
  
**- what your project is about (the aim and the background)**

**Understanding crimes of the past**

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.

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.

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.

**- what your reading list looks like**

**-** [Data Mining - Mark Hall, Ian Witten](http://www.amazon.co.uk/Data-Mining-Practical-Techniques-Management/dp/0123748569/ref=sr_1_1?ie=UTF8&qid=1391176416&sr=8-1&keywords=data+mining)

- Natural Process API: Ling Pipe

- Concordance and co-occurrence

- Mutual Information (transinformation)

- Karen Sparck Jones - inverse document frequency

- Part-of-speech tagging

- Google places vs. Open Street map - gazetteer (list of places)

**- what technical/compsci challenges you'll be facing**

- 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

**- what other (non-technical) challenges you'll have (for example needing   
info from other people, understanding the language of another domain, etc)**

- 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;

**- what the ideal finished result would look like and if there are useful   
intermediate stages where you could stop**

- I could stop after having an algorithm of splitting crimes into different types

- Then I could stop after analysing them depending on age, or sex or place, or year

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

- If I have time I would like to look into making an interactive map of the crimes on years and places;

week 10th

- choice between xml and json to extract text

- taking as string and parsing with dom to xml object

- some issues with xml being more complex: tags and attributes

- encoding and strange characters appearing - tried different encoding types on the inputstream and doc ended up replacing the specific characters with empty string;

- thinking if other parser would be better?

-

**data flow from article text to database and back**

**- get text, split text, take verbs add to array**

**- add array as diff columns in database**

**- get new text, split text, take verbs add to array,**

**- get verbs columns in other array, compare with new articles verbs and make bigger array and add to database**

**questions:**

**how to get text? array of words that get put into url?**

**what words to describe crime? -> read more articles check common words/verbs**

**when are we sending things to database?**

**Decisions to make:**

java or python => java

xml or json => xml

hashmap or array of objects

nosql or sql

weka/ javaml/stanfordnlp/ opennlp => stanford

**Abstract**

**Understanding crimes of the past - a machine learning look into the 19th Century news**

**crime mystery in the 19th Century**

The poster will present my research project that will be an analysis of the data collection provided from the historical newspapers found at the National Library of Wales. The aim of the project is to discover an algorithm that will allow a classification of crime related articles into different types of crime. Its outcome will be a model that can then be applied to the rest of the articles.

The process will involve machine learning techniques and natural language processing techniques such as name entity recognition, POS tagger and tokenizers. Natural language processing is a hard task for a computer but with the right data sets, patterns and variables a machine learning model can teach us a lot about the labels we want to attach to each article. Looking into the process of teaching computers to understand human language is changing the field of artificial intelligence and can improve not only the business/marketing field but also social and medical fields.

The final goal is not only to have an effective model but also to extract scientific data and answer questions such as: What are the words that better describe a crime article? Where the articles in a specific part of the newspaper, on a specific page? Were specific locations more prone to crime activity? Did crime dropped or spiked during the years?

I believe that history is overlooked nowadays by the young generation and it's important to have the data available online and make it exciting to look at and easy to research so presenting it through statistics and info graphs would be the best choice.