Dissertation 1 - Abstract

"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 within the articles and using data mining processes and statistics to present them in an interesting way to the user."

This is how it starts - my final year project at Aberystwyth University. Doesn't seem very clear right? It wasn't for me - I haven't done any machine learning or statistics during my time at this university => Terrifying! But also exciting, I mean it's not like I haven't been through this a couple of times before. I guess I always knew my previous interest in data gathering and analysis during my marketing degree won't fade. I just need to mix this up with a programming language and some web development. Easy?!?

First supervisor meeting: Friday 30 January. "I can do this!" - my mind (unsure and full of questions ) was encouraging me to walk the last pair of stairs in the Computer science department with my heart rate slightly elevated.

Dissertation 2 - First meeting

My supervisor, Amanda welcomed me with an understanding smile .. "It's normal to be nervous, it's the first week!" and just like that my nervousness went out the window.

In the first part of the meeting I wasn't sure which way to go with the amount of data I have. Little by little we discussed all my ideas:

1. Sentiment Analysis ?! - maybe interesting and easy with blogs and tweets but a bit tough on formal newspapers.

2. Interactive Wales Map with all the local news split on domains and mood? - is this really a project for one semester? - should think methodically, in steps that can be ended at any point and still offer a finite product.

3. I got a first introduction into machine learning: articles and domains and attributes to connect them both together.

Maybe think of: specific words frequency (football, tennis - they would appear in sport articles), pair of words, maybe the length of the article or its position in the newspaper (first page? third page?) Is there a pattern?

4. Are only crime articles easier to analyse rather than classifying every article? There could be different types: theft, murder, attempted suicide, property damage, bodily harm. Does the crime rate spiked in specific years? Did it depend on age or sex?

5. Why can't I access the API from my computer? :( It seems like at the moment only computer science machines are allowed to access the API. Still need to try accessing it through SSH. The good thing is that the data on it is a fraction of the one they have on their servers just so we can play with it without disturbing the real user.

Homework:

- Concordance and co-occurrence

- Mutual Information (transinformation)

- Karen Sparck Jones - her study on words used in similar context and her contribution reagrding inverse document frequency

- Part-of-speech tagging

- Machine Learning - Weka (Data Mining - Mark Hall, Ian Witten), Orange, SCIKIT ML

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

A productive meeting and I must admit I was more relaxed than expected (I don't usually feel at ease about discussing my worries and asking questions face to face) .

I was so excited about getting the Data Mining book! After 4 pairs of stairs and 30 minutes of waiting for the nice library lady to decipher my writing and search for the book there was no positive result.

-------------------------

Dissertation 3: Everyone else

The second meeting I got to know the other students that will be working with Amanda and we all discussed our projects. It was interesting to find out that I'm not the only one nervous and with a lot of work ahead. We all are doing research on different domains: natural language processing, machine learning, interactive web programming, data mining, image processing some even need to look outside their domain (biology, insects, DNA, brain connections).

So on Thursday some of the pressure seemed to lift after hearing their ideas, their worries, the parts they are afraid of and how much work we all need to do. I'm not alone!

Friday I got some news that brought back the pressure but not in a bad way: the chair in Digital Collections from The National Library of Wales wants to meet me and talk about how I'll be using their collection. I was excited as I know if I do this right it would be an amazing achievement for me. I must do this right.

Other news:

I can finally access the API from my own computer. They were nice enough to give me my own username and password.

I understand how the URLs work and I can easily query what I need, it's pretty straight forward.

Started coding JAVA methods to extract the data as a String and now looking into how to parse it to an XML object. This will help me with tokenising, tagging and counting and basically getting me used to using the data in any way I want. For not I hardcoded the username and password using the [Authenticator](http://docs.oracle.com/javase/7/docs/api/java/net/Authenticator.html) Class.

At the moment the plan for the week is:

- Coming to university each week day

- Make the feature list and split everything into small tasks

- Play with XML

- Read all journals found on NLP

- Understand Apache OpenNLP

- Think about best way to prepare and hold the articles for data processing

- \*Apply to some jobs (side project)

Dissertation 5: Until now

After being gone for a few days I felt a bit bad going to the meeting with nothing really important to show.

What I have until now:

XML from the National Library API is being parsed using DOM library.

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

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

Part of background document. More exactly 540 words.

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

Dissertation 6: The next stages

Not many weeks left, the plan is: (as discussed in last meeting with supervisor)

1. Get all training data in database:

Subtasks:

- Make list of words the initial search will be done on

- Don't add duplicate articles just because they have multiple words from the above list

- Test if article is split on two pages and connect the article text

- Extract verbs and remove duplicates,

2. Database design and set-up

- Choose database and design structure

- Add all articles to database

- Label the training data manually (>500 articles)

3. Machine learning stage 1: Developing the model

- Research and play with Weka (look at Apache Mahout, Java-ML )

- Develop the model that classifies the articles.

4. Machine learning stage 2: Testing of Model

- Getting new articles to test the model with

- Tweaking the model

5. Machine learning stage 3: Applying the model on the rest of the data.

- verify article is not part of training model

- get final result into final database.

All these should be accompanied by constant documentation writing and regular blog posts. (Even if the coding is really the fun/exciting part for me! )

Next: 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?

Dissertation7: Decisions decisions..

Time is passing by quick and there is no more time for indecisions so what to do when I have a few?

Dissertation 8: Parsing the XML

I am working with a XML of the following form:

<response>

<lst name="responseHeader">

<int name="status"></int>

<int name="QTime"></int>

<lst name="params">

<str name="q"> </str>

<str name="rows"></str>

</lst>

</lst>

<result name="response" numFound="2842" start="0">

<doc>

<str name="PID"> </str>

<str name="ModsArticleID"> </str>

<str name="Region"> </str>

<str name="ArticleID"></str>

<str name="ArticleTitle"></str>

<str name="ArticleAbstract"></str>

<int name="ArticleWordCount"></int>

<str name="PageLabel"></str>

<str name="PagePID"> </str>

<arr name="ArticlePagePID">

<str> </str>

<str> </str>

</arr>

<str name="PageCode"> </str>

<str name="PublicationTitle"> </str>

<str name="TitleCode"> </str>

<str name="TitlePhase"></str>

<date name="IssueDate"></date>

<str name="PublicationPID"> </str>

</doc>