Topic modelling in the archives

Author : Tim Sherratt

Tagged as : archives, invisibleaustralians, topic modelling, White Australia

Date : May 17, 2012

There seems to be a lot of topic modelling going on at the moment. Any why not? Projects like Mining the Dispatch are demonstrating the possibilities. Tools like Mallet are making it easy. And generous DHers like Ted Underwood and Scott Weingart are doing a great job explaining what it is and how it works.

I've talked briefly about using topic modelling to explore digitised newspapers, something that the Mapping Texts project has also been investigating. But I've also been following with interest Chad Black's use of algorithmic techniques, including topic modelling, to look for local variations amidst the legal system of the early modern Spanish empire.

As part of the Invisible Australians project, Kate and I are exploring the bureaucracy of the White Australia Policy. In particular, we're interested in the interaction between policy and practice, between the highly-centralised bureaucracy and the activities of individual port officials. Like Chad, we're interested in mapping local variations — to try and understand the bureaucracy from the point of view of an individual forced to live within its restrictions.

I recently gave a presentation about the project at Digital Humanities Australasia (post coming soon!), and in preparation I decided to try a few topic modelling experiments. They were very simple, but I was impressed by the possibilities for exploring archival systems.

The problem I started with was this. The workings of the White Australia Policy are well documented by records held by the National Archives of Australia. Some series within the archives are specifically related to the operations of the policy — such as those containing many thousands of CEDTs. But there are also general correspondence series created by the customs offices in each state, as well as the Commonwealth Department of External Affairs which administered the Immigration Restriction Act (responsibility was later taken by the Department of Home and Territories and it's successors). These general correspondence series are important, because they often include details of difficult or controversial cases — those that required a policy judgment, or prompted a change in existing practices. But how do you find relevant files within series that can contain large numbers of items?

Series A1, for example, is a correspondence series created by the Department of External Affairs. It contains more than 60,000 items. Past research tells us that amongst these 60,000 files are records of important policy discussions relating to White Australia. But these files tend to be labelled with the names of the people involved, so unless you know the names in advance they can be difficult to find.

Mitchell Whitelaw's A1 Explorer, part of the Visible Archive project, lets you to explore the contents of Series A1 in a easy and engaging way. But while the A1 Explorer provides new opportunities for discovery, it doesn't offer the fine-grained analysis we need to sift out the files we're after. And so… topic modelling.

The process was pretty simple. While I can dip into my bag of screen-scrapers to harvest series directly from the NAA's RecordSearch database, there was already an XML dump of A1 available from data.gov.au. So I extracted the basic file metadata from the XML and wrote the identifiers and titles out to a text file, one item per line. Following the instructions on the website I then loaded

this file into Mallet:

```
/Applications/Mallet/bin/mallet import-
file --input ./A1.txt --output A1.mallet --keep-sequence --remove-stopwords
```

Then it was just a matter of firing up the topic modeller:

```
/Applications/Mallet/bin/mallet train-topics --input ./A1.mallet --output-state ./A1.
gz --output-doc-topics ./A1-topics.txt --output-topic-keys ./A1-keys.txt --num-
topics 40
```

Again, I just followed the examples on the Mallet site.

Once it was finished I opened up A1-keys.txt to browse the 'topics' Mallet had found. The results were intriguing. There are a large number of applications for naturalisation in A1, so it's no surprise that 'naturalisation' figures prominently in a number of the topics. What was more interesting was the way Mallet had grouped the naturalisation files. For example:

naturalization christian hans hansen jensen petersen andersen nielsen larsen christensen johannes jens niels pedersen andreas johansen martin jorgensen

and

naturalisation certificate giuseppe salvatore frank la leo samios spina sorbello leonardo fisher natale patane torrisi barbagallo luka rossi ross

Based on the co-occurrence of names within the file titles, Mallet had created groupings that roughly reflected the ethnic origins of applicants. It makes sense when you think about what Mallet is doing, but I still found it pretty amazing.

Mallet also found clusters around the major activities of the department, such as the administration of the territories. But of most interest to us was:

1 0.55539 passport ah student exemption students lee wong chinese young deserter education sing wing chong readmission son hing chin wife

The Chinese names alongside words such as 'readmission' and 'wife' suggested that this topic revolved around the administration of the White Australia Policy. This was easy to test. In A1-topics.txt was a list of every file in the series and their weightings in relation to each of the topics. I wasn't sure what was a reasonable cut-off value to use in assessing the weightings, but after a bit of trial and error I fixed on a value of 0.7. I then just extracted the identifiers of every file that had a weighting greater than 0.7 for this topic. I used the identifiers to build a simple web page that Kate and I could browse. I also included links back to RecordSearch so we could explore further.

It’s a pretty impressive result. Instead of fumbling with the uncertainties of keyword searches, we now have a list of more than 1,300 files that are clearly of relevance to Invisible Australians. There’s a few false positives and there are likely to be other files that we’ll have missed altogether, but now we have a much clearer picture of the types of files that are included and how they are described.

And that was at my first attempt, simply using the default settings. I’m now starting to play around with some of Mallet’s configuration options to see what sort of difference they make. I’m also keen to try out GenSim, a topic modelling package for Python.

I’m really excited about the possibilities of these sort of tools for analysing the contents of archival descriptive systems, something I mentioned in my Digital Humanities Australasia paper. Much more to come on this I suspect…

Share this:

- Click to email this to a friend (Opens in new window)
- Click to print (Opens in new window)
- Click to share on Twitter (Opens in new window)
- Share on Facebook (Opens in new window)
- Click to share on Google+ (Opens in new window)
-