NSW GOVERNMENT State Archives & Records

Case Study – External Pilot – Machine Learning and Records Management

September 24, 2018

Conclusion of our exploratory adventure

Just over 12 months ago the Digital State Archives team here at NSW State Archives was challenged to explore the application of machine learning to records management with a specific focus on digital disposal. We began by publishing a preliminary research blog which assessed the 'state of art' of the technology, the 'state of play' in terms of its uptake for records management, and which sketched out a plan of action. We identified a need in terms of a general lack of uptake in our jurisdiction and proposed undertaking a series of pilot projects to demonstrate the potential benefits of the technology. Our goals with the pilots were to get hands on experience, to experiment with different algorithms and specially prepared data sets, to assess the potential of the technology, and to share the results with our jurisdiction.

We shared the results of our first pilot project, an internal pilot conducted in November and December of 2017, in this blog post: //futureproof.records.nsw.gov.au/case-study-internal-pilot-machine-learning-and-records-management/. This post describes our second pilot project.

Second Pilot Project

The second pilot project was conducted in June and July 2018. For this project, the Digital State Archives team had the opportunity to work collaboratively with the staff from Corporate and Ministerial Services at the Department of Premier and Cabinet (DPC). The DPC supplied a pre-sentenced corpus of 81 GB of digital documents from their Objective (EDRM system) and XML metadata for over 21,000 pre-sentenced paper files.

This pilot differed from the earlier, internal pilot project in that it: involved an external agency partner; included a much larger labelled (pre-sentenced) corpus; and involved a much more diverse set of disposal authorities and classes. This extra size and complexity made it a good vehicle for validating the first pilot's findings. We decided early on to use the same algorithm (the Multi-layer Perceptron) that was used in the first pilot and to re-use our initial set-up with scikit-learn. The main point of difference with the first pilot was the use of Aspose to extract text from the digital documents which proved more robust and gave us more data for our model.

The data

The total number of digital documents that were classified and extracted from the DPC Objective corpus for use in this pilot numbered 108,064. In preparing the data for the model, the text extraction process started on 27 June 2018 and completed on 5 July 2018: a total of 9 days. During this process a number of the files failed to successfully extract text. This was in line with our expectation that a proportion of files, such as digital images or graph data, would not be suitable for classifying with the model. This left 86,453 usable documents that were loaded into CSV files to run through the model.

The tests

After the first run of the model, we were surprised at suspiciously high estimations of 91% and 97% predicted success. These early figures were unfortunately too good to be true and were caused by a weighting problem. The corpus, we discovered, had a very dominant class: FA254-02.02.02. Almost two-thirds of the corpus (or 62,988 files) was classified with this one class. During this investigation it was also discovered that other classes had too few documents to perform adequately, due to a lack of training data.

In order to make the training set more representative, and therefore make the predictive results more realistic, we determined that it would be necessary to run tests excluding this class. We also decided to test how well this class performed alongside other classes just within FA254 (i.e. excluding other disposal schedules).

Ultimately we ran a series of different tests, using parts of the corpus classified against different classes and schedules, to see how predictive accuracy varied according to disposal class/schedule. Confusion matrices have been produced for a number of the tests that demonstrate how the model performed and where it was lacking. For larger test runs with multiple values, the confusion matrices don't provide useful insights (as they are too confusing!), and so results presented for the larger tests are just the predictive scores.

The following "text" only (i.e. just using text as a feature and not including other features such as Objective metadata) tests were run:

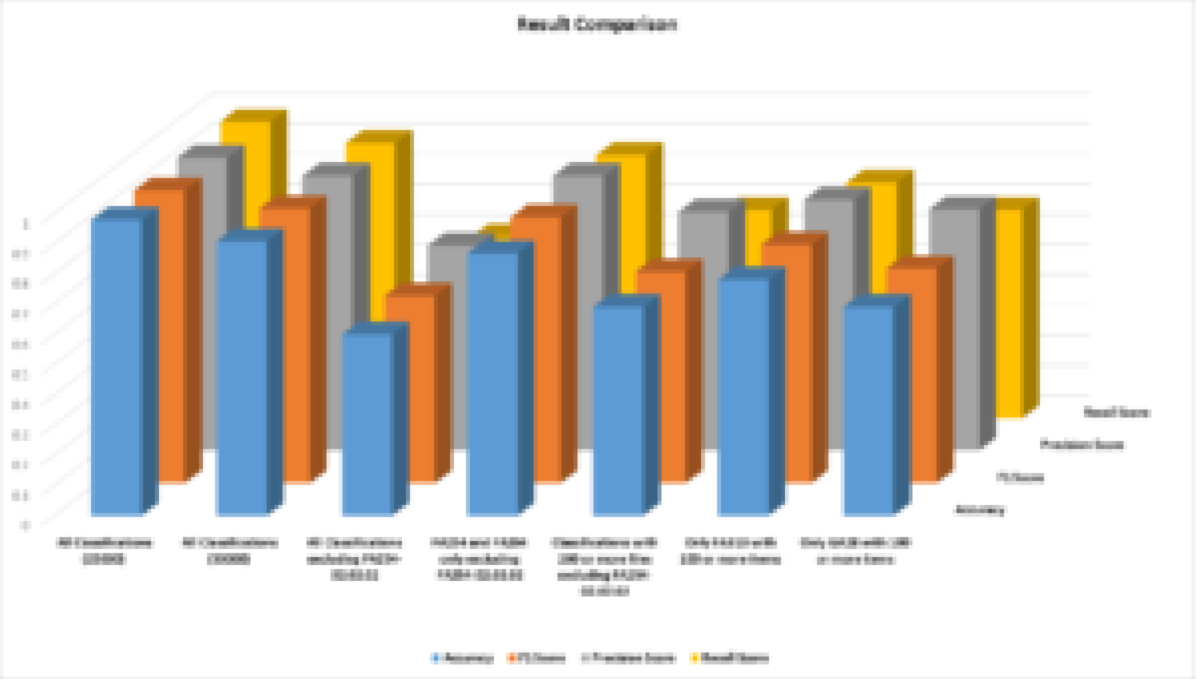
1. Test 1 Whole corpus (without FA254-02.02.02)
2. Test 2 FA 254 (with FA254-02.02.02)
3. Test 3 FA 254 and FA 294 (without FA254-02.02.02)
4. Test 4 FA 313 with at least 100 files per class
5. Test 5 GA28 with at least 100 files per class
6. Test 6 Whole corpus (without FA-02.02.02) with at least 200 files per class
7. Test 7 XML metadata (physical files)

We conducted a final test for the XML encoded metadata of 21,805 physical files (i.e. no digital files that could be text extracted for classification). A sophisticated test would have made use of the different structured fields in the metadata (i.e. title, author, etc. fields could be separated as distinct features), however ours was a rough test that simply applied the same text-based classification model used in the other tests but using the aggregated metadata as the input "text".

The Result

The overall results from the tests within our second pilot gave us new insights into how the model behaves and showed that there is a need for more investigative work, particularly into the ways that feature of the corpus (size and complexity of disposal coverage) affect the results.

Data	F1	Precision	Recall	Sample size
Trial 1. Sample of whole corpus	0.97	0.97	0.98	25,000
Trial 2. Sample of whole corpus	0.91	0.91	0.91	49,999
Test 1. Whole corpus (without FA254-02.02.02)	0.62	0.68	0.60	30,400
Test 2. FA 254 (with FA254-02.02.02)	0.97	0.98	0.98	62,988
Test 3. FA 254 and FA 294 (without FA254-02.02.02)	0.88	0.91	0.87	7,427
Test 4. FA 313 with at least 100 files per class	0.79	0.83	0.78	3,180
Test 5. GA 28 with at least 100 files per class	0.71	0.80	0.69	15,276
Test 6. Whole corpus (without FA254-02.02.02) with at least 200 files per class	0.70	0.79	0.69	25,568
Test 7. XML metadata (physical files)	0.69	0.71	0.68	21,805



Discussion

The primary goal of this external pilot was to test how well the in-house NSW State Archives machine learning model fitted the DPC Corpus. The results, which were broadly comparable with the internal pilot, show that there is definitely promise in the approach.

The extra complexity of multiple disposal schedules and the dominance of a single class in this corpus demonstrate the need to proceed cautiously and ensure that you measure from a level base line.

One of the interesting features of this pilot was the diversity of disposal classes and the clear impact that the different retention and disposal authorities and classes had on precision and recall. Further research into why certain classes performed better that others is recommended. Such research would not only assist in improving machine learning classification but might also inform decisions made in the future about how disposal classes are defined.

A challenge we faced at NSW State Archives was a change in the team's composition (caused by rotation of our graduates) and needing a new staff member, who had not participated in the initial pilot, to pick up that work and extend it. This was a highly complex task for a new resource and highlights the importance of writing well-documented and well-tested code when developing internal software projects.

The production of a machine learning-based classification service would have value for the NSW jurisdiction but it requires an initial input of a significant amount of classified data to deliver good predictive values. At this stage there is still a requirement for human checking and fine tuning to ensure that classifications are sound.

The Potential

At the beginning of our exploratory look at Machine Learning we posed the following questions:

Can machine learning fit within the NSW recordkeeping regulatory environment and what would it look like?

and

Is there an opportunity to operationalise the technology to provide a solution to NSW public offices?

At the end of this project, these questions remain unanswered but worth pursuing.

The two pilot projects we have conducted demonstrate early promise, with a range of results between 70% and 80% accuracy. It was worth noting that only 100 lines of code produced these preliminary results. The code we used is here and is available on https://github.com/srnsw/machine-learning-pilot.

We learnt that expertise needs to be gathered and nurtured when working on machine learning projects and that not one person has it all. There is a need for both specialist ICT/ data science knowledge as well as subject matter experts who understand the content of the data and the nature of the rules that apply to that data.

The need for a large data corpus that is well labelled and representative to train the model stood out as being imperative. This is much more important than choosing particular algorithms or machine learning technologies. In practical terms, this means that we really need to do a lot of digital sentencing before we can benefit from automated sentencing. Where possible we should also share and pool our data, especially where it relates to common disposal authorities such as GA28.

Further assessment and research is required to understand why and how the model is predicting the results that it generates. More human input is needed not only to check the results but also ensure that the data used to train the model is accurate. Bad data produces bad predictions.

Further research is also required into the ways we should adjust our processes and instruments in light of this technology. Our preliminary results could inform an analysis of what types of classes perform weakly and which are strong, and this may eventually may inform future appraisal decisions. How can we develop smarter retention and disposal authorities?

Ethics and bias needs to be part of the discussion too. The application of technology does not remove the need for accountability for retention and disposal decisions. The risk of bias in data sets and/or in algorithms is a social issue which may be even harder to solve. Models do risk perpetuating existing social and cultural biases. When implementing machine learning tools these issues need to be taken into account or risk derailing future developments in this field.

Acknowledgements

I would like to thank the assistance and support from the team at Corporate and Ministerial Services, Department of Premier and Cabinet in this second pilot. I would also like to thank Mohammed Ibrahim (ICT Graduate) who was on placement at this time with the Digital State Archives team.

Author: Glen Humphries