**KILNER — Discovering**

<0 images>

**Discovering and Rediscovering Full Text: Unearthing and Refactoring**

Kilner, K. and Fitch, K.

AustLit contains thousands of full text items ranging from seminal works of 19th- and early-20th-century Australian literature through collections of early science and speculative fiction, to a large corpus of children’s literature, alongside selected criticism and scholarship. In addition, AustLit bibliographical records link outwards to tens of thousands of full text items available online.

This paper presents the results of a project undertaken by the AustLit team in 2014 and 2015 to totally refactor the existing AustLit full text corpus, including a massive expansion of the corpus by identifying and harvesting literary texts published in newspapers in the period covered by the National Library of Australia’s (NLA) database of digitised newspaper available through Trove.1

A number of different formats and digitisation protocols have been used over the past 14 years to build a corpus of works that has the potential to support a range of different use cases. That potential had not been met until the total restructure of the AustLit database and content management system over the past two years provided an opportunity to look again at the material we have and the way we deliver that material to researchers and readers. A major factor in AustLit’s future plans to deliver full text is the NLA’s newspapers database. It offers a valuable opportunity to build our corpus and advance knowledge about the place of literature in culture and reading practices across the 19th and early 20th centuries. Newspapers were the primary form of transmission for literature during the period covered by the NLA’s database; the possibility of identifying and unlocking the literary content in the database thus allows us to support new research into reading culture.

This paper will present the refactored full text system AustLit developers have created to expand utility, readability, and research opportunities. It will also discuss an innovative method of identifying and harvesting poetry from the NLA’s newspapers database.

In July 2014, AustLit contained just over 10,500 links to poems identified within the NLA’s Digitised Newspapers collection. Each of these links had been manually created by a combination of inspired searching for words from a known poem, searching for known literary columns, and systematic browsing through each page of each issue of nominated newspaper titles across specific date ranges.

One of AustLit’s many new research projects is the Colonial Newspapers and Magazines Project, undertaken by researchers at UNSW, Canberra. This project is creating a literary ‘map‘ of Australia’s colonial period by collecting and recording information about the reading habits of Australians before 1900 and linking these findings into AustLit’s data structures. This huge task has been begun by concentrating on three specific years in the 1800s, and whilst producing accurate and near-complete results, the only feasible method of browsing every page of selected titles is extremely labour intensive.

Hence, we started exploring an automated approach to identifying at least some relevant content. As a starting point, we noted the effectiveness of Ted Underwood’s genre identification approaches on 17th- and 18th-century texts digitised by HathiTrust.2 We used his vocabulary information to produce two vocabulary frequency lists: one of all words and one of words found in works of poetry.

We first trained using a naive Bayesian classifier with the ‘all’ and ‘poetry’ vocabulary frequencies, and ran the trained classifier on a training set of newspaper articles identified as poetry and on another set whose genre was unknown.

The known poetry article list was derived from the 10,500 articles linked to as poetry in AustLit. The unknown set was generated by randomly selecting articles from NLA’s digitised newspapers. We found that whilst providing a useful signal, vocabulary alone was not sufficient to reliably classify articles as poetry or not-poetry. Examination of classification failures led us to explore additional signals to add to our classification heuristics:

• Text justification.

• Variations in length of successive lines.

• Apparently rhyming lines.

• Presence of digits in OCRed text.

• Presence of a small number of ‘marker’ words.

Our initial results correctly classified just over 80% of articles associated with AustLit poetry links as poetry. Manual examination of the articles not identified (false negatives) revealed that the vast majority were articles containing a small amount of poetry set within a sea of prose. A significant other group where written using a vocabulary not typical of poetry (words such as ‘proclamation’, ‘neutrality’, and ‘precautions’, which pushed the classifier towards rejection as poetry), and there was another large group of articles with such poor OCR that few words were accurately identified. An error in AustLit linking to the incorrect article was also identified.

We then measured the effect on the classifier of automatically correcting the OCR of articles and found it gave only slight improvements to false positives and negatives, because the predominant reasons for rejection of a known poetry article as non-poetry were not related to correctable OCR.

We then implemented the following set of refinements to the classifier, which lifted our successful classification rate to over 85% whilst keeping ‘false positives’ below 1%:

• Improved rhyming detection heuristics.

• Used article metadata to exclude advertisements.

• Internal article segmentation in an attempt to identify ‘islands’ of poetry contained in predominantly prose articles.

• Use of cues from social tagging and commenting.

We aim to harvest what appears to be vast numbers of poems published in Australian newspapers during the 19th and early 20th centuries and to deliver that full text to AustLit users within an enhanced discovery and reading environment. This project also allows AustLit to expand a project that has neither the funding nor the research staff to build on the initial 10,500 manually created records and links into the newspapers database. While the creation of nuanced, human-derived records is no longer possible for the colonial newspapers project, the hope is that this method will provide it with a data boost by building the AustLit full text corpus and record store with greatly reduced human input, thereby enabling the analytical research into the period’s reading and publishing culture.

**Notes**

1. http://trove.nla.gov.au.

2. See http://tedunderwood.com/2012/07/27/getting-everything-you-want-from-hathitrust/.