Digital Scholarship in the Humanities, 2022, **00**, 1–19
https://doi.org/10.1093/llc/fqac079
Full Paper

OXFORD

Creating specialized corpora from digitized historical newspaper archives

An iterative bootstrapping approach

Joshua Wilson Black [1,2]*

[1]UC Arts Digital Lab, University of Canterbury, Christchurch, New Zealand
[2]New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, Christchurch, New Zealand

*Correspondence: Joshua Wilson Black, UC Arts Digital Lab, New Zealand Institute of Language, Brain and Behaviour, University of Canterbury, Christchurch, New Zealand. E-mail: joshua.black@canterbury.ac.nz

Abstract

The availability of large digital archives of historical newspaper content has transformed the historical sciences. However, the scale of these archives can limit the direct application of advanced text processing methods. Even if it is computationally feasible to apply sophisticated language processing to an entire digital archive, if the material of interest is a small fraction of the archive, the results are unlikely to be useful. Methods for generating smaller specialized corpora from large archives are required to solve this problem. This article presents such a method for historical newspaper archives digitized using the METS/ALTO XML standard (Veridian Software, n.d.). The method is an 'iterative bootstrapping' approach in which candidate corpora are evaluated using text mining techniques, items are manually labelled, and Naïve Bayes text classifiers are trained and applied in order to produce new candidate corpora. The method is illustrated by a case study that investigates philosophical content, broadly construed, in pre-1900 English-language New Zealand newspapers. Extensive code is provided in Supplementary Materials.

1 Introduction

The digitization of historical newspapers allows researchers in the historical sciences to quickly access a massive and rapidly expanding range of primary source materials. Digitization also offers new opportunities for representing, and reasoning about, the content of such archives. These include, for instance, the application of methods for extracting topics or significant relations between terms from large collections of newspaper items (e.g. Smith *et al.*, 2014; Scheirer *et al.*, 2016; Alfano *et al.*, 2018). However, such methods are often impractical for the individual researcher to apply to entire archives. This is especially true when a researcher is interested in a specialized subject which is represented in a small handful of items. In such cases, even if unsupervised methods, such as topic modelling or clustering algorithms, can be applied to a complete archive, it is unlikely that anything of use to the researcher will emerge. To overcome this problem, it is necessary to select some relevant corpus of items without falling into the problems of simple keyword searches. This article presents a general method for generating specialized corpora from newspaper archives digitized using the common METS/ALTO XML standard (Veridian Software, n.d.) and illustrates it with a case study concerning philosophical discourse in early New Zealand newspapers.

The obvious approach to selecting texts from a digital archive of text content is keyword search. However, there is a growing literature raising concerns about the loss of context which often goes along with keyword searches. Indeed, one of the main motivations for using more sophisticated text processing methods is to gain insight into large-scale patterns in a collection of texts and thus to situate particular texts within a wider context. A core idea of this article is that the text processing methods which a researcher hopes to apply to gain insight into a specialized subject as represented in a digital historical newspaper archive can also be used to help generate the kind of corpus required for effective use of the same methods.

The corpus construction workflow presented in this article is a kind of 'iterative bootstrapping' in that it

starts with an initial 'candidate corpus', which is inappropriate for the researcher's purposes, and, by repeating a cycle of corpus exploration, item labelling, classifier training, and the generation of new candidate corpora, produces an appropriate specialized corpus.

This article begins by setting out previous research relevant to the project, including discussions of the methodological shortcomings of keyword search. It then turns to a description of the data used in the case study. The corpus creation workflow is then set out in detail. The results obtained after three iterations of this method within the case study are then explored both qualitatively and quantitatively. All code and data for the project are available through the Open Science Framework project.[1]

2 Previous research

The corpus construction method presented in this article is best understood against the background of methodological discussions of the role of digital archives in the historical sciences. These discussions often appeal to the difference between digital methods and in-person archival research and draw attention both to research practices which are threatened and to new opportunities provided by digital representations of archival material. Sometimes, digital solutions can be found to problem which arise from the loss of in-person research practices.[2]

Keyword search is an obvious feature of work with digitized text material. For example, if interested in the 'Socialist Church' in New Zealand newspapers, one can enter the phrase in the Papers Past search feature and find relevant items.[3] This is not what one would do when confronted with a physical archive of the same material. One might, as Owens and Padilla (2021) emphasize, deploy an explicit sampling strategy in order to find relevant material and understand it within its context (p. 328). Owens and Padilla (2021) thus argue that keyword searches can only produce 'existence proof' arguments (p. 329). That is, they enable the researcher to say that material of a given description is present but not how representative it is. In this vein, Putnam (2016) argues that keyword searching historians risk stringing together multiple unrepresentative items into 'compelling tales without really seeing the terrain they span' (p. 399).[4]

The presentation of keyword search results has also been criticized. Typically, images of original sources are returned with keywords highlighted.[5] This discourages immersive reading and hides the dependence of search results on often-unreliable optical character recognition (OCR) (Hitchcock, 2013, p. 12). Hitchcock (2013) estimates, in the case of the Burney Collection of Eighteenth Century Newspapers, that up to 52% of the apparent content of the archive is unsearchable due to poor OCR (pp. 13–14). One might be tempted, on finding no results for some keyword in a digitized archive, to claim that the keyword 'never' appears in the archive. The inaccuracy of OCR techniques means this is a mistake. In other words, keyword search does not enable 'negative' existence proofs, either.

While the localized, material, nature of traditional archives has benefits which can't be captured by digital archives, digital representations enable forms of research which can help to overcome the limitations of keyword search. Owens and Padilla (2021) endorse a 'source as data' orientation, while Sternfeld (2011) encourages us to attend to the possibilities which arise from digitized archives when understood as 'digital historical representations'. The core idea is that the historical items are transformed in the process of digitization, becoming computer processable information (Nicholson, 2013, p. 64). Nicholson (2013) emphasizes the question 'what can we do now that we couldn't do before?' (p. 63). Text mining and visualization methods are put forward as providing a way to recapture context lost by unreflective keyword search methods (Sternfeld, 2011; Hitchcock, 2013; Nicholson, 2013; Putnam, 2016; Owens and Padilla, 2021).[6]

As Putnam (2016) puts it, 'computational tools can discipline our term-searching' (p. 399). Hitchcock (2013) encourages a shift to 'maps of meaning' and away from 'piles of books' (p. 20). Processing large numbers of items, 'distant reading' in Moretti's (2013) phrase (pp. 47–49), allows for representations of the content of the archive, or a subset of it, as a whole. This allows for keyword searches to be placed within a context and for new representations of the content of the archive as a whole. These representations may then raise new questions for researchers working in either traditional or digital humanities methodologies.[7] Insofar as these methods are not feasible for investigations of specialist topics within large newspaper corpuses, methods for selection of texts which do not fall victim to the problems of keyword search are required.

Widespread adoption of text-mining methods will require clear and open presentations of techniques, their underlying assumptions, and their scope and limits. Gibbs and Owens (2013) argue that 'historians should deliberately and explicitly share examples of how they are finding and manipulating data in their research, with greater methodological transparency' (p. 168).[8] This is part of a more general push towards reproducible research and the use of Supplementary Materials containing relevant code and data (see, e.g. Berez-Kroeker *et al.*, 2017). The need for this kind of methodological articulation is made even more pressing by the fact that many text-mining tools are made by private companies who do not necessarily have

research use at the front of mind.[9] The Supplementary Material to this article is an attempt to satisfy these demands.

Turning to literature on newspapers specifically, poor OCR is one of the major barriers for text-mining historical newspapers. Tanner *et al.* (2009) manually investigated a sample of the British Library's ninenteeth century digital newspaper archive finding a character accuracy of 83.6% and a word accuracy of 78%. They argue that searchability drops rapidly as word accuracy drops below 80% (see also Strange *et al.*, 2014). This constrains the choice of appropriate text-mining methods. Consequently, methods that rely on high-quality sequence data or local context information for words will be less effective than they would be with higher-quality text.[10] For this reason, a text classification algorithm which does not require sequence data will be selected.

Applications of text mining to contemporary newspapers have typically adopted a manual approach to corpus construction or have relied on metadata to categorize newspaper items (e.g. Alfano *et al.*, 2018). But high-quality metadata is not typically present in historical archives. Some methods, especially those which manually find and re-transcribe large collections, require too much labour for a single researcher with a specialized interest (cf. Rubinstein, 2019). Because the aim is to construct a 'specialized' corpus, random sampling will not be effective either (cf. Calude and James, 2011). An automatic method for selecting texts for the corpus is desired. Many methods for text classification have been applied to newspapers. These include clustering methods (e.g. Hagen, 2012) and the use of Naïve Bayes classifiers (Steger, 2013). In both cases, manual labelling of items to train the classifier were required.

Finally, consider the literature concerning the case study. It is first worth noting that history 'of philosophy' has an ambiguous relationship with the historical sciences. Some philosophers are interested in history as a store of concepts or arguments without much concern for their context. Laerke *et al.* (2013) refer to supporters of this approach as the 'appropriationist school' and distinguish them from the 'contextualist school' (p. 1). The latter are, as the name suggests, devoted to understanding the history of philosophy 'for its own sake and in its own terms' (Laerke *et al.*, 2013, p. 1). If the value of an historical piece of philosophical discourse is in its contemporary use, then methodological worries about keyword searching will not apply. If an idea found by keyword search can be made interesting within the contemporary literature, then it doesn't matter whether it was characteristic of arguments at the time or not. However, those who adopt a more contextualist approach are likely to have the same concerns as other historians.[11] This is especially true of those who think that the context of philosophical texts must include public debates ongoing in 'newspapers and flyers, sermons, and caricatures' (Goldenbaum, 2013, p. 76). The methods developed in this article should thus be of interest to contextualist historians of philosophy.

In the context of New Zealand, there is a paucity of research into philosophy before the middle of the twentieth century. Davies and Helgeby (2014) argue that, before this time, 'many of those who had longstanding [academic] chairs published next to nothing' and that the New Zealand philosophical community was small and not well connected (p. 24). However, there is reason to think that a turn to newspaper content may allow a story about philosophy in New Zealand in the nineteenth century. Ballantyne (2012) argues that newspaper was 'the fundamental infrastructure for intellectual life' in colonial Otago and that this was connected with economic limitations which prevented a strong periodical or monograph culture (p. 57).[12] One particularly interesting mode of academic work reported in the newspapers is the public lecture (e.g. Crane, 2013). But, in addition to finding 'academic' philosophy, a turn to newspapers might enable the inclusion of a wider range of voices from journalists and writers of letters to the editor.

Public debates, particularly concerning issues of religion and science, are represented in the newspapers of the time (e.g. Bush, 2018).

Finally, it must be noted that this project works only with English-language newspapers and covers a time period in which there was a thriving newspaper industry among Māori and in te reo Māori (the Māori language).[13] There is research on the Papers Past collection of Māori newspaper (niupepa Māori) resources. This material is absolutely vital for any inclusive, and thus any comprehensive, account of philosophy in Aotearoa New Zealand. It is hoped that the methods applied here will be of use to those studying specialized topics as discussed in niupepa Māori digital archives.[14]

3 Data

In August 2020, the National Library of New Zealand released a large portion of their English-language Papers Past newspaper archive as part of a pilot programme exploring their engagement with open data.[15] For copyright reasons, the release excludes all material from 1900 on.

The data were scanned from microfilm, processed, and run through OCR software.[16] The data are in the METS/ALTO XML format and are organized into compressed tar.gz files each of which contains all of the issues of a given newspaper in a year. There are 79

distinct newspaper titles and 306,538 distinct issues. Each issue has a METS file, which contains the logical and physical structure of the issue as a whole, and an ALTO file for each page. There are 1,471,384 pages of newspaper content in the data set. The ALTO files contain XML elements corresponding to text blocks, to blocks of text blocks and images, for lines within text blocks, and for individual words. These elements are always associated with vertical and horizontal positions, width, and height on the scanned image and sometimes with an explicit indication of text style and size.

4 Method

Given 1,471,384 pages of newspaper content, it is not obvious how to find the very small slice concerned with a given specialized subject. The aim is to create a corpus which will include a sensible collection of items which can be used for both 'close' reading and 'distant' reading. The proposed method is depicted in Fig. 1.

The first step is to carry out preprocessing. This moves from METS/ALTO XML files to a representation of the dataset which contains only the desired material. Once preprocessing is complete, an indefinite number of iterations of a corpus construction loop begins.

The second step is the corpus exploration stage. At this stage, the content of a candidate corpus, perhaps generated through keyword search, is examined to determine whether it includes unwanted items or excludes wanted items. If the resulting corpus is not acceptable, the labelling step is entered. The researcher then labels items according to interest. The processed data may be sampled from again in order to label a range of items which are not of interest or to collect items whose absence has been identified in an earlier iteration of the corpus exploration stage.

As the next step, a classification method is trained on the basis of the new labels. There are various methods which can be applied here and the GitHub repository presents options for those who wish to experiment. In this article, Naïve Bayes classification will be used (see James *et al.*, 2021, pp. 153–158). At this stage, qualitative and quantitative evaluation of the classifier is carried out to determine its overall accuracy, recall, and precision when compared with the labelled items and the kind of items it fails to classify properly. This information conditions any inferences drawn from the resulting corpus and allows improvement of the corpus in a future iteration of the corpus construction loop. An iteration of the corpus construction loop concludes when the new classifier is applied to the dataset to generate a new candidate corpus.

This method is iterative insofar as the process can be repeated multiple times. It is 'bootstrapping' insofar as the results of previous iterations go in to the next iteration and enable the researcher to more successfully target the items they are interested in. By labelling unwanted items picked out by previous classifier, future classifiers can become more discerning. In addition, the results of previous classifiers enable additional items of

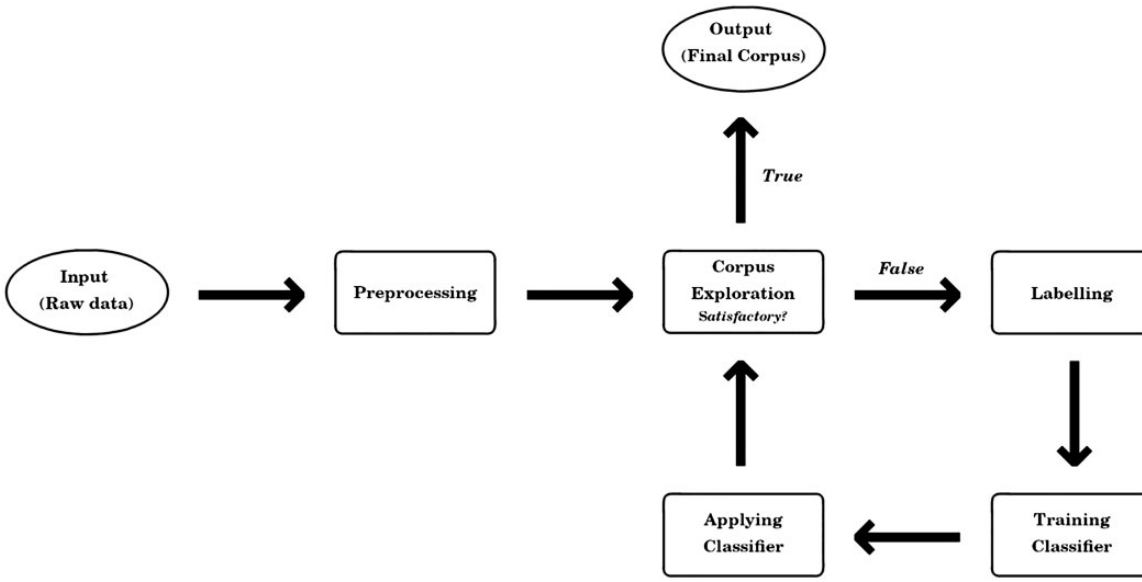


Figure 1. Flow diagram of the corpus construction method

interest to be easily found and labelled. That is, previous iterations are helpful for both including and excluding items. Each stage is now discussed in turn with reference to the case study.

5 Preprocessing

Preprocessing takes raw METS/ALTO XML data and extracts the information to be used for classification. In the case study, the relevant details are the newspaper, issue date, title, and plain text of newspaper items.[17]

The data are processed issue by issue. First, the 'Logical Structure' section of the issue's METS file is processed, collecting the item titles with the 'ARTICLE' attribute and identifiers for each block of text in the item. The label 'UNTITLED' is assigned to items without titles. Text block identifiers contain the page number and a block ID, enabling items which appear across multiple pages to be collected. The article codes are then iterated through, collecting each text block as a string.[18] Each item has an ID assigned to it by combining the newspaper code, the issue date, and the article number from the METS file. For instance, the first article from the *Lyttleton Times* on the twelfth of March 1872 is 'LT_18720312_ARTICLE1'.[19]

Figure 2A shows the distribution of items by region after preprocessing. The Otago and Canterbury, regions of the South Island, massively dominate the dataset. In addition, there are significant contributions from the West Coast, Nelson, and Marlborough, all of which are in the South Island. The North Island is less represented, although Auckland does make the third largest individual contribution. Northland is almost completely unrepresented, and the number of items from Bay of Plenty (1571) is not even visible. Any conclusions drawn from the resulting corpus must be conditioned by the dominance of Otago and Canterbury.

Figure 2B shows, unsurprisingly, that the by-year count of items grows rapidly. While the dataset starts in the 1840s, the great majority of the material is from the 1880s and 1890s.

Because the dataset is too large for most personal computers to keep in memory at once, it is saved to multiple compressed 'slices'. Specifically, the dataset is stored in 26 compressed data frames using the Python Pandas library and its pickling functions (McKinney, 2010). Uncompressed, the raw data take up around 300 GB, while the processed data take up around 45 GB. The average size of the compressed slices is around 300 MB each. The resulting count of items is 7,592,619.

6 Corpus exploration stage

In the corpus exploration stage, a wide range of close and distant reading methods are applied to explore and evaluate a candidate corpus. Methods include random inspection of texts from the candidate corpus; word clouds, concordancing and collocation analysis of keywords

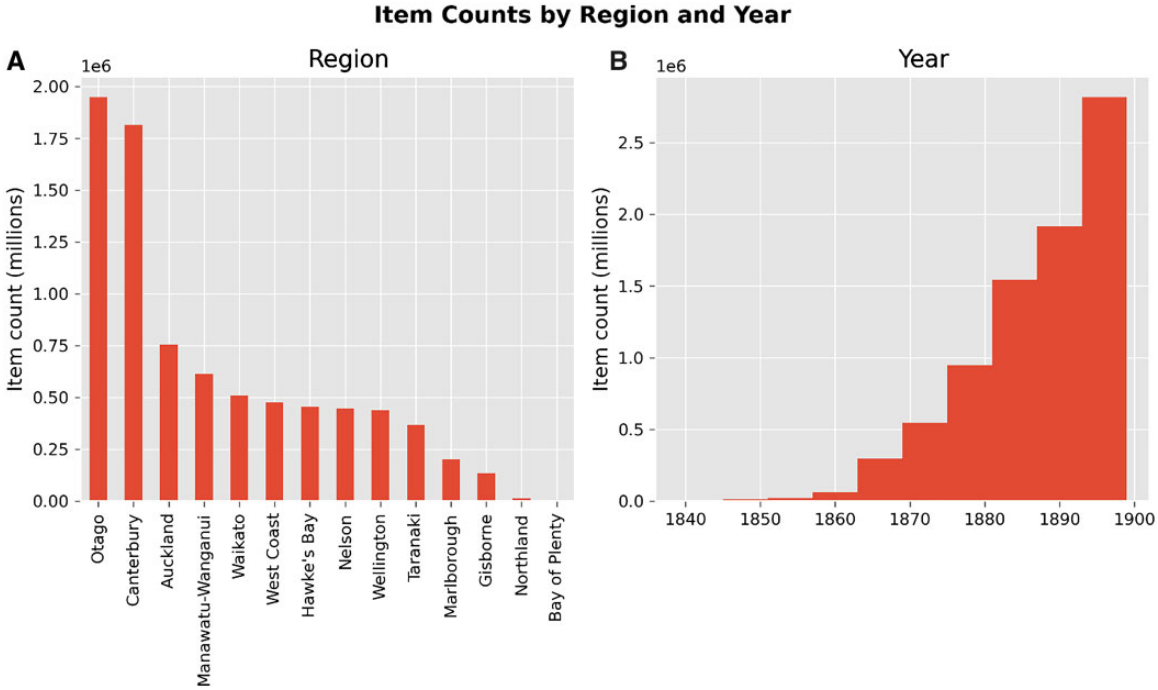


Figure 2. Distribution of article items across (**A**) regions and (**B**) years

using the Python NLTK package (Bird *et al.* 2009); cooccurrence networks implemented using Plotly Dash (Plotly Technologies Inc., 2021) and Latent Dirichlet Allocation (LDA) topic models using Gensim (Řehůek and Sojka, 2010). Each provides a distinct lens on the candidate corpus.

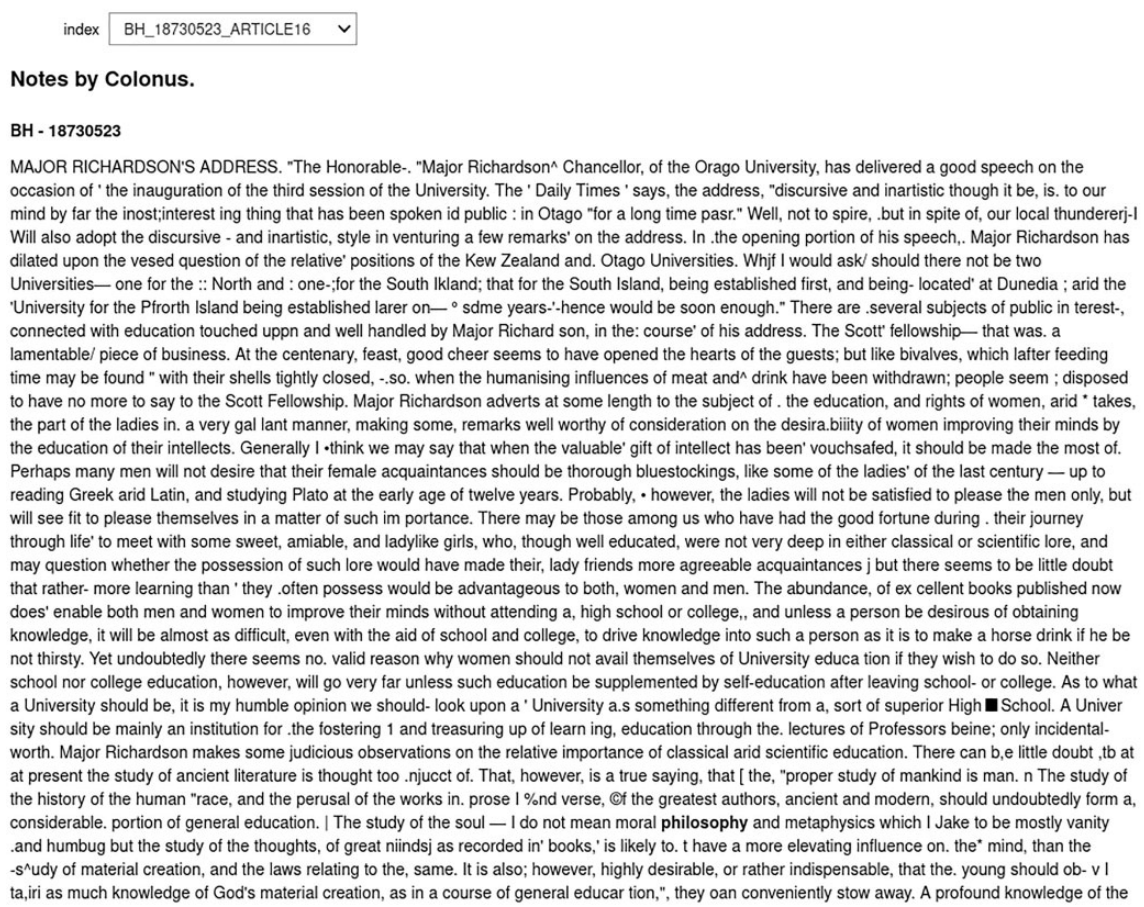
Explaining each of these methods would require something approaching the scope of a text book.[20] In this section, the role of manual inspection and LDA topic modelling will be discussed. The former is best illustrated by the first iteration of the process, while the latter is more useful in later iterations.

The first iteration of the process requires an initial candidate corpus. In the case study, the first candidate corpus is the collection of items which have a match for the case insensitive wildcard keyword search 'philoso*'. This keyword search collects any item which has a word starting in 'philoso' and ending in any other way. This includes, for instance 'philosophy' and 'philosophical'.

Figure 3 shows an item from the first candidate corpus as it appears in the project interface. The dropdown box allows a random sample of items from the candidate corpus to be viewed. The instance of 'philoso*' in the item is in bold. At this stage, the aim is to understand the broad range material in the candidate corpus and to pick out items to label if another classifier is to be trained. For instance, the pictured article is about the role of classical philosophy in education and so should be included in a philosophy corpus.

During later iterations, the results of LDA topic modelling become useful for finding groups of unwanted items. An instance of this in the case study comes during the third iteration of the process. Figure 4 shows the keywords characteristic of 15 topics generated by LDA. In the case study, the first topic was used to pick out advertising content which is not wanted in the corpus.[21] Figure 5 shows items which have Topic 1 representing more than 50% of their content. Inspection of these items allows the researcher to easily find and label unwanted items and thus train a more effective classifier.

The corpus exploration stage ends when a decision is made concerning the acceptability of the candidate


index BH_18730523_ARTICLE16

Notes by Colonus.

BH - 18730523

MAJOR RICHARDSON'S ADDRESS. "The Honorable-. "Major Richardson^ Chancellor, of the Orago University, has delivered a good speech on the occasion of ' the inauguration of the third session of the University. The ' Daily Times ' says, the address, "discursive and inartistic though it be, is. to our mind by far the inost;interest ing thing that has been spoken id public : in Otago "for a long time pasr." Well, not to spire, .but in spite of, our local thunderer j-I Will also adopt the discursive - and inartistic, style in venturing a few remarks' on the address. In .the opening portion of his speech,. Major Richardson has dilated upon the vesed question of the relative' positions of the Kew Zealand and. Otago Universities. Whjf I would ask/ should there not be two Universities— one for the :: North and : one-;for the South Ikland; that for the South Island, being established first, and being- located' at Dunedia ; arid the 'University for the Pfrorth Island being established larer on— ° sdme years-'-hence would be soon enough." There are .several subjects of public in terest-, connected with education touched uppn and well handled by Major Richard son, in the: course' of his address. The Scott' fellowship— that was. a lamentable/ piece of business. At the centenary, feast, good cheer seems to have opened the hearts of the guests; but like bivalves, which lafter feeding time may be found " with their shells tightly closed, -.so. when the humanising influences of meat and^ drink have been withdrawn; people seem ; disposed to have no more to say to the Scott Fellowship. Major Richardson adverts at some length to the subject of . the education, and rights of women, arid * takes, the part of the ladies in. a very gal lant manner, making some, remarks well worthy of consideration on the desira.biiity of women improving their minds by the education of their intellects. Generally I •think we may say that when the valuable' gift of intellect has been' vouchsafed, it should be made the most of. Perhaps many men will not desire that their female acquaintances should be thorough bluestockings, like some of the ladies' of the last century — up to reading Greek arid Latin, and studying Plato at the early age of twelve years. Probably, • however, the ladies will not be satisfied to please the men only, but will see fit to please themselves in a matter of such im portance. There may be those among us who have had the good fortune during . their journey through life' to meet with some sweet, amiable, and ladylike girls, who, though well educated, were not very deep in either classical or scientific lore, and may question whether the possession of such lore would have made their, lady friends more agreeable acquaintances j but there seems to be little doubt that rather- more learning than ' they .often possess would be advantageous to both, women and men. The abundance, of ex cellent books published now does' enable both men and women to improve their minds without attending a, high school or college,, and unless a person be desirous of obtaining knowledge, it will be almost as difficult, even with the aid of school and college, to drive knowledge into such a person as it is to make a horse drink if he be not thirsty. Yet undoubtedly there seems no. valid reason why women should not avail themselves of University educa tion if they wish to do so. Neither school nor college education, however, will go very far unless such education be supplemented by self-education after leaving school- or college. As to what a University should be, it is my humble opinion we should- look upon a ' University a.s something different from a, sort of superior High ■ School. A Univer sity should be mainly an institution for .the fostering 1 and treasuring up of learn ing, education through the. lectures of Professors beine; only incidental- worth. Major Richardson makes some judicious observations on the relative importance of classical arid scientific education. There can b,e little doubt ,tb at at present the study of ancient literature is thought too .njucct of. That, however, is a true saying, that [the, "proper study of mankind is man. n The study of the history of the human "race, and the perusal of the works in. prose I %nd verse, ©f the greatest authors, ancient and modern, should undoubtedly form a, considerable. portion of general education. | The study of the soul — I do not mean moral **philosophy** and metaphysics which I Jake to be mostly vanity .and humbug but the study of the thoughts, of great niindsj as recorded in' books,' is likely to. t have a more elevating influence on. the" mind, than the -s^udy of material creation, and the laws relating to the, same. It is also; however, highly desirable, or rather indispensable, that the. young should ob- v l ta,iri as much knowledge of God's material creation, as in a course of general educar tion,", they oan conveniently stow away. A profound knowledge of the

Figure 3. Manual inspection dashboard

```
{'Topic 1': 'sale, sheep, stock, wellington, sold, street, goods, messrs, price, cure',
 'Topic 2': 'council, bill, act, committee, motion, provincial, system, member, board, session',
 'Topic 3': 'political, party, lord, speech, member, parliament, liberal, perhaps, editor, office',
 'Topic 4': 'court, wellington, district, yesterday, board, street, road, messrs, town, fire',
 'Topic 5': 'school, education, university, professor, class, science, college, history, knowledge, system',
 'Topic 6': 'told, lie, boy, death, saw, room, door, wife, body, doctor',
 'Topic 7': 'woman, love, lady, face, girl, room, dear, miss, heart, wife',
 'Topic 8': 'messrs, committee, club, hall, school, yesterday, board, rev, miss, street',
 'Topic 9': 'gold, trade, labour, value, system, capital, population, amount, working, labor',
 'Topic 10': 'war, prince, king, army, empire, peace, nation, political, emperor, national',
 'Topic 11': 'god, human, truth, religion, spirit, moral, law, faith, love, religious',
 'Topic 12': 'book, miss, author, story, art, play, character, literary, lady, written',
 'Topic 13': 'water, race, white, horse, round, sea, river, black, beat, along',
 'Topic 14': 'captain, south, ship, york, lord, death, united, board, recently, vessel',
 'Topic 15': 'water, earth, scientific, professor, science, air, sun, sea, surface, heat'}
```

Figure 4. Keywords for a 15 topic LDA model of a candidate corpus, as they appear in the Jupyter Notebook interface

	Title	Text
WC_18990503_ARTICLE27	TWO OF A TRADE.	[That two of a trade seldom agree is a common ...
LWM_18731015_ARTICLE35	SULPHURIC ACID FOR DIPHTHERIA SUCCESSFUL.	[Mr Greatbead's remedy for diphtheria is being...
WC_18850131_ARTICLE28	MEN, WOMEN, AND THINGS.	[" A mad world my masters," — Old Play. The er...
WC_18971202_ARTICLE31	UNTITLED	[" Drunkenness is not a sin— simply an excess ...
WC_18990810_ARTICLE26	TUBERCULOUS PHEASANTS	[(Pen Press- Association.), . ' y WELLINGTON, ...
...	...	...
WC_18980209_ARTICLE25	UNTITLED	[Drunkennea? is not aain— •simply an Bxces9 of...
ST_18880731_ARTICLE29	UNTITLED	[A Great Enter prise— The Dr Soule's American ...
ST_18960109_ARTICLE9	"PIE."	[Granite ia quarried in Southern India by burn...
WC_18980413_ARTICLE42	QLD AND NEW WORLD WISDOM.	[Away in the dim mysterious years when the vas...
WC_18980118_ARTICLE33	SANDERSON'S "SCOTCH."	[Established 1816. . Original Blenders of Whis...

716 rows × 7 columns

Figure 5. Items with greater than 50% of words generated by Topic 1, as they appear in the Jupyter Notebook interface

corpus. The first candidate corpus is very unlikely to be acceptable, as it will usually be generated by a simple keyword search. Given this, information about previously trained classifiers will almost always be available when a candidate corpus is evaluated. This information will be returned to within the discussion of the training stage. In addition to this information, investigation of the content of a candidate corpus during the corpus exploration stage helps to reveal whether the connections between terms or topics picked out by text-mining methods are being determined by real associations in desired material or not.

7 Labelling

At the labelling stage, items are labelled for inclusion in the corpus and with supplementary labels which may be useful for model criticism. A supplementary label might reveal, for instance, a class of material which is being improperly classified.

Figure 6 shows a screenshot of the labelling dashboard. This dashboard has a drop-down box for selecting articles to be labelled. Basic information is given, including the newspaper title and issue date, along with a URL to access the full issue from which the item is taken on the Papers Past website. The panel on the right contains all labels. The first is the primary label for the case study, which answers whether the item belongs in a corpus of philosophical discourse or not. The remaining labels are supplementary. In practice, labelling criteria are likely to be refined in the early stages of the corpus construction process.

The labelling used in the case study is instructive insofar as 'philosophical discourse' is a particularly contested idea. This is illustrated by the number of books and articles with the title 'What is Philosophy?' (e.g. Deleuze and Guattari, 1991; Priest, 2006; Agamben, 2017). The approach adopted here is to select items which either make claims about, or appeal to, some conception of ultimate value or reality. For instance, on this conception, an argument for a liberal political outlook on the basis that freedom is the most important value is a philosophical discourse. Similarly, if you argue that divine revelation is a more reliable source of knowledge than scientific investigation, you are engaged in philosophical discourse. This characterization of 'philosophical' does not require sustained argumentation. Evaluating whether such argumentation is present is too labour intensive to be feasible for a single researcher.[22]

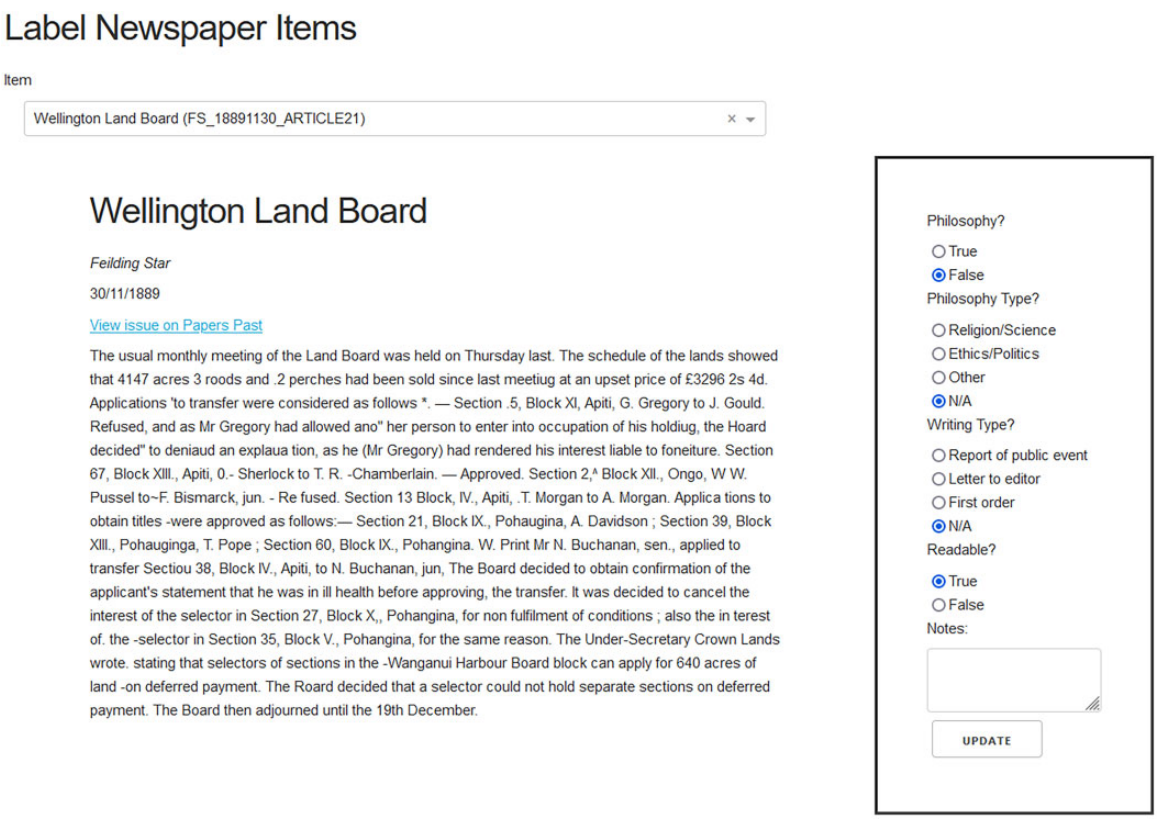


Figure 6. Labelling dashboard

Supplementary labels track sub-topic and genre. Philosophical discourse is found in reports of public events, in letters to the editor, and in 'first-order' pieces. This is a genre label and is useful insofar as a classifier might perform better or worse for different genres. Similarly, sub-topics of philosophy are labelled. In this project, one of the areas of interest is in the relationship between science and religion, particularly in the wake of the publication of Darwin's *On the Origin of Species* in 1859 (Darwin, 2009). Any items relevant to this theme were labelled with 'Religion/Science'. Ethical and political discussions are also tagged along with an 'other' category. In addition to model criticism, these labels may be useful for training an additional classifier and generating a more targeted sub-corpus. Finally, it is indicated whether OCR errors render an item unreadable.[23]

Usually, a good portion of the items to label will be found in the current candidate corpus at the corpus exploration stage. These will include both desired and undesired items. Input from outside the current candidate corpus is also required as a wide range of undesired items is required. This can typically be achieved by randomly sampling the processed dataset. In addition, if it has been determined that desired items are missing from the candidate corpus, then they must also be brought in. If examples are known, they can be collected from the data set using their article identifier code. If not, keyword search is again useful.

8 Training and applying classifier

Many methods for text classification are available. In the case study, a very simple classification method is used: the Naïve Bayes classifier. The labelled data are divided into training and testing portions, using a 75/25 split. That is, 75% of the labelled data is used to train a classifier and 25% of the data to test the classifier once it has been trained. Upsampling is then carried out in order to ensure a 50/50 split between desired and non-desired items in the training set. Classification algorithms are often more effective when given balanced data, even when the actual classes are quite different from one another.

The pipeline for classification is implemented using the Python package Scikit-Learn (Pedregosa *et al.*, 2011). The first step converts strings to bag of words representations, after which the classifier is applied. At each step, multiple hyperparameters can be adjusted (Table 1). These are the parameters of the classifier which are not directly trained. The best settings are found by using k-fold cross-validation. That is, the classifier is retrained k times for each combination of parameters, leaving out a different subset of the training data each time. The classifier which performs best by overall accuracy is then chosen (see James *et al.*, 2021, pp. 203–208).

Once the hyperparameters have been selected, the pipeline is run on the full training set and its performance on the test data is examined. Quantitative evaluation considers the metrics of overall accuracy, precision, and recall. Accuracy is simply the proportion of classifier's predictions which are correct. Precision is the proportion of the items which the classifier selects as desired which are actually desired. If the precision is low, then the classifier produces lots of false positives. The recall is the proportion of the items which are actually desired which the classifier selects as desired. If the recall is low, then there are lots of false negatives. Ideally, precision and recall are balanced, but if a more comprehensive corpus is desired then recall can be prioritized over precision. That is, a higher false positive rate can be accepted in order to reduce false negatives.

At the qualitative evaluation stage, the actual items from the test set which are misclassified are considered. Common themes are looked for in the places where the model fails. Here, the use of supplementary labels is often useful. If all of the false negatives are in the same genre, for instance, then more positive examples from this style may be required to train the classifier in a subsequent iteration. Any patterns found at this stage are helpful when the corpus exploration stage is returned to. The terms which are given high probability within each class are also examined. This gives an idea of what features are being picked out by the model.

Once a model has been trained, it is applied to the dataset as a whole, keeping only those items which the classifier selects. In order to handle a problem which arises for highly composite items, such as editorials in which many topics are discussed, an option is provided to divide the items from the dataset into 'chunks' of a given length and accept the item if any one of its chunks is positively classified. That is, items can be accepted as philosophical if they have a sufficiently large 'chunk' which is identified as philosophical. It is also possible, on the basis of model evaluation, to reduce the probability threshold for considering an item 'philosophical' from 0.5 in order to increase recall. Another iteration of the corpus construction process is then begun by returning to the corpus exploration stage.

9 Results

The results from the case study after three iterations of the corpus construction method are now presented. First, the final corpus and the classifier which produced it are considered. Second, an external validation of the corpus is presented by looking at the newspaper items picked out by historians working on relevant aspects of early colonial New Zealand intellectual life and

Table 1. Model hyperparameters

Hyperparameter	Description	Example Ranges
vect__min_df	Minimum 'count' of items in which a word must appear in order to be included in the dictionary.	2, 5, 7, 10, 15
vect__max_df	Maximum 'proportion' of items in which a word can appear and be in the dictionary.	0.2, 0.3, 0.4, 0.5, 0.6
vect__ngram_range	Determines whether the dictionary should include pairs or triples of adjacent words in addition to single words.	(1, 1), (1, 2), (1, 3)
tfidf__use_idf	Determines whether word counts are scaled by the frequency of words in the corpus as a whole.	True, False
clf__alpha	A smoothing parameter which boosts all term counts in order to avoid giving zero probability to any term in the dictionary.	0.2, 0.5, 0.75, 1

Table 2. Candidate corpus sizes

Iteration	Corpus size
Processed dataset	7,592,619
Iteration 0 ('philoso*' search)	29,647
Iteration 1	239,649
Iteration 2	31,131
Iteration 3	61,252

determining if they are also picked up by this corpus construction method. Finally, an example of the use of cooccurrence networks to derive insight into philosophical discourse in early colonial New Zealand from the corpus is provided.[24]

Table 2 shows the sizes of the corpora at the end of each iteration. The initial candidate corpus was simply the result of a keyword search which matched all occurrences of 'philoso*'. In the first iteration, 248 items were labelled (Table 3) and a smaller candidate corpus was produced. The second iteration further reduced corpus size. At this point, it was found that desired items were being excluded. A concerted effort was to correct this by labelling appropriate articles resulting in a larger corpus after the final iteration of the process. This illustrates the difficulty of balancing selectiveness and inclusiveness.[25]

10 Classifier performance

The model trained at the third iteration of the process is a Naïve Bayes classifier with hyperparameters determined by a grid cross-validation search (Table 4).

The first qualitative evaluation occurs at the level of the model. The items from the testing set which are incorrectly labelled are considered and this is used to determine potential shortcomings to be looked for in subsequent corpus exploration. At the third iteration, the classifier was applied to chunks of 4,000 characters long rather than to full items in order to include composite items containing philosophical discourse. The confusion matrix is presented in Table 5.

The overall test accuracy is 86.6%. This is slightly lower than the raw test accuracy reported before applying the chunking method (87.8%). The difference is in the distribution of false positives and negatives. By allowing multiple chances for a long item to be classified as philosophy, the number of false negatives is decreased at the expense of increased false positives. The resulting recall, the proportion of labelled philosophy items which are correctly classified, is 0.92 (versus 0.87 without 'chunking'). The precision, the proportion of positively classified items which are correctly classified, is 0.78 (versus 0.83 without 'chunking'). Since the aim, in line with the methodological literature reviewed earlier, is to put items 'in context', it is acceptable to improve recall at the expense of precision. Some irrelevant material may be picked up in the corpus, but as long as it contains the majority of desired material and remains computationally tractable, there is no problem.

It is worth examining the false negatives in some detail as part of the qualitative evaluation of the classifier. There are 10 false negatives. Of these, seven are labelled as 'ethics/politics', two are labelled as 'other', and one is labelled as 'religion/science' (Table A2). The failure to pick up discussion of an argument concerning women's franchise by Robert Stout, former Premier of New Zealand is notable. This is the kind of failure to be kept in mind at the corpus evaluation stage.

At this stage, the words which are ranked highly by the model are considered. The Naïve Bayes model simply assigns a probability to each word in the dictionary for each of the classes. The model classifies items by calculating the probability of their collection of terms given each class. By Bayes's theorem, it converts this into a probability that an item is in the desired class or not. Table 6 presents the highest probability terms for philosophy items. These terms indicate a tendency towards disputes over religion and science and education. The model has also picked out the name 'Salmond' (rank 36), and words which might indicate genre, such as references to a 'lecture', 'paper', 'letter', and 'argument'.

Table 3. Count of labels at each iteration

Label	Value	First iteration	Second iteration	Third iteration
Philosophy	True	101	299	502
	False	147	620	642
Philosophy type	Religion/science/ metaphysics	58	140	271
	Ethics/politics	25	94	143
	Other	18	65	84
Writing type	Public event	40	97	145
	Letter	23	69	139
	First order	36	111	187
	Review	2	22	22

Table 4. Final classifier hyperparameters

Hyperparameter	Selected value	Range
vect__min_df	5	[1, 2, 5, 10, 20]
vect__max_df	0.2	[0.1, 0.2, 0.3, 0.4, 0.5]
vect__ngram_range	(1, 2)	(1, 1), (1, 2), (1, 3)
tfidf__use_idf	True	True, False
clf__alpha	1	[0.5, 0.75, 1, 1.5, 2]

Table 5. Confusion matrix for final model

	Classified: not philosophy	Classified: philosophy
Actual: not philosophy	168	34
Actual: philosophy	10	117

It is also worth looking at the words which the model assigns low probability to in non-philosophy items (Table 7). Some words appear in both lists, for example, 'woman' (rank 35 in Table 6 and rank 44 in Table 7). This is important for interpreting these lists. That is, the same word can be given a high probability in both classes. If so, it is not a word which is playing an important role in distinguishing the two classes. Table 7 can be understood as capturing something like the words in a dictionary which most characterize the average newspaper item. The content words here seem to represent government, financial, and agricultural items.

Qualitative and quantitative evaluation of the final classifier suggests that this corpus will be most reliable when making inferences concerning the religion and science theme. This is clear both from looking at the supplementary labels and examining the terms picked out by the classifier.

At the corpus exploration stage, it was decided that the corpus after the third iteration was sufficient. This required testing to ensure that the failure to pick out an item on women's suffrage did not indicate an absence of this material from the corpus. Concordance analysis of a random sample of items indicates that 36% of the 834 items containing the phrase 'suffrage' can be identified by their immediate context as concerning women's suffrage. Cooccurrence networks also revealed that key concepts like 'nature' are being closely connected with issues of gender and law.[26] Additionally, it was decided that the blurry boundary between politics and philosophy was unlikely to be significantly sharpened by further labelling.[27]

11 External validation

The 'external' test of the method presented here is simply whether it picks out the same items picked out in research employing more traditional methods of text selection. It is not expected that other researchers would include this as part of their model criticism. The aim here is to determine whether the method is at least as comprehensive in terms of item selection as 'the standard methods' of historical investigation.

Scholarship on specific debates which ought to be present in the corpus was selected, with a view to picking out the cited newspaper items and determining whether they are included in the corpus. Four articles were selected for this purpose (Ballentyne, 2012; Crane, 2013; Wood, 2014; Bush, 2018). These articles focus on discussions about evolution in colonial Otago, the public lectures of the biologist Thomas Parker, the dispute over William Salmond's pamphlet *The Reign of Grace*, and the public debates of the freethinker William Collins and Methodist minister John Hosking. Relevant items were then selected from the bibliographies of the articles and then looked for within the candidate corpora at each iteration of the process.

Table 8 shows that the initial keyword-search-generated candidate corpus contains only 11% of the articles which were identified from these articles (for full results, see Table A1). The first trained classifier has good performance in picking out the identified items,

Table 6. Top 60 words for 'philosophy' class in final classifier

Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	Moral	16	Spencer	31	Editor	46	Letter
2	Christian	17	Facts	32	Women	47	Argument
3	Education	18	Teaching	33	Lecturer	48	Physical
4	Bible	19	School	34	Sense	49	Cause
5	Professor	20	Evolution	35	Woman	50	Universe
6	Evil	21	Doctrine	36	Salmond	51	Children
7	Theory	22	Laws	37	Tbe	52	Spiritual
8	Christ	23	Idea	38	Divine	53	Earth
9	Darwin	24	Belief	39	Self	54	Future
10	Christianity	25	Mental	40	Aud	55	Principles
11	Scientific	26	Political	41	Modern	56	Force
12	Natural	27	Free	42	Intellectual	57	Soul
13	Faith	28	Love	43	Schools	58	Seems
14	Religious	29	Law	44	Social	59	Phenomena
15	Lecture	30	Book	45	Paper	60	Mere

Table 7. Top 60 words for 'non-philosophy' class in final classifier

Rank	Word	Rank	Word	Rank	Word	Rank	Word
1	Government	16	Members	31	Rev	46	Left
2	000	17	Young	32	Half	47	Milk
3	Zealand	18	Water	33	Dunedin	48	Messrs
4	House	19	Letter	34	England	49	Money
5	Board	20	Evening	35	London	50	Back
6	Year	21	Committee	36	Christ	51	Amount
7	School	22	Body	37	City	52	Motion
8	Tbe	23	Association	38	English	53	Among
9	New Zealand	24	Children	39	Bill	54	February
10	Wellington	25	Colony	40	Miss	55	Tha
11	Received	26	Lord	41	Court	56	Name
12	Mrs	27	Editor	42	Aud	57	Days
13	Auckland	28	South	43	Press	58	Death
14	Company	29	Next	44	Woman	59	George
15	Night	30	Morning	45	Get	60	Came

Table 8. External validation success rate

Iteration	0	1	2	3
Success rate	0.11	0.81	0.58	0.81
Corpus size	29,647	239,649	31,131	61,252

but with a very large corpus size. The second trained classifier then becomes too specific, missing many of the items picked out by the identified items. Finally, the third trained classifier produces a corpus which combines coverage of the identified material in a more targeted way. It is a quarter of the size of the corpus generated by the first trained classifier, while slightly improving on the success rate of the first iteration.

In the final iteration, all of the failed items were composite items. The trick introduced for improving performance with respect to these items is not a complete solution. Other cited items were dropped because they were too short to be classified or because they were advertisements (Table A1). Many of the items not picked up are from Crane (2013), which concerns the public science lectures of Thomas Parker. Often these items do not touch directly on philosophical issues associated with the sciences. The failure of the classifiers on these items helps to reveal where the classifier draws the line between philosophical and non-philosophical items. A report of a pure lecture on astronomy, for instance, is unlikely to be included.

12 Maps of meaning

This article aims to set out a general method rather than to directly answer specific questions about the history of philosophy in New Zealand. However, as in many cases in the digital humanities, the test of a method is in its use to generate insight from the source material. Given this, it is worth briefly illustrating the

use of this corpus in investigating early New Zealand philosophy.

The method introduced above which most naturally answers to the description 'map of meaning' is cooccurrence networks. Figure 7 presents a cooccurrence network for the term 'materialism' generated from the final corpus. Many interesting features of the discussion of materialism in early colonial New Zealand newspapers are clear from this network.

Figure 7 shows, for instance, that uses of the term 'materialism' tend to cooccur with uses of terms for various other '-ism's, such as 'idealism', 'theism', 'atheism', 'Darwinism', 'spiritualism', and 'scepticism'; discussion of the natural sciences with 'organic', 'atoms', and 'evolution'; discussion of various international figures, such as the theosophists Annie Besant and Helena Blavatsky, and the evolutionary figures Charles Darwin and Thomas Huxley; along with the local figure William Salmond.[28] A full discussion of the material in this network, and the investigation of its various features would, of course, be the task of another paper.

The key point, for present purposes is that this network would not have been feasible when applied to the entire corpus of newspaper articles. If the same methods were applied to the whole corpus, a much smaller dictionary would be required. To achieve the level of detail in this association network for 'materialism', a specialized corpus was required.

13 Conclusion

This article has presented a general method for corpus construction from METS/ALTO format archives of historical newspaper content. The method has been

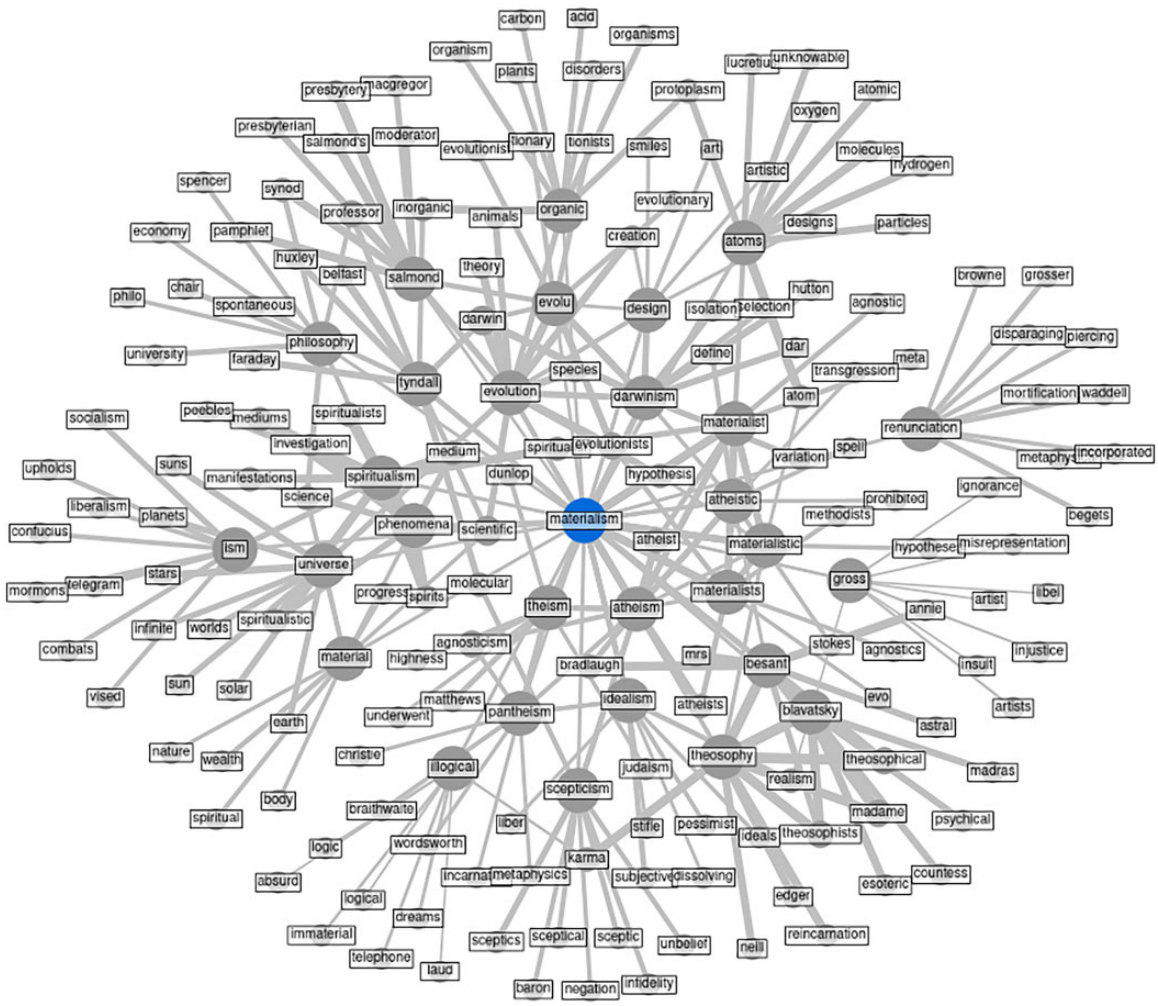


Figure 7. Cooccurence network for 'materialism' in final corpus, with dictionary consisting of words in the NLTK English wordlist supplemented by Spacy-identified proper nouns in corpus

illustrated by means of a case study in which philosophical discourse in early New Zealand newspaper content was investigated.

First, a literature review was provided focusing on the methodological role of keyword searching digital historical archives. Keyword search methodologies risk a loss of the contextual understanding which any researcher who had to wrestle with an in-person, physical archive has to develop. At its worst, as Owens and Padilla argue, researchers are left with the ability to provide mere existence proofs that this or that sort of material is present in an archive. It was seen that an increase in technological sophistication promises to reduce some of the problems of context loss brought in by digital historical archives. In particular, methods which enable the generation of representations of the context around certain key terms or ideas in historical newspaper sources are required.

This article responds to the demand for public, reproducible methods which exploit the opportunities which arise from digital historical representations of archival material in a special case. The sheer size of such archives means that many off-the-shelf text-mining techniques will not be particularly helpful for investigating specialized topics. The core idea of this article is that, while this is true, the same techniques can be deployed to generate a specialized corpus from which humanistic insight can be derived.

Having presented the method at a high level, and pointed to its technological implementation in a series of Jupyter Notebooks, the results achieved in the case study were considered. The challenging interplay between including too much and too little was demonstrated, along with methods for quantitatively and qualitatively examining the resulting classifiers and corpora. This evaluation connects directly to the problem of determining what kind of historical inferences will be licensed by the corpus. In the case study, a tendency to lose material labelled with the 'ethics/politics' label was found, qualifying any generalizations about ethical or political issues from the corpus. However, disputes over science and religion were found to be well covered. An example of a cooccurrence network generated from the corpus was given to indicate the kind of insight which the case study's corpus can provide.

In the case study, three iterations of the method were sufficient to generate a specialized corpus of philosophical writing in early colonial New Zealand newspapers. After three iterations, the method achieved a balance of both selectiveness and accuracy. Accuracy was measured both internally, by splitting data into training and testing data, and externally, by comparing the items selected with those picked out in previous intellectual history research on the time period. It is hoped that this case study, along with the accompanying code repositories, will enable technologically sophisticated researchers in the digital humanities to carry out similar projects in the future. Moreover, by allowing for users to look 'under the hood' at the trained classifiers, readers who are less technologically sophisticated can understand what kind of material might be expected to be included or left out of consideration. That is, this project enables both model construction and model criticism.

Notes

1. See https://osf.io/7crgt/.
2. There is also an existing literature on the principles of corpus construction (e.g. Bauer and Aarts, 2000). I have chosen to focus on the more specific issue of how researchers in the historical sciences interact with digital archives. However, the method set out in this article is consistent with the general principles in the literature on corpus construction. In particular, it sets up a cyclical procedure for improving corpora (cf. Bauer and Aarts, 2000, p. 29).
3. To replicate this search, use the following URL: https://paperspast.natlib.govt.nz/newspapers?query=%22socialist+church%22
4. Further discussions of the problems of keyword search methodology are available (e.g. Bingham, 2010, pp. 229–30). It is also important to note that keyword search methods and online methods remain useful for many purposes. The connections between researchers enabled by chance meetings made online through keyword-search-mediated engagement with the Internet (Leary, 2005) and the ability to quickly 'glance sideways' (Putnam, 2016) into different national archives have generated spontaneous and serendipitous insights (Fyfe, 2015; Ramsay, 2014). Moreover, it is possible to keyword search in more rigorous ways, by, say, reporting search terms and quantifying results (e.g. Nicholson, 2013, pp. 67–7).
5. For example, the 'Socialist Church' example above.
6. It is important to note that digital methods can also be used to 'zoom in' to a specific text (see Froehlich, 2018; cited by Owens and Padilla, 2021). For an example of features of archival research which cannot be captured digitally see Plunkett (2008).
7. For instance, consider Alfano *et al.* (2018), who use text mining and network visualization to investigate values and virtues ascribed to people in (contemporary) newspaper obituaries. Their article argues that certain 'virtues', including, say, sports fandom, have been insufficiently attended to by philosophers, given their prominence in the thought of the wider public.
8. This enables a turn to 'digital hermeneutics', according to which a researcher can ensure that they do not leave 'the computer's assumptions and limitations unarticulated' (Romein *et al.*, 2020, p. 309). To do this exactly what is being done to the data between input and output must be presented. See Koolen *et al.* (2019) for a discussion of data and digital tool criticism in the digital humanities. An

illuminating discussion of some alternative methods for presentation of digital humanities research, as carried out in the *Valley of the Shadow* project is found in Thomas (2004).

9. Putnam (2016) notes the origin of search methods in the desire to connect customers with products and that this is not the appropriate relationship between a historian and their sources (p. 377). Even when products are produced with research uses in mind, their underlying methods need not be well documented. See, for instance, Nicholson's discussion of Google's Ngrams (2013, p. 65).
10. Automated methods for improving OCR with historical newspaper data are being developed, but are outside the scope of this study (e.g. Drobac and Lindén, 2020). One instance of OCR quality affecting research using historical newspapers is provided by Smith *et al.* (2014), who model test reuse with overlapping *n*-grams. They find that the choice of *n* is constrained by poor OCR quality.
11. Laerke (2013) himself conceives of the history of philosophy as dealing properly with clusters of texts around controversies and approvingly quotes Lepenies' characterization of Dilthey's history of philosophy as 'anthropology carried out in the archive' (p. 14). The role of the archive and of the selection of texts 'in context' means that this kind of historian of philosophy has to be attuned to the methodological worries discussed above.
12. Ballantyne's work also emphasizes the materiality of the newspaper industry, following the process from raw materials to finished product.
13. Strictly speaking, there are items in te reo Māori in the dataset, but the majority is in English.
14. Research using the digital niupepa Māori database include Keelan *et al.* (2021), Paterson and Wanhalla (2017), and Whaanga and Wehi (2017). For an account of niupepa Māori, see Paterson (2006). Paterson and Wanhalla (2017) discuss the relative roles of digitized and physical archives in their work in the section 'Note on Sources'.
15. While the pilot project is over, the data are still available as of mid-November 2021 (URL: https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot).
16. The software used for OCR was ABBYY FineReader 8.1.
17. It was decided to focus on only those items classified as 'articles'. This excludes advertising. A fuller study of intellectual culture would probably include both advertising of public lectures *and* the use of advertising space as a publishing platform for philosophically interesting texts (e.g. https://paperspast.natlib.govt.nz/newspapers/SCANT18930215.2.29.2).
18. At this stage, there are other properties of the article which it is possible to extract. For instance, once could extract the physical properties of the article, such as the width of each line. This was not considered relevant for the case study.
19. When applying this method to the original compressed files, 11 were unable to be processed. There were individually decompressed using the Gzip programme, and processed separately. The 11 corrupted files consisted of the *Lyttleton Times* for 1890 and 1891, the *Christchurch Star* for 1883–86, the *Otago Daily Times* for 1898, the *Clutha Leader* for 1886, the *Nelson Evening Mail* for 1889, the *Manawatu Standard* for 1884, and *The Colonist* for 1898. After taking this extra step, only two issues of the *Lyttleton Times* were lost, eight issues of the *Christchurch Star*, five issues of *The Colonist*, and one issue each of the *Clutha Leader, Nelson Evening Mail, Manawatu Standard*, and *Otago Daily Times*.
20. The NLTK text book is a good place to start for concordancing, collocations, and word clouds (Bird *et al.*, 2009; see also Baker, 2006). A good tutorial on cooccurrence networks can be found in Niekler and Wiedemann (2020). The original paper for LDA topic modelling is Blei *et al.* (2003). The Gensim website contains accessible introductions to applying LDA topic modelling with Python (URL: https://radimrehurek.com/gensim/).
21. Removing items tagged as 'advertisement' during the preprocessing stage did not remove all advertising content.
22. Consequently, this way of operationalizing 'philosophical discourse' downplays the core connection of philosophy with *argument* and *reason*. There are text-mining approaches in development which aim to extract argumentative structure from texts (e.g. Lawrence and Reed, 2020). These methods are still in their infancy and there is reason to think they will struggle with the poor sequence data generated by even the best performing OCR in historical newspaper datasets. The development of these methods remains an interesting direction for future research.
23. Originally, the 'Readable' label was included with a thought to train a classifier to distinguish the material whose OCR errors did not, by quantity or quality, make the text unreadable and those which did.
24. A random sample of texts, cooccurrence networks, and cooccurrence scores can be inspected from each iteration of the process at https://newspaper-philosophy.canterbury.ac.nz.
25. A possible addition to the method set out here would be to add quantitative measures of corpus similarity and difference and watch how they change over multiple iterations of the corpus construction process. For a discussion of appropriate measures for this see Kilgarriff (2001).
26. To explore these cooccurrence networks go to https://newspaper-philosophy.canterbury.ac.nz.
27. An anonymous reviewer encouraged a fourth iteration to ensure that debates around woman's suffrage were comprehensively included in the final corpus. This was carried out by adding 47 new philosophical items around the woman's suffrage debate, and, more broadly, use of terms like 'womanhood'. The result continued to struggle with the blurry boundary between 'philosophy' and 'politics' but did include the item missed at the third iteration. The resulting corpus is added endnote to above paragraph: available on the OSF page for this article and can be explored using the project dashboard. This kind of 'single issue' iteration may be a sensible option for 'fine tuning' the corpus for investigation of a specific issue. However, it required the probability threshold to be lowered to 0.45 in order to maintain all of the items from the 'external validation' section, below, and includes more than 100,000 items. This is on the edge of what I am able to work with given the methods presented here and the computational resources currently available to me.
28. This list provides some additional confirmation that the corpus includes voices not traditionally included in the history of philosophy. For instance, Besant has recently been taken up as a neglected female philosopher (Leland, 2021).

Acknowledgement

I would like to acknowledge the helpful engagement of the anonymous reviewers with this work. This material was presented to audiences at the University of Canterbury and the Australasian Association for Digital Humanities 2021 conference. Their feedback is gratefully acknowledged. Finally, I would like to thank Ben Adams, Geoff Ford, James Williams, Christopher Thomson and Karin Stahel for their engagement with, and encouragement of, this project.

References

Agamben, G. (2017). *What Is Philosophy?* Redwood City: Stanford University Press.

Alfano, M., Higgins, A., and Levernier, J. (2018). Identifying virtues and values through obituary data-mining. *Journal of Value Inquiry*, **52**(1): 59–79.

Baker, P. (2006). *Using Corpora in Discourse Analysis*. London: Continuum.

Ballantyne, T. (2012). Reading the newspaper in Colonial Otago. *The Journal of New Zealand Studies*, **12**: 47–64.

Bauer, M. W. and Aarts, B. (2000). Corpus construction: a principle for qualitative data collection. *Qualitative Researching with Text, Image and Sound: A Practical Handbook*, 19–37.

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., *et al.* (2017). Reproducible research in linguistics: a position statement on data citation and attribution in our field. *Linguistics*, **56**(1): 1–18.

Bingham, A. (2010). The digitization of newspaper archives: opportunities and challenges for historians. *Twentieth Century British History*, **21**(2): 225–31.

Bird, S., Loper, E., and Klein, E. (2009) *Natural Language Processing with Python*. Sebastopol: O'Reilly Media.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, **3**: 993–1022.

Bush, M. (2018). 1891: The Collins–Hosking debate, Christchurch. *Public Understanding of Science*, **27**(7): 897–904.

Calude, A. and James, P. (2011). A diachronic corpus of New Zealand newspapers. *New Zealand English Journal*, **25**: 39–52.

Darwin, C. (2009). In Endersby, J. (ed.), *On the Origin of Species*. Cambridge: Cambridge University Press.

Crane, R. (2013). A dangerous visionary? The lectures of the evolutionist T.J. Parker. *The Journal of New Zealand Studies*, **15**: 36–49.

Davies, M. and Helgeby, S. (2014). Idealist origins: 1920s and before. In Oppy, G. and Trakakis, N. N. (eds), *History of Philosophy in Australia and New Zealand*. Dordrecht: Springer, pp. 15–54.

Deleuze, G. and Guattari, F. (1991). *What Is Philosophy*? New York: Colombia University Press.

Drobac, S. and Lindén, K. (2020). Optical character recognition with neural networks and post-correction with finite state methods. *International Journal on Document Analysis and Recognition*, **23**: 279–295.

Fyfe, P. (2015). Technologies of serendipity. *Victorian Periodicals Review*, **48**(2): 261–266.

Froehlich, H. (2018, June). *Distance-Reading the Feminine Landscapes of the Awakening*. https://blog.bham.ac.uk/clic-dickens/2018/06/29/distance-reading-the-feminine-landscapes-of-the-awakening/ (accessed 4 December 2021).

Gareth, J., Witten, D., Hastie, T. and Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R*. Second Edition. New York: Springer.

Gibbs, F. and Owens, T. (2013). The hermeneutics of data and historical writing. In Dougherty, J. and Nawrotzki, K. (eds), *Writing History in the Digital Age*. Ann Arbor: University of Michigan Press, pp. 157–170.

Goldenbaum, U. (2013). Understanding the argument through then-current public debates or my detective method of history of philosophy. In Laerke, M., Smith, J. E. H., and Schliesser, E. (eds), *Philosophy and Its History: Aims and Methods in the Study of Early Modern Philosophy*. New York: Oxford University Press, pp. 71–90.

Hagen, T. M. (2012). Automatic topic classification of a large newspaper corpus. In Andersen, G. (ed.), *Exploring Newspaper Language: Using the Web to Create and Investigate a Large Corpus of Modern Norwegian, Amsterdam, pp. 111–130.*

Hitchcock, T. (2013). Confronting the digital. *The Journal of the Social History Society*, **10**(1): 9–23.

Keelan, T. J. (Ngāti Porou, Ngāti Awa, Tūhoe), Stewart, K. (Te Arawa, Ngāti Raukawa), Te Awekotuku, N. (Ngāti Whakaue, Tūhoe), Waimarie Nikora, L. (Te Aitanga-a-Hauiti, Tūhoe), Edge, K. (Ngāti Maniapoto) and McRae, O. (Ngāti Whakaue, Ngāti Mahuta) (2021) The case of a change in meaning and its impact. *Kōtuitui: New Zealand Journal of Social Sciences Online*, **16**(1): 168–79.

Kilgarriff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, **6**(1): 97–133.

Koolen, M., Van Gorp, J., and Van Ossenbruggen, J. (2019). Toward a model for digital tool criticism. *Digital Scholarship in the Humanities*, **34**(2): 368–385.

Laerke, M., Smith, J. E. H., and Schliesser, E. (2013). Introduction. In Laerke, M., Smith, J. E. H., and Schliesser, E. (eds), *Philosophy and Its History: Aims and Methods in the Study of Early Modern Philosophy*. New York: Oxford University Press, pp. 1–6.

Laerke, M. (2013). The anthropological analogy and the constitution of historical perspectivism. In Laerke, M.,, Smith, J. E. H., and Schliesser, E. (eds), *Philosophy and Its History: Aims and Methods in the Study of Early Modern Philosophy*. New York: Oxford University Press, pp. 7–29.

Lawrence, J. and Reed, C. (2020). Argument mining: a survey. *Computational Linguistics*, **45**(4): 765–818.

Leary, P. (2005). Googling the victorians. *Journal of Victorian Culture*, **10**(1): 72–86.

Leland, K. (2021). 'Friendly to all beings': Annie Besant as ethicist. *British Journal for the History of Philosophy*, **29**(2): 308–26.

McKinney, W. (2010). Data structures for statistical computing in python. In van der Walt, S. and Millman, J. (eds), *Proceedings of the 9th Python in Science Conference, Austin*, pp. 56–61.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Nicholson, B. (2013). The digital turn: exploring the methodological possibilities of digital newspaper archives. *Media History*, **19**(1): 57–73.

Niekler, A. and Wiedemann, G. (2020). *Tutorial 5: Co-occurrence Analysis*. https://tm4ss.github.io/docs/Tutorial_5_Co-occurrence.html (accessed 22 January 2021).

Owens, T. and Padilla, T. (2021). Digital sources and digital archives: historical evidence in the digital age. *International Journal of Digital Humanities*, **1**: 325–341.

Paterson, L. (2006). *Colonial Discourses: Niupepa Māori, 1855–1863*. Dunedin: University of Otago Press.

Paterson, L. and Wanhalla, A. (2017). *He Reo Wāhine: Māori Women's Voices from the Nineteenth Century*. Auckland: University of Auckland Press.

Pedregosa, F., Varoquaux, G. and Gramfort, F. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, **12**: 2825–30.

Plotly Technologies Inc. (2021). *Collaborative Data Science*. https://plot.ly (accessed 5 February 2021).

Plunkett, J. (2008) From optical to digital (and back again), *19: Interdisciplinary Studies in the Long Nineteenth Century*, **6**. https://doi.org/10.16995/ntn.479

Priest, G. (2006) What is philosophy? *Philosophy*, **81**(2): 189–207.

Putnam, L. (2016). The transnational and the text-searchable. *American Historical Review*, **121**(2): 377–402.

Řehůek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Valletta: European Language Resources Association, May 2010, pp. 45–50.

Ramsay, S. (2014). The hermeneutics of screwing around; or what you do with a million books. In Kee, K. (ed.), *Pastplay: Teaching and Learning History with Technology*. University of Michigan Press, pp. 111–120.

Romein, C. A., Kemman, M., Birkholz, J. M., *et al.* (2020). State of the field: digital history. *History*, (**105**): 291–312.

Rubinstein, A. (2019). Historical corpora meet the digital humanities: the Jerusalem corpus of emergent modern Hebrew. *Language Resources & Evaluation*, **53**(4): 807–35.

Scheirer, W., Forstall, C., and Coffee, N. (2016). The sense of a connection: automatic tracing of intertextuality my meaning. *Digital Scholarship in the Humanities*, **31**(1): 204–17.

Smith, D. A., Cordell, R., Dillon, E. M., Stramp, N., and J, Wilkerson. (2014). Detecting and modeling local text reuse. In *IEEE/ACM Joint Conference on Digital Libraries*. doi: 10.1109/jcdl.2014.6970166

Steger, S. (2013) Patterns of sentimentality in Victorian novels. *Digital Studies/Le Champ Numérique*, **3**(2).

Sternfeld, J. (2011). Archival theory and digital historiography: selection, search, and metadata as archival processes for assessing historical contextualization. *The American Archivist*, **74**(2). 544–75.

Strange, C., McNamara, D., Wodak, J., and Wood, I. (2014). Mining for the meanings of a murder: the impact of OCR quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, **8**(1). http://www.digitalhumanities.org/dhq/vol/8/1/000168/000168.html

Tanner, S., Muñoz, T., and Ros, P. H. (2009). Measuring mass text digitization quality and usefulness: lessons learned from assessing the OCR accuracy of the British Library's 19th Century online newspaper archive. *D-Lib Magazine*, **15**(7/8).

Thomas, W. G., III. (2004). Computing and the historical imagination. In Schreibman, S.,, Siemens, R., and Unsworth, J. (eds) *A Companion to Digital Humanities*. Oxford: Blackwell.

Veridian Software. (n.d.) *What Is METS/ALTO?* https://veridiansoftware.com/knowledge-base/metsalto/ (accessed 22 January 2021).

Whaanga, H. and Wehi, P. (2017). Rāhui and conservation? Māori voices in the nineteenth century Niupepa Māori. *Journal of the Royal Society of New Zealand*, **47**(1): 100–106.

Wood, S. A. (2014). The reign of grace: liberalism and heresy in the new world. *Journal of Religious History*, **38**(4): 535–60.

Appendix

Table A1. External validation items

	Iteration				Source	Notes
Article	0	1	2	3		
ODT_18710509_ARTICLE18	False	True	NA	NA	Ballantyne (2012)	Used in training, second and third iteration.
TT_18710720_ARTICLE23	NA	NA	NA	True	Ballantyne (2012)	XML parse error fixed on final iteration.
OW_18730531_ARTICLE11	False	True	True	True	Ballantyne (2012)	
ODT_18760516_ARTICLE22	False	True	True	True	Ballantyne (2012)	
OW_18760520_ARTICLE81	False	True	True	True	Ballantyne (2012)	
OW_18760527_ARTICLE71	False	True	True	True	Ballantyne (2012)	
ODT_18760617_ARTICLE20	False	True	True	True	Ballantyne (2012)	
OW_18780907_ARTICLE28	False	True	True	True	Ballantyne (2012)	
ST_18800924_ARTICLE14	NA	NA	NA	True	Ballantyne (2012)	XML parse error fixed on final iteration.
ODT_18801204_ARTICLE15	False	True	True	True	Ballantyne (2012)	
OW_18820701_ARTICLE45	False	True	True	NA	Ballantyne (2012)	Used in training on final iteration.
OW_18820701_ARTICLE77	False	True	True	True	Ballantyne (2012)	
ODT_18820705_ARTICLE26	False	True	True	True	Ballantyne (2012)	
OW_18820708_ARTICLE61	False	True	True	True	Ballantyne (2012)	
ME_18940605_ARTICLE7	NA	NA	NA	NA	Ballantyne (2012)	Filtered (too short).
ODT_18880705_ARTICLE21	False	True	True	True	Wood (2014)	
ODT_18880906_ARTICLE30	False	True	False	True	Wood (2014)	
ODT_18881101_ARTICLE18	False	True	False	True	Wood (2014)	
ODT_18881102_ARTICLE31	False	True	False	True	Wood (2014)	
ODT_18881103_ARTICLE24	False	True	False	True	Wood (2014)	
ODT_18881108_ARTICLE3	False	NA	NA	True	Wood (2014)	Filtered (too short on first two iterations).
ODT_18810528_ARTICLE13	True	True	True	True	Crane (2013)	
OW_18810604_ARTICLE108	NA	NA	NA	NA	Crane (2013)	Filtered (too short).
OW_18820701_ARTICLE77	False	True	True	NA	Crane (2013)	Used in training on final iteration.
NOT_18840128_ARTICLE15	False	True	True	True	Crane (2013)	
ODT_18840923_ARTICLE8	False	False	False	False	Crane (2013)	Composite.
ODT_18850602_ARTICLE17	False	True	True	True	Crane (2013)	
ODT_18950927_ARTICLE6	False	False	False	False	Crane (2013)	Composite.
ODT_18910723_ARTICLE10	False	False	False	False	Crane (2013)	Composite.
ODT_18820629_NA	NA	NA	NA	NA	Crane (2013)	Advertisement.
ODT_18861120_ARTICLE3	False	False	False	False	Crane (2013)	Composite.
CHP_18911208_ARTICLE19	NA	NA	NA	NA	Bush (2018)	Filtered (too short).
CHP_18911211_ARTICLE9	True	True	True	NA	Bush (2018)	Used in training on final iteration.
CHP_18920119_ARTICLE36	True	True	True	True	Bush (2018)	
OW_18930824_ARTICLE48	False	False	False	False	Bush (2018)	Composite. Content is pure astronomy.
LWM_18920122_ARTICLE36	False	False	False	False	Bush (2018)	Composite.
ESD_18930817_ARTICLE25	True	True	False	True	Bush (2018)	
WOODEX_18950218_ARTICLE10	False	True	False	True	Bush (2018)	
WOODEX_18950218_ARTICLE4	False	True	False	True	Bush (2018)	
Success rate	0.11	0.81	0.58	0.81		

Table A2. False negatives for final classifier

Article	Philosophy type	Writing type	Notes	Papers past link
DTN_18940820_ARTICLE7	Ethics/politics	First order		https://paperspast.natlib.govt.nz/newspapers/DTN18940820.2.6
LT_18831025_ARTICLE34	Other	First order	Only final block counts. A short obituary for a Timaru lecturer, described as a certain kind of philosopher.	https://paperspast.natlib.govt.nz/newspapers/LT18831025.2.36
AS_18820306_ARTICLE31	Ethics/politics	Letter	On the Bradlaugh case, concerning the possibility of non-Christian MPs.	https://paperspast.natlib.govt.nz/newspapers/AS18820306.2.29
DSC_18600731_ARTICLE28	Other	Public event		https://paperspast.natlib.govt.nz/newspapers/DSC18600731.2.21.4
LT_18970507_ARTICLE14	Ethics/politics	Letter	Against the abolition of capital punishment.	https://paperspast.natlib.govt.nz/newspapers/LT18970507.2.14.4
OO_18910214_ARTICLE3	Ethics/politics	First order	On socialism, the need for work, and mechanisation.	https://paperspast.natlib.govt.nz/newspapers/OO18910214.2.5
ESD_18890826_ARTICLE1	Ethics/politics	First order	Reflection on New Zealand character. Not directly relevant apart from mention of Josiah Royce.	https://paperspast.natlib.govt.nz/newspapers/ESD18890826.2.2
CHP_18970329_ARTICLE55	Ethics/politics	Letter	Prison conditions. Not core to what we are after.	https://paperspast.natlib.govt.nz/newspapers/CHP18970329.2.44.1
ESD_18921006_ARTICLE15	Ethics/politics	First order	Stout on women's franchise. Also mentions Herbert spencer.	https://paperspast.natlib.govt.nz/newspapers/ESD18921006.2.15
TS_18800205_ARTICLE25	Religion/science	First order	Not core. Anecdotes about various Greeks, including Aristotle.	https://paperspast.natlib.govt.nz/newspapers/TS18800205.2.23