STAT601 Project Report
Using Data Science Methods to Investigate Philosophical Discourse in Early New Zealand Newspapers
DRAFT

Joshua Black (46086757)

Due: 22/01/21

Abstract

Data science methods provide new opportunities for humanists to explore large-scale text corpora. This report presents DATA601 summer project work with the UC Arts Digital Lab exploring philosophical discourse in New Zealand newspaper content up to 1900 using the recently released National Library Papers Past newspaper open data pilot dataset.

The project has two parts. First, a corpus of texts is constructed for the investigation of philosophical discourse in early New Zealand newspapers. This is carried out using a method of 'bootstrapping' to generate increasingly targeted collections of articles. Second, the value of the resulting corpus is demonstrated by using it to answer two specific research questions. First, it is used to investigate the understandings of the relationship between religion and the natural sciences present in the corpus. Second, it is used to argue that a sizeable minority of articles in the corpus attempt to apply philosophical concepts to the New Zealand situation.

Contents

Introduction

This project has two guiding aims. First, it aims to produce a corpus for the investigation of philosophical writing in early New Zealand newspapers. Second, it aims to illustrate the value of the resulting corpus by using it to answer two specific research questions about philosophical writing in New Zealand newspapers.

This project has been enabled by the release of the National Library Papers Past newspaper open data pilot [18]. This is a large dataset of newspaper content running from 1839-1899.[1] While, the first of the newspaper content which makes up a large part of the Papers Past collection was made available in 2000 [16], not many projects have attempted to use it as *data*.[2] The process of corpus construction reduces this dataset down from 7592619 articles to 31131 articles, which contain material interesting to the historian of philosophy. It is expected that the kind of methods and workflows developed in this project can be used profitably by other investigators to answer their own specific research questions about the Papers Past dataset.

The construction of the corpus was carried out using a 'bootstrapping' approach with a combination of exploratory corpus analysis methods and supervised learning techniques. Beginning from simple keyword searches on the dataset, a series of articles of interest, and a variety of articles which were not of interest, were labelled and then used to train a Naive Bayes classifier. This classifier was then applied to the whole dataset. Further articles were then labelled using the corpus obtained from the classifier and then used to train a more effective Naive Bayes classifier.

Having constructed the corpus, two more specific research areas are considered. First, the attitudes to the relationship between religion and science present in the corpus. Second, the presence of specifically New Zealand content in the corpus.

This report is divided into four sections. The first discusses the humanities motivations of the project and sets out the three research questions which guide the project. The second section

[1] The dataset does not contain any Māori niupepa content.

[2] For instance, not many of the sample projects in the National Library's collection of [19] use Papers Past, and none of those which do use it as data to answer digital humanities research questions.

covers the corpus construction phase, discussing the methods used, issues with their implementation, and the resulting corpus. The third section demonstrates the use of the corpus developed in the previous section by using topic modelling and co-occurrence network analysis to investigate two specific questions about philosophical discourse in early New Zealand newspapers. The fourth section presents a critical discussion of the project, highlighting its successes and the aspects of the project which could be improved. An appendix contains some more detailed tables of results which are referred to in the report.

1 Organisation, Motivation and Research Questions

1.1 UC Arts Digital Lab

This project was carried out within the UC Arts Digital Lab at the University of Canterbury, under the supervision of Geoffrey Ford (Political Science) and Benjamin Adams (Computer Science).

The UC Arts Digital Lab works in the interdisciplinary area of 'digital humanities'. Digital humanities is the application of computational and data science techniques to handle research and material traditionally studied by humanities disciplines. This includes digital archiving and computational analysis of, e.g., texts, audio and video, to gain insight into human culture.

This project falls within digital humanities by using data from a digital archiving project, the National Library's digitisation of New Zealand newspapers, and by applying data science techniques to extract relevant material and analyse it. That is, it combines data science methods with the traditional humanities aim of generating insights about cultural products.

Digital humanities is in its early stages of development in New Zealand. It is hoped that this project will demonstrate some of the potential of the data sources and resources available in New Zealand for novel work in this area.

1.2 Project Motivations

This project picks out and investigates philosophical discourse in early New Zealand newspapers. This section provides some working definitions of important terms, sets out some of the motivations for carrying out the project and presents three research questions which will direct the project as a whole.

For the purpose of this project, 'philosophical discourse' will be discourse which concerns ultimate values or ultimate reality. It will also have a special connection with argumentation. For instance, in many political speeches the concept of 'liberty' is used. Reflection on, and argument for, a particular idea of what liberty means would count as philosophical discourse (in this case, philosophical discourse about ultimate values) (e.g. [22]). Similarly, reflection on whether there is a God is discourse about what is ultimately real (e.g. [27]). Philosophy, in this sense, can be carried out both as an academic subject in the universities by specialists in the field and as a public activity by 'non-professionals'.

Often the history of philosophy is a history of *academic* philosophy. This leaves not much to say about philosophy in New Zealand until the well into twentieth century. Davies notes that, before this time, 'many of those who had longstanding [academic] chairs published next to nothing' and that the Australasian (and especially New Zealand) philosophical community was small and not well connected [13, p. 24]. It has also been noted that academic philosophy in New Zealand has tended to simply follow international trends, with a 'New Zealand philosophy' as something to be hoped for in the future [8, pp. ix–x]. These limitations in the

usual focus on academic philosophy in the New Zealand context encourage the attempt to look for philosophical discourse elsewhere. Early New Zealand newspapers are a plausible place for us to look.[3] Indeed, there is reason to think that the relative lack of intellectual journals in New Zealand at the time, when compared with the UK, means that New Zealand newspaper contained more intellectual content than would be otherwise expected [2, p. 57] [5, p. 37].

The turn to newspaper writing to investigate philosophical discourse in early New Zealand is enabled by the availability of digital humanities and data science methods. The availability of more than seven million articles in digital form allows for the application of text processing methods to highlight content which might otherwise be missed or to allow 'reading' of more articles than it would be feasible for a single researcher to explore. This is, to use Moretti's phrase, 'distant reading', as opposed to 'close reading' [14, pp. 47–49].[4] [Distant reading]

1.3 Research Questions

The project is structured by three research questions, as follows:

1. Can supervised learning methods be used to produce a useful corpus for investigation of philosophical discourse in Early New Zealand Newspapers from the National Library Papers Past Newspaper Open Data Pilot dataset?

2. Can the resulting data set be used to provide insight into how the relationship between religious belief and developments in the natural sciences was understood in early New Zealand?

3. Can topic modelling and co-occurrence analysis on the resulting corpus reveal that philosophical writing in New Zealand newspapers incorporates concerns which are specific to New Zealand?

The first question concerns corpus construction. The criteria for success in corpus building are not straightforwardly quantifiable and require the use of some qualitative judgement. We will examine the resulting corpus to determine if the articles it contains are relevant to the historian of philosophy and test for the presence of articles identified as of interest by previous studies but not used in the process of corpus construction.

The second and third questions have been formulated to demonstrate the value of the corpus by using it. The second question will be answered by focusing on a particularly important issue at the time: the theological and religious significance of the then-new theory of evolution by natural selection.

The third question is motivated by the desire to see whether and to what extent philosophical discourse was built into, or responsive to, issues which arose for the settlers in the New Zealand context.

1.4 Data and Ethics Constraints

The UC Arts Digital Lab imposed no constraints on the use of the data in this project. The data used in the project is all publicly accessible and does not directly discuss any living person. The

[3] As it happens, there are many interesting connections between newspapers and philosophy in Australasia. One of the first newspapers in New Zealand, *The New Zealand Advertiser and Bay of Islands Gazette*, was produced by Barzillai Quaife a Congregationalist minister who went on to be the first professional philosophy teacher in Australia and to publish *The Intellectual Sciences*, a book which has a claim to being the first philosophical monograph produced in Australasia [13, pp. 16–17]. This example is singular, but further encourages the idea that early intellectual life, and philosophy in particular, may be tracable through early New Zealand newspapers.

[4] For a note of warning about treating this material as simply a 'store of information' see [2, pp. 59–60].

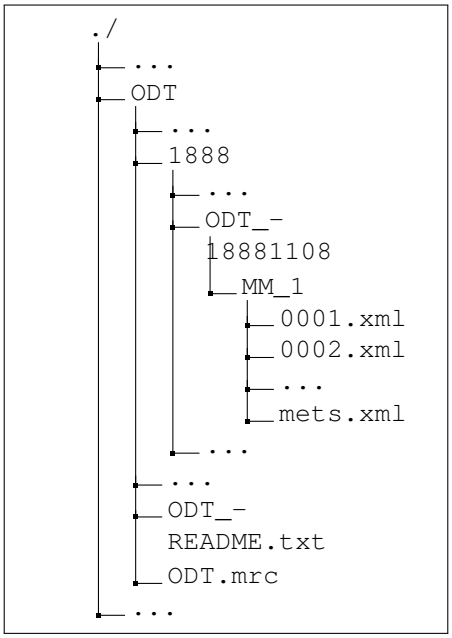

Figure 1: Example directory structure (uncompressed).

National Library states that the material in the dataset is out of copyright in New Zealand [4].

The data does contain some offensive material. For instance, racist descriptions of people. Material of this sort will be avoided in this report, the project presentation, and project poster. A disclaimer has also been added to the project dashboard.[5]

1.5 Data Sources

The data for the project comes entirely from National Library Papers Past newspaper open data pilot [18]. This is a release of newspaper data which was made available August 2020 with the intention of allowing people to experiment with the dataset for up to a year. It contains a subset of the newspaper content from the Papers Past archive, with material from 1839 to 1899 inclusive.

If uncompressed, the dataset is structured into folders by newspaper title, by year, and by issue. Each newspaper title folder can contain multiple 'year' folders, and each year folder can contain multiple issue folders. Each newspaper title contains a MARC ('MAchine-Readable Cataloging') format metadata source file. Each issue of a newspaper contains ALTO ('Analyzed Layout and Text Object') XML files for each page of the issue and a single METS ('Metadata Encoding and Transmission Standard') XML file. ALTO, MARC, and METS are Library of Congress specifications.[6] Figure 1 provides an example for a single issue of the Otago Daily Times.

The Papers Past newspaper dataset is created from microfilm images of individual pages of newspapers. Each microfilm scan is then run through an optical character recognition (OCR)

[5]The dashboard will be discussed later. It is available at `https://www.nz-newspaper-philosophy.herokuapp.com`.

[6]Full specifications for the METS and ALTO files are available at [28]. The MARC files are unnecessary for this project, but details of the MARC specification are available at [12].

algorithm to produce the ALTO files for each page. The article boundaries and corrected titles for each article are contained in the METS file for the issue along with the logical and physical structure of the issue as a whole.[7] At least some of the OCR for Paper's Past was carried out more than 10 years ago using ABBEY FineReader 7.1, released in 2003. Consequently, it is not of the quality which could be reasonably expected now.[8]

The dataset contains 79 distinct newspaper titles, 306538 distinct issues, and 1471384 distinct pages. As each of these corresponds to an XML file, the dataset contains 1778001 distinct XML files. These are delivered from the National Library website in the form of newspaper-year tarballs. That is, each all of the issues for a given newspaper in a given year are collected in a single `.tar.gz` file. On disk, these tarballs take up 315GB. Rather than decompressing the whole dataset, the tarballs were opened one at a time to extract the required information.

2 Corpus Construction

The first task of is to produce a corpus suitable for investigating philosophical writing in early New Zealand newspapers. This section sets out the methods used to construct the corpus (§2.1), the challenges involved in implementing the methods (§2.2), and provides an initial evaluation of whether the resulting corpus is suitable for investigating philosophical writing in early New Zealand newspapers (§2.3).

2.1 Methodology

The process for constructing the corpus consisted of five steps. First, a preprocessing phase, in which the XML files for the corpus were run through, divided into articles, and placed in a Pandas dataframe (§2.1.1). Second, exploration of the dataset to find articles of interest, using keyword searches, concordancing, collocation and co-occurrence analysis (§2.1.2). This step merges with the third step, namely, labelling articles according to a series of hierarchical classificatory labels (§2.1.3). Having labelled a number of articles, we then train a naive bayes classifier using the labelled data, optimising parameters using a grid search method (§2.1.4). The fifth step is to apply the resulting classifier to the whole dataset. This produces a candidate corpus of relevant articles. Having reached this stage, we can return to step two to begin another iteration of the bootstrapping process.[9] In this project, two iterations of the process were carried out.

2.1.1 Preprocessing

The data processing stage begins with newspaper-year tarballs, stored in directories corresponding to each newspaper.[10] These were iterated through, with each tarball decompressed and each issue then processed. The METS file for each issue was first scanned, finding each item tagged 'ARTICLE' within the 'logical structure' tag. This excludes all newspaper items tagged

[7] The National Library provide a useful diagram of the process [17].

[8] This is important when it comes to selection of methods to use with the dataset. In particular, some recent classification methods rely on high quality sequence data which is not possible in the presence of low-quality OCR. These issues will be discussed in more detail (§4).

[9] The term 'bootstrapping' is used to indicate that earlier stages of the process are used in order to generate more satisfactory results at later stages. The process 'pulls itself up by its bootstraps'. The term is widely used, with various more specific meanings, in many disciplines. One prominent example in philosophy is [7].

[10] See 'generate_corpus_df.py' for the full code for this step. Unless otherwise stated, Python scripts and Jupyter notebooks are found in the root of the project's GitHub repository at `https://github.com/JoshuaDavidBlack/NPOD_Philosophy`.

Corpus	Article Count
'philoso*' Corpus	29647
Naive Bayes 1	239649
Naive Bayes 2	31131

Table 1: Article counts for general philosophy corpora.

with 'ADVERTISEMENT'.[11] Each 'div' tag with attribute 'ARTICLE' was then iterated through in order to collect the identifiers of its child 'area' tags. These represent the areas of text which correspond to the article in the scans of the various pages of the issue.

Having collected the articles in the issue, and all of the areas of the physical pages which correspond to each issue, we then iterate through each article again, collecting the text which makes up each block. We collect these as a list of strings, where each string corresponds to an area of the page in which the article appears. In practice, each area corresponds to a paragraph of text.[12]

A unique code is assigned to each article by appending the article number from the issue to the newspaper name and date of issue. So, for instance, an article might have the code 'CHP_-18820107_ARTICLE20', which would indicate that the article comes from the issue of *The Press* from the seventh of January 1882, and that it is the 20th article as numbered in the 'logical structure' contained in the issue's METS file. These codes were used to index a dataframe with the article title and article text as columns.

In order to process this large amount of data efficiently, the Python multiprocessing library was used in order to process multiple tarballs at once. As the whole processed dataset is too large to be conveniently stored in memory, it is stored as eight pickled dataframes, compressed as .tar.gz files. These take up eight gigabytes on disk.[13]

2.1.2 Dataset Exploration

The purpose of the exploration phase of the process is to determine the nature of the developing corpus and to pick out any keyword or articles which are good candidates to be labelled and used to train a classifier. This section sets out some of the main tools used at this stage. The techniques discussed here are, in order, manual inspection of the articles in the developing corpus, concordancing, collocation analysis, and co-occurrence networks.

It is convenient to briefly characterise the three stages in the construction of the corpus now. Rather than beginning with the full dataset, as processed above, the process was started with a subset of the corpus containing regular expression matches with 'philoso*'.[14] We will call this

[11] It was decided that actual philosophical discourse is most likely to appear in full articles. However, as one of the focuses of the project is reports of public events, it is worth noting that advertisements for such events are excluded by this step. If one were wanting, for instance, to track touring public lecturers across the country, this kind of information might be worth keeping and relevant to an understanding of philosophy as it appears in this dataset. Traditional studies of newspaper reports of lectures sometimes appeal to such advertisements (e.g. [5, fn. 13] cites a lecture advertisement in the 29/06/1882 edition of the Otago Daily Times).

A more surprising instance of philosophical material appearing in advertising material is a riff on a recent lecture on evolution in the context of an advertisement a local store [5, pp. 39–40].

[12] These are also referred to elsewhere in the report, and exclusively in the project code, as 'blocks' or 'text blocks'. This is more convenient once we have abstracted away from thinking of the specific areas on a printed page.

[13] This reduction in size is a sign that the methods used in this project use only a small amount of the information contained in the dataset. In particular, we ignore almost all spatial information. The only exception to this is the maintainence of distinct blocks of text.

[14] This matches any string which starts with 'philoso', for instance, 'philosophy' or 'philosopher'. Code for filtering the dataset by regular expression is available in 'keywords_from_corpus.py'. The script 'multithread_keyword_-

1	n with tho university a chair of logic moral	philosophy	and political economy ah along tho church ha
2	own god the cause of ' all causes before all	philosophy	before all systems creeds and divinations be
3	elves to the dangerous charm of specula live	philosophy	and drifting far away from the safe anchorag
4	heaven and earth than are dreamt of in their	philosophy	and if' we should only discover one or two o
5	ish as to try to estab lish a chair of moral	philosophy	in seaeliff as you will observe sir mr adams
6	se you and he had last evening about natural	philosophy	' 'no ghosts ' answered la dy cecil gravely

Table 2: Selected results from 'philosophy' concordance.

corpus the Philoso* corpus. The corpus resulting from the application of the first trained Naive Bayes classifier will be called the NB1 corpus, and the corpus resulting from the second Naive Bayes classifier will be called the NB2 corpus. The sizes of these corpora are presented in Table 1.

Manual inspection of a random subset of a corpus helps to reveal the general kind of material it contains. This might also reveal anything *missing* from the corpus. Inspection of the Philoso* corpus revealed plenty of material of the desired sort, including reports of public events (e.g. 'CHP_18690823_ARTICLE15', on a lecture against Darwinism on the grounds that it goes beyond claims that can be made by the sciences); letters to the editor (e.g. 'CHP_18920926_-ARTICLE10'); and first-order philosophical writing (e.g. 'CL_18790919_ARTICLE27').[15] It was also found to contain various material which we are not interested in. For instance, pieces of serialised fiction which describe a character as 'phliosophical' but don't engage in philosophical discourse (e.g. 'GLOBE_18770315_ARTICLE12'); or use of the word 'philosophical' to mean indifferent or resigned (e.g. 'OW_18920721_ARTICLE7').

To attain a higher level understanding of the content of a corpus, concodancing was carried out using the NLTK package [26]. An entry of a concordance is an occurrence of a given word in context within the corpus. A selected subset of concordance results for 'philosophy', within the Philoso* corpus, is presented in Table 2. We see reference to academic philosophy in (1) and (5); religious involvement in (1) and (2); something which looks like fiction in (6); the idea of philosophy as dangerous in (3); and the use of a well-worn phrase from Hamlet in (4). We also see some of the shortcomings of the OCR, with lots of word breaks present where they should not be and missed letters. For instance, 'seaeliff' should be 'seacliff' in (5).

Collocation analysis also suggested a large amount of relevant material. A collocation is a word whose appearance within a given distance from a search term is significant. The selected window size for collocations with 'philosophy' was five to the left and right of the word. The significance of collocations was ranked by pointwise mutual information score. This, informally speaking, ranks the pair by how much information is given about the occurrence of one word given the appearance of the other. In Table 3 We see philosophy associated with reflection in general in (1); apparent references to the Hamlet quote above in (3), (4), and (6); reference to the sciences in (12); particular schools of philosophy (5) and (8); and some more confusing entries, e.g. (9).

The final exploratory method applied was the use of co-occurrence network analysis. Unlike collocations, which must appear within a window of a given word, we carried out document-level co-occurrence statistics. The approach followed in this project is derived from [1], with modifications to convert the approach from R to Python. Co-occurrence statistics were computed with Numpy [9], and visualised as networks using Dash cytoscapes [11]. In addition to mutual information, co-occurrence scores were calculated with the 'log Dice' metric, which is

search.py' allows for matches with regular expression search terms to be pulled directly from the dataset tarball and returned as a dataframe. This allows the user to avoid generating a dataframe for the whole dataset if it is not useful to them.

[15] See the notebook 'philoso_subset_exp.ipynb' for full code for the exploratory steps using the Philoso* corpus.

Rank	Collocation
1	('philosophy', 'reflectiveness')
2	('synthetic', 'philosophy')
3	('dreamt', 'philosophy')
4	('undreamed', 'philosophy')
5	('baconian', 'philosophy')
6	('undreamt', 'philosophy')
7	('proverbial', 'philosophy')
8	('epicurean', 'philosophy')
9	('philosophy', 'apparel')
10	('axioms', 'philosophy')
11	('matics', 'philosophy')
12	('inductive', 'philosophy')
13	('cosmic', 'philosophy')
14	('foregoing', 'philosophy')
15	('transcendental', 'philosophy')

Table 3: Top 15 Collocations with 'philosophy', window size of 5, ranked by pointwise mutual information.

designed to be consistent across different corpus sizes and insensitive to word frequency [20].[16]

One issue which arises for co-occurrence analysis is the creation of an appropriate dictionary. This is especially important for a large corpus generated by OCR. In such cases, automatic generation of a dictionary is likely to result in a large collection of non-words. A few approaches were used to avoid this. Using the Gensim package [25], it was specified that a word must appear in more than 20 articles and in less than 40% of articles in order to appear in the dictionary.[17] In addition, the tokenised text was further filtered to only include words which appear in the NLTK English words dictionary.[18] The number of resulting words in the dictionary varied by corpus, but this kind of filtering was sufficient to enable the calculation of collocation scores without overwhelming available memory resources. The main constraint here is the size of the 'document-term matrix' and 'term-term matrix' required to calculate the co-occurrence scores. The former is a dictionary size by number of documents matrix containing the number of occurrences of each word in each document, and the latter is a dictionary by dictionary size matrix containing how often a given term appears in a document with another.[19]

One downside of this method of filtering is that the names of entities which appear in the dataset, but which do not appear in an English dictionary, are lost.[20] This is particularly worrying for the historian of philosophy, who may want to pick out particular people or institutions to be the subject of further research. To overcome this problem, two further dictionaries we made by running the corpus through SpaCy [10] to pick out both named entities and proper nouns.[21] These were then ran as the text through the same dictionary creation approach as described above.

Finally, co-occurrences were also calculated using the TF-IDF transformation, which scales

[16]The calculation of log dice and mutual information co occurrences are done by the functions 'log_dice_coocs' and 'mi_coocs' respectively, while the construction of the co-occurrence network is performed by the function 'network_dash'. All three are in the file 'NL_helpers.py'. The Dash visualisation code is most easily inspected within the GitHub repository for the Heroku dashboard at `https://github.com/JoshuaDavidBlack/NPOD_Philosophy_Heroku`.

[17]This code is also contained in 'philo_subset_exp.ipynb'.

[18]The 'wordlist' corpora was used, corresponding to the '/usr/share/dict/words' file on Unix systems. See Chapter 2 of [26].

[19]Code to generate these matrices is included in 'Entity Extraction with Philo Subset.ipynb', 'generate_matrices_propn.py', and 'generate_matrices_entities.py'.

[20]While the wordlist corpus used as an English dictionary has many names, it does not have *all* names.

[21]See code in ''.

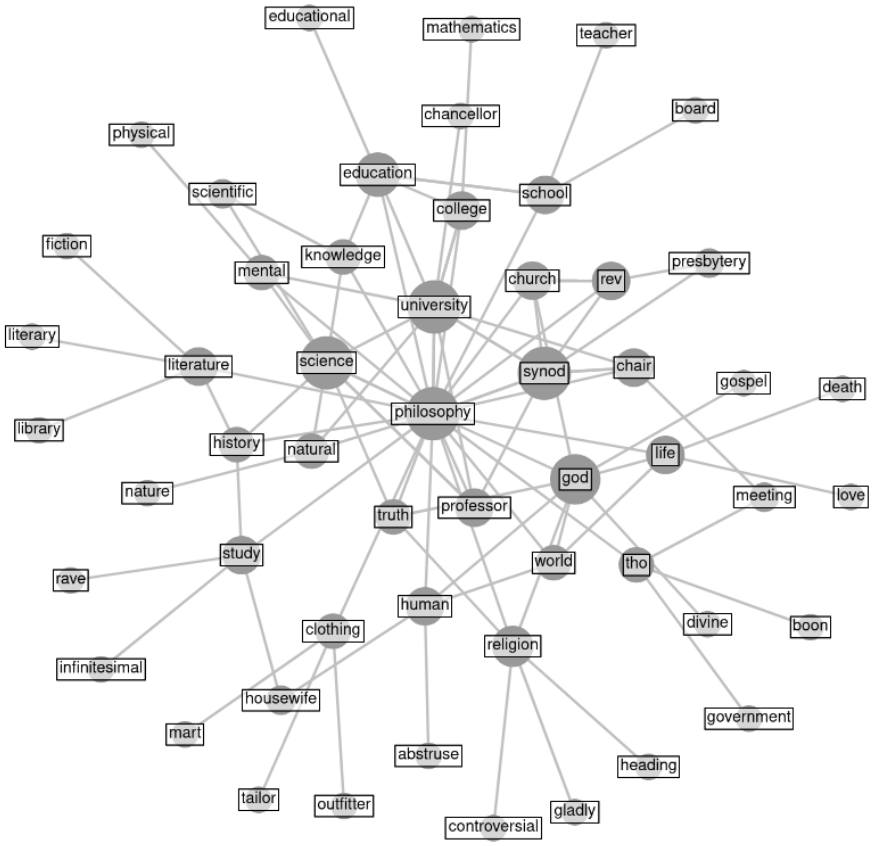

Figure 2: Co-occurrence network for 'philosophy' in 'philoso*' corpus, with TF-IDF transformation, and log Dice statistic. The top 25 co-occurrences for 'philosophy' are shown, with the four top co-occurrences for each of these words also added to the network.

word scores both on the basis of the number of words in the given article and the frequency of the given word in the corpus. This was also done using the Gensim package.

Figure 2 provides an example of a co-occurrence network generated during this phase of the project.[22]. The width of the edges indicate the strength of co-occurrence and the size of the nodes indicates the number of connections the node has. The figure shows the close connections of philosophy both with the academic word and also with Church bodies (see, e.g. 'synod'). There are some surprising words included as well, for instance, 'clothing'.[23]

2.1.3 Labelling Articles

Each of the exploratory steps set out above helped to clarify what kind of philosophical writing we could expect to find, and what kind of material we want to exclude. This understanding

[22]Many more such networks can be produced using the dashboard at `http://nz-newspaper-philosophy.herokuapp.com`

[23]Looking in to this, it seems that an advertisement referring to the 'philosophy of clothing' appended to the end of articles for a clothing company in the *Daily Sourthern Cross* is responsible for this connection, e.g. [23]. This may also explain the confusing entry 'apparel' in the collocation results set out above.

Label	First Iteration Value	Count	Second Iteration Value	Count
Readable	True	247	True	918
	False	26	False	41
Philosophy	True	101	True	299
	False	147	False	620
Philosophy Type	Religion-Science	58	Religion-Science	140
	Ethics-Politics	25	Ethics-Politics	94
	Epistemology-Metaphysics	3	Epistemology-Metaphysics	13
	Other	15	Other	52
Writing Type	Public Event	40	Public Event	97
	Letter to the Editor	23	Letter to the Editor	69
	First-order Writing	36	First-order Writing	111
	Review	2	Other	22
NZ	True	77	True	178
	False	12	False	41

Table 4: Label counts before first classifier was applied (First Iteration) and the second classifier was applied (Second Iteration).

enabled the creation of a collection of labelled articles, stored as a Pandas dataframe.[24] All labelling was carried out at the article level and relied on the investigator's experience in research in the history of philosophy.[25] This puts it in the category of single-annotator corpora rather than those whose labels have been validated by multiple investigators. Five labels were used.

The first label tracks whether an article is readable or not. The criterion for readability was whether the investigator could decipher the sentence without any appeal to the original image. The methods used later in the project use bag of words representations of the texts. Such methods do not rely on having high quality sequences of text. If the investigator was able to pick out from the features on the page that the article is relevant, it is hoped that the features used by the classification and topic modelling algorithms can also pick out the article. In any case in which a user of the corpus is interested in a text which is too garbled to read, enough information is provided for the user to access to the original scan from the Papers Past website and correct the OCR. The readability label is stored as a boolean for all articles in the labelled collection.

The second label tracks whether something is 'philosophy' or not. The criterion for inclusion was, as suggested by the characterisation of 'philosophical discourse' in Section 1.2. If the investigator judged that 'ultimate values' or 'ultimate reality' were being discussed, the article was included. Pieces discussing the practice of philosophical reflection were also included, for instance, material about philosophy provision at the universities. The philosophy label is stored as a boolean for all articles in the labelled collection.

An problem which had to be solved while labelling philosophy articles was the presence of large editorial articles in which multiple subjects are discussed in turn. In such articles, a report on a local fire, a reflection on some philosophical topic, and a bit of society gossip might all be put together. For instance, the article with code 'AS_18860821_ARTICLE40' consists of 28 blocks, of which 5 are dedicated to an argument with a reader over whether suicide is ethical [24]. These five blocks are definitely 'philosophical discourse', but the remaining blocks discuss poetry, a local operatic production and the close of a sitting of parliament. In the current version

[24]The notebook 'Classifying texts.ipynb' contains code used to label articles.

[25]The investigator has a PhD in Philosophy focusing on a particular figure in late-nineteenth century philosophy.

of the labelled dataframe, many articles of this sort are included.[26]

The next label tracks the genre of philosophy present in the article. The two main labels for this were 'r', indicating a discussion of religious belief and its relationship to modern science, understood widely to include the natural sciences, history, and biblical criticism; and 'e', indicating a discussion of ethics or politics. The label 'm' was used for metaphysics and epistemology and the label 'o' was used for 'other'. The metaphysics and epistemology label was designed to be used for purely secular discussion of, e.g., the nature of knowledge or reality. One could, for instance, discuss whether free will is possible given modern scientific results without discussing anything directly to do with religion. However, in practice, almost all discussions of these topics were more readily labelled with 'r'. This label falls under the philosophy label. That is, only those article labelled as philosophical have a 'philosophy type' label.

The fourth label was 'writing type'. This has three values: 'l' indicates a letter to the editor of the newspaper; 'p' indicates a report of a public event, such as a debate, lecture or sermon; 'f' indicates a first-order piece of philosophical discourse and 'r' indicates a book review. Only articles labelled with 'philosophy' were labelled with a writing type. However, this restriction was unnecessary and indeed not helpful if one were to try to train an algorithm to distinguish these writings types. In a future version of the project non-philosophy lectures and letters to the editor should also be labelled.

The final label was 'NZ', which tracks whether the author of the piece is based in New Zealand or not. In practice, almost all articles are tagged as being written by a New Zealand-based author. It was hoped that this tag might help with investigating the 'New Zealandness' of the corpus. However, this was abandoned as the lack of clearly 'non-New Zealand' pieces made it seem implausible that a useful classifier could be trained. The labelled collection also contains any notes which might be relevant to the decision or be interesting to future researchers. Only articles labelled as philosophy and for which the investigator could determine authorship were labelled with an 'NZ' label.

Philosophy articles, which make up a very small proportion of the whole dataset, were found by a few methods. First, some of the articles found in initial exploration of the dataset were added. Second, keyword searches using the results of the more formal exploration set out in the previous section were used. For instance, keyword searching for interesting collocates and co-occurring terms of 'philosophy'. This second step found many non-philosophy articles as well. In addition, a random sample of the processed dataset was included, to ensure a broad range of non-philosophy articles were included.

The distribution of labels in the labelled collection is presented in Table 4.[27] Note that it is harder to find philosophy articles than non-philosophy articles even when using keyword searches and the exploratory methods set out above. Most articles labelled were readable, and the various writing types were well represented. As noted above, the epistemology-metaphysics label was not used much. Material focused on religion is very prominent in the dataset in general and in this labelled collection.

2.1.4 Naive Bayes Classification

Having labelled a set of articles, a classification algorithm was trained. The only label used for the purpose of creating a corpus of philosophical discourse was the 'philosophy' label. In both of the iterations of the process, the labelled collection was divided into testing and training sets.

[26] This is discussed again in Section 2.3 and Section 4.

[27] The version of the labelled dataset used for the first iteration is stored as a pickled Pandas dataframe at 'pickles/classifier_with_text_nb1.pickle'. The version of the dataset used for the second iteration is contained in 'classified_with_text_df.pickle'.

	First Iteration		Second Iteration	
Parameter	Range	Selected	Range	Selected
Minimum documents (count)	5-12	7	[2, 5, 7, 10, 15]	2
Maximum documents (prop)	0.2-0.5	0.2	0.2-0.6	0.4
N-gram range	(1,1)-(1,3)	(1, 1)	(1,1)-(1,3)	(1, 1)
TF-IDF	True/False	True	True/False	True
Stopwords	incl. 'philosophy' (True/False)	False	N/A	N/A
Smoothing parameter	N/A	N/A	[0.5, 0.75, 0.1]	0.5

Table 5: Cross Validation Parameter Space and Selected Value.

In the first run, the philosophy articles were divided with 75 in the training set and 26 in the test set.[28] The training set contained 75 of the non-philosophy articles and test set contained the remainder of the non-phliosophy articles. This was done in order to ensure a 50/50 class balance in the training set, while the higher proportion of non-philosophy in the test set more closely matches the real situation in the dataset as a whole.[29]

In the second iteration of the process, the training set was assigned 75% of each of the philosophy and non-philosophy articles (222 and 465 articles respectively).[30] The remainder of articles were assigned to the test set. To achieve a closer class balance in the training set, random upsampling was applied to the philosophy articles. This was done by randomly doubling a sample of between 20 and 50 philosophy articles, and repeating the process until there were 407 articles assigned to philosophy in the training set.[31]

The Scikit-learn package was used to implement classification algorithms [21]. The pipeline takes the text and converts it into a bag of words representation, implements a TF-IDF transformation, and then feeds the result into a Naive Bayes classifier.

The Naive Bayes classifier was adopted on the basis that it is a simple and easy to train model which performs surprisingly well at text classification [29]. The naive in 'Naive Bayes' indicates that the algorithm treats each of the features fed to it as probabilistically independent. This is obviously not true of our dataset in reality. The appearance of a given word will definitely affect the probabilities of the other words in the dictionary appearing in any case of real writing.[32] This was a particularly appropriate choice given the lack of high-quality sequence data noted above. Naive Bayes operates on a bag of words representation of the document.

In order to determine the optimal parameters for this pipeline, a grid cross validation approach was used. The parameter space considered for each iteration of the process is presented in Table 5, along with the parameters values chosen. In both cases, the training set was divided into five for cross validation. Overall accuracy was used as the cross-validation metric. Precision and recall were also tried as metrics, but balancing the two with overall accuracy seemed to work best. The model selected using overall accuracy does not suffer from either low recall or precision. This will be further discussed in the results section (§**??**).

The majority of the parameters selected for the purpose of the search concern the preprocessing of the articles fed to the classifier. The minimum number of documents containing a given word and the maximum proportion of documents containing it both control the words which

[28]The code for the first iteration is found in the notebook 'NaiveBayes_PhilosophyClassification.ipynb'.

[29]In fact, the test set massively over-estimates the amount of philosophical discourse in the dataset as a whole. For this reason, it was not considered important to follow the usual practice of handling class balance *after* the split between testing and training data.

[30]Code for the second set is found in the notebook 'NaiveBayes_PhilosophyClassification_SecondSet.ipynb'.

[31]The increase in the proportion of non-philosophy in the labelled data in the second iteration meant that it was more convenient to use the usual approach of first dividing the test and training before dealing with class balance.

[32]Strictly, it assumes all features are independent conditional on the class to which the object is assigned. That is, that, e.g., the probability of the appearance of 'Darwinism' given that the article is a philosophy article is independent of the probability that 'evolution' appears given that the article is a philosophy article. This is similarly implausible.

are included in the dictionary used by the classifier. The n-gram range allows for collections of words of length up to three to be considered by the classifier. The TF-IDF transformation has been noted above. It controls for the length of documents and the prevalence of words in the corpus as a whole. In the first iteration, the addition of 'philosophy' to the stop words was considered, as the 'philoso*' corpus was thought to over-emphasise the word. This was abandoned at the second iteration.

Three values of smoothing parameter were considered in the second iteration. A smoothing parameter adds a 'pseudocount' to each of the words in the dictionary to account for the difference between the particular distribution of words in the articles given to the classifier and the real distribution. For instance, it may be that no articles labelled as philosophy in the training data contain the word 'dog'. We would not want to conclude from this that the probability of an article containing 'dog' being philosophy is zero. Consequently, a 'smoothing' value is added to the count to ensure that the probability of any word appearing in a given class is non-zero. The smaller this value is, the closer we are taking the distribution of the articles in the dataset and in the real world to be. One risk is that, if we set this too low, the classifier will over fit the training data. Consequently, values below 0.5 were not allowed.

The use of a Support Vector classifier rather than Naive Bayes was considered, but no improvement on the Naive Bayes approach was found.[33]

In order to 'bootstrap', once a classifier was selected, it was run against the entire dataset, and the resulting dataframe saved as a pickle. At this point, another round of corpus exploration and labelling could be carried out. The motivation for this is that the labelling the non-philosophy picked out from the dataset by an earlier classifier enables later classifiers to more accurately select the relevant material from the dataset.[34]

2.2 Implementation

This section contains some brief notes on the implementation of the above methods.

2.2.1 Hardware and Multiprocessing

The project has been almost entirely carried out on a workstation PC with a AMD Ryzen Threadripper 1950X 16-Core Processor with 32 threads and 32GB of RAM. This setup conditioned the code written for the project in a two main ways.

Firstly, the Python multiprocessing module was used to enable more than one thread to be used at once. This was particularly important for operations involving the whole dataset. Being able to open and process 16 tarballs at once was significantly more efficient than opening one at a time.

Secondly, limitations in RAM required the dataset to be sliced into very small subsets to enable models to be fit to the whole corpus efficiently. Around 70 slices were used, and stored on an SSD to enable faster I/O. This allowed for multiple cores to fit a trained model without having to open up multiple very large slices of the dataset. Similarly, the corpus produced by the first classifier (around 200,000 documents) needed to be reformatted to allow it to be run through a topic modelling algorithm without overrunning the available RAM. This was done by storing the bag of words representation of each document in a '.csv' file on the SSD and

[33] This is further discussed in Section 2.3.1.

[34] An attempt to produce a 'readability' classifier to run before classifying between philosophy and non-philosophy. This is present in 'NaiveBayes_PhilosophyClassification_SecondSet.ipynb'. Good performance was not attained. In practice, almost all non-readable articles are classifier as non-philosophy by the philosophy classifier.

		Predicted	
		False	True
Actual	False	88	9
	True	3	23

Table 6: Confusion Matrix for First Naive Bayes Classifier

feading these documents one at a time to the algorithm without loading the whole corpus into RAM.[35]

2.2.2 From METS/ALTO to Pandas

The tarballs for two years of the *Lyttelton Times*, 1890 and 1891, were corrupted and have been excluded. This is not important for the humanities questions investigated in the following section, but would be if one were trying to compare the number of philosophical articles in different newspapers in aggregate or over time.

A small number of individual issues of newspapers were also excluded on the basis of problems with parsing their ALTO/METS XML. This problem only appeared for a handful of issues, and so has been ignored. Manual inspection also suggests that some articles have also got missing blocks. This is unlikely to be important in the aggregate.

2.2.3 Dashboard Construction: Dash and Heroku

The dashboard available at `https://nz-newspaper-philosophy.herokuapp.com` is hosted on their 'hobby' service and lacks the power to directly compute anything from the corpora available to it. Consequently, precomputed results for various search terms have been added in the form of small dataframes.[36]

2.3 Results

There are two ways of evaluating the corpus construction process. One is simply to look at the classification metrics. The other is to inspect the result of applying the classifier to the dataset as a whole in terms of the research interest of someone interested in philosophy as it appears in early New Zealand newspapers. Each will be considered in turn.

2.3.1 Classification Metrics

Both the first and second classifier have very similar overall accuracy on their respective test sets. With 0.90 for the first classifier and 0.89 for the second classifier. Confusion matrices for these classifiers are presented by Table 6 and Table 7.

This may make it sound like future iterations made the performance of the classifier worse. However, a straightforward comparison between the numbers is not appropriate. The second classifier had many more non-philosophy articles in its test set than the first. Moreover, it was given much finer distinctions than the first classifier. That is, the second classifier had a lot more non-philosophy which looked similar to the desired philosophy articles than did the first classifier. This is because of the 'bootstrapping' approach, where the NB1 corpus was used a

[35] See, especially, the class 'NL_streamed_corpus' in the file 'NL_topicmodels.py'

[36] A Dash cytoscape interface which enables the user to enter their own search terms is available in 'Religion and Evolution in the Rel corpus.ipynb'. This would be run on the users own computer, rather than on the Heroku server.

		Predicted	
		False	True
Actual	False	181	14
	True	15	62

Table 7: Confusion Matrix for Second Naive Bayes Classifier

Category	Count
Composite pieces	4
Ethics (correctly labelled)	7
Ethics (mislabelled)	3
Other (correctly labelled)	1
Other (mislabelled)	0
Total	15

Table 8: Categorised false negatives for second Naive Bayes classifier.

source of new articles for labelling. Given this, many of the non-philosophy articles given to the second classifier were articles that were labelled as philosophy by the first.

Overall accuracy is not the only relevant metric for a binary classifier. The confusion matrices for the two classifiers suggest that a reasonable balance between recall and precision is being achieved. In the case of the first Naive Bayes classifier, we have a recall of 0.88, and a precision of 0.71. In the case of the second Naive Bayes classifier we have a recall of 0.80, and a precision of 0.81. These figures show, at least, that accuracy is not being attained at the expense of either precision or recall.

Another way to consider the difference is to look at the size of the resulting corpora. Applying the first classifier results in a corpus with around 200,000 articles, whereas the applying the second results in a corpus with around 30,000 articles (see Table 1). This shows that the second classifier is at least picking out a smaller proportion of articles to tag as philosophy. However, to determine whether the articles picked are actually the articles which are desired, it will be necessary to turn to more direct inspection of the resulting corpus.

2.3.2 Manual Inspection of False Positives and Negatives

There are two aspects of the classifiers performance which we can check. First, we can consider the false positives and negatives when the classifier is applied to the test set. Second, we can consider the composition of the corpus which results from applying the classifier to the whole dataset.

The false positives and false negatives picked out by the second Naive Bayes classifier will reveal how the classifier is erring on the labelled collection and reveal what kind of articles we should expect to get into the corpus generated by applying the classifier to the entire dataset.[37]

Table 8 presents the false negatives when the second classifier was applied to the test set.[38] Note the presence of four composite pieces. These are pieces where only a small section is 'philosophical'. Note also that the majority of the false positives come from the 'ethics' category, which covers both personal ethics and political issues. Three of these can, on reflection, be seen to be mislabelled. Usually, this is because they only make a passing reference to philosophy, rather than engaging in any philosophical discourse.[39]

[37] We could also consider those articles which are misclassified in the training set. However, for the purposes of getting an idea of the general kind of articles that will be misclassified, the test set should be sufficient.

[38] See Table 1 for a full list.

[39] DTN_18940820_ARTICLE7, for instance, mentions the philosopher John Stuart Mill in passing, but is not otherwise

Category	Count
Mislabelled (religious)	3
Correctly labelled (religious)	5
Edge case (religious)	2
Correctly labelled (other)	3
Unreadable	1
Total	14

Table 9: Categorised false positives for second Naive Bayes classifier.

The false negative results suggest that the classifier is struggling to capture ethics and politics, but not struggling with religious material. They also highlight shortcomings in the labelling scheme adopted in this project. Section 4 will discuss possible improvements to the handling of both composite pieces and the ethics category.

Table 9 presents the false positives when the second classifier was applied to the test set.[40] Where the false negatives are dominated by ethics and politics, the false positives are dominated by religious material. This indicates that the classifier is tending to pick out the desired religious material much better than the desired ethical material. Even so, some cases are, on reflection, either difficult to say whether they ought to have been labelled as philosophy or probably should have been labelled as philosophy.[41] This, as in the case of the false negatives, indicates that the problems for the classifier are in regions where the human labelling of the dataset is also slightly ambiguous. Of the non-religious material, we have a bit of materials science, a list of graduates of Canterbury College and their subjects, and a quack advertisement concerning the 'philosophers stone'. In the latter case, the use of philosophical jargon explains how it gets misclassified, while the other two cases are at least *intellectual* material if they are not philosophical.

2.3.3 Inspection of Resulting Corpora

Looking to the corpora produced by running the Naive Bayes classifiers on the complete dataset encourages the idea that a sensible subset of the dataset is being picked out.[42] For the purpose of this report, a useful way to summarise the content of each corpus is the word cloud.

Figures 3 to 6 display word clouds for a subset of 10000 articles of the original dataset, the 'philoso*' corpus, the corpus resulting from the first Naive Bayes classifier, and the corpus resulting from the second Naive Bayes classifier. Each of these word clouds was generated using the TF-IDF transformation and with dictionaries excluding any word which appeared in more than 20% of the documents and less than 50 documents in total.[43]

We see in Figure **??**, that no particularly 'philosophical' terms are present. The majority are place names, notably Wellington and London, along with terms for reporting on government, business and transportation. Figure 4 looks quite different, although place names are still prominent. It is particularly interesting that 'women' and 'miss' become very prominent. The prominence of God is also striking, and is in line with the observation that religious matters are

philosophical.

[40] See Table 2 for a full list.

[41] One of the edge cases is a report of a public meeting of a group called the White Cross Society, dedicated to improving local moral standard. There are, in the course of a long report, *some* gestures towards the 'ultimate' reasons for adopting their preferred moral framework. This will be discussed again when we return to the question of labelling (§4).

[42] Random samples of 500 articles from each of the corpora derived in this project can be explored at `https://nz-newspaper-philosophy.herokuapp.com/`, using the 'View Texts' tab.

[43] Word clouds were produced using the Python 'wordcloud' package [15]. See 'generate_word_clouds.py'.

Figure 3: Word cloud for random sample of full dataset.

closely intertwined with philosophy in this corpus in general.

The shift to the NB1 corpus (Figure 5) increases the prominence of language about education, we also see words like 'judgement' and 'evidence'. This indicates the selection of intellectual content from the dataset as a whole. Finally, we see in Figure 6, that education and religion maintain prominence, while words like 'lecture' and 'professor' start to come through.

This way of summarising the content of the corpora suggests that the methods deployed here are successfully picking out a meaningful subset of the original dataset which is of intellectual interest.

A slightly more sophisticated way of summarising the material contained in the various corpora we have considered is to run Latent Dirichlet Allocation (LDA) topic models on each.[44] Table 10 presents the results of applying a model with ten topics to the 'philoso*', NB1 and NB2 corpora.

There are a few things worth pointing out about the results in Table 10. Note that the second iteration of the corpus building process seems to have reduced the prominence of legal deliberations. That is, topic two in column one and topic one in column two have no analogue in column three. Note also that the topics in column three show clear evidence of multiple kinds of scientific material, with topic nine indicating discussion of physics and astronomy and topic four suggesting biological and Darwinian material. There is not clear political or ethical topic in the NB2 column. This further motivates worries about how the labelling has worked to emphasise religious material over ethical and political material.[45]

Finally, it is worth noting that, in the second part of this project, a corpus of material about the relationship between religion and the modern sciences is derived by training a Naive Bayes classifier on the 'philosophy type' label and then applying it to the philosophy corpus. To test the comprehensiveness of the religion and the modern sciences corpus, an external validation step is carried out (§3.2.2). This is done by listing the newspaper articles cited in current research on religion and science in early New Zealand and checking whether they are selected by the methods used in this project. The success of this external validation step provides further

[44] See 'subset_topicmodels.py' for details of these models.

[45] Topic eight indicates *some* interest in ethics. It may be that a model with more topics would bring out the ethical content more.

Figure 4: Word cloud for random sample of 'philoso*' corpus.

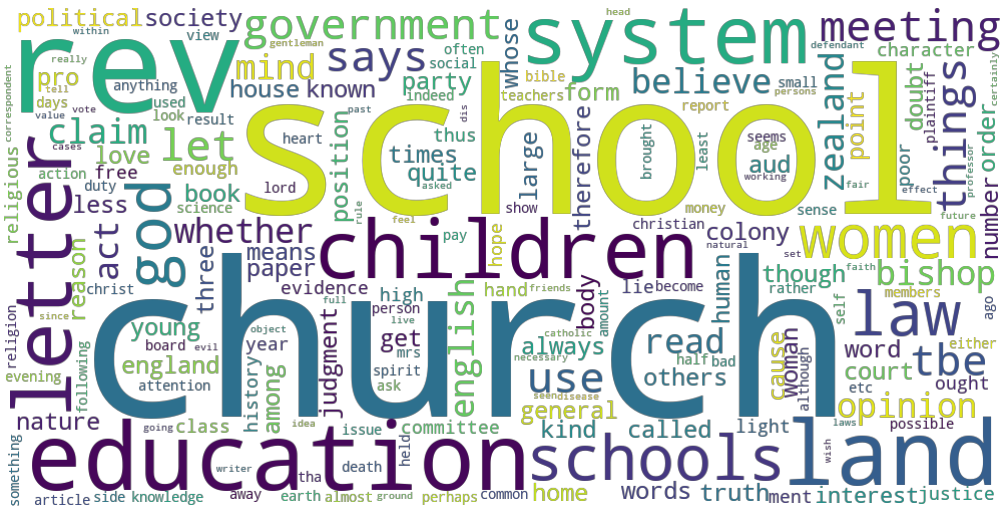

Figure 5: Word cloud for random sample of the first Naive Bayes corpus.

	'Philoso*'	NB1	NB2
1	young, think, mrs, lady, came, eyes, thought, woman, back, miss	claim, judgment, plaintiff, land, defendant, costs, amount, case, sheep, pay	lecture, lecturer, subject, professor, evening, last, christianity, audience, might, lectures
2	lie, case, court, evidence, tin, law, judge, tlie, prisoner, jury	case, letter, sir, editor, say, said, law, question, matter, could	science, nature, tho, evolution, human, thought, religion, matter, natural, scientific
3	english, book, lord, women, love, though, whose, god, character, always	work, book, good, many, first, english, new, science, much, knowledge	christ, jesus, shall, faith, spirit, lord, say, day, know, things
4	school, year, university, education, schools, class, 000, children, best, college	disease, use, life, food, health, many, body, cure, best, every	animals, animal, darwin, species, found, theory, years, two, professor, well
5	disease, health, synod, cure, medical, remedy, medicine, pain, bottle, body	people, england, country, english, world, new, even, power, years, every	tho, church, christian, catholic, faith, churches, religion, christianity, christ, aro
6	water, 000, miles, feet, gold, french, small, days, four, london	school, education, schools, children, system, committee, teachers, present, board, state	spirits, spiritualism, spirit, mind, power, phenomena, body, science, matter, subject
7	science, god, church, nature, human, subject, education, mind, knowledge, religion	church, rev, god, bishop, christ, christian, said, sunday, bible, religion	bible, book, years, history, old, new, tho, work, testament, professor
8	meeting, evening, held, messrs, company, following, year, committee, board, zealand	tho, tha, light, tbe, two, earth, aud, aro, havo, found	love, human, good, religion, every, christ, evil, christian, spirit, power
9	government, land, colony, sir, house, question, zealand, bill, political, state	like, life, said, know, good, never, little, say, day, old	earth, light, sun, heat, matter, theory, water, professor, bodies, stars
10	tbe, aro, havo, club, tha, evening, whioh, night, tlio, next	people, must, new, public, country, much, many, land, good, government	say, bible, church, letter, christian, religion, truth, sir, think, day

Table 10: Ten-topic models for each corpus, with topics summarised by their ten most prominent words.

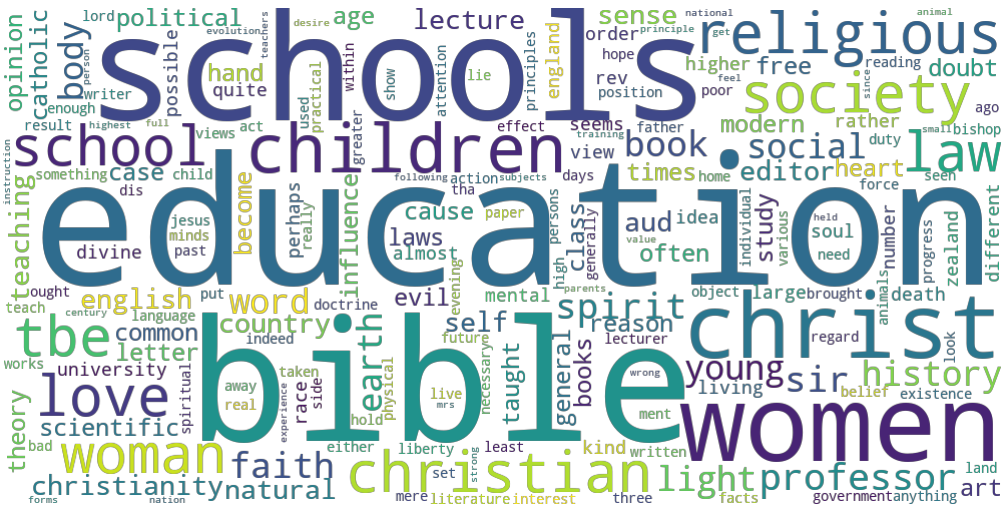

Figure 6: Word cloud for random sample of the final Naive Bayes corpus.

evidence that the methods used in this project are at least picking out a meaningful subset of the desired articles.

All of the evidence from inspection of the corpora produced in this phase of the project suggests that an interesting subset of the original dataset containing material interesting to the historian of philosophy has been generated. If any problem has emerged, it seems that religious matters may be being overemphasised at the expense of political and ethical matters. This will be discussed again in Section 4.

2.3.4 Alternative Classifiers

As noted above the Naive Bayes classifier is a very simple classifier. A more sophisticated method, Support Vector classifiers, was also considered as a replacement for the second Naive Bayes classifier. Like the Naive Bayes classifiers, the support vector classifier creates a decision boundary in the parameter space, in our case the parameter space has as many dimensions as there are words in the dictionary given to the classifier. The support vector classifier draws the boundary by maximising a margin between the classes while given a certain 'budget' for error (the 'tuning parameter'). Support vector classifiers can also be extended by the use of non-linear kernels, which allow for non-linear decision boundaries to be learnt. This is not possible for the Naive Bayes classifier. The kernel can be thought of as a notion of the distance between two observations in an extended space of parameters. For instance, the polynomial kernel uses polynomials of a specified degree for each dimension of the parameter space [46]

A cross validation search method was applied using the same methods as used for the Naive Bayes classifier. Table 11 presents the parameter space and selected values for the classifier. Note that, despite allowing for a non-linear decision boundary, the kernel chosen by the cross validation search was linear. The grid search was carried out using overall accuracy as a metric.

Table 12 presents a confusion matrix for the selected classifier when run on the test set. Note that the classifier is tending towards more false negatives than false positives. The overall

[46] For detail see **[islr]**. Note that Scikit-learn uses 'Support Vector Classifier' to refer to the whole class of linear and non-linear kernels, whereas **[islr]** uses is to refer only to the linear case, using 'Support Vector Machines' for the more general case.

Parameter	Range	Selected
Minimum documents (count)	[2, 5, 7, 10, 15]	2
Maximum documents (prop) 0.2-0.6	0.6	
N-gram range	(1,1)-(1,3)	(1, 2)
Kernel	['poly', 'linear', 'sigmoid', 'rbf']	'linear'
Tuning parameter	[0.5, 0.75, 1, 2, 10]	1

Table 11: Cross validation parameter space and selected values for support vector classifier.

		Predicted	
		False	True
Actual	False	186	9
	True	21	56

Table 12: Confusion Matrix for First Naive Bayes Classifier

accuracy is 0.89, which is about equivalent to the Naive Bayes classifiers trained earlier. The recall and precision are 0.73 and 0.86, respectively. It is likely that these could be balanced more effectively by further refining the classifier. However, it seems unlikely that any greater performance than the Naive Bayes is likely to be achieved. One possible explanation for this is that the shortcomings of the labelling simply make further accuracy unobtainable.[47]

3 Two Questions about Philosophy in New Zealand

We now turn from constructing a corpus for investigating philosophical discourse in New Zealand newspapers to actually using the corpus. The corpus could be used to investigate many specific questions. Recall that, for the purpose of this project, our focus will be on two questions:

1. Can the resulting data set be used to provide insight into how the relationship between religious belief and developments in the natural sciences was understood in early New Zealand?

2. Can topic modelling and co-occurrence analysis on the resulting corpus reveal that philosophical writing in New Zealand newspapers incorporates concerns which are specific to New Zealand?

Figure 6 already suggests that the NB2 corpus contains a lot of material on the status of religion with respect to the natural sciences. It also suggests a plausible option for investigating New Zealand specific content in this corpus: namely, debates over what kind of education system ought to be set up in New Zealand.

This section consists of a discussion of the methods applied for answering these two questions (§3.2) and of their results (§**??**).

3.1 Methodology

3.1.1 Evolution and Religious Belief

The first step towards answering the first question is the further specification of the NB2 corpus using the 'philosophy type' labelling discussed in (§2.1.3) and classification with a further

[47] This will be discussed in more detail later 4.

Religion-Science	**Training (75%)**	**Test** (25%)
True	102	37
False	120	40

Table 13: Class balance of test and training sets for 'religion-science' classifier.

Parameter	Range	Selected
Minimum documents (count)	[2, 5, 7, 10]	5
Maximum documents (prop)	[0.3, 0.4, 0.5]	0.4
N-gram range	(1,1)-(1,3)	(1, 1)
TF-IDF	True/False	True
Smoothing parameter	[0.3, 0.4, 0.5, 0.75]	0.3

Table 14: Cross validation parameter space and selection values for religion-science classifier.

Naive Bayes classifier as in (§2.1.4). Having done this, the corpus exploration methods, of concordancing, collocations, and co-occurrences will be deployed to detect the presence of the main attitudes to evolution sketched out in previous studies the relationship between religion and science. Insofar as these attitudes are detectable or not in the resulting corpus, we have gained some insight into this philosophical topic as it appears in early New Zealand newspapers.

The philosophy type labels come in four categories, which are first transformed into a binary label representing whether an article is tagged with 'religion-science' or is not.[48] Since the labelling is hierarchical, and we are aiming to apply to classifier to the NB2 corpus, we exclude all non-philosophy articles from the labelled dataset.

Having set up this label, we take 75% of the remaining data as training data and 25% as test data. The class balance of the resulting datasets is presented by Table 13. As the classes are roughly even, no class balancing was attempted.

A grid cross-validation search was then applied to select parameters for the classifier (Table 14). Having trained the classifier, it was run on the NB2 corpus and the documents classified as religion-science were stored as the Rel corpus.

No further discussion of the technical details of the methods applied to explore the Rel corpus is necessary here, as these details have been covered in Section 2.1.2.[49] On the humanities side, [**dupree**] and [**gregory**] were used as sources for understanding the material present in the corpus. In particular, their discussion of the broad categories of reaction to evolutionary ideas in scientific and religious circles, and their insistence that those circles were not mutually exclusive. Relevant detail will be given in the results section.

3.1.2 New Zealand Content in the NB2 Corpus

By contrast with Question Two, where there is a clear sense of the kind of discourse we are looking for and where this kind of discourse was built in in the labelling phase, the third question more open-ended and exploratory. As such, an unsupervised method is appropriate. Such methods may reveal content in the NB2 corpus which we were not already aware of. Having found some relevant group of content, the same exploratory methods set out in Section 2.1.2 can again be applied to get closer to the actual texts.

Topic models were used as a convenient unsupervised learning method. These models were trained using the Gensim library.[50] This was done using the 'LdaMulticore' class, modifying

[48] Code for the training of this classifier is available in 'NaiveBayes_Rel_Classification.ipynb'.
[49] See 'Religion and Evolution in the REL corpus.ipynb'.
[50] See 'subset_topicmodels.py'

		Predicted	
		False	True
Actual	False	34	6
	True	6	31

Table 15: Confusion matrix for religion-science classifier.

Figure 7: Word cloud for random sample of the Rel corpus.

only the number of topics parameter. As this is merely exploratory, and no attempt will be made to claim that all of the relevant articles of a certain class are being picked out, there is no need to be sensitive to the details of the parameters of the model.[51]

Having trained these models, the 100 topic model was selected as most convenient to manually inspect. Two topics of interest were picked out by their top-10 keywords, one which contained words in te reo Māori and the other of which contained discussion of the education system and secularism.[52] These topics were then explored to determine the nature of their content.

3.2 Results

3.2.1 Evolution and Religious Belief

Table 15 presents the confusion matrix when the religion-science classifier is applied to the test set. We see an even split between false negatives and false positives, with an overall accuracy of 0.84, and a recall and precision of 0.83. This is roughly in line with the performance of the previously trained classifiers.

Having applied the classifier to the NB2 corpora, we generate a word cloud for the resulting Rel corpus (Figure 7). The indicates some of the main genres of writing we are interested in: public lectures, letters, and perhaps sermons; some of the people that we are interested in: professors, clergy, public lecturers; and some key topics: evolution, creation, and spirituality.

[51] This differs from the use of topic models in [3], where having too many topics would reduce the usefulness of the results as a classification to track developments in the discipline of academic philosophy over time.

[52] Code for this stage of the project is found in 'NZ content.ipynb'.

1	e than the details of mr darwin but this ceaseless	warfare of	the champions of that faith which we at least h
2	ws during his recent visit to america r review the	warfare of	science by andrew dickson whitb ll d president
3	re at the pres byterian church last evening on the	warfare of	science and religion there was a moderately goo
4	in his nbvum organum as quoted by dr white in his	warfare of	science a treatise i commend to the serious att
5	t between science and keligion nor dr white in his	warfare of	science for one moment assumes tbat there is no
6	l continue to make men wiser and purer dr draper's	conflict of	science and religion is a very interesting book
7	untested opinions dr drapers con flict is often a	conflict of	science with science of science adopted by reli
8	ittle as balaam's ass did of hebrew that was not a	conflict of	religion with science but a conflict with nesci
9	ion that there are gratifying signs of decreasing	antagonism between	the church and science while at the same
10	ant interpretations of the bible rest the fancied	antagonism between	scientific progress and religious truth a
11	in papers i have not in my lectures presented any	antagonism between	genesis and geology unless a fair present
12	perfectly true to say that there is and can be no	antagonism between	the church and science whilo at the same
13	extending over an indefinite period the apparent	antagonism between	science and religion can easily be bridge
14	ny dogmatic uit rances now is there any necessary	antagonism between	scriptural teaching and the evolution the
15	lecture with a view to reconciling the theory of	evolution with	the jewish account of the creation and fall
16	idable nature as to the supposed inconsistency of	evolution with	the christian belief as to the person of chr
17	the compatibility or otherwise of the doctrine of	evolution with	religious belief professor salmoud as it see

Table 16: Selected results from 'warfare of', 'conflict of', 'antagonism between' and 'evolution with' concordances.

It is argued by [**moore**] that, what is terms the 'Baconian compromise' between science and religion, according to which the interests of science and religion were distinct, but in harmony with one another, began to lose its hold on many thinkers in the nineteenth century. One reason for this was the development of geological ideas according to which the earth was much older than a certain way of reading of the Bible had suggested, and the introduction of evolutionary theories according to which the various species were not created independently, but evolved from a common ancestor. One manifestation of this was an extreme 'conflict' or 'warfare' view, according to which science and religion are inherently opposed to one another, while others attempted to find new ways of harmonising the two [**dupree**].

Both tendencies are present in concordancing results from the Rel corpus (Table 16).[53] Each of the concordance lines seems to address the antagonism or lack of antagonism between religion and science quite clearly. We see this in the case of evolution specifically in lines (1), (15), (16), and (17).[54] We see discussion in different social contexts, including at churches (3) and public lectures (11). Major international figures are discussed, including Draper and White in(2), (4), (5), (6), and (7). Finally, it is worth noting the name of Salmond, an early New Zealand academic philosophy, who appears in (17) in conn4ection with these issues.

Some interesting co-occurrence networks can also be considered.[55] Four will be briefly discussed here.[56]

Figure 8 presents co-occurrence networks for 'evolution' and 'genesis'. One term comes from the side of the sciences and the other from the side of religion. The network for 'evolution' indicates the interaction with religious and metaphysical issues of creation and of biblical interpretation insofar as it contains 'genesis', 'creation', and 'doctrine'. We also see some of the names which are associated with evolution in the corpus, including, obviously Darwin, but also his advocate Huxley, the popular philosopher Spencer, who is responsible for the phrase 'survival of the fittest' and the local figures Salmond and Parker.[57] The network for 'genesis' on the other hand, highlights some of the important questions of biblical interpretation brought on by

[53]Note that a function was written for the project in order to allow for multi-word concordancing (see 'concordance phrase') in the notebook.

[54]We also see the closely related issue of the interpretation of biblical book of Genesis in light of geology in (11)

[55]The results of collocation analysis are excluded for reasons of space. However, some interesting results are present in 'Religion and Evolution in the REL corpus.ipynb'.

[56]More such networks can be produced using the project dashboard, linked above or 'Religion and Evolution in the REL corpus.ipynb'.

[57]Parker was a popular public lecturer and academic at Otago, himself mentored by Huxley.

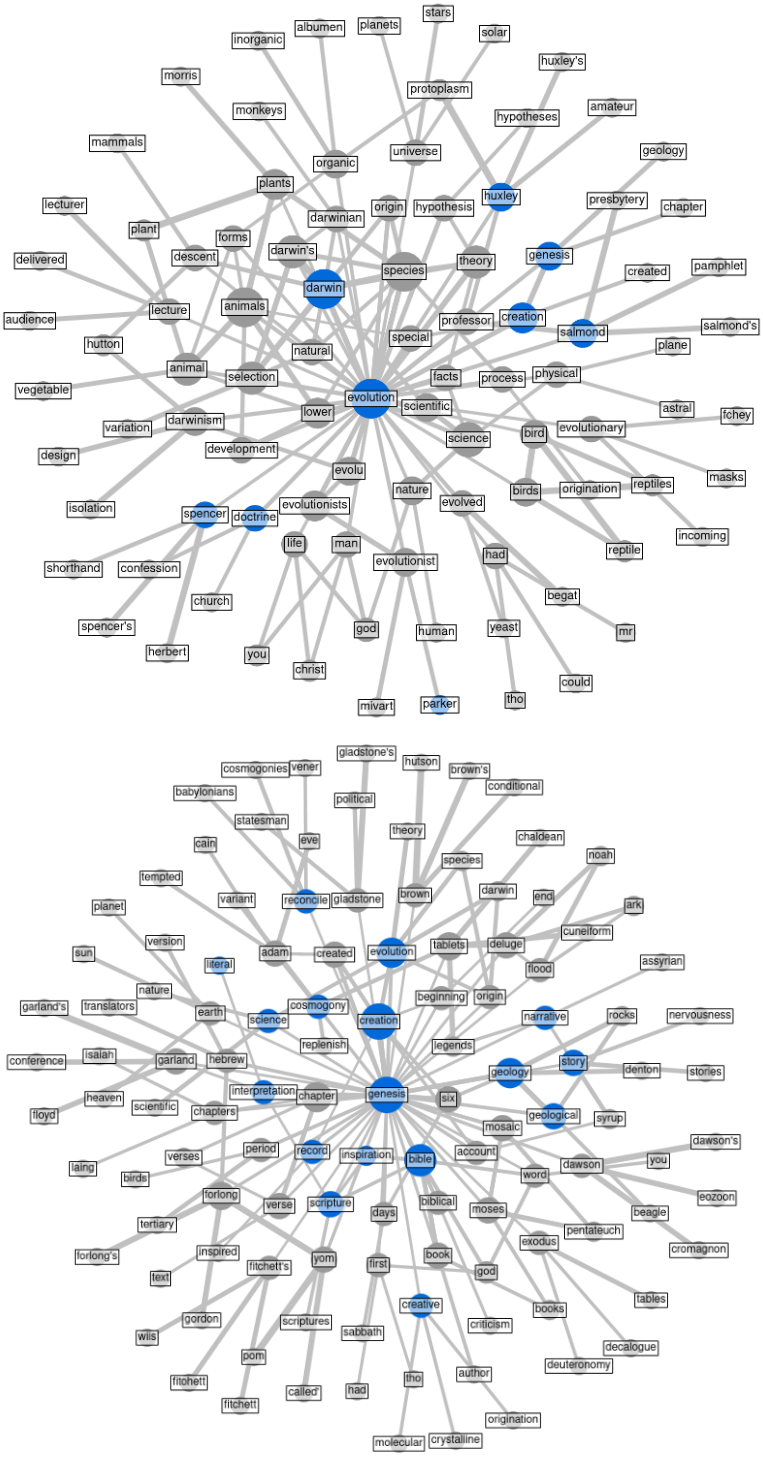

Figure 8: Co-occurrence networks for 'evolution' and 'genesis' in the Rel corpus, with TF-IDF transformation and log Dice statistic. The top 45 co-occurrences for each are shown along with the four top co-occurrences for each of these. Highlighted nodes are to aid discussion.

Membership	Category	Count
In Rel		19
Not in Rel	Composite	8
	Too short	2
	Parsing error	2
	Unexplained	1
Total		32

Table 17: Articles checked for external validation of Rel corpus.

contact with evolutionary ideas. We again see 'evolution' appear, along with 'geology', but also various interpretive options including 'story', 'literal', and 'record'. The inclusion of the word 'reconcile' again suggests some interest in how religion and the results of the natural sciences might be brought together.

Figure 9 reveals some of the background against which the discussions of evolution and religion were being held. In particular, the network for 'human' shows a tight set of interrelationships between spiritual and religious terminology, with the exception of 'animal' and 'science'. There is less evidence in this network of contestation than in the other networks. The network for 'materialism' highlights some of the various alternatives to traditional religious belief that are being discussed in the newspapers at this time and within with evolution would be interpreted. For instance, we see 'positivism', 'atheism', and 'scepticism'. We also see 'theosophy', a spiritual movement which was much discussed at the time.[58]

3.2.2 External Validation of the Rel Corpus

At this point it is convenient to include a further step of corpus evaluation for the Rel corpus which also applies to the NB1 and NB2 corpora. In the course of this project, a few philosophical pieces on the relationship between religion and science were picked out, but not labelled or otherwise used in the process of corpus construction. In addition, a series of articles about philosophically interesting scientific and religious developments which cite early New Zealand newspaper content were also examined. In order to test the process which led up from the raw Papers Past data to the Rel corpus, we can examine how many of these articles end up being selected by our methods and what has happened to the ones who have not.

Three articles have been used for this step. [2] offers an account of colonial intellectual life as present in Otago newspapers. In the course of his discussion, he provides a footnote with 15 references to discussions of the value and significance of evolutionary ideas in the Otago newspapers [2, fn. 53]. The other two papers are much more specific, dealing with the public lectures of Professor Parker, an Otago academic, protegee of Huxley, and public lecturer on biological topics [5] and the debates over a pamphlet entitled *The Reign of Grace* by William Salmond, the second philosophy professor employed in New Zealand [**wood-2014**]. From these articles, ten newspaper mentions of Parker's lectures and six mentions of reports relevant to *The Reign of Grace* are collected. It is expected that the majority of these newspaper articles should be present in the NB2 and Rel corpora.[59]

Table 3.2.2 summarises the result of checking for the presence of each of these articles, and those which were held out from the labelled data set, in the Rel corpus.[60] We see that, the major-

[58]The term 'besant' also appears, for Annie Besant, a travelling theosophical lecturer who toured New Zealand.

[59]Since the Rel corpus is derived by running a classifier on the NB2 corpus, anything in the Rel corpus must also be in the NB2 corpus.

[60]Additional tables are presented in the Appendix for each source in tables 3 to 6.

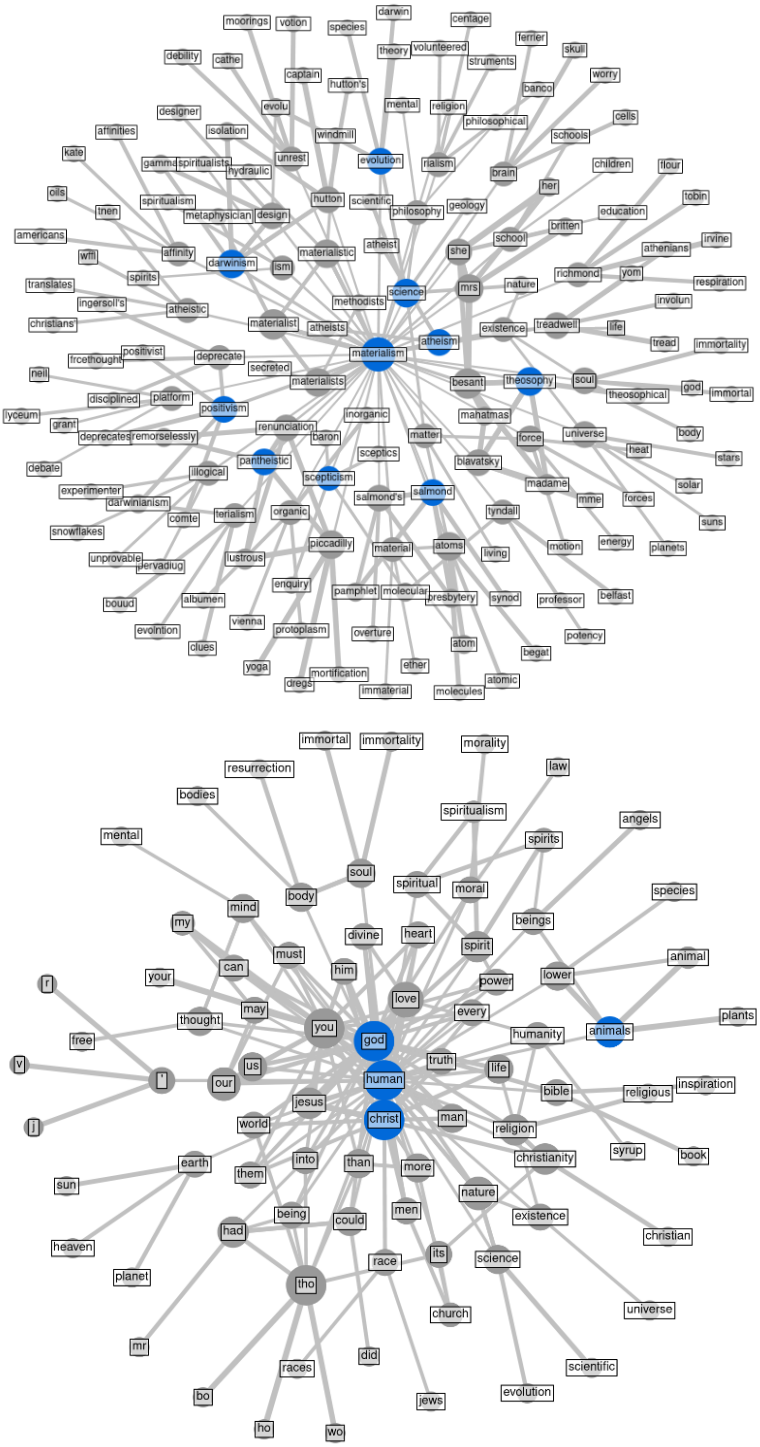

Figure 9: Co-occurrence networks for 'materialism' and 'human' in the Rel corpus, with TF-IDF transformation and log Dice statistic. The top 45 co-occurrences for 'materialism' and the top 50 for 'human' are show, along with the four top co-occurrences for each of these. Highlighted nodes are to aid discussion.

Topic 90	hipi, toa, tohunga, awa, hymns, pole, ahi, ain't, loafer, like.
Topic 93	schools, education, children, state, system, religious, bible, school, secular, religion

Table 18: Two topics of interest from 100-topic model of Rel corpus.

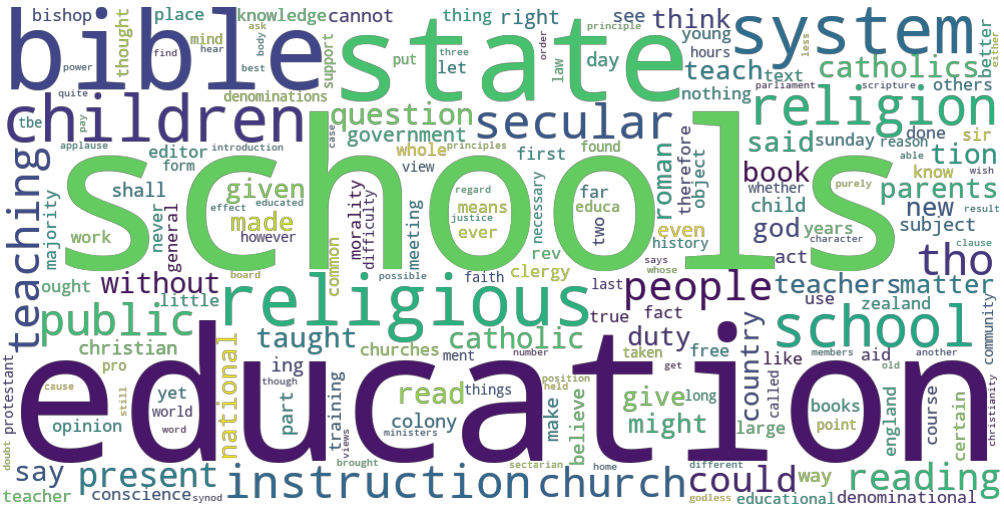

Figure 10: Word cloud for documents with high proportion of Topic 93 in the Rel corpus.

ity are contained in the Rel corpus. This suggests that we are capturing a fairly comprehensive range of the articles about the relationship between philosophy and religion.

Of those articles which are not present in the corpus, the majority are 'composite' pieces, either editorials or reports of meetings covering many topics. The shortcomings of the approach to corpus construction have been noted above and will be discussed again in Section 4. In terms of the comprehensiveness of the Rel corpus, we should worry that we are missing editorial comment, which might take different positions than those of, say, public lecturers [**wood-2014**].

Of the remaining missing articles, two are missing as a result of errors in the process of moving from the raw XML files to pandas dataframes. It is not clear what has caused this in the two cases here, but it is not a case of the supervised learning technique failing. Two further missing articles have been removed because they were taken to be too short to be reliably classified. This leaves one article's absence from the corpus unexplained (see Table 3).

3.2.3 New Zealand Content in the NB2 Corpus

Table 3.2.3 presents the top ten keywords of the two topics picked out for further investigation.[61] Discussion of the New Zealand education system has already been noted, leading to interest in Topic 93. Topic 90's use of te reo Māori encourages the idea that this topic might be assigned to philosophical deliberations unique to New Zealand.

Inspection of documents assigned a high proportion of Topic 90 revealed a focus on hymns and poetry, although this was not entirely clear. As a result, Topic 93 became the focus. Documents assigned a proportion of Topic 93 above 0.4 were selected for further investigation. A word cloud for these articles is presented by Figure 10. In addition to the keywords obviously

[61] The full list of topics and their top-10 keywords for this model is present in 'lda_models/rel_v2_philoso_100.csv'.

connected to the debate over whether education should be secular or not, we see 'zealand' and 'colony' indicating some prominence to discussion of these topics in connection with the New Zealand context.

This New Zealand connection can be confirmed by examining some of the articles high in this topic.[62] The article 'ODT_18800630_ARTICLE24' was found to be particularly revelatory of the context of these discussions. It is a letter to the editor in which the passing of the Education Act of 1877, which established secular education for Pākehā [6], is argued against on the basis of features of New Zealand public opinion and that secular education would undermine the quality of the youth of the nation.

Here, then, is evidence to answer the third question in the affirmative. We can find evidence of specifically New Zealand-based philosophical issues being discussed in the NB2 corpus. This shows that the corpus has some use for investigation of philosophy and philosophical discourse in colonial New Zealand.

4 Discussion

This section reflects on the successes and failures of the project, in turn, and then considers the possibilities for others to use the same methods and workflow for their own projects.

Most importantly, this project shows that a single researcher can quite quickly produce a corpus for the investigation of a particular area of discourse by labelling a small set of articles and using it to train a classifier. This is to answer the first research question with a 'yes'. Moreover, the use of this corpus was illustrated by consideration of two more specific research questions.

Two main shortcomings are worth noting. First, in the discussion of the classifier results, it became clear that the borderline between 'philosophy' and 'non-philosophy' was somewhat blurry, and that this limited the training of the Naive Bayes classifier. It also seemed that ethics and politics content was being left out more frequently than religious content. This could suggest some bias in the labelling carried out by the investigator.[63] Moreover, some of the labels initially set up, notably the 'metaphysics-epistemology' label, were not used much. Given this, it would be useful to establish a more clear set of criteria for labelling articles.[64]

Another problem which has been discussed above, and is related to the issue of labelling, is the presence of composite articles in which a small proportion is 'philosophical'. Many editorials are in this class, and so a whole genre of intellectual activity in early New Zealand newspapers may be being lost from the corpus. A solution to this would be to apply labels at the 'text block' level and then to classify an article as philosophy if it contains a block classified as philosophical. This issue is related to the problem of labelling insofar as one of the difficulties confronted in labelling was whether or not to label composite articles as philosophical.

The project was encouraged by the release of the National Library Papers Past Newspaper Open Data Pilot. However, it seems likely that there is much more material of interest to the historian of philosophy in the National Libraries digitised magazines and journals. However, it is important here to remember Ballantyne's point that New Zealand newspapers were 'the fundamental infrastructure for intellectual life' in Otago, and that this was helped by the economic impossibility of sustainable periodicals and large collections of imported books [2, pp. 57–58].

Turning to the use of this project by other researchers. The publicly accessible dashboard for the project at `nz-newspaper-philosophy.herokuapp.com` enables other researchers

[62] In the 'NZ content.ipynb' notebook, the articles containing the term 'zealand' are explored.

[63] However, note the prominence of 'God' in the Philoso corpus word cloud (Figure 4). This is before any influence of the labelled dataset, and suggests that religious material is just very prominent amongst the intellectual content in early New Zealand newspapers.

[64] Suggestions in this direction are set out in the notebook 'Relabelling.ipynb'.

to produce co-occurrence networks and explore the NB2 and Rel corpora to attempt to generate some insight about philosophical discourse in early New Zealand newspapers. The corpora produced in the project are also available through links of the dashboard for researchers to explore them using their own tools.

This project also provides general patterns which could easily be followed by other researchers. The basic workflow for labelling and training classifiers is not confined to philosophical discourse. This is also true of the steps to move from ALTO/METS XML files to Pandas dataframes and thus into the Python ecosystem. Given the prevalence of the ALTO/METS standard in newspaper digitisation projects, it is likely that the methods used in this project are quite widely applicable.

Conclusion

This report has presented work towards at DATA601 summer project. It was divided into two main tasks: the construction of a useful corpus for investigators in the history of philosophy from the National Library's Papers Past newspaper open data pilot dataset, and the demonstration of some of the uses to which this corpus might be put by digital humanities investigators.

Three research questions were considered, all of which have been answered in the affirmative. First, the project has demonstrated that a useful corpus can be produced using supervised learning techniques with researcher labelling of a small subset of articles and the use of simple Naive Bayes classifiers. Second, it has been shown that insight into the content of discussion about religious belief and the development in the natural sciences, especially evolution, can be obtained from the resulting corpus. Finally, the discovery, by means of topic modelling and the use of simple word clouds, of the presence and prominence of debates over the nature of education and the way it should be organised in New Zealand in the lead up and wake of the 1877 Education Act, showed that New Zealand focused content is present within the corpus.

It is hoped that process of corpus construction set out in this report can be of use to other researchers with other interests in the dataset. It is also hoped that the corpus produced in part one of this project, and the labelled dataset which enabled the training of classifiers, might be of use to other researchers interested in carrying out digital humanities research on philosophy and intellectual life in general in colonial New Zealand.

References

[1] Andreas Niekler and Gregor Wiedemann. *Tutorial 5: Co-occurrence analysis*. 2020. URL: %7Bhttps://tm4ss.github.io/docs/Tutorial_5_Co-occurrence.html%7D. accessed 22/01/2021.

[2] Tony Ballantyne. "Reading the Newspaper in Colonial Otago". In: *The Journal of New Zealand Studies* 12 (2012). ISSN: 2324-3740. DOI: 10.26686/jnzs.v0i12.488. URL: https://ojs.victoria.ac.nz/jnzs/article/view/488.

[3] Brian Weatherson. *A History of Philosophy Journals: Volume 1: Evidence from Topic Modeling, 1876-2013*. 2020. URL: http://www-personal.umich.edu/~weath/lda/.

[4] *Copyright and re-use*. 2020. URL: https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/copyright-and-re-use-papers-past-newspaper-open-data-pilot.

[5] Rosi Crane. "Ä dangerous visionary̋? The Lectures of the Evolutionist T.J. Parker". In: *The Journal of New Zealand Studies* 15 (2013). ISSN: 2324-3740. DOI: 10.26686/jnzs.v0i15.2007. URL: https://ojs.victoria.ac.nz/jnzs/article/view/2007.

[6] *Education Act passed into law*. 2020. URL: https://nzhistory.govt.nz/page/education-act-passed-law (visited on).

[7] Clark N. Glymour. *Theory and evidence / Clark Glymour*. English. Princeton University Press Princeton, N.J, 1980, xi, 383 p. : ISBN: 069107240 0691100772.

[8] Graham Oddie and Roy W. Perret. "Introduction". In: *Justice, Ethics and New Zealand Society*. Ed. by Graham Oddie and Roy W. Perret. Oxford: Oxford University Press, 1992.

[9] Charles R. Harris et al. "Array programming with NumPy". In: *Nature* 585.7825 (Sept. 2020), pp. 357–362. DOI: 10.1038/s41586-020-2649-2. URL: https://doi.org/10.1038/s41586-020-2649-2.

[10] Matthew Honnibal et al. *spaCy: Industrial-strength Natural Language Processing in Python*. 2020. DOI: 10.5281/zenodo.1212303. URL: https://doi.org/10.5281/zenodo.1212303.

[11] Plotly Technologies Inc. *Collaborative data science*.

[12] Library of Congress. *MARC 21 XML Schema*. URL: %7Bhttp://www.loc.gov/standards/marcxml/%7D. accessed 22/01/2021.

[13] Martin Davies and Stein Helgeby. "Idealist Origins: 1920s and Before". In: *History of Philosophy in Australia and New Zealand*. Ed. by Graham Oppy and N. N. Trakakis. Dordrecht, 2014, 15–54.

[14] Moretti, Franco. *Distant Reading*. London, 2013.

[15] Andreas Mueller. *Word Cloud*. 2020. URL: https://amueller.github.io/word_cloud/index.html (visited on).

[16] National Library. *About Papers Past*. URL: https://paperspast.natlib.govt.nz/about (visited on).

[17] National Library. *Data standards*. 2020. URL: https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/data-standards-papers-past-newspaper-open-data-pilot.

[18] National Library. *Papers Past newspaper open data pilot*. 2020. URL: https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot.

[19] National Library. *Projects*. 2020. URL: https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/projects-papers-past-newspaper-open-data-pilot (visited on).

[20] Pavel Rychlý. "A Lexicographer-Friendly Association Score". In: *RASLAN 2008: Recent Advances in Slavonic Natural Language Processing*. Ed. by P. Sojka and A. Horák. Brno: Masaryk University, 2008.

[21] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python ". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[22] "Personal Liberty". In: *Bruce Herald* 30 (3101 1899), p. 3. URL: %7Bhttps://paperspast.natlib.govt.nz/newspapers/BH18990922.2.9.1%7D.

[23] "POLICE COURT.—Saturday. [Before Thomas Beckham Esq., R.M.]" In: *Daily Southern Cross* 29 (5101 1873), p. 3. URL: %7Bhttps://paperspast.natlib.govt.nz/newspapers/BH18990922.2.9.1%7D.

[24] "Random Shots". In: *Auckland Star* (1886), p. 4. URL: %7Bhttps://paperspast.natlib.govt.nz/newspapers/AS18860821.2.33%7D.

[25] Radim Řehůřek and Petr Sojka. "Software Framework for Topic Modelling with Large Corpora". English. In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. http://is.muni.cz/publication/884893/en. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[26] Steven Bird, Edward Loper and Ewan Klein. *Natural Language Processing with Python*. Sebastopol, 2009.

[27] "Theological Debate". In: *Oxford Observer* (1892), p. 3. URL: %7Bhttps://paperspast.natlib.govt.nz/newspapers/OO18921224.2.5%7D.

[28] Veridian Software. *What is METS/ALTO?* URL: %7Bhttps://veridiansoftware.com/knowledge-base/metsalto/%7D. accessed 22/01/2021.

[29] Harry Zhang. "The Optimality of Naive Bayes". In: vol. 2. Jan. 2004.

A Appendix

Article	Title	Philosophy Type	Description
AS_18760720_ARTICLE20	GRANDMAMA ON "WOMAN'S RIGHTS."	Ethics	Correctly labelled
AS_18820306_ARTICLE31	THE BRADLAUGH EPISODE.	Ethics	Correctly labelled
AS_18881121_ARTICLE77	"PROGRESS and AFTERWARDS."	Ethics	Correctly labelled
DSC_18600731_ARTICLE28	Friday, July 13.	Other	Correctly labelled
DSC_18601225_ARTICLE10	LETTERS In Reply to Sir William Martin's Pamphlet on the Taranaki Question. No. 1.	Ethics	Correctly labelled
DTN_18940820_ARTICLE7	The Daily Telegraph. MONDAY, AUGUST 20,1894. THE ROLLESTON BANQUET.	Ethics	Mislabelled (passing reference)
ESD_18890826_ARTICLE1	COLONIALCHARACTERISTICS	Ethics	Mislabelled
LT_18831025_ARTICLE34	TIMARU TALK.	Other	Composite
LT_18970507_ARTICLE14	CAPITAL PUNISHMENT.	Ethics	Correctly labelled
LT_18980920_ARTICLE26	CURRENT TOPICS.	Ethics	Composite
NEM_18800301_ARTICLE9	The Nelson Evening Mail. MONDAY, MARCH 1, 1880. NIHILISM : WHAT IS IT ?	Ethics	Correctly labelled.
NZTIM_18780509_ARTICLE5	UNTITLED	Ethics	Composite
ODT_18830714_ARTICLE20	PASSING NOTES.	Ethics	Composite
OO_18970911_ARTICLE4	BLACKMAILING.	Ethics	Correctly labelled
WI_18470303_ARTICLE5	ORIGINAL CORRESPONDENCE.	Ethics	Mislabelled (passing reference)

Table 1: Full list of false negatives from test set for second Naive Bayes classifier.

Article	Title	Readable	Description
AG_18990504_ARTICLE7	Ashburton Guardian. Megna est Veritas et Prævalebit. THURSDAY, MAY 4, 1899.	True	Correctly labelled (quack advertisement)
CHP_18951209_ARTICLE55	ST. ALBANS WESLEYAN CHURCH.	True	Composite (religious)
ESD_18851028_ARTICLE1	WHITE CROSS SOCIETY.	True	Edge case (religious)
ESD_18891218_ARTICLE59	CHARACTERISTICS OF CHIRISTIANITY.	True	Mislabelled (religious)
ESD_18960111_ARTICLE47	MB RELIGIOUS WORLD 'HUMAN ONENESS.'	True	Mislabelled (religious)
GRA_18960522_ARTICLE19	THE NEW PHOTOGRAPHIC PROCESS,	True	Correctly labelled (materials of warships)
LT_18801016_ARTICLE32	CANTERBURY COLLEGE.	True	Correctly labelled (list of graduations)
LWM_18950614_ARTICLE27	Select Poetry.	True	Correctly labelled (religious)
ME_18870708_ARTICLE32	THE MYSTERY OF INSTINCT.	FALSE	A few philosophical terms are readable.
NEM_18920606_ARTICLE29	THE POPE ON ITS DIFFICULTIES.	True	Edge case (religious)
ODT_18850204_ARTICLE30	THE CATHOLIC CLAIMS T. BIBLEREADING IN SCHOOLS.	True	Correctly labelled (religious)
ODT_18980407_ARTICLE89	THE INFLUENCE OF CHRISTIANITY. TO THE EDITOR.	True	Correctly labelled (religious)
ODT_18981013_ARTICLE51	DEAN FARRAR DEFENDS THEATRE-GOING.	True	Correctly labelled (religious)
OW_18770317_ARTICLE59	BISMARCK.	True	Mislabelled (religious)

Table 2: Full list of false positives from test set for second Naive Bayes classifier.

Article	philoso*	NB 1	NB 2	Rel	
CHP_18921024_ARTICLE53	False	True	True	True	
CHP_18991023_ARTICLE39	False	True	True	True	
CHP_18630117_ARTICLE3	True	False	False	False	Unexplained
WI_18710720_ARTICLE13	True	True	True	True	

Table 3: Articles identified but not in labelled collection.

Article	philoso*	NB 1	NB 2	Rel	Note
ODT_18710509_ARTICLE18	False	True	True	True	Used in training
TT_18710720_ARTICLE23	False	False	False	False	XML parse error
OW_18730531_ARTICLE11	False	True	True	True	
ODT_18760516_ARTICLE22	False	True	True	True	
OW_18760520_ARTICLE81	False	True	True	True	
OW_18760527_ARTICLE71	False	True	True	True	
ODT_18760617_ARTICLE20	False	True	True	True	
OW_18780907_ARTICLE28	False	True	True	True	
ST_18800924_ARTICLE14	False	False	False	False	XML parse error
ODT_18801204_ARTICLE15	False	True	True	True	
OW_18820701_ARTICLE45	False	True	True	True	
OW_18820701_ARTICLE77	False	True	True	True	
ODT_18820705_ARTICLE26	False	True	True	True	
OW_18820708_ARTICLE61	False	True	True	True	
ME_18940605_ARTICLE7	False	False	False	False	Filtered (too short)

Table 4: Articles on evolutionary ideas identified in Ballantyne 2012.

Article	philoso*	NB 1	NB 2	Rel	Note
ODT_18880705_ARTICLE21	False	True	True	True	
ODT_18880906_ARTICLE30	False	True	False	False	Composite (synod report)
ODT_18881101_ARTICLE18	False	True	False	False	Composite (synod report)
ODT_18881102_ARTICLE31	False	True	False	False	Composite (synod report)
ODT_18881103_ARTICLE24	False	True	False	False	Composite (synod report)
ODT_18881108_ARTICLE3	False	False	False	False	Filtered (too short)

Table 5: Articles on 'Reign of Grace' controversy identified in Wood 2014.

Article	philoso*	NB 1	NB 2	Rel	Note
ODT_18810528_ARTICLE13	True	True	True	True	
OW_18810604_ARTICLE108	False	False	False	False	Filtered (too short)
OW_18820701_ARTICLE77	False	True	True	True	
NOT_18840128_ARTICLE15	False	True	True	True	
ODT_18840923_ARTICLE8	False	False	False	False	Composite
ODT_18850602_ARTICLE17	False	True	True	True	
ODT_18950927_ARTICLE6	False	False	False	False	Composite
ODT_18910723_ARTICLE10	False	False	False	False	Composite
ODT_18820629_??	N/A	N/A	N/A	N/A	Advertisement
ODT_18861120_ARTICLE3	False	False	False	False	Composite

Table 6: Material concerning Fletcher's public lectures on evolution identified in Crane 2013.