PHILOSOPHICAL WRITING IN EARLY NEW ZEALAND NEWSPAPERS

Joshua Black

February 5, 2021

UC ARTS
DIGITAL LAB

Overview

Overview

1. Problem:

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results
4. Upshot:

Overview

1. Problem:
    - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
    - corpus construction via self-labelling and supervised learning,
    - corpus analysis with, e.g., co-occurrence networks.
3. Results
4. Upshot:
    - a method applicable for many humanities research questions,

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results
4. Upshot:
   - a method applicable for many humanities research questions,
   - but with shortcomings to be aware of.

PROBLEM

Background

Background

- Organisation:

Background

- Organisation:
  - UC Arts Digital Lab

Background

- Organisation:
  - UC Arts Digital Lab
  - ...part of the digital humanities department.

Background

- Organisation:
  - UC Arts Digital Lab
  - ...part of the digital humanities department.
  - Digital humanities: use of computational (and data science) techniques to achieve insight into cultural products and practices.

Background

- Organisation:
  - UC Arts Digital Lab
  - ...part of the digital humanities department.
  - Digital humanities: use of computational (and data science) techniques to achieve insight into cultural products and practices.
- Myself:

Background

- Organisation:
  - UC Arts Digital Lab
  - ...part of the digital humanities department.
  - Digital humanities: use of computational (and data science) techniques to achieve insight into cultural products and practices.
- Myself:
  - research background in traditional humanities research (in philosophy),

Background

- Organisation:
  - UC Arts Digital Lab
  - ...part of the digital humanities department.
  - Digital humanities: use of computational (and data science) techniques to achieve insight into cultural products and practices.
- Myself:
  - research background in traditional humanities research (in philosophy),
  - with newly developed data science skills.

Humanities Problem

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).
- An explanation: excessive on focus academic publications.

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).
- An explanation: excessive on focus academic publications.
  - ...and on *academic* philosophy.

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).
- An explanation: excessive on focus academic publications.
  - ...and on *academic* philosophy.
- Newspapers as an alternative source:

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).
- An explanation: excessive on focus academic publications.
  - ...and on *academic* philosophy.
- Newspapers as an alternative source:
  - 'the fundamental infrastructure for intellectual life ...

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).
- An explanation: excessive on focus academic publications.
  - ...and on *academic* philosophy.
- Newspapers as an alternative source:
  - 'the fundamental infrastructure for intellectual life ...
  - ...newspapers were ascendant in New Zealand because imported books were expensive and a sustainable local periodical literature was slow to emerge' (Ballantyne 2012, 57–78).

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).
- An explanation: excessive on focus academic publications.
  - ...and on *academic* philosophy.
- Newspapers as an alternative source:
  - 'the fundamental infrastructure for intellectual life ...
  - ...newspapers were ascendant in New Zealand because imported books were expensive and a sustainable local periodical literature was slow to emerge' (Ballantyne 2012, 57–78).
  - Promising: a venue both for the academics and the wider public?

Humanities Problem

- Histories of philosophy in New Zealand don't have much to say before the mid-twentieth century:
  - 'many of those who had longstanding chairs published next to nothing' (Davies and Helgeby 2014, 24).
- An explanation: excessive on focus academic publications.
  - ...and on *academic* philosophy.
- Newspapers as an alternative source:
  - 'the fundamental infrastructure for intellectual life ...
  - ...newspapers were ascendant in New Zealand because imported books were expensive and a sustainable local periodical literature was slow to emerge' (Ballantyne 2012, 57–78).
  - Promising: a venue both for the academics and the wider public?
- An example of the kind of thing we're after:

The following is a brief abstract of the Debate held at the Town Hall East Oxford, on Thursday, 9th.

(Continued from last week.)

6. A simple form of metaphysical argument may be briefly put as follows:—All existence are of two kinds necessary and contingent. By a necessary existence is meant one which never began to be, and can never cease to be. By a contingent existence is meant one which commenced to be and will cease to be. My exist-

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - ‘Big’ by human standards: 1,471,384 pages of content. (315GB compressed)

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - ‘Big’ by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - ‘Big’ by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:
    - all out of copyright, no living people discussed, but …

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - ‘Big’ by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:
    - all out of copyright, no living people discussed, but ...
    - some offensive material present.

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - ‘Big’ by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:
    - all out of copyright, no living people discussed, but ...
    - some offensive material present.
  - https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - ‘Big’ by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:
    - all out of copyright, no living people discussed, but ...
    - some offensive material present.
  - `https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/`
- Data science methods needed to:

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - 'Big' by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:
    - all out of copyright, no living people discussed, but ...
    - some offensive material present.
  - `https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/`
- Data science methods needed to:
  - Find the relevant material (it's a small portion of the data set!)

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - 'Big' by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:
    - all out of copyright, no living people discussed, but ...
    - some offensive material present.
  - `https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/`
- Data science methods needed to:
  - Find the relevant material (it's a small portion of the data set!)
  - Derive insight from it once found (it's still a lot of text!)

Data Science Methods

- Data source: the National Library Papers Past Newspaper Open Data Pilot.
  - A dataset containing the output of OCR for newspaper data in English up to 1900.
  - Made available in 2020 to encourage digital experimentation.
  - ‘Big’ by human standards: 1,471,384 pages of content. (315GB compressed)
  - Ethics:
    - all out of copyright, no living people discussed, but ...
    - some offensive material present.
  - `https://natlib.govt.nz/about-us/open-data/papers-past-metadata/papers-past-newspaper-open-data-pilot/`
- Data science methods needed to:
  - Find the relevant material (it’s a small portion of the data set!)
  - Derive insight from it once found (it’s still a lot of text!)
  - We will engage in ‘distant reading’ (Moretti 2013)

The Project

The Project

1. Corpus construction

The Project

1. Corpus construction
    - Aim: find the relevant material in the dataset.

The Project

1. Corpus construction
   - Aim: find the relevant material in the dataset.
   - Method: labelling articles and training Naive Bayes classifiers following a ‘bootstrapping’ pattern.

The Project

1. Corpus construction
   - Aim: find the relevant material in the dataset.
   - Method: labelling articles and training Naive Bayes classifiers following a ‘bootstrapping’ pattern.
2. Corpus analysis

The Project

1. Corpus construction
   - Aim: find the relevant material in the dataset.
   - Method: labelling articles and training Naive Bayes classifiers following a 'bootstrapping' pattern.
2. Corpus analysis
   - Aim: use the corpus to learn something about philosophical writing in NZ newspapers.

The Project

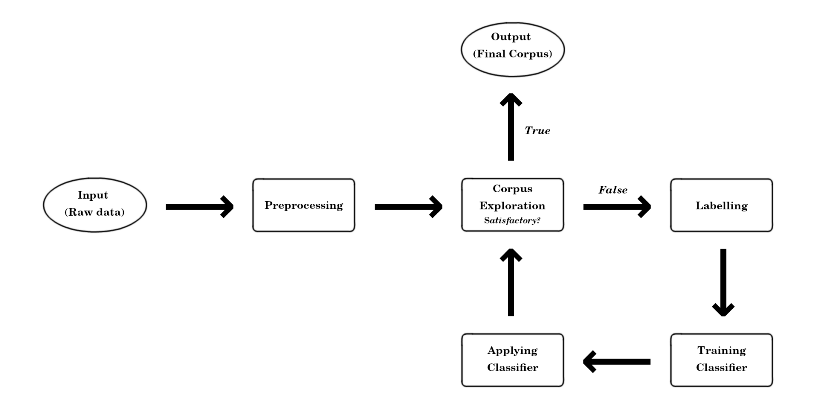
1. Corpus construction
   - Aim: find the relevant material in the dataset.
   - Method: labelling articles and training Naive Bayes classifiers following a ‘bootstrapping’ pattern.
2. Corpus analysis
   - Aim: use the corpus to learn something about philosophical writing in NZ newspapers.
   - Method: many text analysis methods, including concordancing, collocations, co-occurrence networks, and topic modelling.

The Project

1. Corpus construction
   - Aim: find the relevant material in the dataset.
   - Method: labelling articles and training Naive Bayes classifiers following a 'bootstrapping' pattern.
2. Corpus analysis
   - Aim: use the corpus to learn something about philosophical writing in NZ newspapers.
   - Method: many text analysis methods, including concordancing, collocations, co-occurrence networks, and topic modelling.
   - This presentation will focus on co-occurrence networks.

METHOD

Corpus Construction Flow Diagram



Preprocessing (XML $\rightarrow$ Pandas)

Preprocessing (XML $\rightarrow$ Pandas)

- Data format for each issue:

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.

Preprocessing (XML $\rightarrow$ Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)

Preprocessing (XML $\rightarrow$ Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:
  1. iterate through title-year tarballs,

Preprocessing (XML $\rightarrow$ Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:
  1. iterate through title-year tarballs,
  2. for each issue, collect list of articles and corresponding text blocks from METS file,

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:
  1. iterate through title-year tarballs,
  2. for each issue, collect list of articles and corresponding text blocks from METS file,
  3. iterate through ALTO files, collecting text blocks for each article

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:
  1. iterate through title-year tarballs,
  2. for each issue, collect list of articles and corresponding text blocks from METS file,
  3. iterate through ALTO files, collecting text blocks for each article
  4. gather all in Pandas dataframe with each row corresponding to an individual article.

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:
  1. iterate through title-year tarballs,
  2. for each issue, collect list of articles and corresponding text blocks from METS file,
  3. iterate through ALTO files, collecting text blocks for each article
  4. gather all in Pandas dataframe with each row corresponding to an individual article.
- Result:

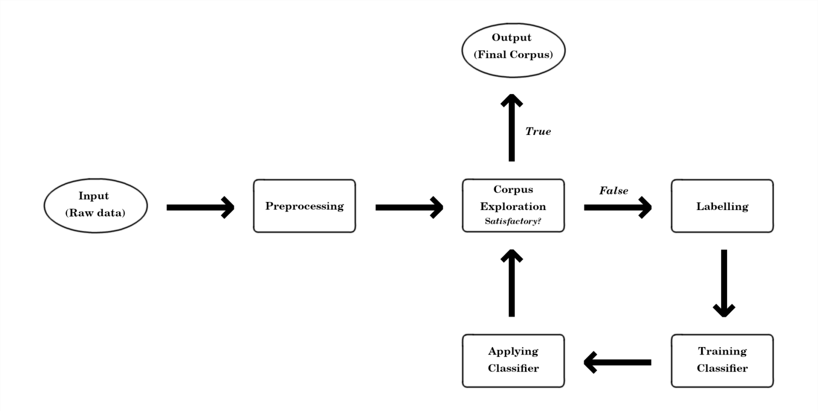
Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:
  1. iterate through title-year tarballs,
  2. for each issue, collect list of articles and corresponding text blocks from METS file,
  3. iterate through ALTO files, collecting text blocks for each article
  4. gather all in Pandas dataframe with each row corresponding to an individual article.
- Result:
  - 7,592,619 distinct articles and their plain text,

Preprocessing (XML → Pandas)

- Data format for each issue:
  - a METS file with structural information and an ALTO file for each page.
- Throw away:
  - Spatial information (e.g. location of each word on page)
  - Items tagged as advertisements.
- Method:
  1. iterate through title-year tarballs,
  2. for each issue, collect list of articles and corresponding text blocks from METS file,
  3. iterate through ALTO files, collecting text blocks for each article
  4. gather all in Pandas dataframe with each row corresponding to an individual article.
- Result:
  - 7,592,619 distinct articles and their plain text,
  - ...stored as eight pickled dataframes (around 8GB total).

Corpus Construction Flow Diagram



Corpus Exploration

Corpus Exploration

- First stage: look at REGEX matched for ‘philoso*’

Corpus Exploration

- First stage: look at REGEX matched for ‘philoso*’
- Lots of methods used to pick out desired and non-desired articles:

Corpus Exploration

- First stage: look at REGEX matched for ‘philoso*’
- Lots of methods used to pick out desired and non-desired articles:
  - inspecting random articles, keyword searches, word clounds, concordancing, collocations, co-occurrence networks …

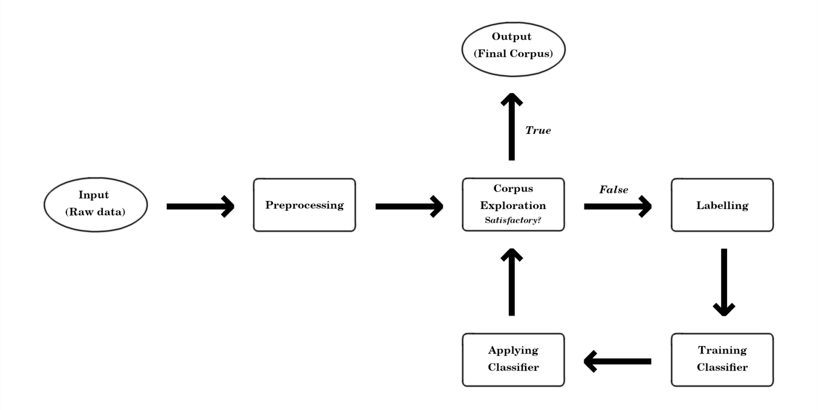
Corpus Exploration

- First stage: look at REGEX matched for ‘philoso*’
- Lots of methods used to pick out desired and non-desired articles:
  - inspecting random articles, keyword searches, word clounds, concordancing, collocations, co-occurrence networks …
  - …more on co-occurrence networks later.

Corpus Exploration

- First stage: look at REGEX matched for ‘philoso*’
- Lots of methods used to pick out desired and non-desired articles:
  - inspecting random articles, keyword searches, word clounds, concordancing, collocations, co-occurrence networks ...
  - ...more on co-occurrence networks later.
- If the corpus contains lots of material that we are not interested in, it is not ‘satisfactory’. If so, we move to the next stage.

Corpus Construction Flow Diagram



Labelling

Labelling

- Label articles in order to train a classifier to find what we are after.

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article ‘philosophical discourse’?

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article ‘philosophical discourse’?
    - A broad definition: does it develop or discuss ideas ‘ultimate reality’ or ‘ultimate value’.

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article 'philosophical discourse'?
    - A broad definition: does it develop or discuss ideas 'ultimate reality' or 'ultimate value'.
    - e.g.: is there life after death, are there multiple sources of knowledge, what is the best way to organise society and why?

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article ‘philosophical discourse’?
    - A broad definition: does it develop or discuss ideas ‘ultimate reality’ or ‘ultimate value’.
    - e.g.: is there life after death, are there multiple sources of knowledge, what is the best way to organise society and why?
    - Some reliance on my own experience in studying 19th century philosophy.

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article ‘philosophical discourse’?
    - A broad definition: does it develop or discuss ideas ‘ultimate reality’ or ‘ultimate value’.
    - e.g.: is there life after death, are there multiple sources of knowledge, what is the best way to organise society and why?
    - Some reliance on my own experience in studying 19th century philosophy.
  - Philosophy type:

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article ‘philosophical discourse’?
    - A broad definition: does it develop or discuss ideas ‘ultimate reality’ or ‘ultimate value’.
    - e.g.: is there life after death, are there multiple sources of knowledge, what is the best way to organise society and why?
    - Some reliance on my own experience in studying 19th century philosophy.
  - Philosophy type:
    - Is it about ethics, the relationship between religious belief and modern thought, metaphysics and epistemology, or other?

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article 'philosophical discourse'?
    - A broad definition: does it develop or discuss ideas 'ultimate reality' or 'ultimate value'.
    - e.g.: is there life after death, are there multiple sources of knowledge, what is the best way to organise society and why?
    - Some reliance on my own experience in studying 19th century philosophy.
  - Philosophy type:
    - Is it about ethics, the relationship between religious belief and modern thought, metaphysics and epistemology, or other?
    - The relationship between religious belief and, e.g., evolution is a very prominent topic at this time.

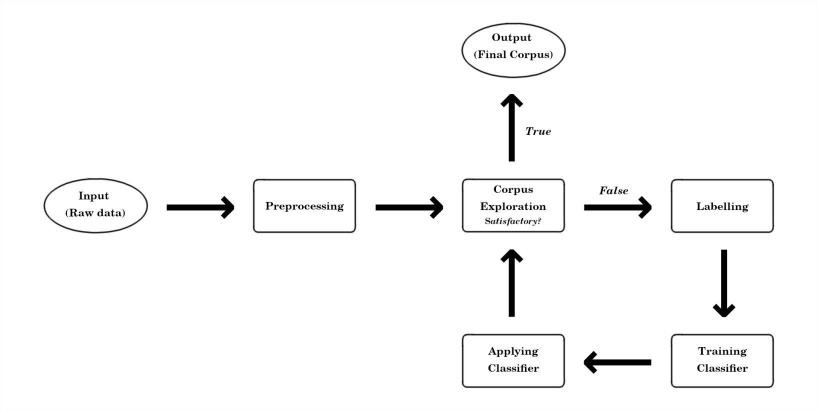
Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article 'philosophical discourse'?
    - A broad definition: does it develop or discuss ideas 'ultimate reality' or 'ultimate value'.
    - e.g.: is there life after death, are there multiple sources of knowledge, what is the best way to organise society and why?
    - Some reliance on my own experience in studying 19th century philosophy.
  - Philosophy type:
    - Is it about ethics, the relationship between religious belief and modern thought, metaphysics and epistemology, or other?
    - The relationship between religious belief and, e.g., evolution is a very prominent topic at this time.
  - Also attempted, but not used: 'Readable', 'Writing Type', 'NZ author'

Labelling

- Label articles in order to train a classifier to find what we are after.
- Two key labels for the project:
  - Philosophy: is the majority of the article ‘philosophical discourse’?
    - A broad definition: does it develop or discuss ideas ‘ultimate reality’ or ‘ultimate value’.
    - e.g.: is there life after death, are there multiple sources of knowledge, what is the best way to organise society and why?
    - Some reliance on my own experience in studying 19th century philosophy.
  - Philosophy type:
    - Is it about ethics, the relationship between religious belief and modern thought, metaphysics and epistemology, or other?
    - The relationship between religious belief and, e.g., evolution is a very prominent topic at this time.
  - Also attempted, but not used: ‘Readable’, ‘Writing Type’, ‘NZ author’
- NB: it is important to ensure that we label a wide range of non-philosophy.

Corpus Construction Flow Diagram



Training and Applying Classifier

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.
- Classifiers tried: Naive Bayes and Support Vector Machines

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.
- Classification algorithms mplemented using Scikit-learn Pipelines.

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.
- Classification algorithms mplemented using Scikit-learn Pipelines.
  1. Text to bag of words (params: size of feature space)

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.
- Classification algorithms mplemented using Scikit-learn Pipelines.
  1. Text to bag of words (params: size of feature space)
  2. TF-IDF transformation

Training and Applying Classifier

- Train a classifier to distinguish 'philosophy' articles from 'non-philosophy'.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.
- Classification algorithms mplemented using Scikit-learn Pipelines.
  1. Text to bag of words (params: size of feature space)
  2. TF-IDF transformation
  3. The Naive Bayes classifier (params: prior on representativeness of labelled set.)

Training and Applying Classifier

- Train a classifier to distinguish ‘philosophy’ articles from ‘non-philosophy’.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.
- Classification algorithms mplemented using Scikit-learn Pipelines.
  1. Text to bag of words (params: size of feature space)
  2. TF-IDF transformation
  3. The Naive Bayes classifier (params: prior on representativeness of labelled set.)
- Grid CV search used to select parameters for the classifier.

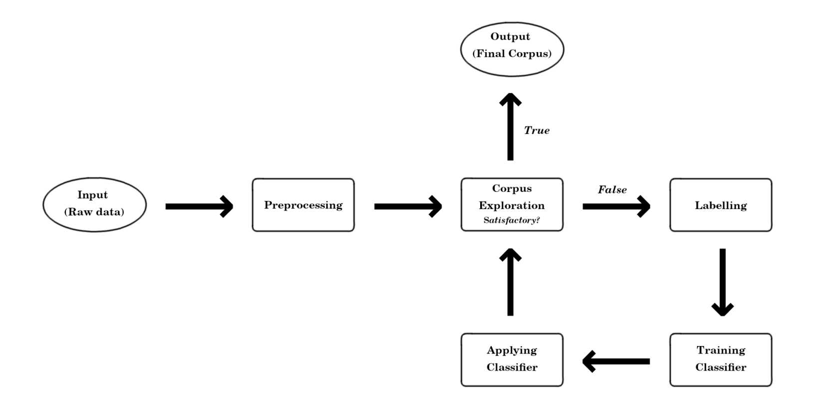
Training and Applying Classifier

- Train a classifier to distinguish 'philosophy' articles from 'non-philosophy'.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.
- Classification algorithms mplemented using Scikit-learn Pipelines.
  1. Text to bag of words (params: size of feature space)
  2. TF-IDF transformation
  3. The Naive Bayes classifier (params: prior on representativeness of labelled set.)
- Grid CV search used to select parameters for the classifier.
- ...with accuracy, recall, and precision tried as measures.

Training and Applying Classifier

- Train a classifier to distinguish 'philosophy' articles from 'non-philosophy'.
- Classifiers tried: Naive Bayes and Support Vector Machines
- Naive Bayes is simple and fast to train, while performing remarkably effectively.
- Training and testing data divided, and training data resampled, as appropriate given the state of the labelled collection.
- Classification algorithms mplemented using Scikit-learn Pipelines.
  1. Text to bag of words (params: size of feature space)
  2. TF-IDF transformation
  3. The Naive Bayes classifier (params: prior on representativeness of labelled set.)
- Grid CV search used to select parameters for the classifier.
- ...with accuracy, recall, and precision tried as measures.
- Trained classifier applied to complete dataset, treating the 'philosophy' articles as a new candidate corpus.

Corpus Construction Flow Diagram



‘Bootstrapping’

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,
  2. having collected a good number ($\sim$200-300), with much higher representation of philosophy than the general dataset,

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,
  2. having collected a good number (∼200-300), with much higher representation of philosophy than the general dataset,
  3. we train and apply a classifier,

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,
  2. having collected a good number ($\sim$200-300), with much higher representation of philosophy than the general dataset,
  3. we train and apply a classifier,
  4. we use the articles classifier as philosophy as a source of new articles to label.

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,
  2. having collected a good number (∼200-300), with much higher representation of philosophy than the general dataset,
  3. we train and apply a classifier,
  4. we use the articles classifier as philosophy as a source of new articles to label.
- NB: after the first classifier has been applied, the new non-philosophy articles added to the labelled collection will have been classifier as philosophy by the previous classifier.

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,
  2. having collected a good number ($\sim$200-300), with much higher representation of philosophy than the general dataset,
  3. we train and apply a classifier,
  4. we use the articles classifier as philosophy as a source of new articles to label.
- NB: after the first classifier has been applied, the new non-philosophy articles added to the labelled collection will have been classifier as philosophy by the previous classifier.
  - ...this means that subsequently trained classifiers can be more selective.

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,
  2. having collected a good number (∼200-300), with much higher representation of philosophy than the general dataset,
  3. we train and apply a classifier,
  4. we use the articles classifier as philosophy as a source of new articles to label.
- NB: after the first classifier has been applied, the new non-philosophy articles added to the labelled collection will have been classifier as philosophy by the previous classifier.
  - ...this means that subsequently trained classifiers can be more selective.
  - ...we need a picky classifier.

‘Bootstrapping’

- The phrase: ‘pull yourself up by the bootstraps’,
- In this case:
  1. starting with nothing, we add articles to our labelled collection,
  2. having collected a good number (∼200-300), with much higher representation of philosophy than the general dataset,
  3. we train and apply a classifier,
  4. we use the articles classifier as philosophy as a source of new articles to label.
- NB: after the first classifier has been applied, the new non-philosophy articles added to the labelled collection will have been classifier as philosophy by the previous classifier.
  - ...this means that subsequently trained classifiers can be more selective.
  - ...we need a picky classifier.
  - Stop when satisfied that the corpus does not contain too much ‘non-philosophy’

Aim 2: Corpus Analysis

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;
  2. compute a document-term matrix (num terms x num documents) for either the BOW or TF-IDF representation;

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;
  2. compute a document-term matrix (num terms x num documents) for either the BOW or TF-IDF representation;
  3. compute a term-term matrix (num terms x num terms), representing co-occurrences in documents of each pair of terms; and

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;
  2. compute a document-term matrix (num terms x num documents) for either the BOW or TF-IDF representation;
  3. compute a term-term matrix (num terms x num terms), representing co-occurrences in documents of each pair of terms; and
  4. given a search term, pass these matrices to a statistical function to return the most closely related words.

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;
  2. compute a document-term matrix (num terms x num documents) for either the BOW or TF-IDF representation;
  3. compute a term-term matrix (num terms x num terms), representing co-occurrences in documents of each pair of terms; and
  4. given a search term, pass these matrices to a statistical function to return the most closely related words.
     - Statistics implemented: log Dice and mutual information.

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;
  2. compute a document-term matrix (num terms x num documents) for either the BOW or TF-IDF representation;
  3. compute a term-term matrix (num terms x num terms), representing co-occurrences in documents of each pair of terms; and
  4. given a search term, pass these matrices to a statistical function to return the most closely related words.
     - Statistics implemented: log Dice and mutual information.
- Networks displayed using Dash cytoscapes:

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;
  2. compute a document-term matrix (num terms x num documents) for either the BOW or TF-IDF representation;
  3. compute a term-term matrix (num terms x num terms), representing co-occurrences in documents of each pair of terms; and
  4. given a search term, pass these matrices to a statistical function to return the most closely related words.
     - Statistics implemented: log Dice and mutual information.
- Networks displayed using Dash cytoscapes:
  - see project dashboard: `nz-newspaper-philosophy.herokuapp.com`

Aim 2: Corpus Analysis

- Corpus analysis $\sim$ the corpus exploration step.
- One method used: co-occurrence networks:
  1. create bag or words or TF-IDF representation of documents;
  2. compute a document-term matrix (num terms x num documents) for either the BOW or TF-IDF representation;
  3. compute a term-term matrix (num terms x num terms), representing co-occurrences in documents of each pair of terms; and
  4. given a search term, pass these matrices to a statistical function to return the most closely related words.
     - Statistics implemented: log Dice and mutual information.
- Networks displayed using Dash cytoscapes:
  - see project dashboard:
    `nz-newspaper-philosophy.herokuapp.com`
- NB: term-term matrices can get very large. Various methods to control the size of the dictionary were employed.

RESULTS

Corpus Construction (Size Reduction)

Corpus Construction (Size Reduction)

Classifiers become more selective:

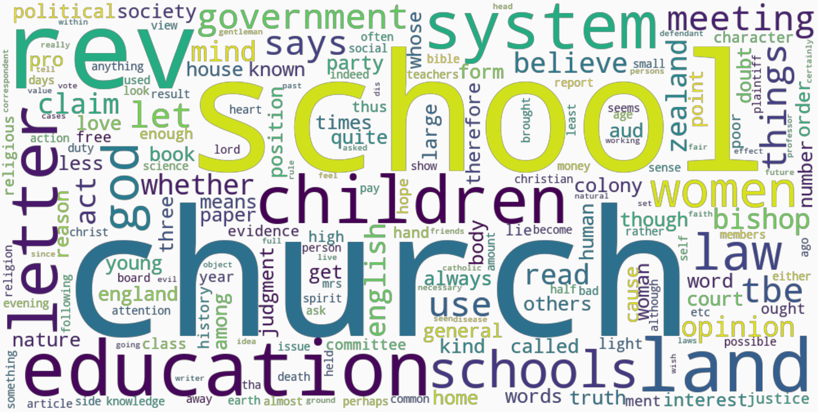
Corpus	Article Count
Processed dataset	7592619
(Step 0) 'philoso*' Corpus	29647
(Step 1) Naive Bayes 1	239649
(Step 2) Naive Bayes 2	31131

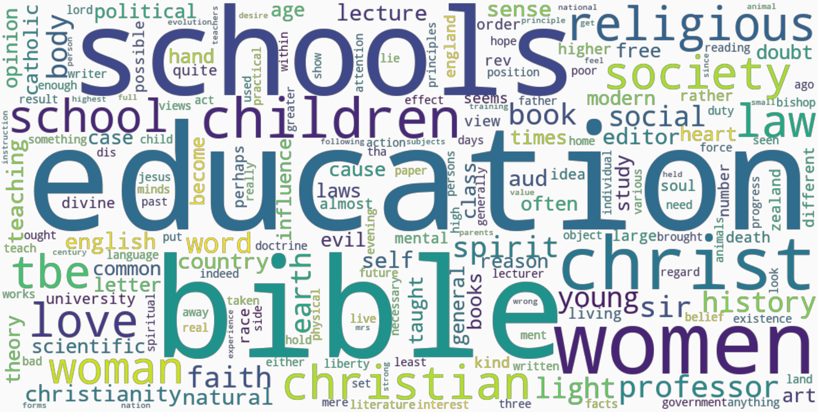
Table: Article counts for processed dataset and general philosophy corpora.

Word Cloud: Sample of Processed Dataset


race
hours
literary
press
sense
policy
university
yesterday
labour
board
influence
correspondent
committee
college
chief
love
return
gold
writer
important
gone
bank
study
prince
christian
article
whioh
messrs
action
books
faith
line
philosophers
professor
looked moment
religion
works
parliament
schools
mother
stated
cost
native
saturday
makes
monday
george
christ
fire
miles
rev
american
district
speech
council
happy
twenty
worth
god
yes
paris
likely
turned
melbourne
live
family
states
husband
north
cases
science
horse
knowledge
appears
association
evidence
married
railway
turn
change
subjects
road
ladies
duty
seemed
looking
religious
sunday
lord
feet
living
knew
air
river
health
bishop
object
boy
special
national
earth
natives
play
author
real
captain
reading
progress
story
dead
girl
trade
person
eyes
black
book
bill
working
editor
service
french
dunedin
west
lecture
paid
towards
scientific
saw
notice
tha
spirit
certainly
stand
aro
justice
british
women
miss
motion
island
dear
run
laws
sure
hard
ten
wish
local
colonial
white
modern
recently
father
king
bible
property
language
interesting
report
art
hia
queen
hall
member
child
club
questions
colonies
regard
philosophical
fellow
police
otago
condition
articles
judge
died
force
chair
value
charge
united
york
education
named
minister
replied
written
auckland
address
hair
alone
wellington
feel
social
german
occasion
france

WORD CLOUD: STEP 1




education
bible
schools
children
school
religious
society
christ
women
christian
law
social
book
love
woman
tbe
professor
history
light
faith
natural
christianity
young
sir
spirit
self
earth
word
country
english
letter
common
scientific
university
teaching
catholic
opinion
body
political
lecture
sense
order
hope
rev
position
poor
free
higher
doubt
modern
rather
duty
bishop
editor
heart
times
home
view
seems
father
effect
act
views
result
enough
writer
possible
hand
quite
age
lord
desire
evolution
teachers
national
principle
get
animal
reading
feel
since
ago
small
seen
force
lie
show
within
practical
used
greater
attention
principles
england
case
something
child
dis
instruction
following
action
subjects
days
tha
cause
paper
laws
almost
class
persons
generally
aud
idea
value
often
individual
study
various
held
soul
need
number
progress
zealand
different
become
perhaps
really
influence
jesus
minds
past
divine
ought
teach
century
language
put
doctrine
evil
mental
parents
object
large
brought
death
animals
high
evening
reason
lecturer
regard
look
indeed
future
necessary
taught
general
books
wrong
living
belief
existence
works
theory
spiritual
taken
away
real
experience
race
side
physical
hold
live
mrs
ment
bad
either
liberty
strong
least
kind
written
set
mere
literature
interest
three
facts
government
anything
land
art
forms
nation

Labelling

Label	Step 1 Value	Count	Step 2 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: Label counts at Step 1 and Step 2.

Classifier Performance (Step 2)

Classifier Performance (Step 2)

		Predicted	
		False	True
Actual	False	181	14
	True	15	62

Table: Confusion Matrix for Second Naive Bayes Classifier

- Accuracy: 0.89
- Precision: 0.81
- Recall: 0.80

Corpus Performance

Corpus Performance

- Metrics are not bad, but not great either.

Corpus Performance

- Metrics are not bad, but not great either.
- Inspection of the false positives and negatives reveals:

Corpus Performance

- Metrics are not bad, but not great either.
- Inspection of the false positives and negatives reveals:
  - prevalence of ‘composite articles’ in the false negatives, where only one or two bits are philosophical; and

Corpus Performance

- Metrics are not bad, but not great either.
- Inspection of the false positives and negatives reveals:
  - prevalence of ‘composite articles’ in the false negatives, where only one or two bits are philosophical; and
  - prevalence of ‘edge cases’ and mistaken labelling in the false positives.

Corpus Performance

- Metrics are not bad, but not great either.
- Inspection of the false positives and negatives reveals:
  - prevalence of ‘composite articles’ in the false negatives, where only one or two bits are philosophical; and
  - prevalence of ‘edge cases’ and mistaken labelling in the false positives.
- Conclusion: the performance of the classifiers is being limited by the quality of the labels.

Corpus Analysis

Corpus Analysis

- Many methods were used at this stage (see project report)

Corpus Analysis

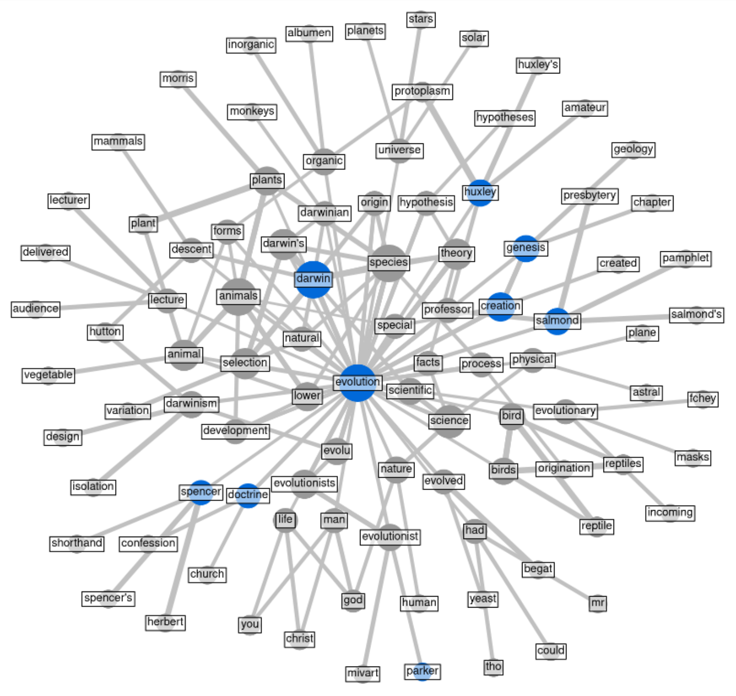
- Many methods were used at this stage (see project report)
- The following two slides contain:

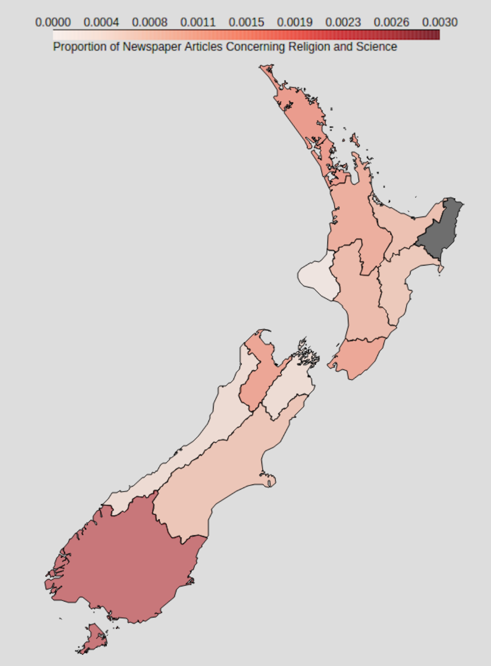
Corpus Analysis

- Many methods were used at this stage (see project report)
- The following two slides contain:
  1. an example of a telling co-occurrence network, and

Corpus Analysis

- Many methods were used at this stage (see project report)
- The following two slides contain:
  1. an example of a telling co-occurrence network, and
  2. a choropleth revealing prominence of religion and science discourse in different regions.


stars
planets
albumen
solar
inorganic
huxley's
morris
protoplasm
monkeys
amateur
hypotheses
mammals
universe
geology
organic
plants
huxley
lecturer
origin
hypothesis
presbytery
chapter
darwinian
plant
forms
darwin's
genesis
delivered
descent
theory
species
created
pamphlet
darwin
animals
lecture
professor
creation
audience
salmond
salmond's
special
hutton
natural
plane
animal
selection
facts
process
physical
vegetable
evolution
scientific
astral
lower
fchey
darwinism
variation
evolutionary
bird
science
development
design
masks
evolu
nature
reptiles
birds
origination
evolved
isolation
evolutionists
spencer
doctrine
incoming
life
man
had
reptile
evolutionist
shorthand
confession
begat
church
spencer's
god
yeast
mr
human
herbert
you
christ
could
mivart
parker
tho


0.0000 0.0004 0.0008 0.0011 0.0015 0.0019 0.0023 0.0026 0.0030
Proportion of Newspaper Articles Concerning Religion and Science

UPSHOT

Positives

Positives

1. Single investigator, self-labelled corpus production using METS/ALTO digitised newspaper files is feasible.

Positives

1. Single investigator, self-labelled corpus production using METS/ALTO digitised newspaper files is feasible.
   - Generalisability: METS/ALTO is the standard for newspaper digitisation, so the same methods could be applied in other countries.

Positives

1. Single investigator, self-labelled corpus production using METS/ALTO digitised newspaper files is feasible.
   - Generalisability: METS/ALTO is the standard for newspaper digitisation, so the same methods could be applied in other countries.
2. The corpus produced at the corpus construction stage shows potential for research into early New Zealand philosophy.

Shortcomings

Shortcomings

1. Problem 1:

Shortcomings

1. Problem 1:
   - Many articles at the time were made up of lots of distinct bits (especially editorials).

Shortcomings

1. Problem 1:
    - Many articles at the time were made up of lots of distinct bits (especially editorials).
    - Since the classifier loses many of these articles, the resulting corpus is not fully representative of philosophical discourse in early NZ newspapers.

Shortcomings

1. Problem 1:
    - Many articles at the time were made up of lots of distinct bits (especially editorials).
    - Since the classifier loses many of these articles, the resulting corpus is not fully representative of philosophical discourse in early NZ newspapers.
    - Possible solution: label text blocks rather than articles.

Shortcomings

1. Problem 1:
    - Many articles at the time were made up of lots of distinct bits (especially editorials).
    - Since the classifier loses many of these articles, the resulting corpus is not fully representative of philosophical discourse in early NZ newspapers.
    - Possible solution: label text blocks rather than articles.
2. Problem 2:

Shortcomings

1. Problem 1:
   - Many articles at the time were made up of lots of distinct bits (especially editorials).
   - Since the classifier loses many of these articles, the resulting corpus is not fully representative of philosophical discourse in early NZ newspapers.
   - Possible solution: label text blocks rather than articles.
2. Problem 2:
   - Labelling criteria were insufficiently clear.

Shortcomings

1. Problem 1:
    - Many articles at the time were made up of lots of distinct bits (especially editorials).
    - Since the classifier loses many of these articles, the resulting corpus is not fully representative of philosophical discourse in early NZ newspapers.
    - Possible solution: label text blocks rather than articles.
2. Problem 2:
    - Labelling criteria were insufficiently clear.
    - Better labelling might improve classifier performance.

Shortcomings

1. Problem 1:
    - Many articles at the time were made up of lots of distinct bits (especially editorials).
    - Since the classifier loses many of these articles, the resulting corpus is not fully representative of philosophical discourse in early NZ newspapers.
    - Possible solution: label text blocks rather than articles.
2. Problem 2:
    - Labelling criteria were insufficiently clear.
    - Better labelling might improve classifier performance.
    - Possible alternative: start with easier distinctions (e.g. is the article a report of a public lecture?), then move to subject matter distinctions.

Overview

Overview

1. Problem:

Overview

1. Problem:
    - to gain insight into philosophical writing in early New Zealand newspapers

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results:

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results:
   - An interesting collection of articles for digital humanities research,

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results:
   - An interesting collection of articles for digital humanities research,
   - with indications that labelling could be improved,

Overview

1. Problem:
    - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
    - corpus construction via self-labelling and supervised learning,
    - corpus analysis with, e.g., co-occurrence networks.
3. Results:
    - An interesting collection of articles for digital humanities research,
    - with indications that labelling could be improved,
    - and ability to reveal features of philosophical discourse about relationship between religious belief and then-new scientific ideas.

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results:
   - An interesting collection of articles for digital humanities research,
   - with indications that labelling could be improved,
   - and ability to reveal features of philosophical discourse about relationship between religious belief and then-new scientific ideas.
4. Upshot:

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results:
   - An interesting collection of articles for digital humanities research,
   - with indications that labelling could be improved,
   - and ability to reveal features of philosophical discourse about relationship between religious belief and then-new scientific ideas.
4. Upshot:
   - a method applicable for many humanities research questions,

Overview

1. Problem:
   - to gain insight into philosophical writing in early New Zealand newspapers
2. Method:
   - corpus construction via self-labelling and supervised learning,
   - corpus analysis with, e.g., co-occurrence networks.
3. Results:
   - An interesting collection of articles for digital humanities research,
   - with indications that labelling could be improved,
   - and ability to reveal features of philosophical discourse about relationship between religious belief and then-new scientific ideas.
4. Upshot:
   - a method applicable for many humanities research questions,
   - but with shortcomings to be aware of.

Links

- Dashboard: `nz-newspaper-philosophy.herokuapp.com`
- GitHub (full project): `github.com/JoshuaDavidBlack/NPOD_Philosophy`
- GitHub (dashboard): `github.com/JoshuaDavidBlack/NPOD_Philosophy_Heroku`