THE STATISTICS OF TOPIC MODELLING

A thesis submitted in partial fulfilment of the requirements for

the Degree

of **Master of Science in Statistics**

in the Department of Mathematics and Statistics

by **Rebecca Katherine Abey**

University of Canterbury

2015

Contents

List of Illustrations

Figures

Tables

Acknowledgements

This research project would not be possible without the collaboration between the Department of Mathematics and Statistics, and the Department of Digital Humanities, at the University of Canterbury. As my thesis borders between both statistics and digital humanities working with both parties has been extremely helpful and I can only hope that my research is the beginning of many future collaborations between the two departments.

I would first like to thank Jennifer Brown and James Smithies for their supervision. Jennifer provided me with excellent feedback on getting the structure of my thesis right and had good ideas of what to add in where. James provided me with good feedback on tailoring this to the humanities, and for providing good discussions on what to include in this project.

I would also like to thank David Bourget and David Chalmers from PhilPapers for allowing me the use of their extremely large database to analyse. I hope this project provides some interesting findings.

Many thanks to Lauren for helping me get my coding in R working. I am not the strongest code writer out there but Lauren helped me fix any errors that came up and helped check that everything was in working order.

A huge thank you to Tim David and François Bissey for introducing me to the idea of using a supercomputer for part of my research. They were extremely helpful and while I did not end up using a supercomputer I still thank them for their time and effort in explaining how the process works.

And finally, to Richard who supported me through the entire year of writing my thesis, and who put up with my constant barrage of questions.

Abstract

This research project aims to provide a clear and concise guide to latent dirichlet allocation which is a form of topic modelling. The aim is to help researchers who do not have a strong background in mathematics or statistics to feel comfortable with using topic modelling in their work. In order to achieve this, the thesis provides a step-by-step explanation of how topic modelling works. A range of tools that can be used to perform a topic model analysis are also described. The first chapter gives an explanation of how topic modelling, and (more specifically), latent dirichlet allocation works; it offers a very basic explanation and then provides an easy to follow mathematical explanation. The second chapter explains how to perform a topic model analysis; this is done through an explanation of each step used to run a topic model analysis, starting from the type of dataset through to the software packages available to use. The third section provides an example topic model analysis, based on the Philpapers dataset. The final section provides a discussion on the highlights of each chapter and areas for further research.

A Selected Glossary

Anomaly Detection – The process of detecting data points which do not fit within an expected pattern or other items within a dataset.

Association Rule Learning – A process that extracts if-then statements from a set of data. It looks for relationships such as if x then y.

Attribute – A piece of information that states the properties of a field or tag within a database.

Classification – The task of predicting the label or class of a given data point with unknown labels.

Clustering – A process in data mining where data points are separated into particular groups.

Conjugate – Any of a set of numbers that satisfy the same irreducible polynomial.

Corpus – A collection of written texts.

Correlated – Having a mutual relationship or connection where one thing depends on another.

Exponential Family Distribution – A set of probability distributions of a specific form

Generative Probabilistic Process – A process in which observable data is generated using random probabilities.

GUI – Graphical User Interface, a computer interface that allows users to connect to the interface using graphical icons.

Humanities – Subjects that study the human experience such as philosophy, literature, history, religion, languages, art and classics.

Iterations – Repetitions of a process until a desired result is achieved.

Parallel Computing – A form of computation where several calculations are performed simultaneously.

Parameter – A constant or variable term in a function that determines the specific form of the function, but not its general nature.

Posterior Inference – An inference made after the relevant information is taken into account.

Probabilistic Model – Statistical model that provides an estimate based on historical data of the probability of an event occurring again.

Regression – A measure of the relationship between the mean of a variable and corresponding values of other variables.

Salmonella Pulse-Field Gel Electrophoresis – A method of detecting salmonella in patients.

Simplex – A space on which a series of points are found.

Sparsity – How spread out or scattered a distribution is. For example, how many beetles in a distribution over beetles tend to have high positive probability.

Summarisation – A process for finding a compact description of a dataset.

Tweets – Post on social media website twitter, consisting of up to 140 characters.

Introduction

Digital Humanities

Digital humanities is the humanities in the digital age (Piez, 2013). It combines the traditional humanities subjects such as philosophy, history, art, linguistics, literature, archaeology and music with tools from disciplines such as data mining, statistics, text mining, digital mapping and information retrieval (Liu, 2013). There is much debate about the precise definition of digital humanities (Svensson, 2012). Two goals commonly described for what the digital humanities should be are as follows:- the first of these is to study digital media and the cultures and cultural impacts of digital media and to design and make digital media (Piez, 2013); and the second is to bring the Humanities into the digital age through digitisation of text, and using computational tools to analyse these texts. Digitisation is the transformation of media such as text, sounds, images and data from electronic devices into computer files (Brynjolfsson & McAfee, 2014). In recent years projects such as Google Books have set out to digitise large libraries of books to allow access to users around the globe and to preserve information.

Large datasets of digitised content offer significant opportunities for humanities researchers. To realise these opportunities, advances in statistical analysis to account for the complex nature of the content are needed. An understanding of these methods and their application by researchers in the digital humanities is also required. The most significant area of change in statistics relevant to analyses of large datasets of digitised content is the field of data mining. Data mining is the collective term for exploring large datasets using various techniques to find patterns in data. It incorporates many fields of academia including machine learning, statistics and database systems. The aim of data mining is to analyse large datasets consisting of thousands to millions of attributes and data points (Zaki & Meira, 2014). Data mining uses six types of analysis: clustering, classification, regression,

anomaly detection, association rule learning, and summarisation (Fayyad, Piatetsky-Shapiro, & Smyth, 1996).

Text mining or text analysis is one specific area of data mining. It not only covers analysis of large volumes of text such as novels, academic journal articles and newspaper clippings, it also covers emails, tweets and blog posts. Any type of text file can be used in text mining (Dean, 2014). There are several techniques within the area of text mining for analysing text. One of the more recent developments in this area is topic modelling. This is a new area of research and one specifically designed for analysis of large datasets of digitised content.

Topic Modelling

Topic modelling is a form of text analysis used to explore relationships between words within a document where the words are grouped together to form topics. The earliest work on topic modelling is by Papadimitriou, Tamaki, Raghavan, and Vempala (1998), and Hofmann (1999). The technique was further developed by Blei, Ng, and Jordan (2003). There are a variety of different methods for topic modelling, using different sampling algorithms for word selection and topic creation. Examples of topic models include latent semantic analysis. This method is the most basic and looks at the frequency of words within a document and creates topics based on the frequencies of words occurring in each document. (Steyvers & Griffiths, 2007). Latent dirichlet allocation is another basic topic model. It groups words together based on how likely they are to appear in a document together (Blei et al., 2003). Correlated topic models explore the correlation of words to other words within a document. Topics are created based on the strength of correlations between words (Blei & Lafferty, 2007). Explicit semantic analysis adds words from a document to a matrix based on frequency and creates topics based on the frequency of co-occurrence between words (Egozi, Markovitch, & Gabrilovich, 2011). Topic modelling can be used in many different academic domains including both

the humanities and sciences. Anyone using text documents in their research could have a potential use for topic modelling.

The amount of data available on the Internet is vast and will only increase over time. Topic modelling provides an easy way to process large amounts of information efficiently. It also allows for individual search topics to be discovered. Edward Y. Chang is a research director at Google and is currently working on implementing topic modelling into Google's search engines. This will allow for a better exploration of Google's databases (Dickman, 2014).

A recent example of the use of topic modelling in science includes the work on topic modelling for Cluster Analysis of Large Biological and Medical Datasets, (Zhao, Zou, & Chen, 2014). In their work, they assessed whether topic modelling is useful for biology and medicine. They analysed three different datasets for Salmonella pulse-field gel electrophoresis, lung cancer, and breast cancer and compared other data mining techniques to topic modelling. Their goal was to assess whether topic modelling gave them a better answer to a particular problem they were trying to solve for each dataset. The analysis found that topic modelling gave them a better result than the other data mining techniques. They concluded that topic modelling is beneficial for sorting through large sets of medical data with slightly better precision than other data mining methods (Zhao et al., 2014).

While topic modelling was created for computer science and analysing scientific data, it is also a useful tool for use in the digital humanities. One interesting area of analysis is poetry. Poetry uses very few words but can have several motifs and themes underlying them. As part of her PhD dissertation on literature, Lisa Marie Rhody (2012) analysed highly figurative poems using topic modelling. A particular example she used was analysing Anne Sexton's *Starry Night*. Rhody found topics were formed that made sense on an intuitive level; words such as 'death', 'life', 'heart', 'dead', 'man' and 'soul' were grouped together. However, none of the topics produced consisted of figurative language topics. For example, the word 'darkness' could be used figuratively to describe uncomfortable or terrifying surroundings rather than the absence of light. Her research suggests that

topic modelling algorithms can produce similar topics to humans on an intuitive level, but on a metaphorical level there is not much interpretation. It suggests that while humans can analyse and interpret poetry to discuss themes, algorithms for topic modelling can produce similar themes but some of the more metaphorical themes may be lacking (Rhody, 2012). In other words, humans can apply emotions to create topics that are based on more of an emotional response rather than just words that relate together in a semantic way.

Historian, Cameron Blevins (2010) ran a topic model analysis on Martha Ballard's diary. Martha Ballard was an American midwife and healer who kept a diary of her daily life from 1785-1812. Laurel Ulrich (2010) published a book entitled *A Midwife's Tale*, exploring the themes of Ballard's life through her diary entries. Blevins then analysed Ballard's diaries using topic modelling to compare how the algorithm fared against Ulrich's work. The results showed that the topic modelling analysis produced much the same results as Ulrich's research did. This suggests that historians can utilise topic modelling for quickly analysing the contents of documents.

Philpapers

Philpapers is an online database of philosophy literature. It was founded in 2009 by David Bourget and David Chalmers. The Philpapers database contains journal articles, books, personal websites and open access archives. It is maintained by a large community of philosophy professionals and students; at the time of this writing there were over 5000 contributors. The database contains over one million research books and articles (Bourget & Chalmers, 2015). This thesis will be carrying out a topic modelling analysis using the PhilPapers dataset. The dataset consists of the bibliographic information contained within the PhilPapers database and will be described in more detail in a later section. The PhilPapers database can be viewed online at http://philpapers.org.

Research Objectives

This thesis aims to investigate topic modelling in the form of latent dirichlet allocation to provide a guide for its use in digital humanities research. It aims to fill a gap in knowledge that currently exists within the literature. This thesis aims to inform digital humanities researchers how latent dirichlet allocation works as it is an important tool for research.

The objectives of this thesis are:

- To provide an explanation of what latent dirichlet allocation is and the mathematics behind it. Latent dirichlet allocation was chosen because it is the most commonly used method of topic modelling and it is the easiest to understand.
- To present an overview of software packages available and methods of conducting a topic modelling analysis.
- To investigate the PhilPapers database to provide them with analyses about the database and possibly provide them with ideas for further research.
- To produce an example of a topic model analysis to show how latent dirichlet allocation can be used to analyse large volumes of text.
- To evaluate topic modelling as a method, to explore limitations, and to decide what role topic modelling might play as part of the digital humanities research tools.

Thesis Structure

Chapter Two offers a non-mathematical explanation of what topic modelling is and how latent dirichlet allocation (LDA) works. It then provides a comprehensive explanation of the mathematics behind LDA, followed by a discussion on what a topic is.

Chapter Three provides a description of how to conduct a topic modelling analysis. It begins with a discussion about the type of dataset to use and how the data should be treated. This is followed by an overview of software packages to use for topic modelling and an evaluation of each one. It concludes with suggestions for analysing the data.

In chapter Four an example of a topic model analysis on the PhilPapers dataset is presented. This chapter provides an overview of the dataset and how it was treated, the software used for the analysis, examples of different methods of analysing and presenting the data, and some conclusions about these findings.

The final chapter is a discussion about what topic modelling is useful for and how it should be used for research in various fields, including humanities subjects.

Literature Review

The current literature agrees that topic modelling is a useful tool for use in the digital humanities (D. Blei, 2012, pp. 8-11; Goldstone, 2014; Underwood, 2012a). Along with this clear statement is the message that an understanding of how the topic modelling algorithms work is needed in order to use the tool successfully (D. Blei, 2012; Schmidt, 2012; Weingart, 2012). There is however, a lack of published literature on understanding topic modelling algorithms for those who may not have a strong mathematical background. The Journal of Digital Humanities offers some published articles about topic modelling in the digital humanities, however, it is a new type of journal where submissions have a two-step process. Blog posts are voted for by others in the digital humanities community. Then, an editorial committee peer reviews them, rather than going through the rigorous peer review process.

Digital humanities is a recent field of research and the term itself is still under debate (Vanhoutte, Nyhan, & Terras, 2013). There are many different articles arguing the definition of digital

humanities. Kirschenbaum (2012) provides an overview of different interpretations of digital humanities. While there is a vast amount of literature outlining what the digital humanities is. There are two main focuses for digital humanities; the first is bringing the humanities into the digital age (Liu, 2004, 2012, 2013) and the second is studying the digital media and cultural aspects of digital media (Piez, 2013). There is also a strong focus within the literature on representation of information using technology (Flanders, 2009; Liu, 2004). Rosenbloom (2010) diverts from these discussions by taking a different approach in which he discusses what scientific domain the digital humanities belongs to. He argues for digital humanities to be in the domain of computer science.

Blei (2012), Templeton (2011) and Weingart (2012) provide a discussion of the role topic modelling plays in digital humanities. The literature agrees that knowing how the topic models are produced is important (Schmidt, 2012; Weingart, 2011, 2012). However, the literature targeted to digital humanities researchers consists mainly of blog posts (M. L. Jockers, 2011; Templeton, 2011; Underwood, 2012a). Some lack any mathematical detail and explain the concepts behind topic modelling in literary terms (Chen, 2011; M. L. Jockers, 2011). Others provide a mathematical explanation but tend to avoid in-depth mathematical explanations (Riddell; Underwood, 2012a).

There are a large number of blog posts discussing examples of applications of topic modelling to digital humanities. This includes but is not limited to areas such as literature (Rhody, 2012; Underwood, 2012b), history (Blevins, 2010), media studies (D. Kim & Oh, 2011) and archaeology (S Graham, 2012). However, there are very few published journal articles outlining ways of applying topic modelling to the digital humanities (Goldstone, 2014; Mimno, 2012; Underwood, 2011).

There is a large amount of digital humanities literature discussing how to interpret topics for digital humanities research. This includes explanations on the semantic meanings of topics (Schmidt, 2012; Song, Pan, Liu, Zhou, & Qian, 2009; Underwood, 2012c) and strategies for how to interpret the results for analysis (Chang, Gerrish, Wang, Boyd-graber, & Blei, 2009; Posner, 2012). There are also a large selection of papers outlining the mathematics of topic models, with Heinrich (2009) and Blei

(2009) providing extensive mathematical explanations of generic topic models. There are different types of topic model so the majority of the papers focus on one or two types of topic models, Blei and Lafferty (2009) provide a mathematical overview of three of these topic models.

The most basic types of topic models are latent semantic analysis and latent dirichlet allocation and, there are a large collection of papers discussing these models (D. M. Blei, 2012; Blei et al., 2003; Grün, Hornik, & Grün, 2014; Steyvers & Griffiths, 2007). While these papers outline latent dirichlet allocation extensively, a solid understanding of mathematics and statistics is required to understand them. LDA was created by David Blei and the majority of literature explaining it is written by Blei with others (D. Blei, 2012; Blei, 2009; D. M. Blei, 2012; Blei et al., 2003). Latent dirichlet allocation can be developed into more refined versions to suit specific purposes or to increase the precision of the model (Asuncion, Welling, Smyth, & Teh, 2009). A large number of papers have been produced describing various alterations and creations of new topic modelling algorithms (Blei & Lafferty, 2006, 2007; Doyle & Elkan, 2009; Mcauliffe & Blei, 2008; Rosen-Zvi, Griffiths, Steyvers, & Smyth, 2004).

The mathematics of hierarchical dirichlet processes to alter different types of topic models are discussed in a variety of papers. One specific area of alteration being discussed is the changing of hierarchical dirichlet processes, different methods of developing these have been created and explained by Teh, Jordan, Beal, and Blei (2006), Wallach, Minmo, and McCallum (2009) and Blei, Griffiths, and Jordan (2010).

Topic modelling for science applications has been a topic of discussion in many academic journals (G. Kim, Park, & Jang, 2014). This occurs in various fields of science such as biology (Zhao et al., 2014), general science (Blei & Lafferty, 2007), climate change (G. Kim et al., 2014) and computer science (Dickman, 2014).

What is Topic Modelling?

Topic modelling is a generic term for several different algorithms that create topics based on documents. The topic modelling algorithm used in this thesis was latent dirichlet allocation, therefore all explanations are of latent dirichlet allocation. It can be used in a variety of ways. To get a good understanding of how topic modelling works some simple examples of topic modelling will be discussed first. Topic models allow for discovering topic in a corpus. Often researchers will have a large collection of texts to read through. To get a good idea of the contents of a corpus, a topic model can provide a good overview.

The evolution of topics over time can be modelled to explore trends in language and ideas. The graph in Figure 1 shows four topics discovered in a selection of science fiction novels. The novels chosen for the analysis are listed in Appendix A. The usage of each topic has been modelled over time.

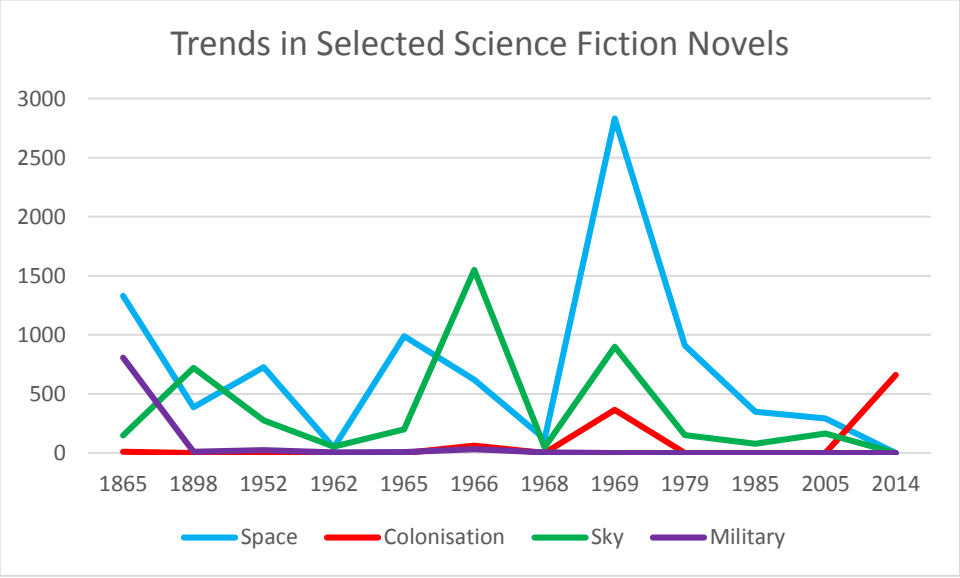

Figure 1. Graph of Trends in Selected Science Fiction Novels over time

Connections between topics can be modelled through the use of word clouds (M. Jockers, 2014). The word cloud below was created using a sample of newspaper and magazine articles from various

publications, all written by Professor Brian Cox. Professor Brian Cox is a professor of astronomy and physics at the University of Manchester (Cox & Forshaw, 2010). A basic topic model analysis on the articles, created a word cloud (Figure 2) based on the topics.

Figure 2. An illustration of a word cloud

This word cloud shows the most prevalent words are universe, billion, higgs and space. Word clouds represent the relative frequency with which words appear in a given text. The larger the word, the more frequently it has been used in the text (Feinberg, 2014). There is some debate over whether or not word clouds are actually useful tools. Most people tend to agree they do not have a large degree of research value (Harris, 2011). They can be used to get an overall sense of which words are the most prevalent throughout the topics being modelled.

Another way to show connections between topics is with networks (Weingart, 2011). Figure three shows a small network of topics from the same topic model analysis and how they can relate to each other. This diagram was created using four of the topics from the PhilPapers dataset.

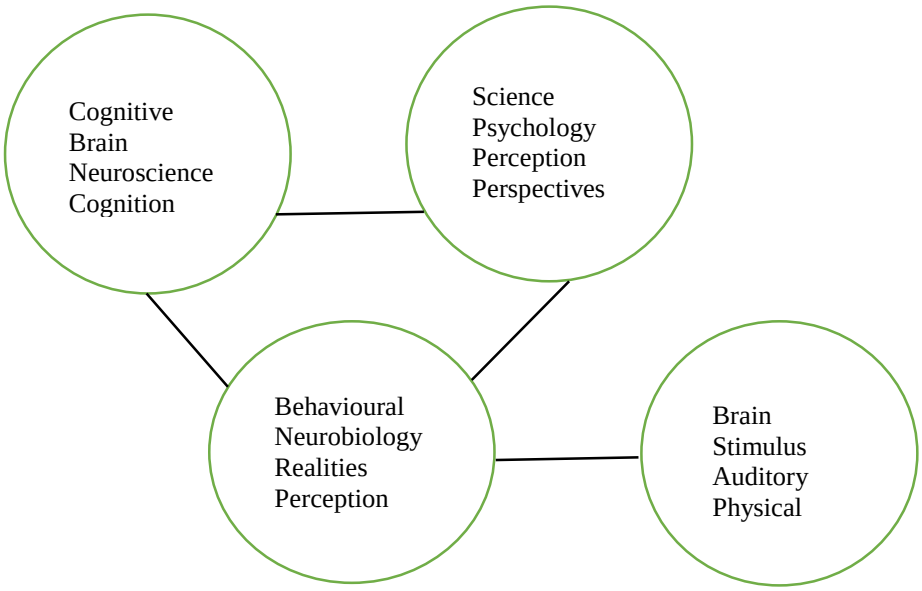

Figure 3. A network diagram of a selection of topics

Topic modelling can be used to annotate images (Blei & Lafferty, 2009). When given an image the topic modelling algorithm can return a list of words relating to that image. When run through a topic modelling program figure 4 could produce a list of words such as cat, door, house, feline, black.

Figure 4. Image of a cat on a garden fence.

A Step-by-Step Introduction

There are many examples of why topic models might be useful for humanities research. It can be used as a heuristic tool to search through large corpora in a small amount of time. The example below looks at the works of H. P. Lovecraft to show a possible use for topic models as well as explaining the topic modelling process at the same time. Literary analysis is one area of the humanities that can benefit from topic modelling. The work of specific authors can be analysed to search for patterns and themes in the texts. Take for example, the works of H. P. Lovecraft. Lovecraft's stories have been used to illustrate the use of topic modelling. The first thing done in topic modelling analysis is to cut every word from Lovecraft's story into single words. Any punctuation marks and words not worth analysing such as 'the', 'to', and 'of' are discarded. The remaining words are placed into a large container. The container is then shaken up, as the order of words does not matter.

It is inferred from the data that there must be seven topics in total being used by H. P. Lovecraft. Seven smaller containers are placed around the larger container. One at a time a word is removed from the larger container and placed into one of the seven smaller containers. These seven smaller containers are the topics. This step is repeated until all of the words have been placed into the seven containers.

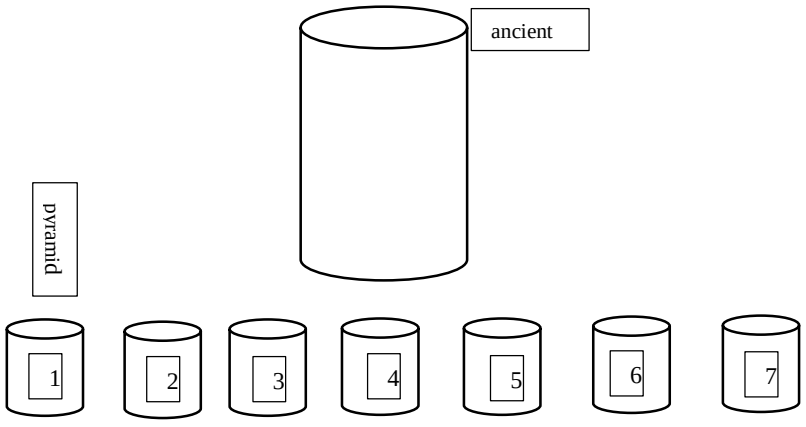

Figure 5. An example of words being taken from the large container and sorted into smaller containers.

The containers are then examined and appear to be random groupings of words. The next step is to compare the words with the complete texts. Each word of the collected stories of H. P. Lovecraft is examined and compared to the words inside the containers. The first words selected from the book is

‘magic’. A decision is made about how much of topic one, as it appears in container number one, is present in the book. Simultaneously, the container that contains the most ‘magic’ is decided. The decision to do this is based on the semantic similarity and the distance between semantically similar words. This task is carried out for each word in the text, swapping words into other containers until they appear to form somewhat coherent topics. This process is repeated until all seven topics are represented accurately in its respective container (M. Jockers, 2014, pp. 163-165).

The topics in this case are:

1. pyramid sphinx khephren temple rope abdul egypt rock pyramids cairo descent dream reis gateway plateau arabs gizeh guide heads

2. great time found long things made place left thought strange stone half years men back ancient curious began day

3. tomb door vault vacant home aged ye ll box coffin daughter iron abandoned speak slope fastened portal slab family

4. crumbling expressed noon farm nerves sharp pale school rolled impressed basis beginning spent caught orders bewildered judge winter standing

5. lake land camp danforth feet city antarctic ice sea mountains sculptures world range vast specimens snow earth mountain plane

6. whateley armitage wilbur dunwich hill glen dr ye night earth whippoorwills whateleys hills frye aout bishop cattle boy house

7. ward curwen willett charles dr mr ye joseph house doctor youth pawtuxet allen room man library town weeden father

These topics can then be analysed as they show patterns in the texts. Comparisons between the topic model and a researcher’s own ideas can be discussed.

Latent Dirichlet

Latent dirichlet allocation is a Bayesian inference model created by David Blei (2003). It associates each document with a probability distribution over topics, where topics are probability distributions over words.

A probability distribution is an equation that links every possible outcome of a random variable with the probability that the event will occur. For example, in a fair coin flip, there are two possible outcomes:- heads or tails. Heads is represented by 1 and tails by 0. There is also a random variable labelled X, with two possible outcomes represented by X:0 or X:1. The probability distribution of X or P(X=x), is:

P(X=0) = 0.5

P(X=1) = 0.5

There are two probability distributions used in topic modelling, the first being the probability distribution over topics. This is the group of topics that are most likely to be used in a specific document. An example of such is:

$\mu'_{1topic1} = 0.25$

$\mu'_{1topic2} = 0.25$

$\mu'_{1topic3} = 0.25$

$\mu'_{1topic4} = 0.25$

The second probability distribution is the probability distribution over words. These are the words most likely to be found in a specific topic. For example, if I was looking at astronomy papers the probability distribution of one topic could be:

$\lambda'_{1elliptical} = 0.35$

$\lambda'_{1orbit} = 0.25$

$\lambda'_{1satellite} = 0.3$

A Bayesian inference model is a method used to calculate the probability of an event occurring, given the observed data. It combines common sense assumptions and the outcomes of previous related events. The model runs through several iterations of assigning words to topics to improve the model. The more iterations done in this way, the more the model will accurately reflect the topics present in a corpus (Dickman, 2014). The data is treated as observations arising from a generative probabilistic process that includes hidden variables. A generative probabilistic process is a process for randomly generating observable data. Hidden or latent variables are not directly observed but are inferred using posterior inference. Posterior inference is where the hidden variables are estimated based on relevant background evidence. In the case of latent dirichlet allocation, topics are hidden variables and are inferred from the words in the documents (Blei, 2009). An important point of latent dirichlet allocation is that documents consist of multiple topics. Latent dirichlet allocation is used to decide which topics are being discussed in a specific document, based on the analysis of a set of documents already observed (Steyvers & Griffiths, 2007).

The Generative Model

We want to express latent dirichlet allocation as a generative probabilistic process. To do this an assumption must be made that there are a number of topics related to a collection of documents. The example below shows how the generative process works. It is on a smaller scale than it would be if done on a computer. Generally, hundreds to thousands of documents would be analysed. In this example, there are two paragraphs of text. Each paragraph is treated as though it is a single document.

The passages used are from Lewis Carroll's *Alice's Adventures in Wonderland*. For the purpose of this example, the whole text of *Alice's Adventures in Wonderland* is the corpus and the paragraphs are

documents. A selection of words from the text have been highlighted. It is important to remember that creating topics by hand is different to creating topics using topic modelling software. Judgments will be made differently in each instance. The algorithm searches through every word in a document and assigns each word to a topic based on the probability of it occurring together with other words in a document. To illustrate this, the passage below has a selection of words highlighted where the different colours represent different topics.

"Alice was not a bit hurt, and she jumped up on to her feet in a moment: she looked up, but it was all dark overhead; before her was another long passage, and the White Rabbit was still in sight, hurrying down it. There was not a moment to be lost: away went Alice like the wind, and was just in time to hear it say, as it turned a corner, 'Oh my ears and whiskers, how late it's getting!' She was close behind it when she turned the corner, but the Rabbit was no longer to be seen: she found herself in a long, low hall, which was lit up by a row of lamps hanging from the roof."
(From Chapter One, Down the Rabbit Hole)

Using this example, we assume there are a number of topics relating to the document collection. Three of these topics are listed below. There are also a number of topics that are not listed but also exist. For example, there could be an additional 47 topics used throughout *Alice's Adventures in Wonderland*. Each of these topics is a distribution over terms in the vocabulary. It also needs to be assumed there is a fixed vocabulary, in this instance it is every word in the text. Each topic is a distribution over the fixed vocabulary. Individual topics contain different words with different frequencies of occurrence. For example, if it is assumed the words are in order of their frequency of occurrence, as listed below, there exists a topic such as feet, sight, hear… The algorithm for latent dirichlet allocation will do this for each document in the selected corpus to create a list of words that relate to each other.

From the passage, the highlighted words can be sorted into topics. Some topics may stand out quite strongly while the meaning of others may be more subtle; this can be especially true when analysing

novels or poetry rather than scientific documents. For the purposes of this example the topics have been labelled; ‘body’, ‘time’, ‘light’, and ‘clothing’.

Feet	0.01	Moment	0.02	Dark	0.01
Sight	0.01	Still	0.01	Lit	0.01
Hear	0.01	Time	0.01	Lamps	0.01
…		…		…	

Figure 6. The top three words for each topic

Every topic contains a probability of how likely a word is to appear in that specific topic. Every word in the vocabulary is assigned a probability of occurrence within each topic, for each topic. A large number of these words will have probability close to zero. Figure six is a list of the top three words for each topic in the example paragraph. There can be times where different topics show the same word as having the highest probability. For example, a topic about rivers may use the word ‘bank’, and a topic about finance may also use the word ‘bank’. To reiterate the idea that different documents exhibit different topics a second example has been created using a second paragraph of *Alice’s Adventures in Wonderland*.

“After a time she heard a little pattering of feet in the distance, and she hastily dried her eyes to see what was coming. It was the White Rabbit returning, splendidly dressed, with a pair of white kid gloves in one hand and a large fan in the other: he came trotting along in a great hurry, muttering to himself as he came, “Oh! The Duchess, the Duchess! Oh! won’t she be savage if I’ve kept her waiting!’ Alice felt so desperate that she was ready to ask help of any one, so, when the Rabbit came near her, she began, in a low, timid voice, ‘If you please, sir—‘. The Rabbit started violently, dropped the white kid gloves and the fan, and scurried away into the darkness as hard as he could go.” (From Chapter Two, The Pool of Tears)

The above paragraph consists of the green and blue topics from the previous example, and a new topic highlighted in pink. To create a generative process for each document, a distribution over topics must

be selected. A distribution over topics is a distribution over the 50 elements as discussed earlier. Assuming *Alice's Adventures in Wonderland* consists of 50 topics, this distribution has 50 possible values. For simplicity and as a visual aid, each of these values is colour coded by its topic. From the first example paragraph, the distribution over topics has green with some probability, yellow with some probability, and blue with some probability. This is the first step to generating a document.

The next step is to repeatedly draw a colour from this distribution. Assume there is a colour wheel with each of the 50 different coloured topics on it. The wheel is spun and lands on green. It is then checked which topic is referred to as the green topic and a word is chosen from that topic. The wheel is spun a second time and lands on blue. A word from the blue distribution is selected. This process continues until a distribution of topics has been created for a document. An example of what this looks like is illustrated in figure seven.

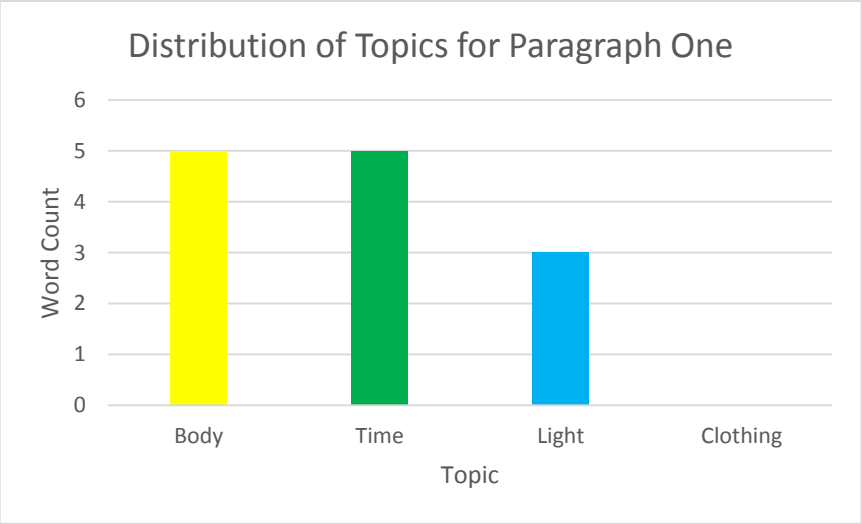

Figure 7. Graph of the distribution of topics for paragraph one.

The process of choosing a colour is representative of a topic. Selecting a word from the corresponding distribution is implicitly assuming the order of words does not matter. The graph of the distribution of the first paragraph shows it is made up of topics body, time and light. The topics body and time have equal weighting whereas the topic light appears less frequently in the text. This example shows the generative process for a single document. The same process would be repeated for another document. Each document will have its own distribution over topics.

Figure eight is an example of a possible distribution over topic for the second paragraph of *Alice's Adventures in Wonderland*. The topic clothing appears more frequently than the other topics.

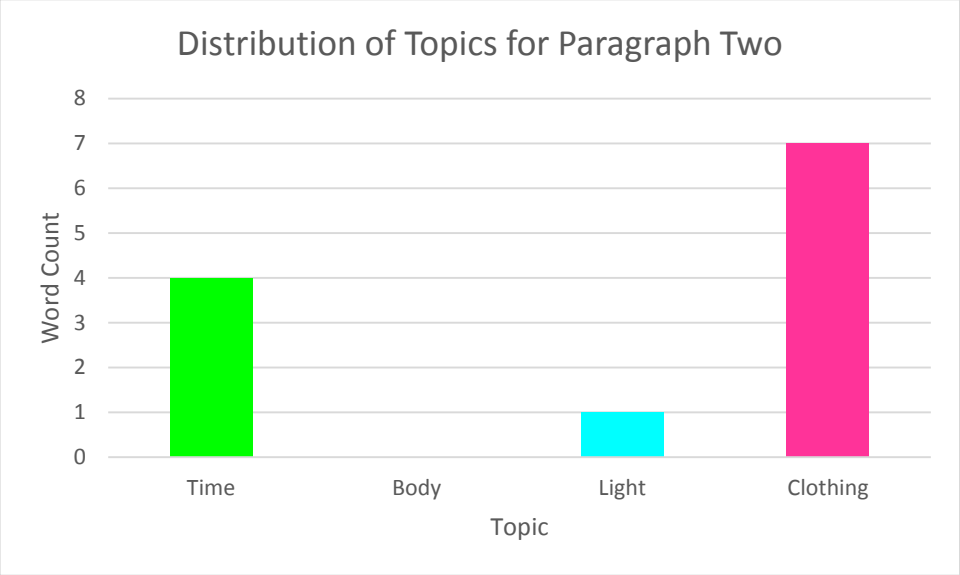

Figure 8. A graph of the distribution of topics for paragraph two.

The Posterior Distribution

The generative model provides a general idea of how topic modelling works. However, we only observe the documents and do not have access to the full complete list of topics and their distributions. The goal is to infer the underlying topic structure, using the observable documents in a given corpus (Blei & Lafferty, 2009). Two things are needed to be able to infer the underlying topic structure. First, the topics that generated the documents are found. Secondly, for each document, the distribution over topics associated with that document must be found. The posterior distribution is a conditional distribution of all the hidden variables based on the observations, which in this case are the words in the documents. The next step is to find an algorithm that will compute the posterior distribution.

David Blei (2003) has created a graphical model to represent how each variable relates to other variables. Figure nine is a slightly modified version of Blei's (2003) diagram.

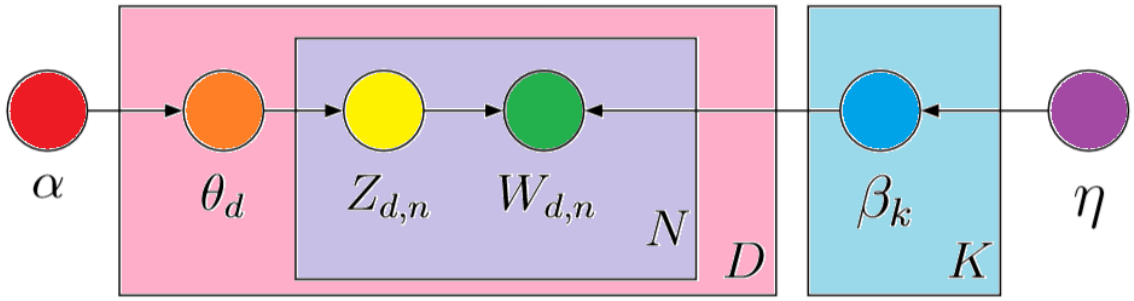

Figure 9. Graphical model of the parameters of a dirichlet distribution

Each piece of the diagram is a random variable. The circles are nodes and the rectangles are plates. The red node, α, is a Dirichlet parameter and will be explained in more detail later on. The largest rectangle, shaded in pink, is the document plate, *D* and it represents the corpus. Within the corpus is, θ_d, in orange, which represents the topic proportions for each document. In the context of the example earlier that used a paragraph from *Alice's Adventures in Wonderland*, θ_d would be the colours, (-for example, the green, yellow and blue topics). There is one of these for each document, *D*. The lavender plate, *N*, refers to each word in a document, *D*. For each word, *N*, is the yellow node, $Z_{d,n}$, the topic assignment. In the context of the *Alice's Adventures in Wonderland* example, again $Z_{d,n}$ represents the colour drawn from the colour wheel. The arrows in the diagram show that $Z_{d,n}$ depends on θ because it is drawn from a distribution with parameter θ. If θ has probability between green, yellow and blue, then $Z_{d,n}$ could be blue and it is drawn from that particular θ. There is a *Z* value, for every word, in the document and in the corpus. The green node, $W_{d,n}$ is the observed word. It is the only observed random variable in the entire model. The observations are a collection of words arranged by document. $W_{d,n}$ depends on both $Z_{d,n}$ and all the ß's. $W_{d,n}$ refers to the nth word in the dth document. $ß_k$, the blue node, is a topic where each ß is a distribution over terms and there are *K*, of these terms. From the previous example, *K* is equal to 50 as there were 50 topics in *Alice's Adventures in Wonderland*. So $ß_{32}$ is some distribution over words and ß is on the vocabulary simplex, which is the space of all possible distributions. It is assumed that ß comes from a Dirichlet

distribution. The purple node, η, is the topic hyperparameter which we will be explained in more detail later (Steyvers & Griffiths, 2007).

The dirichlet distribution is an exponential family distribution over the simplex. The simplex is a space of positive vectors that sum to one. An exponential family distribution is a set of probability distributions based on a specific set of definitions (Andersen, 1970). From figure nine, the red node, α consists of K positive values that parameterise the dirichlet distribution. The density function is described as:

$$p(\theta \mid \vec{\alpha}) = \frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)} \prod_i \theta_i^{\alpha_i - 1}.$$

Where a density function is a function used to describe the likelihood of a random variable to take on a given value (Gentle, 2009, p. 29).

$$\prod_i \theta_i^{\alpha_i - 1}$$

is a product of each component to the power of α_i-1, multiplied by

$$\frac{\Gamma\left(\sum_i \alpha_i\right)}{\prod_i \Gamma(\alpha_i)}$$

which is the normalising function, this is a constant which makes sure that the whole equation integrates to one, where $\prod$ is a symbol referring to a product of a sequence of terms. For example, $\prod_{i=1}^{3} i$ = 1 x 2 x 3 = 6

$\sum$ is a symbol that refers to a sum of a sequence of terms. For example,

$\sum_{i=1}^{3} i$ = 1 + 2 + 3 = 6

The Dirichlet distribution is conjugate to the multinomial distribution. Given a multinomial observation, the posterior distribution of θ is still a dirichlet. The parameter α controls the mean shape and sparsity of θ. The dirichlet is used twice in this model; the topic proportions are K dimensional dirichlet and the topics are V dimensional dirichlet. A multinomial observation is part of

a multinomial distribution where there are *n* independent trials, each of which leads to a success for exactly one of *k* categories, with each category having a given fixed success probability, the multinomial distribution gives the probability of any particular combination of numbers of successes for the various categories.

Latent Dirichlet Allocation

Latent dirichlet allocation is a mixed membership model that also incorporates latent semantic analysis into the model. Latent semantic analysis is a type of text analysis for exploring relationships between a set of documents and the terms they contain. This is done by creating a matrix of word counts per paragraph. A mathematical technique known as singular value decomposition is used to reduce the number of rows in the matrix. Words are compared and assigned values according to similarity (Landauer, Foltz, & Laham, 1998). A mixed membership model is a model where data are grouped and modelled with a mixture. The mixture components are shared between all groups and the mixture proportions are varied between groups. For a latent dirichlet allocation the documents are groups, with single words *Z*, as the observations. The components are distributions over the vocabulary, θ and topic proportions, which represent how much of each document reflects each pattern. The fitted components form topics, $ß_k$, and the proportions describe how each document describes each of these topics (D. M. Blei, 2012). The mixed membership assumption is more appropriate to use than the alternative choice of a simple finite mixture model. From a collection of documents the per-word topic assignment $Z_{d,n}$, the topic proportions, θ_d, and the per-corpus topic distributions $ß_k$ are inferred. The expectation of these parameters, given the words, is used to perform certain tasks such as document similarity comparisons, information retrieval or classification.
The exact posterior cannot be computed. Approximate posterior inference algorithms have been created to estimate the posterior. These include mean field variational methods, collapsed Gibbs sampling, collapsed variational inference and expectation propagation.

Approximate Posterior Inference

Assuming the topics $ß_{1:k}$ are fixed and that they are already known. The posterior distribution is the conditional distribution of the hidden variables given the observations. The posterior for one document has the hidden variables, topic assignment and topic proportions. The per-document posterior is the conditional posterior distribution of one set of topic proportions and the topic assignments for that document, given the words of the documents (Blei et al., 2003). This can be described mathematically as:

$P(\theta|W_{1\text{-}N}) =$

$$\frac{p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}{\int_{\theta} p(\theta \mid \alpha) \prod_{n=1}^{N} \sum_{z=1}^{K} p(z_n \mid \theta) p(w_n \mid z_n, \beta_{1:K})}$$

$P(\theta|W_{1\text{-}N})$ is the joint distribution of the hidden variables divided by the marginal probability. The denominator is the sum of N^k dirichlet integral terms. To compute this, the sum is moved out to the front of the equation and the integral over θ gets moved over to get N^k dirichlet integrals. The dirichlet is conjugate to the multinomial but N^k cannot be computed so an approximate posterior inference is used (Steyvers & Griffiths, 2007). Gibbs sampling was used for this thesis.

Gibbs Sampling

Gibbs sampling was developed in 1876 by Josiah Willard Gibb as a means of determining the energy states of gasses at equilibrium. The method he used was modelled as a Bayesian posterior distribution in 1990 by Alan Gelfand and Adrian Smith (Bolstad, 2012). It defines a Markov Chain whose stationary distribution is the posterior of interest. Independent samples are collected from the stationary distribution to approximate the posterior. This produces a Monte Carlo estimate of an expectation using independent samples from the posterior (Heinrich, 2009).

A Markov Chain is a sequence of random variables such as X_1, X_2, X_3 … that have a present state. Future and past states are independent of the present state. More formally,

$P(X_{n+1}=x \mid X_1=x_1, X_2=x_2, \ldots, X_n = x_n) = P(X_{n+1} = x \mid X_n = x_n)$, if both conditional probabilities are defined to be greater than zero. The possible values of X_i form a computable set, S, the state space of the chain (Givens & Hoeting, 2012, pp. 14-17). An example of this is given below.

Suppose a chocolate bar company, Company A, controls 20% of the chocolate bar market. Suppose they hire a market research company to predict the effects on an aggressive advertising campaign. They conclude someone using their brand will continue to do so 90% of the time. Someone not using their brand will change to their brand with 70% probability. This is represented as a Markov chain as follows (A is the brand of chocolate bar the company sells and A' is another brand):

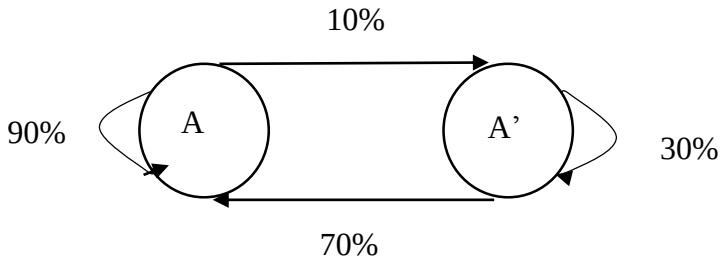

Figure 10. Markov chain

Markov chains can then be made into matrices where calculations can be performed (Meyn & Tweedie, 2012, pp. 24-29). They are important for generating sequences of random numbers to reflect complicated desired probability distributions through a process called Markov chain Monte Carlo. This allows for a wide range of posterior distributions to be simulated and their parameters found numerically (Gentle, 2009, pp. 313-318).

Markov Chain Monte Carlo Methods are a class of computational algorithms for sampling from a probability distribution based on constructing a Markov chain with the desired distribution as its equilibrium distribution. The state of the chain after a number of steps is used as a sample of the

desired distribution (Andrieu, De Freitas, Doucet, & Jordan, 2003). Each iteration improves the quality of the model.

Random walk methods make up a large subclass of Markov chain Monte Carlo methods, Gibbs sampling is a type of random walk Monte Carlo Method. This entails a group of "walkers" moving around randomly when approximating a multi-dimensional integral. At each point where a walker steps, the integrand value at that point is counted towards the integral. The walker then steps tentatively around the area looking for the next place to step. It should have a high contribution to the integral recorded (Givens & Hoeting, 2012, pp. 188-191).

In Gibbs sampling the space of the Markov chain is the space of the possible configurations of the hidden variables. The chain is run iteratively, sampling from the conditional distribution of each hidden variable, given the observations and the current state of the hidden variables. Once a chain has been embedded into the model, samples are collected at a lag to approximate the posterior. The lag is used to ensure the samples are independent (Blei, 2009).

We define n of the topic assignments, $n(Z_{1:N})$, as the counts vector. If a document contains ten words and three of them are assigned to topic fifteen, four of them are assigned to topic ten and the remaining three are assigned to topic eleven, this can be expressed as a table.

15	10	11
3	4	3

Table 1. Topic assignments for a document.

Using the terminology, $N_{15} = 3$, is a way of collating the topic assignment. A simple Gibbs sampler for latent dirichlet allocation can be written as:

$(\theta, Z_{1,n})$

$p(\theta|Z_{1:N}, W_{1:N})$

Where $(\theta, Z_{1,n})$ are the observations and $p(\theta|Z_{1:N}, W_{1:N})$ is the current state of the hidden variables. It is a dirichlet distribution so can also be expressed as:

$\text{Dir}(\alpha + n(Z_{1:N}))$

Where $\text{Dir}(\alpha + n(Z_{1:N}))$ is a dirichlet density with the parameter alpha, plus the counts $Z_{1:N}$. The Z's are assumed to be drawn from θ. θ, given a set of Z's is a dirichlet with parameter α, plus the counts.

Each Z needs to be sampled again. $p(Z_i|Z_{-I}, W_{1:N}, \theta)$ is proportional to the joint distribution of all parameters. This can be expressed mathematically as $P(Z_i|Z_{-I}, W_{1:N}, \theta) \propto p(Z_i|\theta)p(W_i|ß_{1:K}, Z_i)$. $p(W_i|ß_{1:K}, Z_i)$ is renormalized to get the conditional of Z_i given θ and Z_{-i}. A very simple Gibbs Sampler iterates between these two steps. First, θ is sampled given the current state of the Z's. Each Z is then sampled conditionally to the words. If the topic assigns zero probability to a word, that topic will have zero posterior probability under Z. θ will not include that Z in a count. To improve this technique collapsed Gibbs sampling is used (Bolstad, 2012).

In a Gibbs Sampler the state of the Markov chain is θ and $Z_{1:N}$. By making the state space smaller it will converge faster. To do this we can integrate out the topic proportions from the sampler to get a collapsed Gibbs sampler. In a collapsed Gibbs sampler, the state of the Markov chain is the collection of topic assignments. $Z_{1:N}$ and Z_i given the other Z_i's $(Z_i|Z_{-i})$ is iteratively sampled, integrating out θ. The equation can be expressed as:

$$p(Z_i|Z_{-i}, W_{1:N}) \propto p(W_i \mid ß_{1:K}, Z_i) \prod_{K=1}^{K} (n_k(Z_{-i}) + \alpha)$$

where $(n_k(Z_{-i})$ is the number of times topic k has been observed in the collection of topic assignments Z_{-1}. To determine the probability of topic nine, is to count the number of times topic nine was

observed in the other *Z*s. The convergence of the chain is assessed by monitoring the log probability of the state in observations (Blei, 2009).

A basic way of writing the Gibbs Sampling algorithm is as follows:

1. Set k=0 and choose $x^{(0)}$.

2. Generate $x_1^{(k+1)}$ conditionally on $x_2^{(k)}, x_3^{(k)}, \ldots, x_d^{(k)}$,

Generate $x_2^{(k+1)}$ conditionally on $x_1^{(k)}, x_3^{(k)}, \ldots, x_d^{(k)}$,

…

Generate $x_{d-1}^{(k+1)}$ conditionally on $x_1^{(k)}, x_2^{(k)}, \ldots, x_d^{(k)}$,

Generate $x_d^{(k+1)}$ conditionally on $x_1^{(k)}, x_2^{(k)}, \ldots, x_{d-1}^{(k)}$.

3. If: convergence has occurred then:

deliver $x = x^{(k+1)}$;

Else:

Set $k = k + 1$ and go to step 2.

(Gentle, 2009, p. 317)

What is a Topic?

It is important to consider what exactly a topic is and how it can be interpreted, a matter of some debate in the literature. In general, topics give an overview of the themes present in the texts being studied, the topics provide a summary of the contents of the dataset.

There is some discussion about whether the term ‘topics’ is even the right word to use for the groups of words produced by the topic model analysis. The word ‘topic’ may not always be the right word to use for the categories that are produced by topic modelling. It is suggested that ‘discourse’ might be better, as topics are not always unified semantically (Schmidt, 2012). For example, ancestry, logic, stable, science, theory, bin, mathematics, century, and epistemology would not be considered semantically coherent. Each of the words have a meaning but it would be difficult to assign them to a

specific topic. This list seems more to be a discourse of subjects found within philosophy. The topics could be known more as discourses as they tend to be digress from subject to subject without having a clear topic in mind.

The meaning of a topic can change depending on the type of data being modelled. Non-fiction such as scientific journal articles have more clearly defined topics than those from humanities subjects. For example, running a topic model analysis on a journal article from biology may produce a topic that shows a group of words linked by a concept or subject category. An example of this may be the topic: 'color patterns pattern guppies males bright result fig frequencies good', which one could surmise is a topic about colour patterns in male guppies. Modelling humanities data such as fiction, drama and poetry does not tend to produce such coherent topics(Schmidt, 2012). This is likely to be dependent on what kind of humanities subject is being studied. There will be more coherent topics in some domains such as archaeology or philosophy than there would be in analysing fiction. This suggests that when one is analysing their data researchers need to be aware of the type of data being analysed and how this should affect the way a topic is interpreted.

How to do a Topic Model

The purpose of a topic model is to extract potentially useful information from documents (G. Kim et al., 2014, pp. 71-79). This can be done in a number of ways, depending on the dataset. There are different ways of analysing text for different types of data; therefore it is important to know how to run a topic model analysis to achieve the best and most relevant results.

Dataset

Topic modelling is best used as a heuristic tool for sorting through large corpora of text. It is best suited to large datasets ranging from 500 to 1000 documents or 10,000 to 1,000,000, depending on the type of documents being analysed. For example, a collection of literary novels would be smaller in number than a collection of emails (Underwood, 2012d). Documents can be any length such as 10 words to ten million words. The format of the data is important and depends on the software being used. The majority of software packages require documents to be in a text file format such as .txt, .json, .xml, with some allowing for the use of .pdf files. Other programs require the documents to be converted into a document-term matrix before being analysed, while others will do this automatically. Examples of the types of documents analysed include but are not limited to: literary novels, emails, journal articles, genetics data and culinary recipes.

Cleaning up the data is one of the most important steps when conducting an analysis. This allows for a more extensive knowledge of the dataset and will lead to a smoother process when conducting the analysis. There are several different steps for pre-processing the data that depend on the type of software being used, there are often guidelines accompanying the software, about how much pre-processing needs to be done. This can range from removing any punctuation, to putting the dataset into blocks of around 1000 words each. The first thing to do when cleaning the data is to remove any white spaces within sentences, and depending on the type of dataset, removing information such as

publishing years, time stamps, and any cover pages containing publisher or uploader information (G. Kim et al., 2014). Page numbers should also be removed.

Text segmentation can be a useful tool to search for information about a particular dataset. It can also be important for a smooth analysis as software programs such as Mallet, prefer the data to be separated into smaller sections. Depending on the type of data being used, segmentation can be done in different ways. For example, literary novels can have themes present throughout their entirety but can also have themes present in only one or two chapters. Partitioning novels into smaller parts can result in themes being found that would not if the text was in larger parts. There is a general consensus that blocks of text should be around 1000 words each (Underwood, 2012a).

A list of stop words is important for removing extraneous words. Software packages often have a standard list of stop words such as to, the, of, and, etc. Depending on the data being used, creating a context specific list is a good idea. This can be done by creating a .txt file of words to be removed so they do not hinder the analysis. Personal judgment and familiarity with the dataset should be used here. Not all words can be found by manually going through each word and adding any that appear to be extraneous. Post-processing of topic words can also be done to remove anything not included in the list of stop words (Blei, 2009).

Software

Once the data has been tidied up, a software program needs to be found. There are several different text analysis programs available for use and they all use different methods of topic modelling. Each one also requires different skill levels for use. Below is a summary of some of the more popular software programs.

Mallet

Mallet is a java-based program, utilising different types of text analysis. This includes natural language processing, topic modelling and document classification (Marwick, 2013). It uses Gibbs Sampling and Latent Dirichlet Allocation to create its topic models. Mallet can be run on Windows and MacOS. It runs using the command-line structure to allow for a more extensive use of its features. This involves typing commands manually into a console, rather than using a mouse to press buttons or navigate menus (Shawn Graham, Weingart, & Milligan, 2012). A familiarity with the command-line is useful but there are excellent tutorials online for how to do a topic model analysis using Mallet.

For anyone uncomfortable with using a command-line system there is a graphical user interface version of Mallet's topic modelling tool. This was developed by David New and his team. It is available for download through Google project hosting. The download page has a simple explanation accompanied by images, detailing how to use the tool. It appears to accept whole datasets without partitioning whereas Mallet requires the data to be partitioned into smaller files.

R

R is a free to download, statistical software package. It runs on Windows, MacOS and UNIX systems. R has different packages available to download and install for different purposes. It is command-line based but there are packages such as RStudio that provide a graphical user interface for running an analysis. There is also a large library of literature on how to use R and run different packages. There are three packages capable of doing topic modelling analysis. They are mallet, topicmodels, and lda.

Mallet

The mallet package in R is based on the Mallet software package but rather than being capable of running a large selection of text analysis algorithms it can only do topic modelling analysis. It provides an interface to the Java implementation of latent dirichlet allocation (Mimno, 2015).

topicmodels

The topicmodels package can create Latent dirichlet allocation and correlated topic models. It uses Gibbs Sampling or Value Estimation Modelling as a fitting tool for the models. At the time of writing this R version 2.15.0 or later is needed to run the package (Grün et al., 2014).

lda

lda is a package dedicated to latent dirichlet allocation. It uses collapsed Gibbs sampling as the fitting method. As this package focuses only on latent dirichlet allocation, there are a large number of extra options to tailor the model for more specific results (Chang & Chang, 2010).

Gensim

Gensim is a python based program that allows for more customisation. The website has a comprehensive tutorial on how to run a topic model analysis. It runs latent semantic analysis, latent dirichlet allocation and random projections. The preferred file formats are plain text files (Řehůřek & Sojka, 2011).

LDA-C

LDA-C is a C implementation of latent dirichlet allocation software. It was one of the earlier software programs available so does not have as many options to tailor the analysis. It runs a basic latent dirichlet allocation model (Chaney & Blei, 2012).

GibbsLDA++

GibbsLDA++ is a C and C++ implementation of latent dirichlet allocation. It uses Gibbs sampling for fitting the model. GibbsLDA++ is also an older program so it does not have as many features as more recent programs. There is little room for customisation (Phan & Nguyen, 2007).

Running an Analysis

Depending on the software packages some analysis steps may be done automatically, others may need to be done manually. All packages require the number of topics and the number of iterations to be determined before an analysis can begin.

The first step for analysis is determining the number of topics. There are a variety of different ways this can be done. Some software will have a default number of topics that can be altered, others can determine the number of topics through analysis. Determining the number of topics manually can be done in various ways. Often a common sense judgment or estimate is made or the optimal number of topics can be calculated by splitting the data into training and test data to use tenfold cross validation and applying perplexity (G. Kim et al., 2014). Datasets usually have between 100 and 400 topics but depending on the size of the data this can range from 50 to 1000 topics. In general, the larger the dataset, the more topics may be incorporated into the analysis.

Once the number of topics has been chosen, the number of iterations should be selected. This is the number of times the computer will run through the model and make improvements. The selected number of iterations is dependent on the computing power and time available. The more iterations done on a model, the more accurate the model will be.

When the number of topics and iterations have been chosen the model is ready to be built. Once the model is done an output will be produced. This will be different for each software package. The

majority of programs will return a list of topics and their weighting. The weighting is the probability of the topic appearing in the dataset. A mallet output will produce a list of topics such as:

39 0.57246 humans intentional Bayesian benefit explanatory …

Where 39 is the topic number, 0.575246 is the weighting of the topic which is known as the Dirichlet parameter.

There will also be a topic composition list, detailing which documents contain which topics. Using the same example as above, this topic will look like:

#doc	Name	Topic	proportion		
0	file:/C:/mallet/target_data4/1.txt	83	0.636619	74	0.096913
1	file:/C:/mallet/target_data4/10.txt	2	0.415437	74	0.223652
2	file:/C:/mallet/target_data4/11.txt	2	0.418625	74	0.224302
3	file:/C:/mallet/target_data4/12.txt	2	0.34013	74	0.228621

Table 2. An example of a topic model output

In this example the first document, #doc 0, is 63.67% topic 83 which means that 63.67% of meaningful words are relating to topic 83, and 9.69% topic 74. For example, document 0 could be a sports article consisting of 63.67% horse racing and 9.69% statistics.

Once an output has been produced, there are a variety of different ways to further analyse and explore the data. Some programs will output a list of the weightings of each topic for each document. This allows for a more comprehensive understanding of the themes present in each text. If documents are sorted in chronological order the output can be graphed to see changes or patterns of themes over time. Similarities between documents can be analysed too.

Analysis of a Topic Model

The purpose of this chapter is to provide an example of a topic model analysis. It is also to analyse the PhilPapers dataset to search for any interesting patterns and to provide insight into the contents of the dataset.

Dataset

The dataset used for this analysis consists of abstracts and bibliographic information from the PhilPapers database. It is a .json file consisting of 1,100,861 entries. The database is relatively new so has not been analysed or explored extensively.

There were several steps taken to clean up the data. The first thing done was a search through the contents of the dataset to get an idea of its contents. The dataset consists of bibliographic information of philosophy journals. The focal point for the creation of topics for analysis was on the abstracts of each journal article. The majority of entries are written in English but there were several journals written in other languages such as German, Spanish and French. These were retained in the dataset to assess their effect on the topic modelling process. The next thing done was to remove any symbols or punctuation marks. Numbers were also removed. A large number of entries contained Roman numerals but due to the size of the dataset, they were left in to be removed after the topics had been created. The text was segmented into 52 smaller documents to allow for a smoother analysis on some software programs.

Mallet has a standard list of stop words available. Due to the nature of the dataset a large number of stop words were added to ensure a smooth modelling process including pubinfo, www, and http. To get a good idea of words needing removed from the dataset a few test topic models were run. The initial test contained topics such as:

org journal doi jstor ethics type stable sici philosophy article dx pdf abstract id volume year categories

This provided a list of words to remove from the analysis. As the dataset consisted of multiple languages common words in Spanish, German, French and Italian were added to the list of stop words. Appendix B is a comprehensive list of the stop words used.

Software

When trying to find the best software program to conduct a topic modelling analysis two different programs were trialled. Their performance is analysed below.

R

The first program used was R. At the time of testing the software programs access to the PhilPapers dataset was unavailable so an XML file consisting of bibliographic information of twelve science fiction novels were used. These are listed in Appendix A.

Using the Mallet package in R requires installation of the package rJava and the package xml to use an XML file. Considerable time may need to be spent managing dependencies in R before work can start. At the time of writing, R version 3.1.2 is not compatible with XML files, making it unsuitable for this project.

The package lda in R has issues with assigning alpha values and requires a specific type of alpha value to be used creating some limitations with analysis. Once the alpha value has been assigned, the topic modelling analysis can be run to produce a coherent topic model.

Mallet

When researching into using Mallet an example of running mallet through R was discovered. It provides a way to implement a graphical user interface and to use some of the other packages available in R for further analysis. The coding required to run Mallet through R is reasonably simple

to do and only requires alterations to the system settings to run effectively. The initial topic model created using this software provided topics such as:

consciousness mind content null intentionality externalism philosophy mental meaning language qualia knowledge

This suggests the topic model was created successfully as it has produced a coherent topic about philosophy of mind.

The graphical user interface of Mallet developed through Google Project Hosting was also used. There was no need for any debugging or alteration of settings to get working. Running a topic model analysis using this software is as simple as pushing a button. To run a topic model analysis on this software, the documents do not require segmentation as it can handle processing the entire dataset in one document. This software was used for the majority of this analysis.

Running an Analysis

To determine the number of topics to produce, a number of different topic models with varying amounts of topics were run. Using Mallet, through R, the maximum number of topics created was 150. Due to the size of the dataset a larger number of topics was preferred and for this dataset the graphical user interface version of Mallet was used because it produced a model consisting of 300 topics.

When selecting the number of iterations to run, the maximum number of iterations possible was the goal. The Mallet software through R could run through 100 iterations to produce a maximum number of 50 topics, and again the user interface of Mallet provided a maximum of 200 iterations to produce 300 topics.

Once the topic model had been created and a list of topics produced, the topics were sorted through to search for anything of interest and to get a general overview of the content of the PhilPapers dataset.

There is a lot of information to be found in the topics and a variety of different ways to portray and explore the information, below is a selection of methods used to explore the data.

An Overview of the Dataset

Once a topic model of 300 topics was created, the topics were sorted into different categories. This was done by assigning a category label to topics with titles that described the words being used. For example, the topic

god religion divine evil existence sophia theism arguments argument miracles attributes religious atheism plantinga creation topics omnipotence goodness omniscience theistic

was assigned to the religion category. When sorting through the topics some were removed from the analysis as they appeared to be nonsensical and came from words that were not put into the list of stop-words. An example of this type of topic is:

net publication acta element researchgate biotheoretica giorgio agamben clrjames przyrodoznawstwo yadda pl analytica icm bwmeta humanistyka james bd origin cejsh

Topics related to publishing were also removed. These came from the bibliographic information aspect of the dataset and did not include a lot of philosophical content so were not analysed. An example of this type of topic is:

book work philosophers introduction important written thinkers key writings scholars major range including edition wide published offers original philosopher comprehensive

After removing these topics a total of 37 topic labels remained and from them a graph of the topic counts was created.

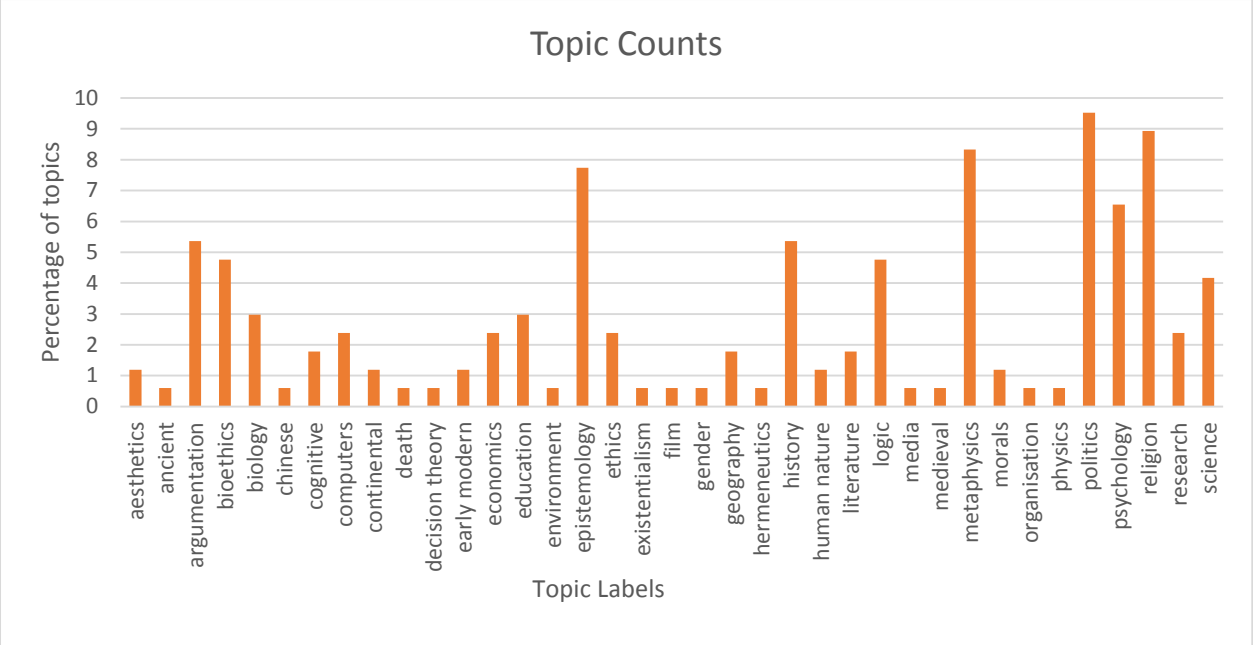

Figure 11. Graph of the frequency of topics displayed in the topic model analysis

This graph gives an indication of the content of the PhilPapers dataset. It shows the percentage of each topic present in the PhilPapers dataset. A pie chart of the ten most prevalent topics was also created.

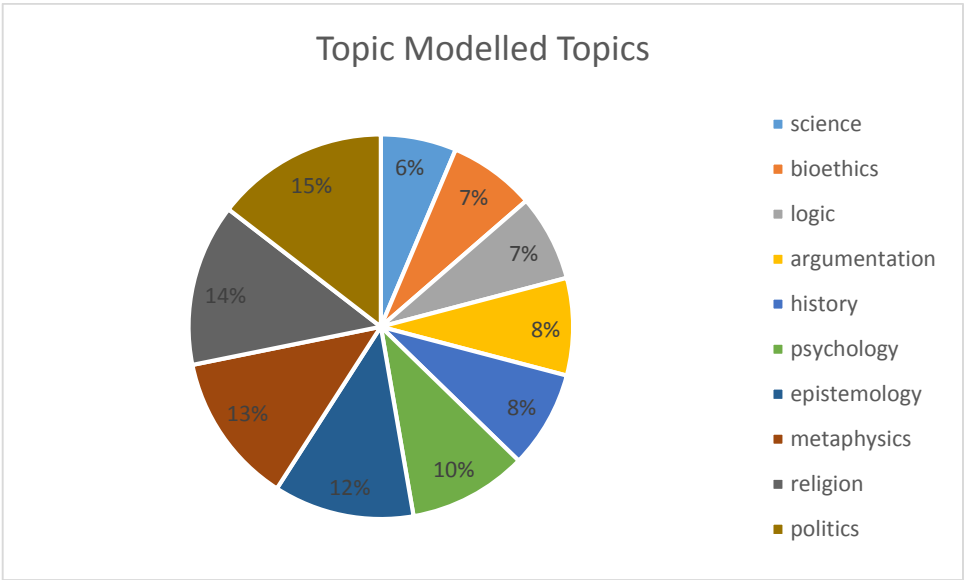

Figure 12. Graph of the top ten topics displayed in the topic model analysis.

The two graphs show that the most common topics found in the dataset are politics, religion, metaphysics, epistemology and psychology. This suggests that the majority of journal articles on the PhilPapers website are related to these topics. To explore this idea further, a comparison of the contents of the dataset with the categories used by the PhilPapers team was conducted.

Their philosophy papers are grouped into seven main categories, which are: Metaphysics and epistemology; Value Theory; Science, Logic, and Mathematics; History of Western Philosophy; Philosophical Traditions; Philosophy Misc.; and Other Academic Areas. Within these categories, there are a total of 40 subcategories. Below is a graph of the percentages used for the PhilPapers categories.

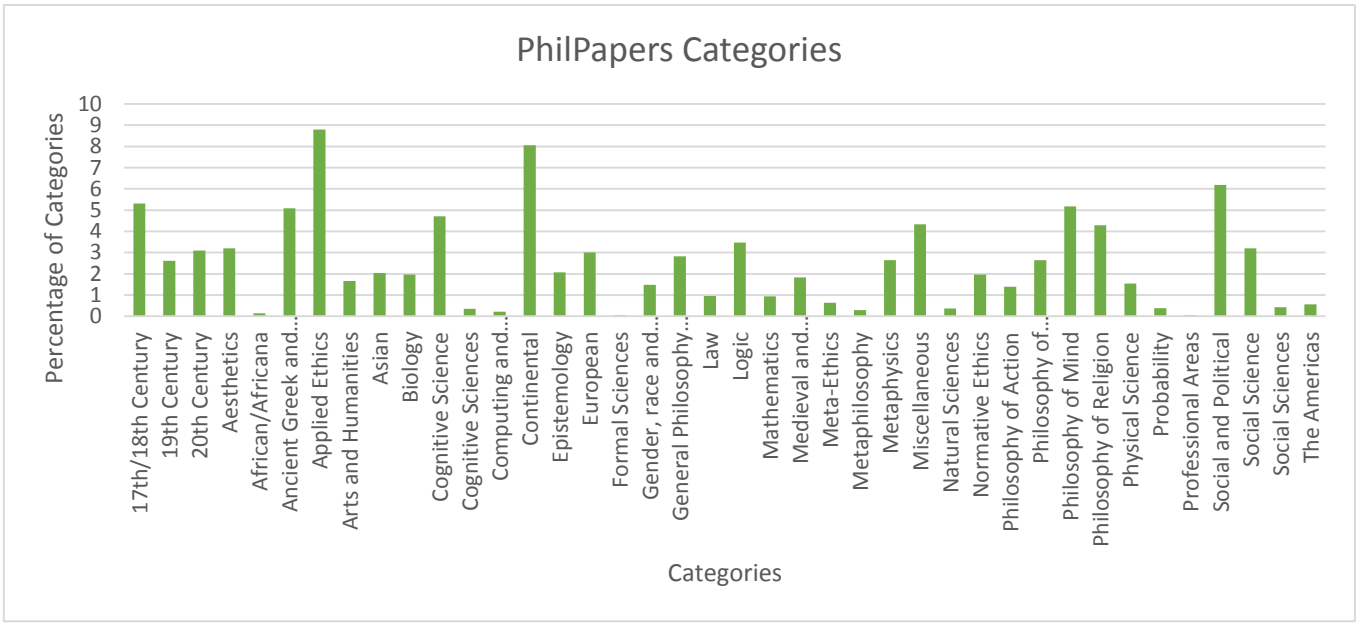

Figure 13. Graph of the frequency of categories found on the PhilPapers website.

A pie chart was also created to illustrate the top ten category labels found in the PhilPapers database.

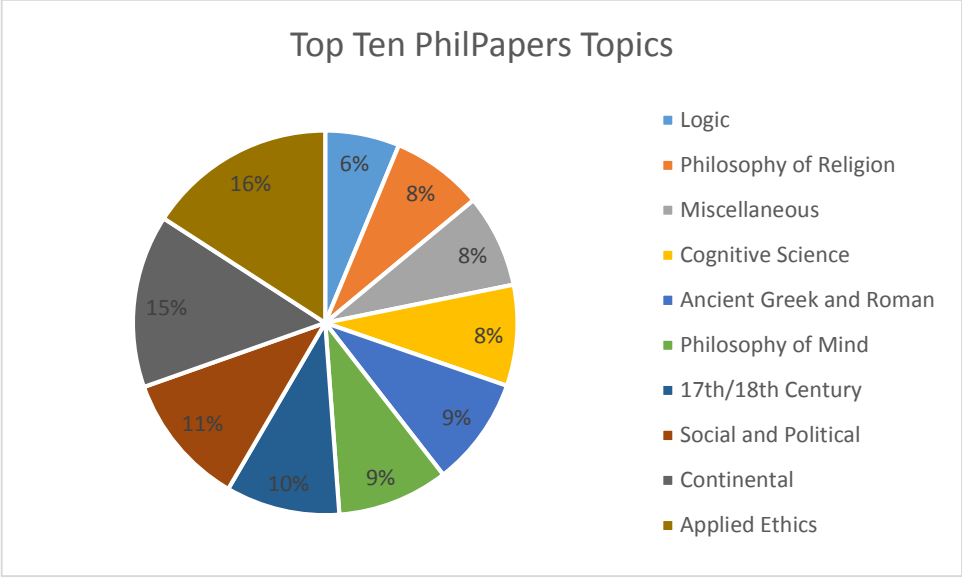

Figure 14. Graph of the top ten categories found on the PhilPapers website.

These two graphs suggest that the most prevalent categories are applied ethics, social and political, continental philosophy and 17th/18th century philosophy. There are some similarities between the PhilPapers dataset analysis and the category labels on the PhilPapers website. For instance, both have logic as a top ten category and they have similar percentages at 6% and 7% each. The psychology label makes up 10% of the topic labels in the PhilPapers dataset whereas philosophy of mind has a similar percentage of 9% for the category labels for the website.

The topic model analysis of the dataset was different to the actual categories of the PhilPapers database for a few reasons. The first of which is that the topics were assigned one label only, whereas more than one category is assigned to papers from the PhilPapers database. The philosophy of mind and cognitive science categories are very similar and would contain some papers that are the same. Another reason is that the category history contains many different historical periods such as 17th and 18th century philosophy, whereas the PhilPapers database lists these categories separately. The database itself is constantly being updated and added to so the contents of the database now may be quite different to the contents of the dataset at the time it was first acquired. Another important reason as to why the contents were different is that the topic model analysis is a model built on a sample of the dataset as a whole and will give an estimate of the contents of the dataset. A model created from a sample of the data may not give a completely accurate picture of the exact contents of the database itself. If another topic model was run in the exact same way, the topics produced would not be identical to this analysis.

Foreign Languages

An area of interest was that of how topic modelling worked with datasets containing entries of multiple languages and how the topics produced would incorporate mixtures of languages, if it did at all. The majority of topics created were in a single language. The only instances of multi-language topics tended to be one or two English language words appearing in a topic in another language. For example,

el del los las una tica filosof sobre su di entre para como der thought da index por des

is a topic of mostly Spanish words but has the English words ‘thought’ and ‘index’ incorporated into it.

The topic below contains mostly German words and with some help from Google Translate, appears to be about political history.

colm kandidaten heilsgeschichte kayser debemus hochschulschriften reva waraa strasbourgeois

Topics being presented in a single language can be explained by the way the topic modelling algorithms work. If an article written about domestic cats used the word cat to refer to cats and another article used the word feline instead, a topic would be produced that contained both of these words. This is because other words relating to cats such as whiskers, would appear in each article. Generally, articles are written in one language so all words related to each other will be in the same language. For example, an article written in French about 17th century philosophy would consist of words only in French, the author is not likely to scatter words from another language throughout the paper. The only similar words would generally be the names of people being referenced in the paper. This is shown in topics where the names of prominent figures are also listed as a word in the topic. For example:

mathematics id logic reasoning frege knowledge russell mathematical theory bertrand unknown probability concept full

contains the names of Frege and Russell who studied and produced prominent papers on mathematics and logic.

Philosophy of Action

While searching through the categories on the PhilPapers website there was an unfamiliar category label entitled Philosophy of Action. As this was an unfamiliar term a topic model analysis was run on a selection of papers to see if this would help determine what Philosophy of Action is. This was done as a faster way to discern the information than reading through a collection of articles on the

Philosophy of Action. A small topic model returning a list of around ten topics was created. Two of the topics are listed below:

man acts character actions pleasure nature regard giving great bad called voluntary objects courage virtues deficiency proud pleasures anger act end appetite indulgent delight reason small sake angry pained praised ignorance money thing spend painful exceeds truth extremes confidence sources greatest easily general cases prodigal terrible indulgence defect feel blamed

action moral actions possibility conception hare freely fails question control common kind understanding responsible thought called kinds acting terms calls source understood choice capacities suggestion contrast agents desires capacity rationality caused rorty explains aim generally concerned assume rational debates intellectual involves argument deny blameworthy phenomena michael temptation remain evaluative spite

These two topics provide some interesting keywords and suggest the philosophy of action could be about the actions of man. The topics suggest the philosophy of action discusses human nature and things that man does. There are also terms relating to intentions and rationality, a discussion of free will may also be covered under the philosophy of action. To check whether these conclusions are accurate the literature on what the philosophy of action is was researched, and it was found that the philosophy of action is about the actions of man, including discussion on free will, motivation, personal reason, actions, and intentionality (Wilson & Shpall, 2012).

Analysis of Single Topics

When searching through the list of 300 topics, there were quite a few interesting topics that could be further explored on an individual level. Some topics were a lot easier to assign labels to whereas others took some thought. Below is a selection of some of the topics that seemed easy to place into categories and an explanation of why they were placed into those categories, and a selection of topics that were not so easy to place.

consciousness science cognitive ancestry psychology mind conscious free unconscious cognition neuroscience action intelligence artificial processes retrieve brain

At first glance this topic is clearly about cognitive science. There is a question of whether or not this topic is worth analysing because it is so obviously about cognitive science. It is beneficial as it suggests the topic model analysis worked, but it does not provide any interesting insight into the contents of the dataset itself. Some topics are good to look at, such as this one, as part of the whole dataset, whereas some other topics, are interesting to evaluate on an individual level.

press google university ethics essays science theory philosophical social ed john language oxford nature religion African world david practice

This topic, appears to be the opposite of the previous one as it does not appear to belong to one idea. However, the topic modelling algorithm placed these words together to form a topic. This could be due to random chance where these words just happen to have high probability of occurring together because of other words that relate to these words. A lack of knowledge or understanding by the researcher could be another reason as to why this topic appears nonsensical. Deeper analysis could be done on this topic, however it is at the discretion of the researcher to decide how relevant a topic is to their analysis. In this instance the topic was discarded from the overall analysis.

feminist gender race sexuality feminism women hypatia sexual sex normative approaches topics hyp orientation varieties miscellaneous racism female racial politics

This topic is quite interesting as it features a list of keywords relating to gender issues. Obvious words such as ‘feminist’, ‘gender’, ‘women’, and ‘feminism’ make this clear. There are some words that do not make immediate sense such as ‘miscellaneous’, ‘varieties’ or ‘hyp’. Not every word in a topic can be justified as there can be some words that were not removed earlier that can appear in topics. At first glance the word ‘hypatia’ does not appear to fit, however, upon researching the word

it is discovered that Hypatia was a female Greek mathematician and philosopher (Petta & Colavito, 2004).

things people time place story make made day back find don read continent world long call fact man end life

This topic appears to be particularly interesting; at first glance the words do not have an obvious topic. It seems clear that the words do fit together semantically. When putting it into context of a philosophy topic, it does not seem to be about an obvious philosophy category. This is the sort of topic that can provoke large amounts of discussion and is up to the discretion of the researcher to decide whether it fits with the model to be analysed or not. In this case the topic appears to be the story of man and words relating to the life of man.

classical classics ancestry arts humanities paper ancient greek roman isbn press class university cased plates cloth dm london studies

This topic appeared to be of interest solely because it contained 'cloth' and 'plates'. However, on closer inspection this topic appears to be about publishing and any topics relating to publishing were removed from the analysis.

good love function functions teleology architecture friendship bad teleological antique grecque internationale kernos pluridisciplinaire eros goodness friends passion sacrifice heart

This topic is of interest as it needs further research into some of the words before its meaning can be fully discerned. There appear to be words relating to emotions such as 'good', 'love', 'friendship', 'passion', and 'heart'. There are also references to Greek culture such as Eros, the Greek god of love; 'kernos' is a type of Greek pottery, and 'grecque' is the French word for Greek. The topic appears to be about Greek philosophy with a particular emphasis on the passions and emotions.

There are several different ways of analysing a topic model and displaying the information. This chapter looked at ways of getting an overview of the dataset as a whole, exploring how foreign languages appear in topics, and information retrieval to extract specific information about a certain area of the dataset. Individual topics can also be analysed to see if there are any unique or interesting details in the dataset.

Discussion

Topic modelling is a useful tool for analysing large collections of text. It can be used for a variety of different purposes in many different areas of research. Topic modelling was predominantly used in scientific disciplines but in recent years has been increasingly used in the humanities. The research has not progressed far with very few published academic articles on the use of topic modelling in the humanities being produced. One of the reasons for this may be due to a lack of understanding of how the topic modelling algorithms work. Another reason could be that it can be difficult to see how topic modelling can be used for humanities research.

Understanding of how topic modelling algorithms work is important for using it effectively. Knowing how it works can provide ideas for research and an understanding of how to interpret topics. There is some debate about how topics can be interpreted which was discussed earlier. Depending on the research area, topics should be interpreted accordingly. For instance, scientific disciplines have more of a rigid and obvious set of topics. This can be said for some humanities subjects such as history, archaeology and philosophy. However, they can also have some topics that require more thought and background knowledge to fully interpret. Subjects such as literature analysis would tend to produce topics more about themes found in the novels themselves. Generally, experts in their field of research would have the skills necessary to interpret topics produced in their field.

Topic modelling has many different applications. The LDA model was developed independently of topic modelling, and has been used in population genetics analysis (Lonsdale et al., 2013). One application used by genetic researchers who are interested in modelling people as being mixtures of their various ancestry. Analysis is conducted to provide information about how much of a person's ancestry can be predicted through genetics (Pinoli, Chicco, & Masseroli, 2014). New ways to incorporate topic modelling into research are being discovered on a regular basis. Researchers in the humanities should have incentive to go out and explore topic modelling and see how it can apply to

their own fields of research. Topic modelling can be seen as an exciting new tool for analysis of humanities data.

The field of topic modelling is so large that this thesis has only touched upon its surface. There are several other areas of topic modelling to be explored. Trends and changes in popularity of topics over time can be used to further explore the PhilPapers dataset. This could be done by creating topic models based on different time periods such as 17th century philosophy and looking at the patterns in topics over time. This could be done to provide an overview of the general philosophy topics being discussed in each time period or individual topics could be analysed over time, such as the topic religion, and graphs could be made to show in what time periods religion was most strongly discussed and whether it is being discussed more or less today than in other periods in history.

Other areas of topic modelling would also be a worthwhile area to explore. This thesis used latent dirichlet allocation for its analysis but there are many other types of topic modelling algorithms available to use. One area of research would be to compare how different algorithms perform and whether or not the topics are significantly different from each other. There are specific topic models tailored to create more precise topics and to run more iterations smoothly, and it would be interesting to compare these algorithms. There are also topic modelling algorithms that focus on specific kinds of topic modelling, for instance, the Author-Topic Model is tailored to exploring the relationships between authors and the topics used by specific authors (Rosen-Zvi et al., 2004). For a large dataset with a lot of different data points, an exploration of the different types of topic modelling algorithms could be conducted to explore various attributes of the dataset.

Using a supercomputer to run a topic model analysis was another potential area of research. There is more computing power available and so there is potential to create a more accurate topic model with a higher number of iterations. Comparisons between the results of a topic model run on a standard PC and those run on a supercomputer could be done. This could look at the advantages in terms of

computational speed of using a supercomputer and whether the topics produced are sufficiently different from each other to warrant the use of a supercomputer for analysis.

Topic modelling is a useful tool for exploring large corpora of text. It can be used in a variety of academic fields and can be a useful tool in any academic researcher's toolkit. Together with a selection of other text mining tools it can provide a good understanding of large corpora of text. Once an understanding of how topic modelling algorithms work and how topic models can be used for analysis it can have many applications for humanities research.

Bibliography

Andersen, E. B. (1970). Sufficiency and exponential families for discrete sample spaces. *Journal of the american statistical association, 65*(331), 1248-1255.

Andrieu, C., De Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine learning, 50*(1-2), 5-43.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). *On smoothing and inference for topic models.* Paper presented at the Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence.

Blei, D. (2012). Topic modeling and digital humanities. *Journal of Digital Humanities, 2*(1), 8-11.

Blei, D. M. (2009). *Topic Models.* Paper presented at the Machine Learning Summer School, Princeton University.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84.

Blei, D. M., Griffiths, T. L., & Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM), 57*(2), 7.

Blei, D. M., & Lafferty, J. D. (2006). *Dynamic topic models.* Paper presented at the Proceedings of the 23rd international conference on Machine learning.

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, 17-35.

Blei, D. M., & Lafferty, J. D. (2009). Topic models. *Text mining: classification, clustering, and applications, 10*(71), 34.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research, 3*, 993-1022.

Blevins, C. (2010). Topic Modeling Martha Ballard's Diary. *Online: http://historying.org/2010/04/01/topic-modeling-martha-ballards-diary*.

Bolstad, W. M. (2012). Gibbs Sampling and Hierarchical Models *Understanding Computational Bayesian Statistics* (pp. 235-263): Wiley.

Bourget, D., & Chalmers, D. (2015). About Philpapers. 2015, from http://philpapers.org/help/about.html

Brynjolfsson, E., & McAfee, A. (2014). The Digitisation of Just About Everything *The second machine age: Work, progress, and prosperity in a time of brilliant technologies* (pp. 57 - 70): WW Norton & Company.

Carroll, L. (1865). *Alice's Adventures in Wonderland*: Project Gutenberg.

Chaney, A. J.-B., & Blei, D. M. (2012). *Visualizing Topic Models.* Paper presented at the ICWSM.

Chang, J., & Chang, M. J. (2010). Package 'lda': Citeseer.

Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., & Blei, D. M. (2009). *Reading tea leaves: How humans interpret topic models.* Paper presented at the Advances in neural information processing systems.

Chen, E. (2011). Introduction to latent dirichlet allocation. *Webseite(02. Mai 2012) http://blog. echen. me/2011/08/22/introduction-to-latent-dirichlet-allocation*.

Cox, B., & Forshaw, J. (2010). *Why Does E= mc2?: And Why Should We Care?* : Da Capo Press.

Dean, J. (2014). *Big Data, Data Mining, and Machine Learning : Value Creation for Business Leaders and Practitioners* Retrieved from http://canterbury.eblib.com.au/patron/FullRecord.aspx?p=1687540

Dickman, J. (2014). Topic Modeling Explained: LDA to Bayesian Inference. *Tech Talk.* 2014, from http://www.tdktech.com/tech-talks/topic-modeling-explained-lda-to-bayesian-inference

Doyle, G., & Elkan, C. (2009). *Accounting for burstiness in topic models.* Paper presented at the Proceedings of the 26th Annual International Conference on Machine Learning.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS), 29*(2), 8.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine, 17*(3), 37.

Feinberg, J. (2014). Wordle.

Flanders, J. (2009). The productive unease of 21st-century digital scholarship. *Digital Humanities Quarterly, 3*(3).

Gentle, J. E. (2009). *Computational statistics* (Vol. 308): Springer.

Givens, G. H., & Hoeting, J. A. (2012). *Computational Statistics* (pp. 14-17). Retrieved from http://canterbury.eblib.com.au/patron/FullRecord.aspx?p=1120265

Goldstone, A. (2014). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us.

Graham, S. (2012). Mining a Day of Archaeology. *Electric Archaeology. Online in: http://electricarchaeology. ca/2012/07/09/mining-a-day-of-archaeology/(last accessed: 24 March 2014*.

Graham, S., Weingart, S., & Milligan, I. (2012). Getting Started with Topic Modeling and MALLET. *The Programming Historian, 2*.

Grün, B., Hornik, K., & Grün, M. B. (2014). Package 'topicmodels'.

Harris, J. (2011). Word clouds considered harmful. *Nieman Journalism Lab*.

Heinrich, G. (2009). A generic approach to topic models *Machine Learning and Knowledge Discovery in Databases* (pp. 517-532): Springer.

Hofmann, T. (1999). *Probabilistic latent semantic indexing.* Paper presented at the Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval.

Jockers, M. (2014). *Text Analysis with R for Students of Literature*: Springer International Publishing.

Jockers, M. L. (2011). The LDA Buffet is Now Open; or, Latent Dirichlet Allocation for English Majors.

Kim, D., & Oh, A. (2011). Topic chains for understanding a news corpus *Computational Linguistics and Intelligent Text Processing* (pp. 163-176): Springer.

Kim, G., Park, S., & Jang, D. (2014). Technology Analysis from Patent Data Using Latent Dirichlet Allocation *Soft Computing in Big Data Processing* (pp. 71-80): Springer.

Kirschenbaum, M. (2012). What is digital humanities and what's it doing in English departments? *Debates in the digital humanities, 3*.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes, 25*(2-3), 259-284.

Liu, A. (2004). Transcendental data: Toward a cultural history and aesthetics of the new encoded discourse. *Critical Inquiry, 31*(1), 49-84.

Liu, A. (2012). The state of the digital humanities A report and a critique. *Arts and Humanities in Higher Education, 11*(1-2), 8-41.

Liu, A. (2013). The meaning of the digital humanities. *pmla, 128*(2), 409-423.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., . . . Young, N. (2013). The genotype-tissue expression (GTEx) project. *Nature genetics, 45*(6), 580-585.

Marwick, B. (2013). R2Mallet.r. 2015, from https://gist.github.com/benmarwick/4537873

Mcauliffe, J. D., & Blei, D. M. (2008). *Supervised topic models.* Paper presented at the Advances in neural information processing systems.

Meyn, S. P., & Tweedie, R. L. (2012). *Markov chains and stochastic stability*: Springer Science & Business Media.

Mimno, D. (2012). Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage (JOCCH), 5*(1), 3.

Mimno, D. (2015). Package 'mallet' *Packages*.

Papadimitriou, C. H., Tamaki, H., Raghavan, P., & Vempala, S. (1998). *Latent semantic indexing: A probabilistic analysis.* Paper presented at the Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems.

Petta, A., & Colavito, A. (2004). *Hypatia: Scientist of Alexandria*: Lampi di Stampa.

Phan, X.-H., & Nguyen, C.-T. (2007). GibbsLDA++: AC/C++ implementation of latent Dirichlet allocation (LDA): Technical report.

Piez, W. (2013). Something Called Digital Humanities. *Defining Digital Humanities: A Reader*, 187.

Pinoli, P., Chicco, D., & Masseroli, M. (2014). *Latent Dirichlet Allocation based on Gibbs Sampling for gene function prediction.* Paper presented at the Computational Intelligence in Bioinformatics and Computational Biology, 2014 IEEE Conference on.

Posner, M. (2012). Very Basic Strategies for Interpreting Results from the Topic Modeling Tool. Retrieved from http://miriamposner.com/blog/very-basic-strategies-for-interpreting-results-from-the-topic-modeling-tool/

Řehůřek, R., & Sojka, P. (2011). Gensim–Python Framework for Vector Space Mo delling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*.

Rhody, L. M. (2012). Some Assembly Required: Understanding And Interpreting Topics in LDA Models of Figurative Language.

Riddell, A. B. A Simple Topic Model (Mixture of Unigrams).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). *The author-topic model for authors and documents.* Paper presented at the Proceedings of the 20th conference on Uncertainty in artificial intelligence.

Rosenbloom, P. S. (2010). Towards a conceptual framework for the digital humanities. *Digital Humanities Quarterly, 6*.

Schmidt, B. M. (2012). Words alone: dismantling topic models in the humanities. *Journal of Digital Humanities, 2*(1), 49-65.

Song, Y., Pan, S., Liu, S., Zhou, M. X., & Qian, W. (2009). *Topic and keyword re-ranking for lda-based topic modeling.* Paper presented at the Proceedings of the 18th ACM conference on Information and knowledge management.

Steyvers, M., & Griffiths, T. (2007). Probabilistic topic models. *Handbook of latent semantic analysis, 427*(7), 424-440.

Svensson, P. (2012). Envisioning the digital humanities. *Digital Humanities Quarterly, 6*(1).

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the american statistical association, 101*(476).

Templeton, C. (2011). Topic modeling in the humanities: An overview. *Maryland Institute for Technology in the Humanities Blog*.

Ulrich, L. T. (2010). *A midwife's tale: The life of Martha Ballard, based on her diary, 1785-1812*: Vintage.

Underwood, T. (2011). LSA is a Marvellous Tool, but Literary Historians May Want to Customize it for their own Discipline. *The Stone and the Shell, Oct, 16*(2010), 421.

Underwood, T. (2012a). Topic Modeling Made Just Simple Enough. Retrieved from http://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/

Underwood, T. (2012b). A Touching Detail Produced by LDA ... Retrieved from http://tedunderwood.com/2012/03/25/a-touching-detail-produced-by-lda/

Underwood, T. (2012c). What kinds of "topics" does topic modeling actually produce. *The Stone and Shell*.

Underwood, T. (2012d). Where to start with Text Mining. Retrieved from http://tedunderwood.com/2012/08/14/where-to-start-with-text-mining/

Vanhoutte, E., Dr, Nyhan, J., Dr, & Terras, M., Dr. (2013). *Defining Digital Humanities : A Reader* Retrieved from http://canterbury.eblib.com.au/patron/FullRecord.aspx?p=1426876

Wallach, H. M., Minmo, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter.

Weingart, S. (2011). Topic Modeling and Network Analysis. Retrieved from http://www.scottbot.net/HIAL/?p=221

Weingart, S. (2012). Topic Modeling for Humanists: A Guided Tour. *The Scottbot Irregular, 25*.

Weir, A. (2014). *The Martian: A Novel*: Crown/Archetype.

Wilson, G., & Shpall, S. (2012). Action. 2015, from http://plato.stanford.edu/entries/action/

Zaki, M. J., & Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*: Cambridge University Press.

Zhao, W., Zou, W., & Chen, J. J. (2014). Topic modeling for cluster analysis of large biological and medical datasets. *BMC bioinformatics, 15*(Suppl 11), S11.

Appendices

Appendix A: Science Fiction Novels

Adams, D. (2014). *The hitchhikers guide to the galaxy*. DavCom.

Asimov, I., & Hockridge, K. T. (1958). *The currents of space*. Panther Books.

Card, O. S. (1991). Ender's Game. 1985. *Author's Definitive Edition." NY: Tor*.

Clarke, A. C. (2010). *2001: a space odyssey*. Hachette UK.

Dick, P. K. (1990). *We Can Remember It for You Wholesale* (Vol. 2). Citadel Press.

Herbert, F. (2003). *Dune*. Penguin.

Piper, H. B. (1987). *Little Fuzzy*. Penguin.

Scalzi, J. (2007). *Old Man's War*. Pan Macmillan.

Verne, J. (1970). From the Earth to the Moon. 1865. *Translated by J. and R. Baldick. Dutton, New York*.

Weir, A. (2014). *The Martian: A Novel*. Crown/Archetype

Wells, H. G. (2003). *The war of the worlds*. Broadview Press.

Wyndham, J. (2010). *Chocky*. Penguin UK.

Appendix B Stop Words

a
able
about
above
according
accordingly
across
actually
after
afterwards
again
against
all
allow
allows
almost
alone
along
already
also
although
always
am
among
amongst
an
and
another
any
anybody
anyhow
anyone
anything
anyway
anyways
anywhere
apart
appear
appreciate
appropriate
are
around
as
aside
ask
asking
associated
at
available
away
awfully
b
be
became
because
become
becomes
becoming
been
before
beforehand
behind
being
believe
below
beside
besides
best
better
between
beyond
both
brief
but
by
c
came
can

cannot
cant
cause
causes
certain
certainly
changes
clearly
co
com
come
comes
concerning
consequently
consider
considering
contain
containing
contains
corresponding
could
course
currently
d
definitely
described
despite
did
different
do
does
doing
done
down
downwards
during
e
each
edu
eg
eight
either
else
elsewhere
enough
entirely
especially
et
etc
even
ever
every
everybody
everyone
everything
everywhere
ex
exactly
example
except
f
far
few
fifth
first
five
followed
following
follows
for
former
formerly
forth
four
from
further
furthermore
g
get
gets
getting
given
gives
go

goes
going
gone
got
gotten
greetings
h
had
happens
hardly
has
have
having
he
hello
help
hence
her
here
hereafter
hereby
herein
hereupon
hers
herself
hi
him
himself
his
hither
hopefully
how
howbeit
however
i
ie
if
ignored
immediate
in
inasmuch
inc
indeed
indicate
indicated
indicates
inner
insofar
instead
into
inward
is
it
its
itself
j
just
k
keep
keeps
kept
know
knows
known
l
last
lately
later
latter
latterly
least
less
lest
let
like
liked
likely
little
look
looking
looks
ltd
m
mainly

many
may
maybe
me
mean
meanwhile
merely
might
more
moreover
most
mostly
much
must
my
myself
n
name
namely
nd
near
nearly
necessary
need
needs
neither
never
nevertheless
new
next
nine
no
nobody
non
none
noone
nor
normally
not
nothing
novel
now
nowhere
o
obviously
of
off
often
oh
ok
okay
old
on
once
one
ones
only
onto
or
other
others
otherwise
ought
our
ours
ourselves
out
outside
over
overall
own
p
particular
particularly
per
perhaps
placed
please
plus
possible
presumably
probably
provides
q

que
quite
qv
r
rather
rd
re
really
reasonably
regarding
regardless
regards
relatively
respectively
right
s
said
same
saw
say
saying
says
second
secondly
see
seeing
seem
seemed
seeming
seems
seen
self
selves
sensible
sent
serious
seriously
seven
several
shall
she
should
since
six
so
some
somebody
somehow
someone
something
sometime
sometimes
somewhat
somewhere
soon
sorry
specified
specify
specifying
still
sub
such
sup
sure
t
take
taken
tell
tends
th
than
thank
thanks
thanx
that
thats
the
their
theirs
them
themselves
then
thence
there

thereafter
thereby
therefore
therein
theres
thereupon
these
they
think
third
this
thorough
thoroughly
those
though
three
through
throughout
thru
thus
to
together
too
took
toward
towards
tried
tries
truly
try
trying
twice
two
u
un
under
unfortunately
unless
unlikely
until
unto
up
upon
us
use
used
useful
uses
using
usually
uucp
v
value
various
very
via
viz
vs
w
want
wants
was
way
we
welcome
well
went
were
what
whatever
when
whence
whenever
where
whereafter
whereas
whereby
wherein
whereupon
wherever
whether
which
while
whither

who	bsl	http
whoever	books	thing
whole	citation	year
whom	sagepub	editors
whose	cfm	misc
why	cfid	edu
will	abstract	jhu
willing	id	hs
wish	org	ase
with	journal	pub
within	www	pdf
without	links	pubinfo
wonder	dx	bf
would	article	fulltext
would	issue	misc
x	title	type
y	springer	collection
yes	link	unknown
yet	jstor	philmat
you	editors	xiv
your	pages	cftoken
yours	de	lc
yourself	tb	psi
yourselves	span	pdfplus
z	pdcnet	os
zero	phd	manuscript
doi	frompage	xcii
http	sici	nki

xxii	le	imuse
discourses	cgi	conf
jhu	blackwell	add
volume	rft	die
author	file	der
authors	springerlink	das
acm	url	la
hph	revue	le
html	und	ein
jr	revues	eine
fj	quarterly	einen
null	cr	della
philosophy	chapter	el
journals	pp	las
categories	review	los
cambridge	revmetaph	des
wiley	fnd	di
content	oi	du
jme	amp	php
abs	eds	filozoficzne
displayabstract	categories	index
la	rft	scielo
php	show	der
oxfordjournals	cpsem	pom
es	info	xue
bies	fmt	svc
onlinelibrary	ver	oom
en	val	ojs

des	casopis	al
tandfonline	aise	del
printsec	traduzione	
full	scholarworks	
di	hdl	

Appendix C R Code

Code adapted from Github user benmarwick's example.

```
# Set working directory
dir <- "C:\\" # adjust to suit
setwd(dir)

# configure variables and filenames for MALLET
## here using MALLET's built-in example data and
## variables from http://programminghistorian.org/lessons/topic-modeling-and-mallet

# folder containing txt files for MALLET to work on
importdir <- "C:\\mallet\\sample-data\\web\\en"
# name of file for MALLET to train model on
output <- "tutorial.mallet"
# set number of topics for MALLET to use
ntopics <- 20
# set optimisation interval for MALLET to use
optint <-  20
# set file names for output of model, extensions must be as shown
outputstate <-  "topic-state.gz"
outputtopickeys <- "tutorial_keys.txt"
outputdoctopics <- "tutorial_composition.txt"

# combine variables into strings ready for windows command line
cd <- "cd C:\\mallet" # location of the bin directory
```

```
import <- paste("bin\\mallet import-dir --input", importdir, "--output", output, "--keep-sequence --
remove-stopwords", sep = " ")
train  <- paste("bin\\mallet train-topics  --input", output, "--num-topics", ntopics, "--optimize-
interval",  optint, "--output-state", outputstate,  "--output-topic-keys", outputtopickeys, "--output-doc-
topics", outputdoctopics, sep = " ")

# setup system enviroment for R
MALLET_HOME <- "c:/mallet" # location of the bin directory
Sys.setenv("MALLET_HOME" = MALLET_HOME)
Sys.setenv(PATH = "c:/Program Files (x86)/Java/jre1.8.0_31/bin")

# send commands to the Windows command prompt
# watch results scroll by in R console...
shell(shQuote(paste(cd, import, train, sep = " && ")),
      invisible = FALSE)

# inspect results
setwd(MALLET_HOME)
# outputstateresult <-
outputtopickeysresult <- read.table(outputtopickeys, header=F, sep="\t")
outputdoctopicsresult <-read.table(outputdoctopics, header=F, sep="\t")

# manipulate outputdoctopicsresult to be more useful
dat <- outputdoctopicsresult
l_dat <- reshape(dat, idvar=1:2, varying=list(topics=colnames(dat[,seq(3, ncol(dat)-1, 2)]),
                                  props=colnames(dat[,seq(4, ncol(dat), 2)])),
            direction="long")
library(reshape2)
```

```
w_dat <- dcast(l_dat, V2 ~ V3)
rm(l_dat) # because this is very big but not longer needed


# write reshaped table to CSV file for closer inspection
write.csv(w_dat, "topic_model_table.csv")
# find the location of that CSV file
# should pop open a window of the folder
# where the CSV is
shell.exec(getwd())
```