Extracting Philosophical Topics from Reddit Posts via Topic Modeling

Evan M. Gertis

September 30, 2021

Abstract

The purpose of this study was to develop a system for extracting philosophical arguments from social media content. In this paper we examine a reddit data that contains the top 1,000 all-time posts from the top 2,500 subreddits, 2.5 million posts in total. From this dataset we examine a sample of 100 posts. In this paper we aim to address a specific question. What are the topics generated by Latent Dirichlet Allocation after removing the top 1000 most common English words from our data set. Our hypothesis is that the topics extracted from the data set will contain philosophical topics.

1 Introduction

What is the history of topic modeling? An early topic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998. [Steyvers] Another one, called Probabilistic latent semantic indexing (PLSI), was created by Thomas Hofmann in 1999.[David M. Blei] Latent Dirichlet allocation (LDA), perhaps the most common topic model currently in use, is a generalization of PLSI developed by David Blei, Andrew Ng, and Michael I. Jordan in 2002, allowing documents to have a mixture of topics.[Blei]

LDA is a generative probablistic model that assumes that each topic is a mixture over an underlying set of words, and each document is a mixture of a set of probabilities. For example, if we take M documents consisting of N words and K topics then the model uses these parameters to train the output.

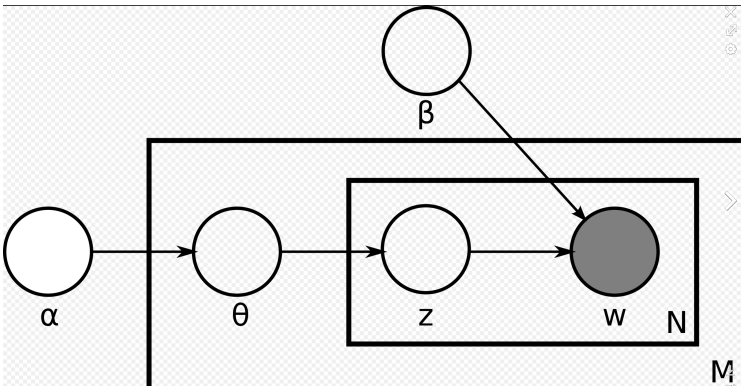


Figure 1: LDA in plate notation

Using the figure fig above. We can describe LDA. with the following parameters.

- K number of topics.
- N number of words in the document.
- α the Dirchlet-prior concentration parameter of the per-document topic distribution.
- β the same parameter of the per-topic distribution.
- $phi(k)$ word distribution for topic K.
- $\theta(i)$ the topic distribtuion for document i
- $z(i,j)$ the topic assignment for word $w(i,j)$
- $w(i,j)$ the j word in the ith document

In this list ϕ and θ are the dirchlet distributions, z and w are the multinomials.

The α parameter is known as the dirchlet prior concentration parameter. It represents document-topic density. With a high alpha, documents are assumed to be made up of more topics and result in more specific topic distribution per document.

The β parameter is a prior concentration parameter that represents topic-word distribution. With a high beta, topics are assumed to be made up of most of the words and result in a more specific word distribution per topic.

In this paper we present an practical application topic modelling. Our justification is that there is need for papers that describe practical implementations of a topic modeling system. In our research we combine the fields of philosophy and text mining in attempt to extract meaningful philosophical topics from conversations on reddit.

2 Literature review

What is the history of topic modeling? An early topic model was described by Papadimitriou, Raghavan, Tamaki and Vempala in 1998. [Steyvers] Another one, called Probabilistic latent semantic indexing (PLSI), was created by Thomas Hofmann in 1999.[David M. Blei] Latent Dirichlet allocation (LDA), perhaps the most common topic model currently in use, is a generalization of PLSI developed by David Blei, Andrew Ng, and Michael I. Jordan in 2002, allowing documents to have a mixture of topics.[Blei] The inspiration for this work was to find a way to apply topic modeling to a dataset. The first paper, [Michael Röder, 2015], presents a method for applying topic modeling. This paper was used to develop the gensim library. The second paper explains how topic modeling enables us to organize and summarize archives that would not be achievable by human annotation. The third paper uses three topic models to evaluate two models. They provide an in depth explanation of LDA analysis.

1. Exploring the Space of Topic Coherence Measures by Michael Röder and others.
2. Reading Tea Leaves: How Humans Interpret Topic Models by Jonathan Chang and others.
3. Surveying a suite of algorithms that offer a solution to managing large document archives by David M. Blei

3 Reading Tea Leaves: How Humans Interpret Topic Models by Jonathan Chang and others.

3.1 Authors

Jonathan Chang, Jordan Boyd-Graber, Sean Gerrish, Chong Wang, David M. Blei

3.2 Overview

The authors of this paper discuss machine probabilistic models for unsupervised analysis of text. They focus on providing a predictive model of future text and a latent topic representation of the corpus. These studies show that they capture aspects of the model that are typically undetected by previous measures of models based on held-out likelihood. The authors of this paper discuss machine probabilistic models for unsupervised analysis of text.

4 Exploring the Space of Topic Coherence Measures

4.1 Authors

Andreas Both, Alexander Hinneburg

4.2 Overview

The authors of this paper present a framework that constructs existing word based coherence measures as well as new ones by combining elementary components. They present combinations of components which outperform existing measures with respect to correlation to human ratings. The authors suggest that results can be transferred to further applications in the context of text mining and information retrieval.

5 Surveying a suite of algorithms that offer a solution to managing large document archives

5.1 Authors

David M. Blei

5.2 Overview

The authors of this paper explain how topic modeling enables us to organize and summarize electronic archives at a scale that would be impossible by human annotation. The example shown in the paper took 17,000 articles from science magazines and identified 100 topics. The application described by this paper show how topic modeling provides an algorithmic solution to managing, organizing, and annotating large archives of texts.

6 Theory

Our hypothesis is that by removing the most common 1000 English words from 100 reddit posts we will be left with philosophical topics that describe the reddit community.

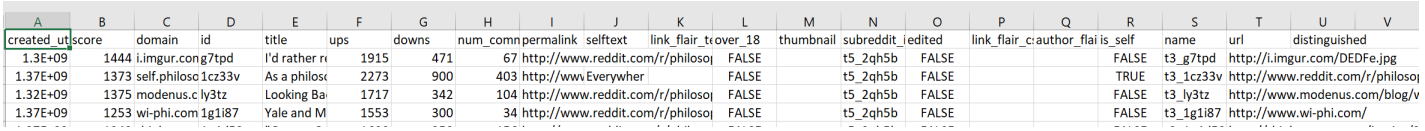
6.1 Designing An Effective Topic Modeling System

The system that we've designed in this study follows the recommended steps described by Jiawei Han.[Jiawei Han] Han describes the knowledge discovery process in 7 steps.

- Data Cleaning (remove noise and inconsistent data)
- Data Integration (reading data from multiple data sources)
- Data Selection (selecting data that is relevant for analysis)
- Data Transformation (data are transformed and consolidated into forms appropriate for mining by performing summary or aggregation operations)
- Pattern Evaluation (identifying truly interesting patterns representing knowledge based on interestingness)
- Knowledge Presentation (visualization and knowledge representation)

6.2 Data Cleaning

The initial reddit dataset contained 21 columns. For our purposes we were only interested in the contents of the title and selfText. The justification for this is that we were only interested in the topics discussed with in the posts. We remove the rest of the metadata columns. We also removed all punctuation and converted titles to lower case.



A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
created_ut	score	domain	id	title	ups	downs	num_comn	permalink	selftext	link_flair_t	over_18	thumbnail	subreddit_i	edited	link_flair_c	author_flai	is_self	name	url	distinguished	
1.3E+09	1444	i.imgur.con	g7tpd	I'd rather r	1915	471	67	http://www.reddit.com/r/philosoj			FALSE		t5_2qh5b	FALSE			FALSE	t3_g7tpd	http://i.imgur.com/DEDFe.jpg		
1.37E+09	1373	self.philosc	1cz33v	As a philosc	2273	900	403	http://www	Everywher		FALSE		t5_2qh5b	FALSE			TRUE	t3_1cz33v	http://www.reddit.com/r/philosoj		
1.32E+09	1375	modenus.c	ly3tz	Looking Ba	1717	342	104	http://www.reddit.com/r/philosoj			FALSE		t5_2qh5b	FALSE			FALSE	t3_ly3tz	http://www.modenus.com/blog/v		
1.37E+09	1253	wi-phi.com	1g1i87	Yale and M	1553	300	34	http://www.reddit.com/r/philosoj			FALSE		t5_2qh5b	FALSE			FALSE	t3_1g1i87	http://www.wi-phi.com/		

Figure 2: Compact display of the reddit dataset display

6.3 Data Integration

The dataset used in this study came from the all-time top 1000 posts, from the top 2500 subreddits by subscribers, pulled from reddit between August 15–20, 2013.[umbrae, 2017] We used a third party library, pandas, to load the data into our system. Once the data is loaded into the system the dataframe object can be used throught the program for data manipulation.

6.4 Data Selection

Our topic modeling system was build with python. We used a 3rd party library, pandas, to load our dataset. Using this library we were able to visualize the columns of data for analysis before and after prepossessing the records.

6.5 Data Transformation

We used the 3rd party library, gensim, to build bigram and trigram models from our selected data. A bigram or digram is a sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words.[gensim] We used the gensim library to automatically detect common phrases – aka multi-word expressions, word n-gram collocations – from a stream of sentences. Bi-agrams and trigrams should start with an n-gram. An n-gram is a contiguous sequence of n items from a given sample of text or speech. Bigrams are two words frequently occuring in the document. Trigrams are three words frequently occuring in the document. After constructing the bigram and trigram models for the dataset we used a 3rd party library, nltk, to construct our set of stop words. Stop words are any word in a stop list (or stoplist or negative dictionary) which are filtered out (i.e. stopped) before or after processing of natural language data (text).[XPO6, 2009] The stop words used in our research come from a list of the 1000 most common words used in the English language.[deekayen, 2021] After we removed the stop words from the dataset we constructed new bigrams from the new data. The next step in the transformation process is lemmatization. Lemmatization is used to enhance the system's accuracy by returning the base or dictionary form of a word..[Rania Albalawi1 and Benyoucef2]

6.6 Pattern Evaluation

We evaluate the topic coherence via LDA analysis. To accomplish this we used the gensim library to develop an LDA model. We then developed a CoherenceModel processes an LDA model constructed from the lemmatized data, a corpus dictionary developed from the lemmatized data, and coherence measure. For our purposes we used the default coherence measure that was mentioned in [Michael Röder, 2015], The topic coherence model used in gensim follows the implementation of the four stage topic coherence pipeline as described by [Michael Röder, 2015].

We evaluated our coherence score using a range of 2 to 11 topics and a step size of 1. Based on trial and error we determined appropriate ranges for $\alpha = [0.01, 0.3]$ and $\beta = [0.01, 0.3]$.

In our analysis we iterate through our validation corpuses. Then for each k topic we evaulate our model for each value of α for each each value of β. This gives us a diverse set of results that utilizes a range values for α and β.

6.7 Knowledge Presentation

The goal of this study was to determine what topics were left after removing common the 1000 most common English words[deekayen, 2021]. As stated by Blei, we can use topic models to organize, summarize, and help users explore large corpora. In our study we used a 3rd party library for Data Visulization, pyLDavis. This library uses the data generated from our model to create an html file which we can then use to examine our topics.

- Selected Topics
- The Interopic Distance Map
- The top 30 modst relevant terms for each topic
- The top 30 Most Salient Terms

7 Research Design

Our hypothesis stated that by removing the most common 1000 English words from 100 reddit posts we will be left with philosophical topics that describe the reddit community.

The questions that we ask in this research are the following:

1. What topics are we left with after removing the 1000 most common English words?
2. Do these topics represent philosophical relationships?

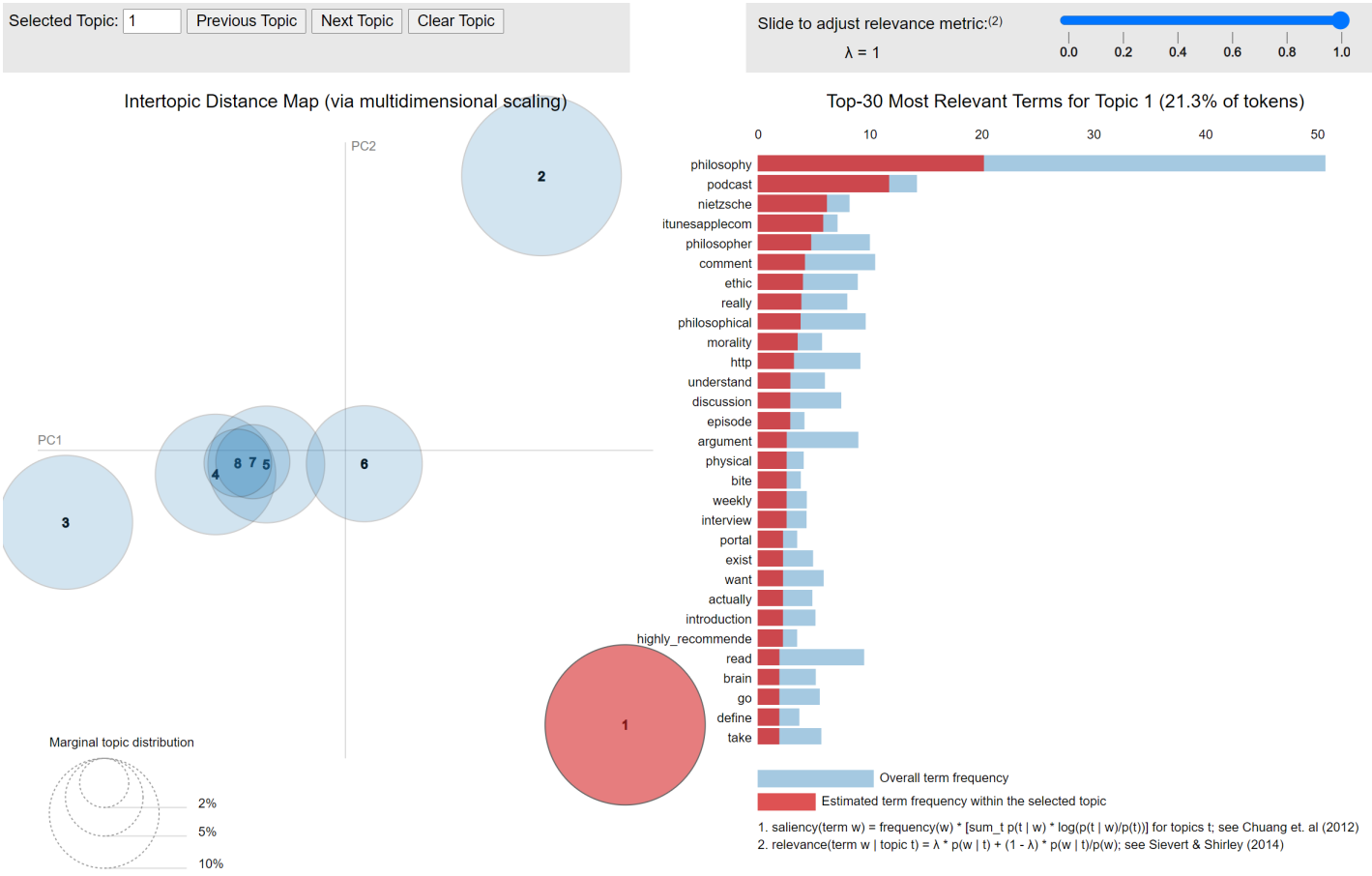


Figure 3: Data Visualization

By removing common words from the dataset that we expected to remove the ambiguity of some of the arguments and thoughts being shared within the reddit community at the time. The word 'ambiguous', at least according to the Oxford English Dictionary, is ambiguous: it can mean uncertainty or dubiousness on the one hand and a sign bearing multiple meanings on the other. Ambiguity has been the source of much frustration, bemusement, and amusement for philosophers, lexicographers, linguists, cognitive scientists, literary theorists and critics, authors, poets, orators and pretty much every other being who uses language regularly to communicate.[Sennet, 2021] By removing overly used words we expect that the context of of a thought or argument would become more clear. As William Russ from Payne College asks Is Truth Relative to Meaning? There is a further potential source of confusion about truth that might be worth addressing at this point. Words and sentences can be used in lots of different ways. Even if we are not being inventive with language, there is lots of vagueness and ambiguity built into natural language.[Payne]. Most words are ambiguous: a single word form can refer to more than one different concept. For example, the word form "bark" can refer either to the noise made by a dog, or to the outer covering of a tree. This form of ambiguity is often referred to as 'lexical ambiguity'.[Rodd, 2021]

8 Conclusion

In this study we analyzed a dataset of 100 reddit posts using Latent Dirchlet Allocation. The hypothesis that aimed to address, what are the topics extracted from a dataset of 100 reddits after removing the top 1000 most common English words? Using the data visualization We can see that one of the most frequent terms for each topic is philosophy. Based on the initial results from this analysis it seems that philiosophy is a frequently discussed on reddit. This is not a concrete conclusion. Much work needs to be done to develop in the area a parameter tuning. Also, the limitated amount of data analyzed present a problem. With a larger dataset and more computing power a better analysis could be performed. We have created a live demonstration of the Data Visualization it is hosted on github. Our contribution to the field is the implementation of the LDA analysis and the Data Visualization that was created from our analysis.

References

D.; Lafferty Blei. *A correlated topic model of Science*. Applied Statistics 1.

John D. Lafferty David M. Blei. Dynamic topic models.

Probabilistic Topic Models David M. BLEI. Surveying a suite of algorithms that offer a solution to managing large document archives. *ACM*, 2012.

deekayen. 1-1000.txt. https://gist.github.com/deekayen/4148741, 2021.

gensim. https://radimrehurek.com/gensim/models/phrases.html.

Jian Pei Jiawei Han, Micheline Kamber. *Data Mining*. Morgan Kaufman.

Jordan Boyd-Graber Jonathan Chang and David M. Blei Sean Gerrish, Chong Wang. Reading tea leaves: How humans interpret topic models. *ACM*, 2009.

Alexander Hinneburg Michael Röder, Andreas Both. Exploring the space of topic coherence measures. *ACM*, 2015.

W. Russ Payne. An introduction to philosophy. *Bellevue College*.

Pantelis Leptourgos Joshua G. Kenney Stefan Uddenberg Christoph D. Mathys Leib Litman Jonathan Robinson Aaron J. Moss Jane R. Taylor Stephanie M. Groman Philip R. Corlett Praveen Suthaharan, Erin J. Reed. Surveying a suite of algorithms that offer a solution to managing large document archives. *Nature*, 2021.

Tet Hin Yeap1 Rania Albalawi1 and Morad Benyoucef2. Using topic modeling methods for short-text data: A comparative analysis.

Jennifer Rodd. Department of experimental psychology, university college london. *Nature*, 2021.

Adam Sennet. Ambiguity, the stanford encyclopedia of philosophy. *The Stanford Encyclopedia of Philosophy*, 2021.

Tom Steyvers, Mark; Griffiths. Probabilistic topic models. *Handbook of Latent Semantic Analysis (PDF). Psychology Press.*

umbrae. Reddit top 2.5 million. https://github.com/hanzohan/reddit-top-2.5-million.git, 2017.

XPO6. List of english stop words. https://xpo6.com/list-of-english-stop-words/, 2009.