# Using Latent Dirichlet Allocation & Word2Vec to measure the relationships between major philosophical works

Bobby Becker
Tulane University
rbecker3@tulane.edu

**Abstract**

A small number of pivotal works in philosophy—including those by Plato, Kant, Aristotle, and others—receive the vast majority of attention and criticism from scholars of philosophy. These works are usually incredibly interconnected, as most major philosophers directly influenced subsequent philosophers and engaged with prior ones. Due to this, Natural Language Processing tools offer exciting opportunities to help scholar analyze those texts' complex relationships to each other. However, despite their recent advancements in ability and accessibility, applying Natural Language Processing algorithms to philosophical texts is still relatively niche within the philosophy discipline. This project illustrates how basic NLP algorithms can uncover thematic relations between major philosophical works. This project both visualizes these relations and presents a potential application for them through a web-interface. This application also presents wider potential for how these NLP algorithms can be used in conjunction with LLMs to give more advanced queries into specific pieces of text.

## Introduction

This project analyzes the corpus of 43 major philosophical texts from thinkers who have had an immense amount of influence on Western thought in all disciplines, especially in the fields of mathematics, politics, physics, psychology, ethics, logic, economics and, of course, philosophy. Among these works include those by Plato, Aristotle, Kant, Descartes, Leibniz, Marx, Freud, Nietzsche, Bertrand Russell, William James, Fredrich Hegel, Ralph Waldo Emerson, David Hume, Baruch Spinoza, Mitchell Montaige, and Blaise Pascal. While the influence of these thinkers is hard to overstate, it is also difficult to quantify; the majority of analysis into their level of influence within philosophy departments is done qualitatively by individuals who read and interpret their works.

Since each major philosopher influences and engages with other writers, we can observe their influence by analyzing the number of and the level of influence that texts have which reference their work (much like how research papers show their influence by their number of citations.) On Wikipedia, for example, it is commonly observed that if a person clicks the first link on each

page, they will soon get to the Wikipedia page on philosophy, which is an easy way to concretely illustrate the field's influence. Similarly, NLP tools can be used to uncover more subtle, mathematical relations between philosophical works that would normally be impossible for individuals to observe.

In this project, we take advantage of two of the most foundational NLP algorithms to analyze these major works: latent Dirichlet allocation—a Baysian network which generates the most important themes of any given passage of text—and a word2vec model, which generates a vector representation of each word based on the word's relation to all other words in the corpus used to train the model.

Using these algorithms in tandem, we created two software artifacts: first, our Jupyter notebook shows how latent Dirichlet allocation can be used to generate words which represent the topics of each philosophical work. For each text, these topic-representing words are then vectorized and averaged together, creating a unique vector representation to represent the work. This vector representation can visualize the relationships between works through Principle Component Analysis—in a method akin to how to vector representations are commonly used to derive semantic relationships between words.

In addition, we built a demo web-interface to illustrate a potential application of this method; this web-interface offers an exciting new possibility for building software systems which combine LLMs and traditional NLP algorithms to query more detailed and accurate information on specific text documents.

## Previous Work

The paper "LDA Topic Modeling: Contexts for the History & Philosophy of Science", published in 2020, spends much of its focus on addressing the common criticisms of topic modelling in the humanities. When used without rigor, they point out out topic modelling algorithms can often give 'superifical', 'blunt', or unreliable information on historical texts. To given legitimately valuable information, the authors suggest using multiple different types of models tailored for each specific document analyzed.

In "Domain-Specific Evaluation of Word Embeddings for Philosophical Text using Direct Intrinsic Evaluation", the authors evaluate the effectiveness of word embeddings trained on the texts of a single philosopher to capture domain-specific meanings of philosophical terms. Traditional language models trained on vastly more texts often do not accurately represent the nuanced language found in humanities texts, including philosophy; the researchers here compare six models trained on specific philosophers to see which one best aligns with expert judgments in tasks related to synonym detection and coherence. While this study is a great example of how Word Embeddings can be studied on the texts of individual philosophers, it is aimed at creating a generative model, and the authors did not use their research to do any comparative analysis between the philosophers studied.

In the paper, “What Is This Thing Called *Philosophy of Science*? A Computational Topic-Modeling Perspective, 1934–2015” the authors analyze the articles covered from the *Philosophy of Science* journal using unsupervised topic-modelling algorithms to identify the rise and fall of specific research topics.

In “Measuring Philosophy in the First Thousand Years of Greek Literature,” the author demonstrates how LDA topic modeling can be used to automatically identify philosophical passages within a large corpus of the first thousand years of Greek literature, sourced from the Open Greek and Latin group and the Perseus Digital Library. By combining qualitative analysis, topic modeling, and expertise in Classics, the study devises three numerical scores to evaluate texts on "good and virtue," "scientific inquiry," and their overall "philosophicalness."

In “Eight journals over eight decades: a computational topic-modeling approach to contemporary philosophy of science”, the authors apply computational text-mining and topic-modeling algorithms to the full-text content of nearly 16,000 articles from eight major philosophy of science journals, spanning from the 1930s to 2017, to trace the evolution of the discipline's research themes. Through this analysis, 25 research themes and 8 thematic clusters were identified, illustrating the dynamic changes in the philosophy of science's research agenda over eight decades and how each journal uniquely contributed to these thematic developments. My research uses similar algorithms, but in an attempt to measure the development of specific pivotal figures who have contributed to the development of science and philosophy, not of journals.

Most of these works are fundamentally different from mine in one of two ways: first, in the first two papers, the authors primarily investigate the effectiveness of the tools in question. Their investigations were largely intended to *argue for* and *analyze* the topic modelling itself, whereas this research tries to apply topic modelling to analyze the texts in question. The following three papers apply topic modeling to much larger collections of textual corpuses; they investigate large numbers of articles to find thematic clusters and themes of works to find bigger picture trends. This project is, in a sense, less democratic; we focus on applying the algorithms to the most notable philosophical texts to try to uncover interesting relationships between them.

## Approach

Our project uses the library Gensim for the LDA and Word2Vec model, NLTK for tokenization and preprocessing, and Project Gutenberg to collected the texts. All 43 texts are used to train the Word2Vec model, and for each textual corpus, LDA is used to generate 3-6 words to represent the text. The vectors of those 3-6 words are then averaged using cosine similarity to create a vector to represent the entire text. Those vectors are then compared to each other: in a matrix, we list how related each work is to each other work in our dataset.

Using the sklearn and matplotlib libraries, we run Principle Component Analysis on the 100 dimensional vectors to reduce them to 3 dimensions and to 2 dimensions; this allows us to

visualize the different works' relationships to each other work just as vectorized words are used to visualize their semantic relationships to other vectorized works. To take advantage of this, we also ran PCA on our Word2Vec model's representations of philosophers and philosophical themes. Those vector representations are then visualized alongside the works which we analyzed.

Our web application takes advantage of this combination of LDA and Word2Vec model to create a query system of Plato's dialogues. For the application, we divided 8 of Plato's dialogues into 500 passages, which we then used LDA to generate 3-6 words to represent the passage. Much like did before, we then take the vector representations of those words using a Word2Vec model and average them using cosine similarity, creating a unique vector for each passage. When the user gives an input on our application, their input goes through the same process: an LDA models generates 3-6 words to represent their query which are then vectorized and averaged. Our application then determines which passage of Plato most represents their query; as a final layer, that passage is given to a GPT-3.5 model, which is instructed to find the most relevant portion of the passage. That portion is then shown to the user, as well as a citation of it (generated by another GPT-3.5 model.)

**Results: Data Representation & Visualization**

In one block of our code, our code determines how related each work is to all 43 other works in the dataset. For this information to be useful, the majority of the list should be expected—as that demonstrates that it has an accurate representation of how the works compare to each other according to the intitution of someone familiar with philosophy. For example, our model says that the works in the dataset most similar to "Kant's Critique of Pure Reason" are as follows:

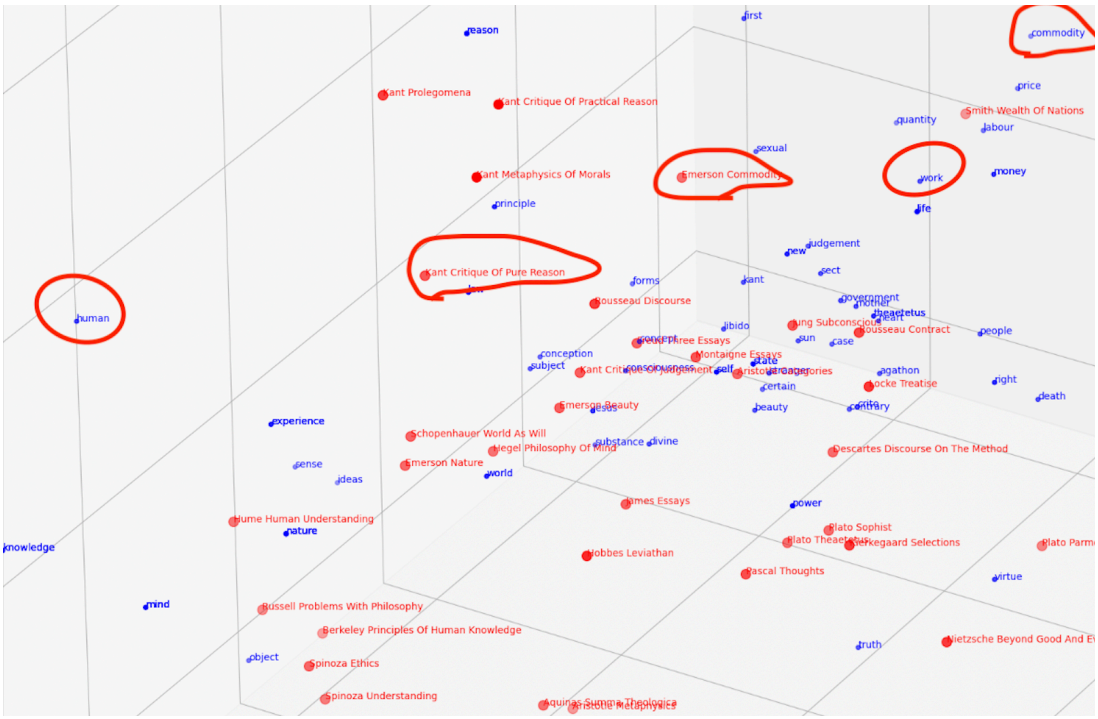
1. Schopenhauer: The World As Will and Representation
2. Kant: Prolegomena to Future Metaphysics
3. Kant: Critique Of Judgement: 0.9775988832149689
4. Ralph Waldo Emerson: Essay on Beauty
5. Ralph Waldo Emerson: Essay on Nature
6. Mitchell Montaigne: Collection of Essays
7. Emerson: Essay on 'Commodities'
8. Rousseau: Discourse on the Origin of Inequality
9. Hume: "An Essay Concering Human Understanding"
10. Hegel's "Philosophy Of Mind"

Results 1, 2, 3, 9, and 10 are quite expected: Schopenhauer and Hegel were direct followers of Kant who both directly cited their influence from him in their main works. The questions raised in Hume's metaphysics and epistemology sparked Kant's critical project, and Kant's Prologomna to Future Metaphysics and Critique of Judgement are built directly from the ideas explored by his Critique of Pure Reason.

The positions of Emerson's essays is less obvious; while Emerson was influenced by Kant, his work is far less related to his (intuitively) than philosophers such as Berkeley, Kierkegaard,

Leibniz, or Descartes. Since we can look into the words which LDA generated to represent the topics of each text, we can try to get a better sense on why our model thinks Emerson's work has a strong relationship to Kant's:

The words used to represent Kant's Critique of Pure Reason from LDA are "Reason", "Experience", and "Conception." The three words used to represent Emerson's essay on Beauty are "Beauty," "Nature," and "Forms", and the three words used to represent his essay on Nature are "Nature," "New," and "Mind." Fair enough—those topics seem similar enough to Kant's Critique of Pure Reason to warrant its placement. However, the three words used to represent Emerson's essay on Commodity are "Commodity," "Human," and "Work"—words that should place the essay far closer to the works of Marx. So, what's going on here? PCA can give us an idea:




In this PCA, we see that the word vectors for "Human," "Commodity," and "Work" are each, actually, quite far from "The Critique of Pure Reason." However, since they are of opposite distances of the Critique, they average out to being quite close to it.

Problem solved right? Well, if we run the LDA model again to get 6 words to represent each book, "Emerson's Commodity" is still considered quite close to Kant's Critique of Pure Reason. Looking at Emerson's essay, I discovered that this isn't a flaw in the model, but that on the contrary, the essay has strong thematic similarities to Kant's work. Here's a short passage:

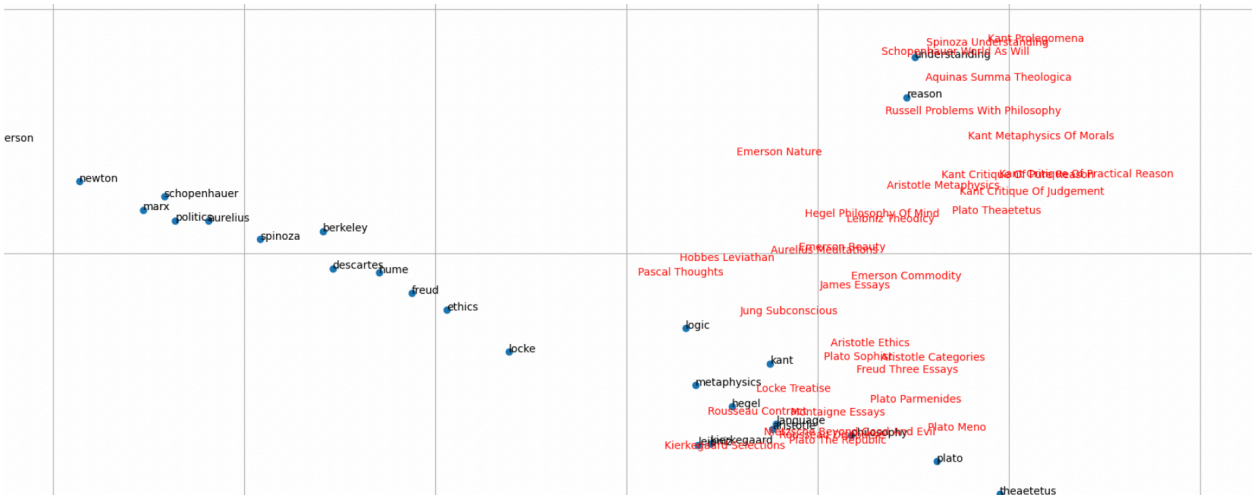
> "Under the general name of Commodity, I rank all those advantages which our senses owe to nature. This, of course, is a benefit which is temporary and mediate,

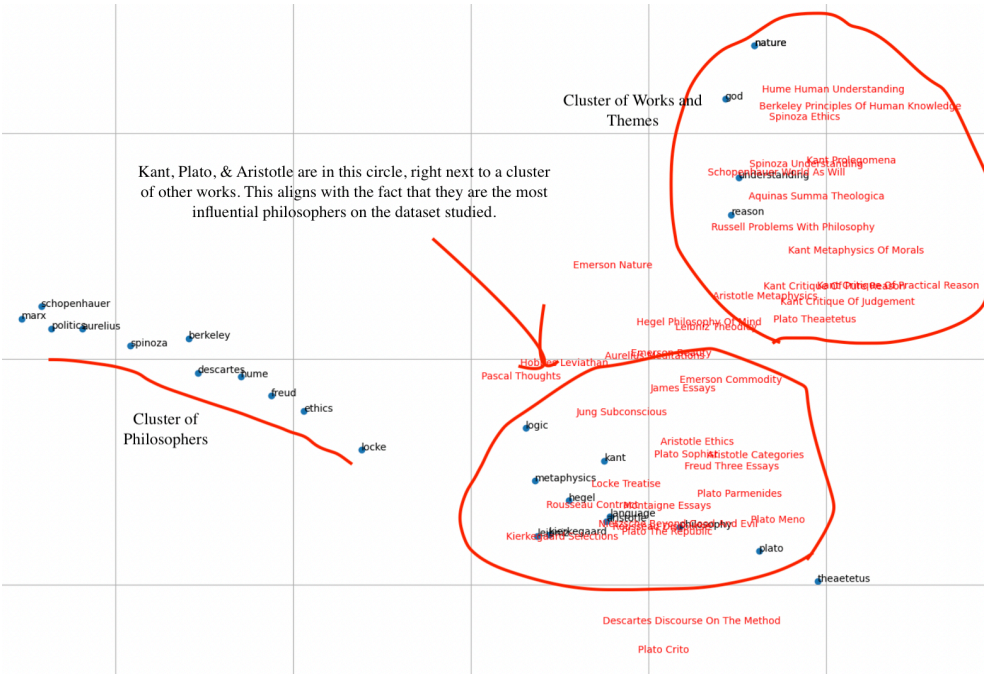
not ultimate, like its service to the soul. Yet although low, it is perfect in its kind, and is the only use of nature which all men apprehend."

Emerson's use of "commodity" was in a very abstract sense unrelated to its common use today in economics. Since our Word2Vec model is trained on a limited number of works, it was able to assign 'commodity' a vector value based on the way it was used. Since, as the quote above shows, Emerson uses the term as a tool to describe his metaphysics and epistemology (the main topics of Kant's Critique of Pure Reason), our model correctly assigns Emerson's essay as very similar to Kant's work.

By tweaking different parameters on our Juptyer notebook file, we can uncover an essentially infinite amount of these investigations. Sometimes, when looking into the reasons why books are determined to be related to other books, we will find the limitations of our model; other times, like in this case, it seems to be able to uncover thematic similarities that are unintituive but truthful.

Here are some general observations on the relationship between the vectors:






**Results: Demo**

Our web application demonstrated a promising architecture for how traditional topic modelling and Word2Vec models can be used in tandem with LLMs to give more advanced queries of a specific text. When the user inputs their interests are in “love as it relates to friendship” the app gives this Plato passage from The Symposium:

## Matching Results

**Passage:** I fancied that he was seriously enamoured of my beauty, and I thought that I should therefore have a grand opportunity of hearing him tell what he knew, for I had a wonderful opinion of the attractions of my youth. In the prosecution of this design, when I next went to him, I sent away the attendant who usually accompanied me. Well, he and I were alone together, and I thought that when there was nobody with us, I should hear him speak the language which lovers use to their loves when they are by themselves, and I was delighted.

**Citation:** This passage is from Plato's "Symposium."

Make Another Query

While more experimentation and research is needed, this simple architecture could potentially be designed to fit many different corpuses of texts, which could allow a user to query to find specific passages of different philosophical works.

The prototype created in this project is, of course, very limited. When given the user input “I’m interested in “politics and questioning authority”, the app gives this passage:

## Matching Results

**Passage:** For he makes me confess that I ought not to live as I do, neglecting the wants of my own soul, and busying myself with the concerns of the Athenians; therefore I hold my ears and tear myself away from him.

While that passage is somewhat related to politics, there is, without a doubt, many dozens of passages with the dialogues of Plato which match the query more strongly.

## Conclusion & Reflection

This was truly a fantastic project for my own learning experience, and I am incredibly proud of the data visualizations and the demo. However, I still feel like I am scratching the surface of how NLP algorithms could be applied to philosophy; it is very difficult to determine if similar methods to what I did already exist (I'm assuming something close to my query system does, because it seems like it has too much potential to be discovered in an undergraduate project, but I have not found anything like it yet.) There is a vast expanse to explore in this niche, both in terms of the texts that can benefit from this type of analysis and the algorithms that can be used.

In the future, there is potential to explore:
→Different scopes of text (models which just analyze the works of a single philosopher, or of a movement.)
→BERT models and more advanced NLP architectures
→Different methods for how LDA and Word2Vec could be combined.

In the future, I would also like to do a better job at more deeply researching this area, so that I can engage more directly with the current methodologies of applying NLP to the humanities. The papers which I read did not get too specific into the exact topic modellings used, and this detail is essential so I can make sure that I'm actually doing something novel.

While I'm excited about this technology, I've also gained an understanding on why it has received criticism in scholarly circles: it is incredibly difficult to determine objective benchmarks in determining the ability of these models. Ultimately, it seems like the data analysis plays a secondary role to my own intuition: if works are rated as similar to each other but don't seem to be, then I label the model as 'wrong' and adjust it. In that sense, isn't this just a tool to confirm my own biases? One way which we can benchmark the quality of these NLP algorithms are if they are *useful*—that is how search engines and LLMs prove their effectiveness. For that reason, continuing to develop this architecture for a querying system could be a logical next step.

There is, also, a ton of potential for analyzing this stuff through the philosophical methodology; while I tried to keep this paper as close to the 'computer science' style of writing as possible, it was hard not to dip into explore deeper, more unknowable questions about the nature of using data to represent semantic relations. Does this data show more information the text itself? Or, does it simply reflect our understanding of them? Or does it give more insight into the mathematical models? All of these are absolutely fascinating questions.

## References


Foster, J.C. (2020). Mapping the Modern History of the Philosophy of Religion with Machine Learning. *The Macksey Journal*. Retrieved from https://mackseyjournal.scholasticahq.com

Malaterre, C., Lareau, F., Pulizzotto, D., & St-Onge, J. (2021). Eight journals over eight decades: a computational topic-modeling approach to contemporary philosophy of science. *Synthese*. Springer.

Köntges, T. (2020). Measuring philosophy in the first thousand years of Greek literature. *Digital Classics Online*. Retrieved from journals.ub.uni-heidelberg.de

Chandra, R., & Ranjan, M. (2022). Artificial intelligence for topic modelling in Hindu philosophy: Mapping themes between the Upanishads and the Bhagavad Gita. *PLOS ONE*. Retrieved from journals.plos.org

Allen, C., & Murdock, J. (2020). LDA topic modeling: Contexts for the history & philosophy of science. Retrieved from philsci-archive.pitt.edu