Web Crawling Linguistic Philosophies for Targeted Search Results and Creation of a Natural Language Generating Agent

Dave Lobue

Northwestern University

MSDS 453: Natural Language Processing

October 4, 2020

Abstract:

The contributions of analytic philosophers to our current understanding of language have formed the basis for much of our semantics-interpretation modeling today. In order to further our collective understanding in this area, the beginnings of a focused language processing agent based on these philosophies and related concepts has been developed. This is superior to traditional search engines today in that focused questions can yield specific answers by returning hyper-targeted search results. This research also will contribute to the eventual development of a text-generation agent, which can function as an interactive student tutor or provide novel perspectives on these concepts.

Introduction

The field of natural language processing has progressed immensely in recent years largely driven by advances in open source shared programming, widely available text databases, and cost reductions in GPU computing infrastructures. This has enabled major leaps forward in the field of artificial intelligence and in particular text generation, but fundamental concepts of 'language creation' have roots dating back thousands of years to the ancient Greek philosophies of Plato and Aristotle. More recently, the linguistic turn of contemporary philosophy contributed a modern perspective of language as it pertains to meaning and understanding intermingling with the field of mathematics. Major contributions from this era include works by Ludwig Wittgenstein, Gottlob Frege, and Bertrand Russell, who's philosophies still influence language interpretation and processing today. To better understand the theories of these philosophers and learn from their works, it is important that their perspectives not only be accessible but integrated with modern AI so that they can again contribute to novel philosophies of language.

Text generation based on a narrowly defined but rich corpus of linguistic philosophers is the end goal of this research. The best method to achieve this is to process all written documents by this set of philosophers and build deep neural network models in order to generate predictive text 'learned' from these subjects. However, since most of these texts are not freely available to scrape online, The Stanford Encyclopedia of Philosophy has one of the largest repositories of philosophers, philosophies, and related materials available for free and will be used as a starting corpus for this objective. Novel text generation based on these sources not only has the potential to deepen our understanding of language but can also make immediate contributions in the academic world in the form of a student tutor or reference agent for those seeking answers to specific questions within this branch of philosophy. Students, teachers, or anyone with interest can benefit from a focused chat bot of this nature that can readily retrieve, reference, or contribute perspective on analytic language, understanding, or meaning based questions.

Methods

In order to achieve these objectives, this research can be broken out into three distinct components: (1) Data Collection, (2) Information Distillation, and (3) Generative Modeling. This paper focuses on the first two topics, as the deep learning phase is yet to be conducted and thus can only be generalized without providing specific examples or learnings from this corpus.   The text data sources that comprise the current corpus originate from the Stanford Encyclopedia of Philosophy due to its detailed and expansive collection of philosophical works. This is an excellent starting point since topics are clearly cataloged with direct links to related entries provided on each article page. As a result, the web crawler can begin its collection from a central node (in this case 'Ludwig Wittengstein' was used as a starting point) and all related philosophers, topics are directly linked from this node. The crawler is currently limited to the plato.stanford.edu domain but can be easily scaled to crawl external sites as numerous journals and reference domains are directly linked from each encyclopedia article. However, with this expansion brings several challenges in information source relevance, to be addressed later.

The Scrapy framework was used to build the focused crawler for data collection and since each article within the Stanford Encyclopedia of Philosophy has multiple subtopics, text parsing was completed using a combination of the BeautifulSoup package, Regular expression operations, and traditional Python string parsing methods. The decision to sub-index each article entry topic was made because it enables a more detailed hierarchical structure, making information retrieval more targeted. For example, the second-tier node of 'Mental Representation' as the article states, "is, arguably, in the first instance a theoretical construct of cognitive science" (Pitt, 2020) and thus given it's dense content is broken out into individual sub-documents based on table of contents <h2> tags such as: (1.1) Representational Theory of Mind, (1.2) Propositional Attitudes, and (1.3) Conceptual and Nonconceptual Representation.

Given the breadth of topic coverage within this dense field of study, information retrieval efficiency and accuracy are critical to building a useful query-based engine. The Python built program includes a query prompt input operation where users are asked to enter questions about specific topics. Using these user generated inputs, the program searches all previously externally retrieved documents using TF-IDF vectors returning results based on cosine similarity distances. The 5 most relevant results are retrieved, and the user is presented with the top document name, the matching keywords, and an option to view the full text, if desired. To ensure query retrieval accuracy, each document is cleaned and tokenized prior to vectorization through a process including: punctuation stripping (regex), alpha numeric filtering, stopword removal (using NLTK), and vocabulary filtering generated from the full corpus of words with a minimum occurrence = 3. From this methodology, link crawling a depth of level 1 (links followed) results in: 97 docs from 19 subjects with vocabulary: 10,207 and tokens: 3,676 and for level 2: 1,049 docs from 160 subjects with vocabulary: 34,360 and tokens: 14,118.

Results

Given the specialized nature of this crawler, the queries tested and in general should be focused around general topics of language, semantics, language of mathematics, logic, etc. but can also be useful for more broad philosophical categories of religion, science, or metaphysics, insofar as they relate to the field of analytic philosophy. For example, from the query prompt “What would you like to learn about today?” by entering the input response: “Can language provide a formula of meaning?” returns a top hit from the article “Propositional Function” subsection “Montague Semantics” containing the top 5 keyword identifiers: [(‘montague’, 9), (‘formula’, 8), (‘expressions’, 7), (‘semantics’, 5), (‘expression’, 5)]. After accepting the program’s user-prompt to view the full text, the article begins:

> “…One theory in which propositional functions do good work is Montague semantics, developed in the late 1960s by Richard Montague. In order to understand Montague's method we need to understand lambda abstraction. For the formula A(x) we read the expression λx[A(x)] as a predicate expression. It extension (in a given possible world) is the set of things that satisfy the formula A(x). Lambda abstractors are governed by two rules, known as α-conversion and β-reduction: (α-con) A(a) (a formula with a free for x) can be replaced by λx[A(x)]a. (Pietroski, 2015)

Clearly this is a relevant search hit, which also goes on to discuss the meanings of expressions and beginnings of a formula for sentence structure. By comparison, the same query posed to Google search returns the following top 5 hits:

(1) Wikipedia.org “Formulaic language”: https://en.wikipedia.org/wiki/Formulaic_language
(2) Wikipedia.org “Language of mathematics”: https://en.wikipedia.org/wiki/Language_of_mathematics
(3) ELT article: “Real Language through poetry”: https://academic.oup.com/eltj/article-abstract/57/1/19/514700?redirectedFrom=PDF
(4) Dictionary Definition for “Formula”: https://www.learnersdictionary.com/definition/formula
(5) Merriam Webster Definition for “Formula”: https://www.merriam-webster.com/dictionary/formula

If a student or researcher were looking for a quick answer to this question without having to go through multiple subsequent links, the targeted search engine would be superior choice.

Accepting the current crawler is based solely on the Stanford Encyclopedia of Philosophy, this performs no better than the query search of the source site itself (insofar as the index algorithms are comparable). Expanding the scope of the crawler to include academic journals, full text documents (e.g., Tractatus Logico-Philosophicus is available in full on http://www.gutenberg.org/ebooks/5740) and other relevant sources from external domains will deepen the scope and value of this particular agent.

Literature Review

There are several philosophy based web indices available, notably https://philindex.org/, https://philpapers.org/, or https://www.ergophiljournal.org/ although these and those similar do not have the same specialized focus as this NLP agent. Aggregators such as these would function more as a contributing source, since the focus of this particular bot is to provide hyper-focused responses to specific (or broad) philosophical questions and not solely return entire articles on a subject.

Conclusions

A focused crawler of this nature can provide immediate benefits to the academic and research communities. Inevitably, opening the gates of a philosophy crawler will reach topics of Plato and Aristotle, which can in turn lead to subjects of nearly all academic disciplines known today. As the crawler expands to additional domains (including those listed in the literature review), it is critical that additional classification and document scoring be incorporated to maintain the relevancy and intention of this language focused agent. These operations should be incorporated into the early/mid pipeline operations to identify relevant and high value nodes and discard those paths that lead to under-indexed documents on TF-IDF scoring.

References:

Pietroski, Paul, "Logical Form", *The Stanford Encyclopedia of Philosophy* (Spring 2016 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2016/entries/logical-form/>.

Pitt, David, "Mental Representation", *The Stanford Encyclopedia of Philosophy* (Spring 2020 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2020/entries/mental-representation/>.