Digging by Debating (DbyD):

From Big Data Text Repositories to Argument Analysis

- The InPhO Group, Indiana University: Colin Allen with Robert Rose, Jaimie Murdock, Jun Otsuka, Doori Lee, Indraneel Dey
- The Cyberinfrastructure Network Science Center, Indiana University: Katy Börner with Robert Light
- The International Centre for Public Pedagogy (ICPuP), University of East London (UEL): Andrew Ravenscroft with Simon McAlister
- ARG:dundee, University of Dundee: Chris Reed with John Lawrence
- Centre for Digital Philosophy, University of Western Ontario: David Bourget
From Big Data to Argument Analysis

Linking massive datasets to specific arguments, where ‘text is data’

Project Goals

● Uncover and represent the key argumentative structures of digitized documents from a large philosophy/science corpus;
● Allow users to find and interpret detailed arguments in the broad semantic landscape of books and articles, and to support innovative interdisciplinary research and better-informed critical debates

Data Sources

Stanford Encyclopedia of Philosophy, HathiTrust Collection, PhilPapers

4 Levels of Analysis: Macro (Sci/Phil Maps) to Micro (detailed arguments)

1. Visualizing points of contact between philosophy and the sciences
2. Topic modeling to identify the volumes/pages ‘rich’ in a chosen topic;
3. Identify and map key arguments; apply a novel analysis framework for propositions and arguments;
4. Sentence modeling to get back to HathiTrust materials
Mapping Philosophy in the Sciences
(analysis level 1)

- **UCSD Map of Science:** generated using more than 12M papers and their references from Elsevier’s Scopus and Thomson Reuters’ Web of Science (25,000 journals), see [http://sci.cns.iu.edu/ucsdmap](http://sci.cns.iu.edu/ucsdmap)
- Shows **554 subdiscipline nodes** aggregated into **13 color-coded disciplines**.
- Overlaid are citations made by the **Stanford Encyclopedia of Philosophy** to visualize the impact of the sciences on philosophy.
- **Node sizes** scale from 0 (no circle) to 43. Highest numbers are in **Humanities, Earth Sciences, and Math & Physics**.
LDA Topic Modeling: Bayesian method generates set of "topics" – probability distributions over terms in the corpus
- Every topic contains every term – different probabilities in the different topics
- The number of topics is a user-selected parameter

Finds the set of topics best able to reproduce term distributions in the documents

Documents may be whole volumes, chapters, articles, single pages, even individual sentences – modelers' choice
LDA Topic Modeling: Bayesian method generates set of "topics" – probability distributions over terms in the corpus

- Every topic contains every term – different probabilities in the different topics
- The number of topics is a user-selected parameter

- Finds the set of topics best able to reproduce term distributions in the documents
- Documents may be whole volumes, chapters, articles, single pages, even individual sentences – modelers' choice
1. Identifying arguments from rated pages (currently human/manual, but developing algs for automation)

2. Mapping of key arguments with OVA mapping tool:
   - Provides a formal framework, or ‘lens’, for investigation and comparative analysis, e.g. role, structure, status
   - More indirectly: meaning, importance and context
Back to the Sentences
(analysis level 4)

(P2 A0) [Bethe says] Every statement that another being possesses psychic qualities is a conclusion from analogy, not a certainty; it is a matter of faith

['every', 'statement', 'another', 'possesses', 'psychic', 'qualities', 'conclusion', 'analogy', 'certainty', 'matter', 'faith']
Back to the Sentences
(analysis level 4)

What did we get? -- tokenized sentences (word lists) followed by original text...

```
In [22]: #tok_sents
In [23]: orig_sents

Out[23]:
['Every statement that another being possesses \n psychic qualities is a conclusion from analogy, not a certainty; it is a matter of faith.',
'If any consciousness \n accompanies it, then the nearest human analogy to such \n consciousness is to be found in organic sensation, and these, \n as has just been said, must necessarily be in the human mind \n wholly different in quality from anything to be found in an \n animal whose structure is as simple as the Amoeba’s."
',
'; learning, 208, \n \n 214.',
'On the other \n\n1).',
'Dytiscus, 86.',
'Burnett, 126, 170.',
'Willem, 130, 192.',
'Caterpillars, 192, 196.',
'Murbach, 107, 158.',
'Fancy, for example, one of us entering a \n room in the dark and groping about among the furniture.',
'This, of \n course, does not refer to the power to judge distance.',
'Again, a bodily structure entirely unlike our own \n must create a background of organic sensation which renders \n the whole mental life of an animal foreign and unfamiliar to \n us.',
'She dispenses of the psychic learning by experience theory of Nagel \n by saying that the only experience upon which the animal \n could reject the filter paper must be experience that it is not \n good for food.',
'We speak, for example, of an “angry” wasp.',
'All traces of the “learning” \n \n “learning” \n \n \n give evidence of the power to judge distance.'
]
```
DbyD: Conclusions so far...

Project Achievements - *proof-of-concept and loose integration of key components*:

1. Method for visualizing points of contact between philosophy and sciences
2. Method for Text selection from Big Data using multi-scale modeling techniques
3. Identified, represented and mapped key arguments about topics (OVA) and devised a novel framework for investigation and comparative analysis
4. Used sentence-level Topic Modeling to ‘go back’ to the texts to find similar propositions to those mapped (investigating context)

Future work

- From loose integration to usable tools, e.g. linked to Philpapers
- To understand and incrementally construct argument maps (mark-up interface) AND automated extraction and mapping
- Funding from future DiD or similar Big Data initiatives (*‘open data’?*)

diggingbydebatings.org