Breaking Barriers in Research Projects: BeagleTM, a Powerful Python-based Text Mining Tool for Visual Discovery in Scientific Literature

Oliver Bonham-Carter, Ph.D.
Dept of Computer and Information Science, Meadville, PA

ALLEGHENY COLLEGE

https://www.cis.allegheny.edu
obonhamcarter@allegheny.edu
Presented at PyCon 2024

Project Objectives

This project presents

- A study by text analysis of ethically inclined language in the literature of Bioinformatics research, and in related disciplines
- Results are amassed by use of the BeagleTM text analysis project
- We create networks to visualize relationships between articles of the literature that share common threads of language relaying ethical themes

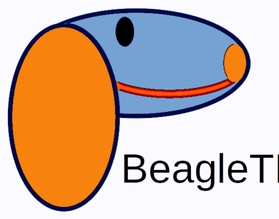

Figure: 1. BeagleTM is a robust, supervised, text analysis software written in Python that uses *bag of words approach.*

Methods

Corpus

- Data: a corpus created from the non-commercial publication archives of National Center for Biotechnology Information (NCBI) `https://www.ncbi.nlm.nih.gov/`
- Over 19 GB of textual data: articles originating from over 2000 respected publishers (i.e., Science, Nature, Elsevier, IEEE, ACM and others) of diverse subjects in science

Keywords

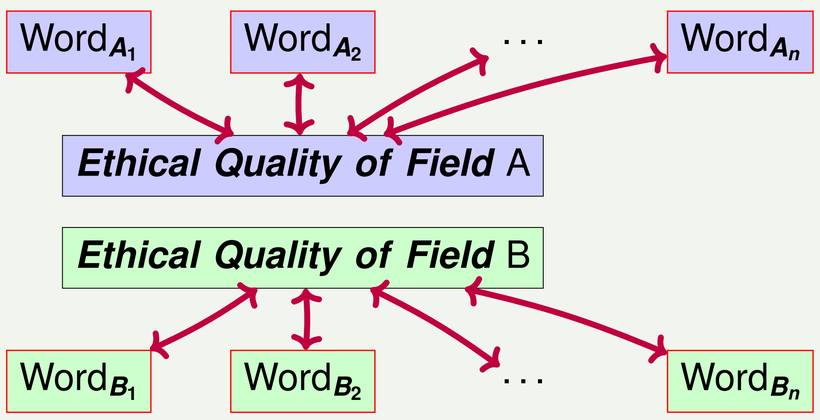

Figure: 2. **Determining relevant keywords to extract discussion of ethics in research.** Keywords are selected from two disciplines; *A* and *B*. We note that each discipline has own lexicon of words relating to ethical details. To capture the nuances of each area, our curated list of keywords concerns a host of simple definitions of ethical issues in articles.

- The keywords were chosen to be *generic* indicators of conversational ethics in scientific articles
- **Disciplines**; *Biology*, *Bioinformatics*, *Computer Science*, and *Informatics*, and **Terms** relating to ethical conduct in research (i.e., *ethics*, *liability*, *responsibility*, and similar)

Parsing for Keywords in Literature to Create Networks

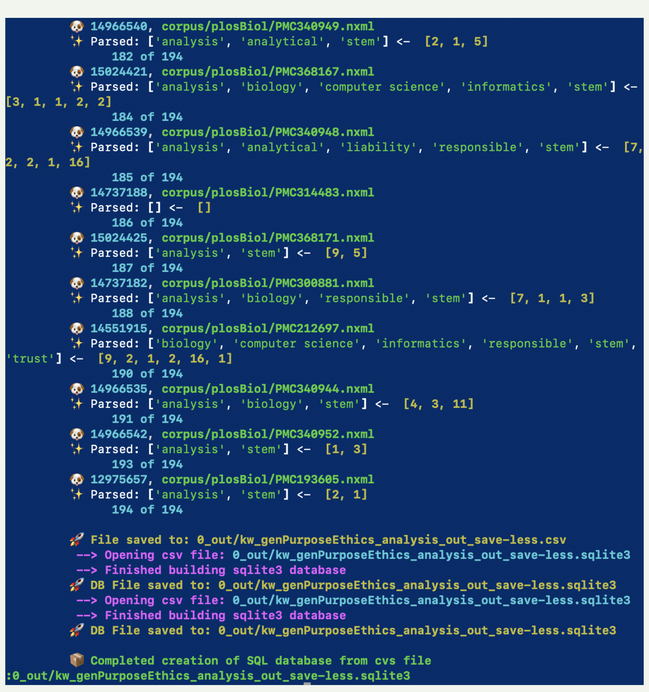

Figure: 3 Left; we used BeagleTM to parse all articles in the corpus created from PubMed archives for specific keywords. Articles in which keywords were found were used to build Relationship Networks to visualize connections of related ideas. Right; BeagleTM's network creater and browser.

- Python resources: Poetry (Parsing), Streamlit (Network Creation and Browsing), NetworkX, PyVis and Plotly (Network creation) and others

Methods Continued

Relationship Networks

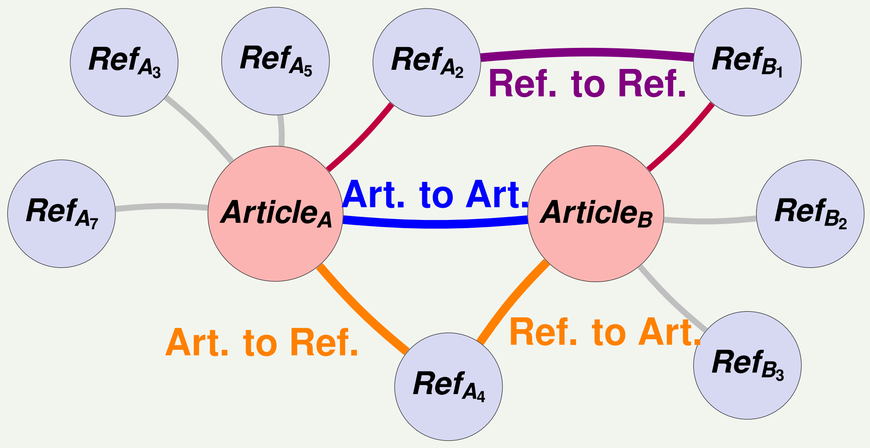

Figure: 4. Relationship networks. The larger (red) nodes represent the articles in which keyword occurrences were found. The smaller (blue) nodes represent the supporting documents comprising the article's bibliography. The edges connecting nodes imply that one publication is citing the other.

There are three types of connections to observe in these Relationship Networks

- **Type 1**: *Reference* to *Reference*
- **Type 3**: *Article* to *Article* (Strongest)
- **Type 2**: *Article* to *Reference* to *Article*

Results

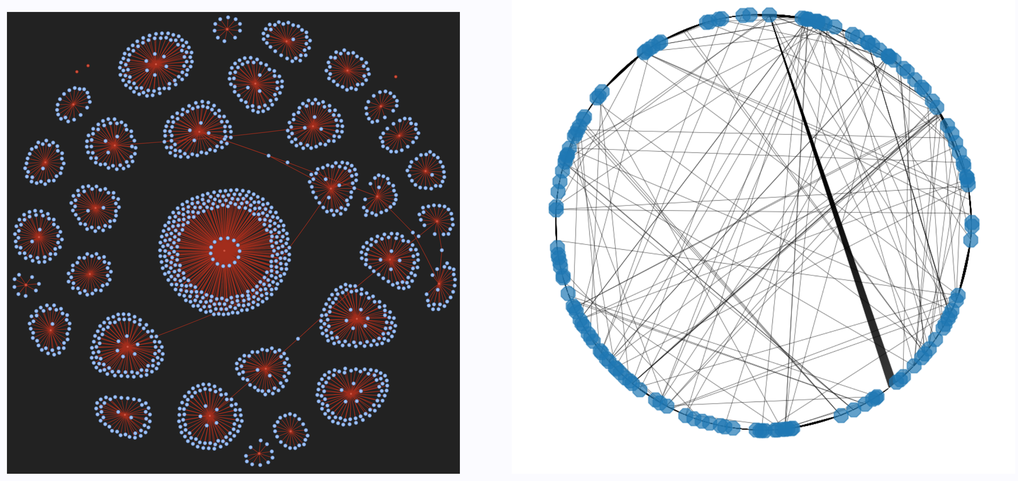

Figure: 5. **Left: A Relationship Network**: Groups represent articles (red nodes) sharing edges with bibliography references (blue nodes) according to common language stemming from contextual keywords. **Right: A general view of connectivity**: Nodes indicate articles, and the edges indicate common keywords. Here, all nodes must have the *same set of keywords* to be included the network, likely indicating common ideas.

Conclusions

- By studying the keyword content in a discipline's articles, one may study the spread of word usage to infer a spread of ideas
- Articles (generally) that reference others having ethical inclinations, appear to also contain similar ethical language

References

- BeagleTM2: `https://github.com/developmentAC/beagleTM2`
- Bonham-Carter, Oliver. "Text Analysis of Ethical Influence in Bioinformatics and Its Related Disciplines." Future of Information and Communication Conference. Cham: Springer Nature Switzerland, 2024.