Tracing the Genealogies of Ideas with Sentence Embeddings

Lucian Li
PhD Student
School of Information Science
University of Illinois, Urbana-Champaign
zilul2@illinois.edu

Abstract

Detecting intellectual influence in unstructured text is an important problem for a wide range of fields, including intellectual history, social science, and bibliometrics. A wide range of previous studies in computational social science and digital humanities have attempted to resolve this through a range of dictionary, embedding, and language model based methods.

I introduce an approach which leverages a sentence embedding index to efficiently search for similar ideas in a large historical corpus. This method remains robust in conditions of high OCR error found in real mass digitized historical corpora that disrupt previous published methods, while also capturing paraphrase and indirect influence.

I evaluate this method on a large corpus of 250,000 nonfiction texts from the 19th century, and find that discovered influence is in line with history of science literature. By expanding the scope of our search for influence and the origins of ideas beyond traditional structured corpora and canonical works and figures, we can get a more nuanced perspective on influence and idea dissemination that can encompass epistemically marginalized groups.

1 Introduction

In Darwin's Plots, (Beer, 2009) examines Darwin's influence on literature as a complex and reciprocal system. Beer identifies in Darwin's writings not only the influence of naturalists and geologists like Lyell, but also the stylistic and lyrical influence of Wordsworth, Coleridge, and Milton. Proceeding onwards, Beer delves into a close reading of how Darwinian metaphors, themes, and worldviews emerge in the works of George Eliot and Thomas Hardy, both correspondents of Darwin who wrote extensive commentaries and reactions to the Origin of Species.

As Beer's work shows, there are connections between intellectual figures and avenues for the spread of ideas not possible to observe except through deliberately interdisciplinary efforts. But scholars cannot have expertise in every field and every potential author; experts with training in dozens of subfields and time to read hundreds of thousands of books are in short supply.

Computational methods can enable analysis across some of these boundaries. In this paper, I present a novel method to detect intellectual influence across a large corpus. Taking advantage of the unique affordances of large language models in encoding semantic and structural meaning while remaining robust to paraphrasing, we can search for substantively similar ideas and hints of intellectual influence in a computationally efficient manner. Such a method allows us to operationalize different levels of confidence: we can allow for direct quotation, paraphrase, or speculative similarity while remaining open about the limitations of each threshold.

I apply an ensemble method combining General Text Embeddings (GTE), a state-of-the-art sentence embedding method described in (Li et al., 2023) optimized to capture semantic content while also retaining aspects of style and vocabulary choice. I vectorize sentences from a corpus of roughly 250,000 nonfiction books and academic publications from the 19th century for instances of ideas and arguments appearing in Darwin's publications. This functions as an initial evaluation and proof of concept; the method is not limited to detecting Darwinian ideas but is detecting similarities on a large scale in a wide range of corpora and contexts

2 Related Literature

Previous attempts to quantify and detect intellectual influence have taken three overall directions: topic modelling, text reuse detection, and word sense similarity. Studies using topic models generally compare topic distributions across documents

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 9–16
November 16, 2024.

or subdocuments. They can capture a zeitgeist of themes and shifting focus but lack granular focus on specific claims. (Rockmore et al., 2018) uses topic models to trace the genealogy of national constitutions. In (Barron et al., 2018), the authors measure K-L divergence of the Topic Distributions of French Revolutionary speeches. In general, these approaches are generally more effective in a limited context with a controlled set of topics and a high likelihood of influence between documents in the corpus. However, changes in topic distribution may reflect high level shifts in societal context rather than direct influence.

Text reuse methods focus on high confidence detection of exact quotation. They can detect one form of direct influence with near certainty but are more limited to paraphrasing and indirect influence. (Funk and Mullen, 2018) and (Smith et al., 2015) both search a large corpus for direct quotations while using a mix of computationally intensive corrections to remain robust to OCR errors. While direct quotation detection ensures high confidence, it necessarily only captures a very limited range of potential influence, excluding similarities in language use, indirect quotation, and similar claims. The n-gram alignment problem is also highly computationally intensive, and requires extensive resources to apply to large corpora.

Finally, approaches focused on detecting similarity and changes in word sense (for example, comparing diachronic embeddings of how concepts like 'justice' evolved over time) can capture stylistic and discursive influence. (Soni et al., 2021) studies Abolitionist newspapers uses word2vec word embeddings. Other approaches, such as (Vicinanza et al., 2023) use language models such as BERT to measure stability and innovation in word senses. However, these findings can be very difficult to interpret across entire vocabularies and are unable to capture any changes in content or argumentation. The influence they capture is also highly speculative; stylistic changes may reflect wider shifts in language use instead of direct interactions.

My proposed method attempts to synthesize text reuse and word sense embedding methods. By evaluating claims on the sentence level, we can gain a granular understanding of specific ideas, while also remaining open to abstract similarities in meaning or structure. Specialized sentence embeddings language models have demonstrated improved effectiveness in encoding semantic meaning in general evaluation tasks as compared to standard BERT and Word2Vec embeddings (Reimers and Gurevych, 2019). Sentence embeddings have been applied to the task of detecting citation and plagiarism in general academic literature in (Alvi et al., 2021) and (Lagopoulos and Tsoumakas, 2021) as well as encoding documents specific to disciplinary subfields in (Chen et al., 2019). I selected GTE vectorization because of the lower computational demands of the GTE-small model and its higher performance in evaluation metrics to other sentence embedding methods.

Finally, the subword tokenization strategy used by BERT and more recent language models was demonstrated in (Nguyen et al., 2020) to be resilient to OCR error. Real large scale historical datasets, such as HathiTrust's digitized book collection, have extensive OCR error, averaging 7% and up to 20% character error depending on the scan quality and time period (Jiang et al., 2021).

Previously published word embedding and topic model approaches are heavily impacted by OCR character error, and while some text reuse approaches mitigate OCR error through machine learning correction, these tend to be highly corpus specific.

3 Dataset

To evaluate my method, I constructed a dataset based around authors active in 19th century academic societies in the British Empire. I curated a list of journals based on secondary readings (Pal, 2014) (Barton, 1990) as well as prior knowledge about the period. This is not meant to capture comprehensively all academic publications in the 19th century, but rather to gather a representative cross section of the most active members of this community. Below is a list of the journals scraped:

- General:
  - Royal Society
  - Royal Institution
  - Cambridge Philosophical Society
- Chemical:
  - (London) Chemical Society
- Medical:
  - (London) Medical and Chirurgical Society
- Biological:

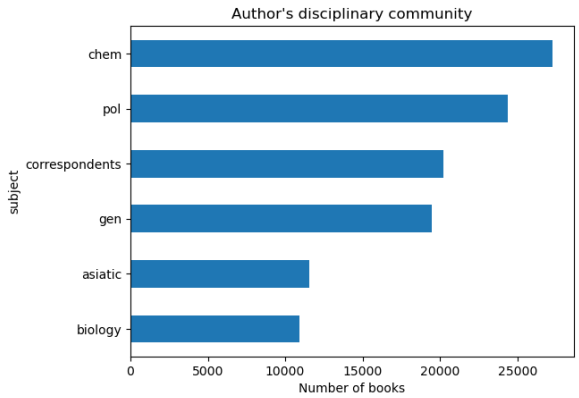

Figure 1: Distribution of disciplinary community of author (extensive overlaps between classes)

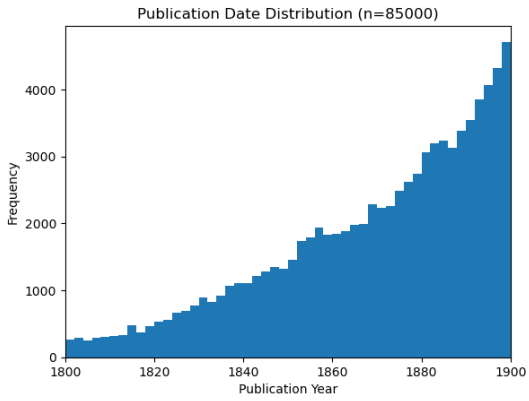

Figure 2: Distribution of books by year

  - Linnean Society
  - Zoological Society
  - Entomological Society
- Geographical:
  - Geographical Society
  - (Royal, Calcutta, American) Oriental Society
- Political and social scientific:
  - The Economist
  - Westminster Review
  - Edinburgh Review

I grouped these societies into proto-disciplines such as biology, geology, chemistry, and politics/social science. I constructed a supplementary dataset of books by Darwin's correspondents using letters from the Darwin Correspondence Project.[1] Author names were extracted from downloaded proceedings using Spacy's NER utility. 250,000 books by the 1.000,000 identified potential authors were downloaded as digitized texts from the Internet Archive and Project Gutenberg. Metadata about the books used are available in this csv. [2] I also used the Project Gutenberg editions of Darwin's Origin of Species and Descent of Man and Herbert Spencer's Principles of Sociology and Principles of Biology for a comparative sample.

4 Method

4.1 Preprocessing

I performed sentence tokenization per book using NLTK. Overly short documents (<1000 characters) and sentences (<45 characters) were removed because short documents tended to be either mislabelled, or consist mostly of images that could not be accurately converted to text. Short sentences tended to not contain enough information for a coherent argument, or represented formatting, index, and table of contents elements. No further text cleaning was performed; the BERT base of the model used for vectorization uses case and punctuation markings to encode meaning.

4.2 Model selection and finetuning

I used the General Text Embeddings (GTE) model, a BERT based approach fine-tuned with internet text and specific entailment tasks to capture semantic meaning. For a preliminary set of randomly sampled books, GTE embeddings were generated for each sentence using the GTE-small model implemented in the sentence-transformers Python package. GTE-small was selected due to memory and computational power constraints.

For fine tuning, I randomly sampled pairs of books to generate 1,000 pairs of sentence similarity scores. I inspected the pairs to label the accuracy of the score. If the sentences were similar due to purely coincidental factors (for example, transitional phrases like "I go on to argue" or "it should be obvious"), I assign a score of -1. If they have a missed similarity (i.e. making the same argument) but have a score that does not meet the threshold, they are assigned a score of 1. Otherwise, if the score if correct, the fine tuning score was left the same. It is difficult to determine the effectiveness of this method in resolving false positive matches across the broader corpus due to the lack of labelled data, but it successfully removed all instances of the hand annotated false positives from future matches.

[1] https://www.darwinproject.ac.uk/
[2] https://uofi.app.box.com/file/1412863623947

I used these pairs to fine tune the GTE-small model using cosine embedding loss in Hugging Face.

4.3 Search

No additional hyperparameter changes were performed. Using my fine tuned version of the GTE-small model, I generated sentence embeddings of each sentence in the corpus. From these vectors, I used FAISS (Douze et al., 2024) to create rapidly searchable cosine indices for every sentence in the corpus. For further analysis, I used thresholds of >0.85 cosine similarity (speculative and low confidence), >0.90 cosine similarity (indirect/medium influence) and >0.95 (high confidence and direct quotation). All code for the project are available in this GitHub repository. [3]

5 Findings

5.1 Robustness to OCR error

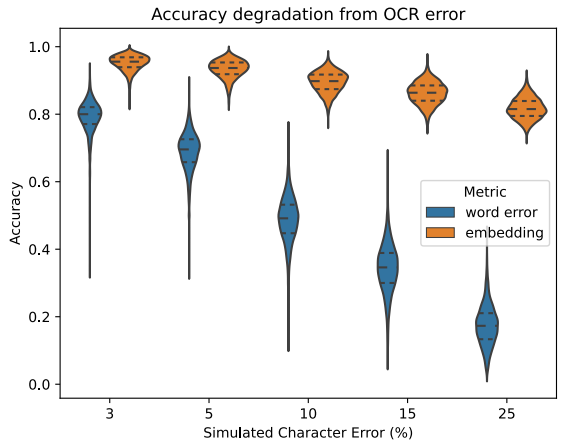

Figure 3: Error rates on simulated corrupted data with artificially permuted characters. The blue distributions show the word error rate at different levels of character error. The orange shows the cosine similarity in sentence embeddings between the original and corrupted text.

First, I evaluate the impact of low quality OCR on the performance of this method. Figure 3 demonstrates its robustness. I took a subset of the corpus consisting of human transcribed books from Project Gutenberg and simulated character error by randomly permuting parts of each sentence with a random character, integer, or empty string. I compared the sentence embedding representation of this new string with that of the original string, and found that there was generally very little decrease in cosine similarity. Even at 10% character error, roughly 90% of sentences will still be captured above the 0.85 cosine similarity threshold.

Conversely, word accuracy is highly sensitive to increased character error. At the most frequent CER of 7% in real scanned corpora, between 30-50% of words are corrupted. At higher CER levels, the overwhelming majority of words are lost. For dictionary based approaches, like text reuse, word embeddings, and topic modelling, this creates extensive accuracy issues.

This method is more robust to OCR error than dictionary based approaches at all CER levels, suggesting that in applications where transcription error is expected, this method will generally preserve more signal accuracy.

5.2 Validation against historical ground truth

Because annotated data does not exist for the very messy corpus of scanned 19th century books, I conducted evaluation against historical ground truth. I take the set of sentence embeddings for selected books by Darwin, Herbert Spencer, and 2 randomly sampled books published in the same year as *Origin of Species*. Based on academic consensus about Darwin (Mayr, 1995), and documented evidence about his correspondents, we would expect Darwin's publications to display more similarity with his intellectual circle and in certain disciplinary communities (geology, natural history) vs (chemistry and orientalist circles).

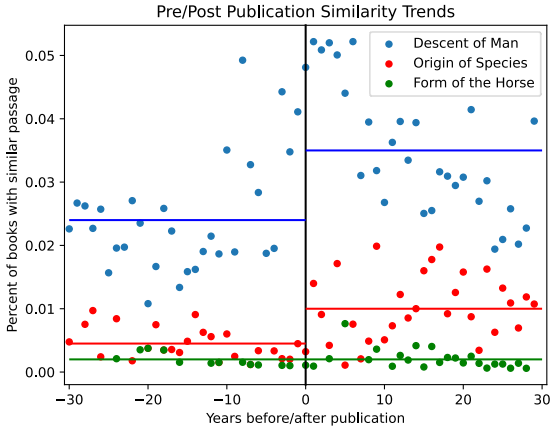

Figure 5: Similarity with books published before and after.

Also, we would expect the impact of Darwin's books to display prescience, i.e., that they exert more influence on future publications than they re-

[3] https://github.com/lucianli123/darwin-novelty

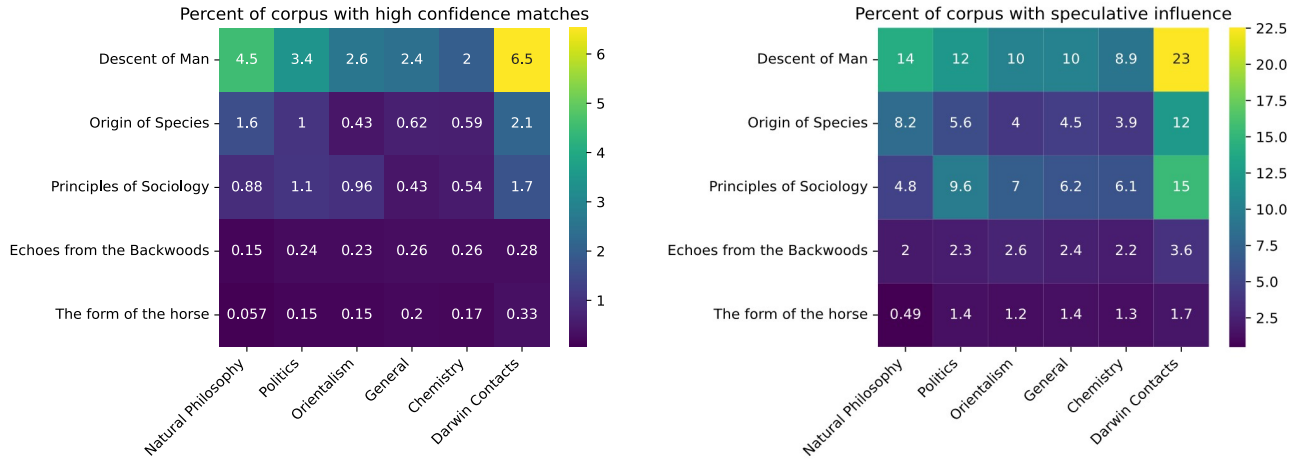

Figure 4: Percent of books post-publication with any detected influence. The last two books are randomly sampled books published in the same year as Origin of Species and included as a baseline comparison.

ceive from past ones. This would be consistent with the idea that Darwin's publications revolutionized attitudes and ideas.

When we plot the influence over time (Figure 4), we see the method's sensitivity to shifts in the overall discourse. Each point represents the similarity of books published each year and the colored lines represented the average similarity of all pre- and post- publication books. In red, the Origin of Species (1859) draws from a handful of primarily geological and biological sources prepublication, but radically shifts the overall discourse. In blue, the Descent of Man (1871) engages more with discourses across a diverse range of disciplines as well as the evolutionary ideas already introduced in the *Origin*. *Descent* has relatively more connections to previously published works, coming from Darwin's main thesis already existing in the discourse community. However, it likewise radically shifts the discourse in the corpus. Both of Darwin's major works proved innovative, as they drew less from previously published texts while exerting significant influence on future texts. The randomly sampled book in green shows that the effect is not likely due to corpus wide factors.

We can get a more detailed view of influence in specific disciplinary communities in Figure 5. The rightmost column of both Figure 5 heatmaps show the overrepresentation of Darwinian influence in books by Darwin's correspondents (people with documented interactions with Darwin). As a further confirmation, we can see more influence from Darwin's books in Biology and Geology than Chemistry or Political Theory. Even when the confidence threshold is lowered and more speculative matches are allowed, the same patterns persist. While this may not give us confidence that all examples of influence are being detected, we can at least be more sure that the distributions of detected influence reflect some kind of underlying historical pattern.

The very low levels of influence detected for the "control" books in Figures 4 and 5 gives us some confidence in the resilience of this method against excessive false positives. Individual false positive matches do not result in a book level false signal, as books we expect to be obscure have extremely few matches across the corpus.

5.3 Close reading

In table 1, we see examples of sentences at each similarity threshold. In the first example, we can see that the method detects direct quotation at high confidence while remaining robust to OCR errors and minor structural and punctuation changes. The second example shows the ability of the method to identify cases of paraphrase with very limited shared word use. These two statements make the same claim, but only share a limited number of words. Corpus based approaches will likely fail to capture the similarity in ideas in this case.

Lastly, we see the ability of the method to capture speculative matches across genres. The first quote is from *Origin of Species* and the second book is from George Eliot's *Middlemarch*. In this quote, we can see Darwin's metaphor of the web of life echoed in Middlemarch. Eliot uses the same metaphor to describe a complex network of human

Sentence 1	Sentence 2	Cosine Similarity
Would it be believed, that the larvae of an insect, or fly, no larger than a grain of rice, destroy some thousand acres of pine-trees, many of them from two to three feet in diameter, and a hundred and fifty in height?	Would it be believed, says Wilson, the ornitholog-ist, ' that the larvs of an in-sect, or fly, no larger tliaan a grain of rice, should, destroy some thousand ncres of pine trees, many of uiem two or three feet in diameter, and one himdred and fifty feet high.	High confidence (0.97)
I have called this principle, by which each slight variation, if useful, is preserved, by the term natural selection, in order to mark its relation to man's power of selection.	The expression "natural selection" was chosen as serving to indicate some par-allelism with artificial selection–the selec-tion exercised by breeders.	Medium confidence (0.92)
I have so much to do in unraveling cer-tain human lots, and seeing how they were woven and interwoven, that all the light I can command must be concentrated on this particular web, and not dispersed over that tempting range of relevancies called the universe.	We shall never disentangle the inextricable web of affinities between the members of any one class; but when we have a distinct object in view, and do not look to some unknown plan of creation, we may hope to make sure but slow progress.	Speculative influ-ence (0.85)

Table 1: Selected examples of sentence pairs with similarity scores

relationships. She draws the same conclusion as Darwin: that in a highly complex situation, we must focus on the particular rather than the general. We know this isn't random chance because literary scholars like Beer have examined the correspon-dence between Eliot and Darwin, but that relies on a whole infrastructure of experts in the papers of both authors. countless lesser known examples that subject area specialists haven't focused on study-ing remain unknown. The speculative matches in discovered here include extensive false positives, cases where stylistic or structural similarities don't suggest true influence, but may allow for the dis-covery of previously unknown influence.

6 Limitations

Because of the training process for GTE, semantic similarity is the main component in calculating embeddings. This captures the spread and influence of specific claims, but is much weaker in terms of metaphor, stylistic similarity, and influence in argument structure and construction.

For future work, I plan to create an ensemble ap-proach generating AMR graphs (Opitz et al., 2021) or knowledge graphs from the structure of each in-dividual argument. Then, graph embeddings can be generated through a neural network based approach like (Wang et al., 2018).

There are also issues with false positive matches, particularly in terms of generic and stock sen-tences used as transitions or argumentative sign-posts rather than conveying a specific claim. First, once larger scale results are discovered across the corpus, highly frequent sentences across the corpus can simply be removed. I plan to train a relatively simple BERT based model to detect false positive matches, especially because there are commonly appearing stock phrases that account for a large percentage of false positives.

However, false negatives are likely impossible to adequately evaluate or completely remove. To identify with complete confidence all instances of false negatives, the entire corpus must be examined and annotated. The preliminary results presented in this paper suggest that the matches discovered by the method roughly approximate what we expect from historical research. While this is no guarantee against false positives and negatives, it suggests that the proportion of false negatives and positives is not dramatically skewed. But we must remain aware that this method is not able to comprehen-sively identify all influence, but instead discovers previously unknown avenues of research.

7 Conclusion

7.1 Future directions

This method allows for a hypertextual exploration of any given text. As shown in Figure 6, it can

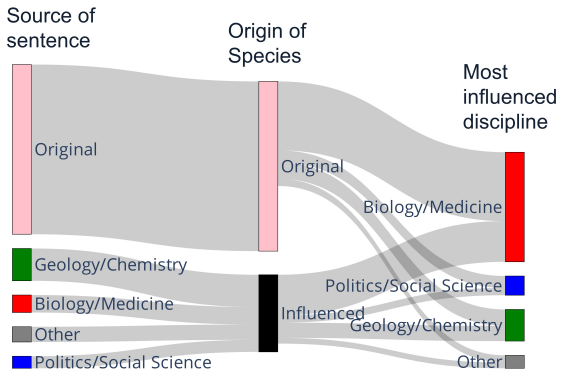

Figure 6: Origins and post-publication influence of each statement in *Origin of Species*

trace all occurrences of a specific claim over time in the corpus. Imagine an edition of the Origin of Species where a reader can click each sentence and receive information on where that argument appeared pre-publication. They would be able to observe the heavy influences from geology, as well as Darwin's own original observations based on his travels. A reader would then be able to look forward and see which fields each statement resonated with and the context for how they read and interpreted sections differently – in our Darwinian case, similarities and differences in ways eugenicists read the Origin compared to botanists. I hope to collaborate and make these enhanced editions available to a wide range of humanist scholars. By enabling researchers to perform more comprehensive searches for the origins and impacts of claims in their subjects and texts of interests, I hope to open additional avenues for interesting research.

Now imagine this on a larger scale: instead of arguments from the Origin, all arguments in the corpus. Would we be able to find common features of ideas which gained wider traction or leapt across disciplinary boundaries? My future work will focus on larger scale patterns in this corpus, with particular focus on the generalizable qualities of ideas or authors whose ideas gained influence beyond their disciplinary communities.

7.2 Reflection

Traditional narratives of discovery and invention valorize the contributions of individual geniuses - almost exclusively wealthy men from metropolitan societies. While historians of science have challenged this paradigm, the types of sources currently available for historical resource have limited practical moves toward reform. Dependence on personal papers and close reading of related works limits the potential scale and representativeness of these efforts; at some point, it becomes impossible to read the hundreds of thousands of now unknown publications. Even Beer's incisive work ultimately limits itself to Anglo-American literature and canonical authors. Responsible use of potentially destabilizing new AI technologies, keeping in mind their gaps and exclusions, can radically reshape our view of genealogies of ideas and influence and suggest previously unexplored possibilities for further exploration.

This mode of analysis has the potential to uncover connections between the work of hundreds of thousands of authors, among them women explorers and scientists, interlocutors from colonized peoples, and simply those whose ideas and contributions have been forgotten in the present. These ideas are as much part of the patchwork of intellectual life in the 19th century as those of Darwin or Herbert Spencer or Charles Lyell. Taking a wider view has the potential to reinvent the history of science.

8 Acknowledgements

Thank you to Professor Ted Underwood for discussion about the intellectual goals and evaluation metrics of this project, as well as help reading and editing earlier drafts.

Thank you as well to the anonymous reviewers for helpful comments and points of clarification.

References

Faisal Alvi, Mark Stevenson, and Paul Clough. 2021. Paraphrase type identification for plagiarism detection using contexts and word embeddings. *International Journal of Educational Technology in Higher Education*, 18(1):42.

Alexander TJ Barron, Jenny Huang, Rebecca L Spang, and Simon DeDeo. 2018. Individuals, institutions, and innovation in the debates of the french revolution. *Proceedings of the National Academy of Sciences*, 115(18):4607–4612.

Ruth Barton. 1990. 'an influential set of chaps': The x-club and royal society politics 1864–85. *The British journal for the history of science*, 23(1):53–81.

Gillian Beer. 2009. *Darwin's plots: Evolutionary narrative in Darwin, George Eliot and nineteenth-century fiction*. Cambridge University Press.

Qingyu Chen, Yifan Peng, and Zhiyong Lu. 2019. Biosentvec: creating sentence embeddings for

biomedical texts. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–5. IEEE.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library. *Preprint*, arXiv:2401.08281.

Kellen Funk and Lincoln A Mullen. 2018. The spine of american law: Digital text analysis and us legal practice. *The American Historical Review*, 123(1):132–164.

Ming Jiang, Yuerong Hu, Glen Worthey, Ryan C Dubnicek, Boris Capitanu, Deren Kudeki, and J Stephen Downie. 2021. The gutenberg-hathitrust parallel corpus: A real-world dataset for noise investigation in uncorrected ocr texts.

Athanasios Lagopoulos and Grigorios Tsoumakas. 2021. Self-citation analysis using sentence embeddings. *arXiv preprint arXiv:2105.05527*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Ernst Mayr. 1995. Darwin's impact on modern thought. *Proceedings of the American Philosophical Society*, 139(4):317–325.

Thi Tuyet Hai Nguyen, Adam Jatowt, Nhu-Van Nguyen, Mickael Coustaty, and Antoine Doucet. 2020. Neural machine translation with bert for post-ocr error detection and correction. In *Proceedings of the ACM/IEEE joint conference on digital libraries in 2020*, pages 333–336.

Juri Opitz, Angel Daza, and Anette Frank. 2021. Weisfeiler-leman in the bamboo: Novel amr graph metrics and a benchmark for amr graph similarity. *Transactions of the Association for Computational Linguistics*, 9:1425–1441.

Eszter Pal. 2014. Scientific societies in victorian england. *Review of Sociology*, 20:85–111.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Daniel N Rockmore, Chen Fang, Nicholas J Foti, Tom Ginsburg, and David C Krakauer. 2018. The cultural evolution of national constitutions. *Journal of the Association for Information Science and Technology*, 69(3):483–494.

David A Smith, Ryan Cordell, and Abby Mullen. 2015. Computational methods for uncovering reprinted texts in antebellum newspapers. *American Literary History*, 27(3):E1–E15.

Sandeep Soni, Lauren Klein, and Jacob Eisenstein. 2021. Abolitionist networks: Modeling language change in nineteenth-century activist newspapers. *arXiv preprint arXiv:2103.07538*.

Paul Vicinanza, Amir Goldberg, and Sameer B Srivastava. 2023. A deep-learning model of prescient ideas demonstrates that they emerge from the periphery. *PNAS nexus*, 2(1):pgac275.

Zhichun Wang, Qingsong Lv, Xiaohan Lan, and Yu Zhang. 2018. Cross-lingual knowledge graph alignment via graph convolutional networks. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 349–357.