RI:Small: Beyond Titles and Abstracts: Text + Links are Better Together

Project Summary

Overview

The need: Recommender systems and systems for assigning papers to reviewers tend to focus on titles and abstracts, but when we started EMNLP in 1990s, we used references to decide who should review which paper. In our collection of 200M documents from Semantic Scholar [1], there are more papers in the citation graph without abstracts (46M) than vice versa (34M). Links have been super-important in websearch. There is an opportunity to take more advantage of citations (and citing sentences) in information retrieval and NLP tasks such as **topic modeling** and **document summarization**. Firth's *you shall know a word by the company it keeps* [2], has had considerable influence on Natural Language methods for processing words and phrases (PMI [3], Word2vec [4, 5], BERT [6]). Firth's approach has been generalized from words and phrases to topics and documents, using methods such as link mining [7, 8], node2vec [9], graph learning [10] and graphical neural nets (GNNs) [11]. GNNs, Specter [12] and LinkBERT [13] use links at training time to improve models of short texts (512-subwords), but these methods do not capture longer texts and links at inference time.

Approach: Prior work tends to encode documents as vectors in just one way; we propose multiple embeddings, f_1, f_2,..., to capture multiple perspectives such as titles, abstracts, full text and citations. Cosines of two vectors, $cos(f(i), f(j))$, denote similarity of i and j, where i, j can be documents and/or "topics" [14, 15, 16, 17]). Cosines of vectors based on text (e.g., bags of words [18, 19, 20], BERT [6, 12, 21]) denote word similarity, whereas cosines based on node2vec [9, 22] (spectral clustering [23] of citation graph, G) can be interpreted as $dist(i, j)$, distances in G.

Intellectual Merit

Documents that are similar in at least one way, tend to be similar in other ways, as well. We refer to this assumption as Cos-Dist: $cos(f_1(i), f_1(j)) \sim cos(f_2(i), f_2(j)) \sim dist(i, j)$. Multiple representations create opportunities for error detection/correction. If an abstract is missing and/or corrupted, the citation graph can be used as a workaround. Moreover, building on [24, 25], Cos-Dist opens opportunities to generalize results from relatively well understood Linear Algebra to recent advances in deep nets, since some embeddings are based on Linear Algebra and others are based on deep nets. In addition, diversity over representations creates opportunities for diverse computing environments. For example it has been popular to use GPUs with GBs of RAM for deep nets, but CPUs with TBs of RAM may be preferable for embeddings based on Linear Algebra. When we have TBs of RAM, we can compute (in RAM) the SVD of a large citation graph with 200M papers (nodes) and 2B citations (edges).

Some Use Cases

Recommender systems and systems for assigning papers to reviewers tend to focus on titles and abstracts, but when we started EMNLP in 1990s, we used references to decide who should review which paper. Links have been super-important in websearch. There is an opportunity to take more advantage of citations (and citing sentences) in information retrieval and NLP tasks such as **topic modeling** and **document summarization** in tasks such as the following:

1. Search, Rank Retrieval and Relevance Feedback: given a query (text and/or documents with relevance labels): find documents in the collection that are similar to the query.
2. Recommender systems: What should I read? What should I cite? Many systems focus too much on relevance, and not enough on credibility. It is better to recommend a highly cited seminal paper than a paper that is buzz-word compliant, but not cited or even published.
3. Paper Reviewer Matching [27, 28, 29]: conferences, journals and funding agencies use reviewing platforms [30, 31, 32]. Many of these platforms use software to suggest who should review what. There are opportunities to improve assignments, as well as evaluation benchmarks. When we started EMNLP, we assigned papers to tracks by hand. Many systems use titles and abstracts, but we used references. Reviewers that are cited in the references are more likely to be qualified and sympathetic with the topic. We fear that widespread deployment of poorly tested programs may be teaching authors to write incremental papers that can be reviewed by unqualified and unsympathetic reviewers [33].
4. Summarize collections of documents: Compare and contrast one or more collections of documents with similar authors, venues, topics, etc. Much of the work on summarization focuses on a single paper. Suppose we want to compare and contrast collections. How do ACL papers differ from SIGIR papers? How do papers in the recent conferences compare with previous conferences? Compare and contrast papers by the Bengio brothers. How does Sammy's work differ from Yashua's? How does their work differ from Geoff Hinton? Yann LeCun?

Opportunity: Beyond Titles And Abstracts

In natural language processing (NLP) and information retrieval (IR), it has become standard practice to represent words, phrases and documents as vectors using methods such as: Latent Semantic Indexing (LSI) [34], Word2vec [4], BERT [6] and Specter [12]. Many of these methods were inspired by Firth's famous quote from the 1950s: *You shall know a word by the company it keeps* [2]. Levy and Goldberg [35] established a connection between more recent methods such as Word2vec and older methods such as our work on PMI (pointwise mutual information) [3].

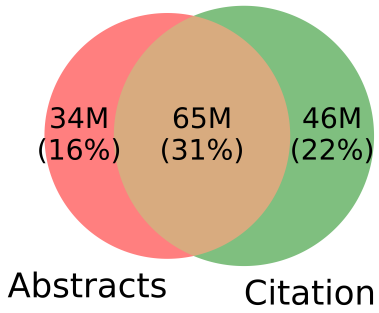


Figure 1: Venn diagram of 208M papers in Semantic Scholar [1]. Proposed method will expand coverage into green.

This proposal advocates methods that take advantage of additional features beyond titles and abstracts. Abstract-based methods fail when abstracts are missing and/or corrupted. The proposed method is intended to expand coverage into the green region of Figure 1, where abstracts are missing. Note that the green region is larger than the red region. That is, there are more papers in the citation graph without abstracts (green), than vice versa (red). There are an additional 63M papers (30%) missing both abstracts and citations. Perhaps one can match on titles, though this proposal will focus on papers in Figure 1.

f	$cos_S(q,c)$	$cos_P(q,c)$	**Cites**	**Paper**
f_S	0.984	0.827	0	Learning of Social Representations [36]
f_S	0.809	0.951	2	Topic-aware latent models for representation learning on networks [37]
f_S	0.797	0.947	4	SimWalk: Learning network latent representations with social relation similarity [38]
f_S	0.783	0.033	0	Deep Representation Learning on Complex Graphs [39]
f_P	0.771	0.999	6007	node2vec: Scalable Feature Learning for Networks [40]
f_P	0.711	0.998	3632	LINE: Large-scale Information Network Embedding [41]
f_P	0.664	0.997	1025	A Comprehensive Survey of Graph Embedding: Problems, Techniques, and Applications [42]
f_P	0.711	0.996	1157	metapath2vec: Scalable Representation Learning for Heterogeneous Networks [43]

Table 1: Approximate n-best matches for query, q = [44]: "DeepWalk: online learning of social representations" [45]. Four candidates, c, are shown for f_S (Specter), followed by four for f_P (Proposed). Cosines, cos_S and cos_P, are shown for all 8 candidates. f_P candidates (below the line) have more citations.

f	**Rank**	$cos_S(q,c)$	$cos_P(q,c)$	**Citations**	**Paper**
f_S	1	0.794	0.957	3	What Language Model to Train if You Have One Million GPU Hours? [46]
f_S	2	0.779	0.976	117	On the Stability of Fine-tuning BERT: Misconceptions, Explanations, and Strong Baselines [47]
f_S	3	0.777	0.961	1	Emergent Properties of Finetuned Language Representation Models [48]
f_P	1	0.615	0.992	139	RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models [49]
f_P	2	0.750	0.992	49	The Radicalization Risks of GPT-3 and Advanced Neural Language Models [50]
f_P	3	0.542	0.992	52	Persistent Anti-Muslim Bias in Large Language Models [51]

Table 2: Approximate n-best matches for query, q = [52]: "On the Dangers of Stochastic Parrots..." [53], a paper on Responsible AI aspects of deep nets. All six candidates are about deep nets, but f_P candidates have more citations, and they are closer to Responsible AI.

This proposal will make use of both text and citations at inference time. A number of systems such as Specter and Graphical Neural Nets (GNNs) [11] take advantage of citations when fine-tuning, but not at inference time. The goal of these fine-tuning processes is to produce a model, f, that will be applied at inference time to a string, s, typically titles and abstracts. For now, assume f is a BERT-like model such as Specter that inputs strings, s (titles and abstracts), and outputs vectors of length $K = 768$, where K is the number of hidden dimensions of the final layer of the model. The similarity of two documents, d_1 and d_2, is simply the cosine of the two vectors: $cos(f(d_1), f(d_2))$. We will generalize this approach to take advantage of multiple representations (and additional properties) so estimates of similarities will succeed even when abstracts (and other

properties) are corrupted and/or missing at inference time. That is, we assume document ids, d, are associated with various properties that are often available (but not always):

- text: titles, abstracts, tl;dr (too long; didn't read), full text, and
- context (links): citations, citing sentences
- more: authors, venues, fields of study

In this way, the proposed approach:

1. expands coverage into the green region (with missing abstracts),
2. enables error detection, which is important when abstracts are corrupted (Table 3), and
3. enables queries over strings (titles and abstracts), links (citations) and more.

Rank Retrieval And Recommendation Systems

Tables 1-2 show examples of two embeddings, f_S and f_P, in a rank retrieval application. Suppose we want to recommend papers to read that are similar to input queries, q:

- Table 1, q =[44]: "DeepWalk: online learning of social representations" [45]
- Table 2, q = [52]: "On the Dangers of Stochastic Parrots..." [53]

Both tables show recommendations from Specter (above the line), followed by recommendations from the proposed method (below the line). Approximate nearest neighbors (ANN) [54, 55] is used to find n-best matches in both f_S embeddings as well as f_P embeddings.

Many recommendation systems focus on relevance, but in addition to that, we should also consider credibility. Most papers are not cited much, if at all. If possible, we should recommend highly cited papers. The recommendations in Tables 1-2 are all on deep nets, but the proposed method recommends papers with more citations.

As mentioned above, we are interested in both IR (Information Retrieval) and NLP (Natural Language Processing). The query system illustrated in Tables 1-2 can not only be used for IR, but we believe, that it can also be useful for NLP tasks such as topic modeling and document summarization. Documents near the query tend to be on similar topics. This may not be surprising with Specter embeddings, f_S, given the literature on combinations of BERT and topic modeling [56, 57, 58, 59]. But it appears that the proposed method is also finding documents on similar topics. Thus, if one is interested in studying the use of text and other NLP features to study NLP applications such as topic modeling and summarization, one can use the proposed method to create training materials. For the purpose of constructing materials to study NLP questions, it is desirable to avoid NLP features in the construction of the training materials. The proposed method avoids NLP features by replacing those features with features from the citation graph.

Hybrid combinations of both methods are better than either by itself. Multiple representations introduce possibilities for error detection and error correction. The data from Semantic Scholar [1] is extremely useful, but it is not free from errors. In Table 1, the cos_P score for candidate [39] is too low (because of omissions in the citation graph). Table 3 will show some examples where cos_S are too high because of issues with abstracts.

Error Detection

In Table 3, we illustrate an opportunity to take advantage of multiple perspectives to detect errors. Most of the abstracts in Semantic Scholar are correct, but there are a few documents in Semantic Scholar where the abstract field was incorrectly replaced with some boilerplate from JSTOR such

f	**Rank**	cos_S	cos_P	**Citations**	**Paper**
f_S	7	0.777	0.097	15	The Value of Natural Sounds [60]
f_S	8	0.766	0.089	16	Re-Examining the Foundations [61]
f_S	9	0.757	0.106	8	Facts and Faith in Biblical History [62]

Table 3: Multiple embeddings create opportunities for error detection. Normally, $cos_S \approx cos_P$, but in these three cases, $cos_S \gg cos_P$. These cos_S are large and misleading because of corrupted abstracts in the query [63], "On stress and linguistic rhythm" [64] and the candidates [60, 61, 62]. The smaller cos_P scores are more credible because these errors have no impact on cos_P.

as: *JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive...* When this error happens to the query, then many of the candidates are also likely to suffer from the same error, as illustrated in Table 3, where the error impacted the the query [63] as well as the candidates [60, 61, 62]. These errors produce large (misleading) cos_S, but these errors can be detected by comparing cos_S with cos_P, because cos_P does not depend on abstracts, and is therefore robust to corrupted abstracts.

Notation

1. Let $N \approx 200$M be the number of documents in the collection.
2. Let $E \approx 2$B be the number of citations in the collection.
3. Let K be the number of hidden dimensions (typically, $K = 768$ for models based on BERT, and $K = 280$ for models based on Linear Algebra).
4. Let f be a model and M be an embedding. Models, f, can be applied at inference time to novel inputs, or they can be rows of M. Embeddings, M, are large matrices: $M \in \mathbb{R}^{N \times K}$, where $f(d_i)$ is the row in M for document d_i. That is, $f(d_i) = M[d_i,]$.
5. Let f^{-1} be the inverse of f. f maps document ids to vectors ($f(d_i) = M[d_i,]$), and f^{-1} maps vectors to document ids ($f^{-1}(M[d_i,]) = d_i$). We can implement f^{-1} with approximate nearest neighbors (ANN) [65, 54, 55], so f^{-1} can be applied to vectors that do not match any of the rows in M exactly.
6. The subscripts S and P will be used as necessary to distinguish f_S from f_P, and M_S from M_P. The subscript S refers to Specter, and the subscript P refers to the proposed model, which is based on ProNE [22]. More subscripts will be introduced for additional embeddings.

Multiple Representations

Our project is intended to show that multiple representations are better together, as illustrated in Figure 2. The next section will introduce the Cos-Dist assumption, where cosines in embeddings can be used to estimate distances in graphs, and vice versa. We view embeddings, M, and graphs, G, as different representations of documents, somewhat analogous to frequency and time domains in speech, where filtering can be implemented as multiplication in the frequency domain, or convolution in the time domain. So too, under Cos-Dist, one can estimate document similarities with cosines in the embedding space and distances in graphs. Embeddings are convenient for estimating cosines, and finding approximate nearest neighbors, though graphs are more compact. Different representations have different advantages and disadvantages.

Graphs are more compact because their size depends on E (nonzero edges), unlike embeddings, which are stored as dense matrices, $M \in \mathbb{R}^{N \times K}$. Graphs are more compact because $K \gg E/N$.

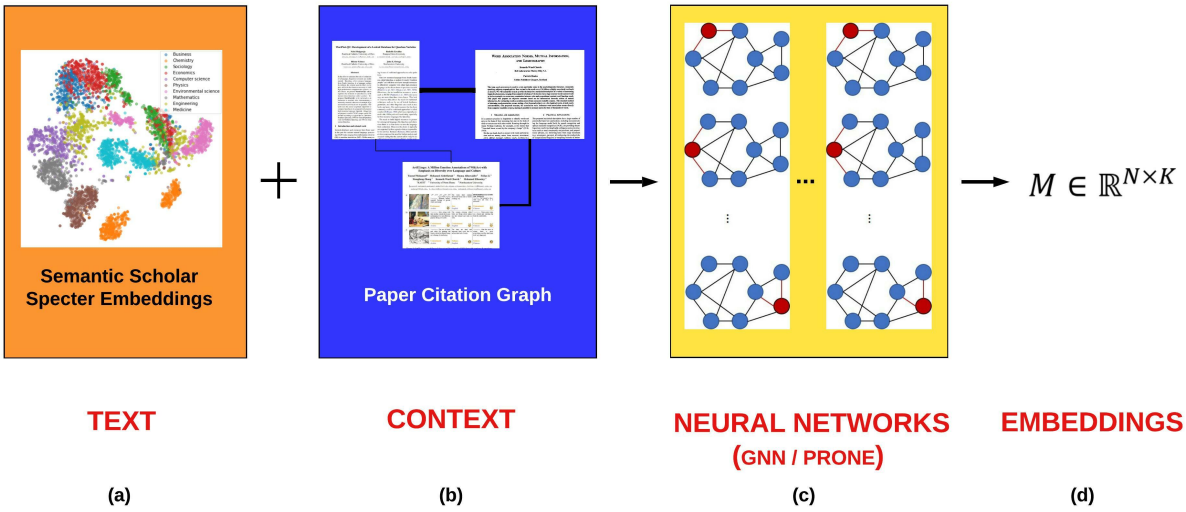


Figure 2: Embeddings are typically based on text (titles and abstracts). We will create multiple embeddings based on combinations of text and context (links such as citations and citing sentences).

Recall that $K \gg 100$ and $E/N \approx 10$. We will refer to E/N as the average citation rate (about 10 because $N \approx 200$M papers and $E \approx 2$B citations).

We will experiment with many embeddings, in addition to M_S and M_P. One such embedding is based on citing sentences [66, 67, 68], which are like anchor text [69] in web search. We expect citing sentences to be helpful when subsequent literature introduces contemporary terminology. Consider, for example, Turing's seminal paper [70], which introduced what is now known as a "Turing Machine," though of course, Turing did not use that term in his paper. Citing sentences can be more useful than primary sources for appreciating contributions in contemporary contexts. We have already run Specter on a billion citing sentences.

Cos-Dist: Multiple Views Of Similarity

Prior work tends to encode documents as vectors in just one way. We propose multiple embeddings, f_1, f_2,..., to capture multiple perspectives such as titles, abstracts, full text and citations. Let $cos(f(i), f(j))$ denote the similarity of i and j, where i, j can be strings, documents and/or "topics" [14, 15, 16, 17]. Cosines of vectors based on text (e.g., bags of words [18, 19, 20], BERT [6, 12, 21]) denote word similarity, whereas cosines based on node2vec/ProNE [9, 22] (spectral clustering of citation graph, G) can be interpreted in terms of $dist(i, j)$, distances in G.

Documents that are similar in at least one way, tend to be similar in other ways, as well. We refer to this observation as Cos-Dist: $cos(f_1(i), f_1(j)) \sim cos(f_2(i), f_2(j)) \sim dist(i, j)$. Diversity over representations opens many opportunities:

1. Interpretability (Cos-Dist): Similarity of documents can be estimated as cosines in embeddings, $cos(f(i), f(j))$, or distances in G. In other words, embeddings can be viewed as an alternative representation of graphs. Both representations have advantages and disadvantages. Graphs are compact, but embeddings are convenient for computing cosines and ANNs.
2. Redundancy: if we have two embeddings, f_1 and f_2, then we have three redundant estimates of document similarity: (a) $cos(f_1(i), f_1(j))$, (b) $cos(f_2(i), f_2(j))$ and (c) distances in G. Redundant estimates create opportunities for error detection and error correction.

3. Cos-Dist holds for embeddings based on Linear Algebra as well as embeddings based on non-linear deep nets, creating opportunities to generalize results from relatively well understood Linear Algebra to recent advances in deep nets, as suggested in [24].
4. Diversity over representations introduces opportunities for diverse computing environments. For example it has been popular to use GPUs with GBs of RAM for deep nets, but CPUs with TBs of RAM may be preferable for embeddings based on Linear Algebra. When we have TBs of RAM, it becomes feasible to compute the SVD of a large citation graph in RAM. Standard recipes for training deep nets with mini-batches can be viewed as an external memory algorithm. Because external memory is slower than RAM, CPUs with TBs may be preferable to GPUs with GBs, at least in some cases.

Figure 3 provides some evidence for Cos-Dist. We created pairs with a random walk. We started with nearly 1M document ids, selected at random. From each of these, we randomly selected one of its references. From there, we continued the walk by randomly selecting one of its references, and so on. For each of these pairs, when the cosine is available, we report the length of the shortest path on the x-axis, and cos_S (and cos_P) on the y-axis. More than 20% of the pairs are omitted from the plot for Specter because those cosines are often unavailable (due to missing abstracts).

For both Specter and the Proposed method, we find that cosines are negatively correlated with path lengths, though the pattern is stronger for f_P because that method uses an optimization criterion that is closer to Cos-Dist. We believe, however, that Cos-Dist holds for most reasonable embeddings of documents because they are all estimating similarities of the same documents.

Time Invariance And Incremental Updates

It is natural to model a paper as cast in stone when it is published, but we prefer to view the literature as a conversation, somewhat like social media. For example, while Turing's paper [70] has not changed since it was published in 1936, there is a large (and growing) body of work that builds on that contribution. The value of a paper to society is a combination of the primary source plus audience appreciation (subsequent literature and secondary sources).

Some embeddings mentioned above evolve over time (M_P), and some do not (M_S). The value of a paper combines time-invariant contributions from authors (M_S), with contributions from the audience (M_P) that accumulate over time.

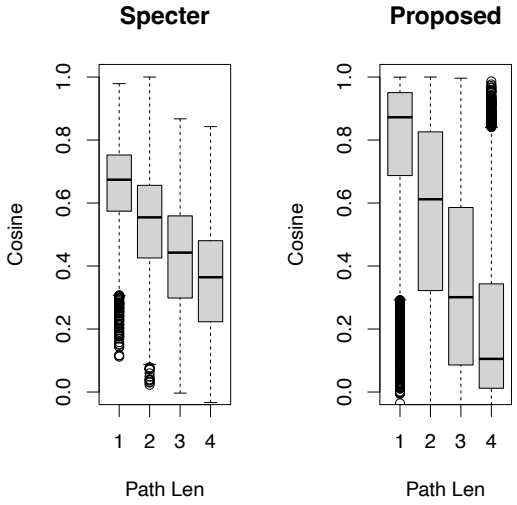
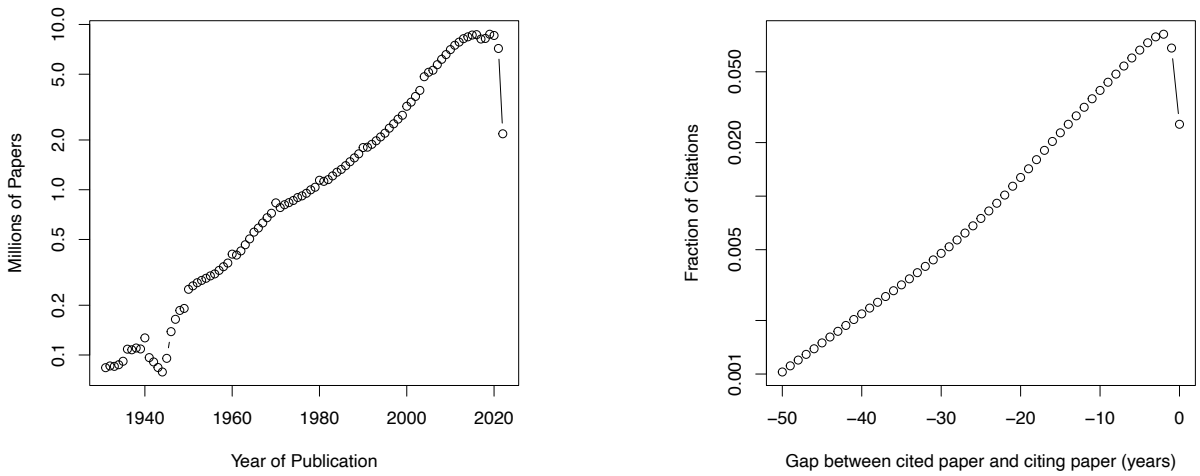


Figure 3: *Cos-Dist*: Cosines in embeddings can be interpreted as the length of shortest path in G: $cos(f(i), f(j)) \sim pathlen(i, j)$.

The next section will describe how to compute embeddings from the citation graph. That solution is relatively straightforward, but it takes nearly a week to compute (on our hardware). Given G's rapid growth (see Figure 4), it is highly desirable to support incremental updates. One suggestion for incremental updates is to compute embeddings for overlapping sets. Suppose we have a set of papers published before t_1, and a set of recent updates published after t_2. Let M_1 and M_2 be embeddings for the two sets. We can use the Orthogonal Procrustes Problem [71, 72] to estimate a rotation



(a) Papers: Exponential growth.

(b) Most citations refer to recent publications.

Figure 4: Exponential growth of citations graph in [1] over time. Warning: counts for more recent years may be incomplete. Note: y-axes are on log scales.

matrix, $R \in \mathbb{R}^{K \times K}$, so vectors in M_2R will be compatible with vectors in M_1. Suppose we construct M_1 and M_2 to have a large overlap (where $t_2 \ll t_1$), then we can select rows from M_1 and M_2 for the overlapping papers. Let O_1 and O_2 be the vectors for the overlap. Use the Orthogonal Procrustes Problem to find an R that minimizes $|O_1 - O_2R|$. In this way, we can support incremental updates. That is, we can compute M_2 without recomputing M_1.

ProNE

The ProNE method, f_P, is based on the citation graph, G. We start by downloading G from [1]. G is stored as a $N \times N$ Boolean matrix, $M_G \in \{0,1\}^{N \times N}$, where $N \approx 200$M is the number of nodes (papers) in G. If (and only if) paper d_i cites paper d_j, then $M_G[d_i, d_j] = 1$. This matrix is sparse because there are only $E = 2$B citations (nonzero edges). G might appear to be a huge matrix but it requires just 7.8 GBs when stored in scipy [73] as an npz file because G is extremely sparse.

We then run the ProNE [22] method in nodevectors [74] to produce an embedding, $M_P \in \mathbb{R}^{N \times K}$, where N is the number of nodes (papers) in G, and K is the number of hidden dimensions. ProNE is a variant of node2vec [9]. Unlike Specter and many variants of node2vec, ProNE is based on Linear Algebra and spectral clustering, and does not use deep nets.

The first step in ProNE is called *prefactorization*. It uses a memory internal SVD procedure to factor a normalized version of G as UDV^T, where $U, V \in \mathbb{R}^{N \times K}$. While G is sparse, U and V are not. Since we have TBs of RAM, it is feasible to compute this SVD in RAM.

After prefactorization, ProNE uses spectral propagation to compute the embedding M_P so cosines in M_P can be interpreted in terms of distances on G. For efficency, ProNE uses Chebyshev expansion to avoid explicit Eigen decomposition. While the Chebyshev expansion is efficient, it makes a number of copies of large matrices in memory, each of which are the size as U. The total memory requirement is about 5 times larger than U. The output embedding, M_P, is also a dense matrix of the same size as U.

ProNE takes considerable time and space. It takes nearly a week to compute M_P on a CPU

with 2 TBs of RAM with $K = 280$. The time complexity is: $O(NK^2 + E)$, but in our case, we can simplify that to $O(NK^2)$ because $NK^2 \gg E$. Increasing K improves the approximation that cosines on M are related to distances on G. Of course, increasing K also increases computational costs (time and space). One of the work items for this proposal is to make it easier to compute M_P:

1. Improve time and space constants,
2. Increase K, and evaluate benefits and costs of doing so, and
3. Provide support for incremental updates to keep up with exponential growth in Figure 4.

Diversity over representations creates opportunities for diverse computing environments. It has been popular to use GPUs with GigaBytes (GBs) of RAM for deep nets (such as BERT), but CPUs with TeraBytes (TBs) of RAM may be preferable for embeddings based on Linear Algebra (such as ProNE). We typically train deep nets (such as GNNs) by grouping edges into mini-batches. Each mini-batch updates model parameters in GPUs. This process is repeated for a few epochs, and therefore, time complexity grows linearly with the size of training set. With ProNE, we run the SVD in TBs of RAM. TBs of RAM costs about the same as GPUs, but TBs of RAM make it feasible to consider algorithms beyond linear time. With sufficiently large memories, it becomes feasible to sort the training data, or to run SVD on it. Many methods in Linear Algebra require more than linear time.

Embeddings, M, are convenient for estimating similarity. Similarity of two documents (and/or topics), $M[i,]$ and $M[j,]$, is computed with cosines: $cos(M[i,], M[j,])$. Approximate nearest neighbors (ANN) can be used for rank retrieval, as will be discussed in the next section.

Title	**tl;dr (too long; didn't read)**
On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? [52]	Recommendations including weighing the environmental and financial costs first, investing resources into curating and carefully documenting datasets rather than ingesting everything on the web, and carrying out pre-development exercises evaluating how the planned approach fits into research and development goals and supports stakeholder values are provided.
RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models [49]	It is found that pretrained LMs can degenerate into toxic text even from seemingly innocuous prompts, and empirically assess several controllable generation methods find that while data- or compute-intensive methods are more effective at steering away from toxicity than simpler solutions, no current method is failsafe against neural toxic degeneration.
The Radicalization Risks of GPT-3 and Advanced Neural Language Models [50]	GPT-3 demonstrates significant improvement over its predecessor, GPT-2, in generating extremist texts and its strength in generating text that accurately emulates interactive, informational, and influential content that could be utilized for radicalizing individuals into violent far-right extremist ideologies and behaviors.
Persistent Anti-Muslim Bias in Large Language Models [51]	GPT-3, a state-of-the-art contextual language model, captures persistent Muslim-violence bias, demonstrating that it appears consistently and creatively in different uses of the model and that it is severe even compared to biases about other religious groups.

Table 4: tl;dr summaries from Semantic Scholar [1] for some papers from Table 2.

Document Summarization

Tables 1-2 suggest another NLP application: summarization. In addition to titles (as in Tables 1-2), it would be desirable to produce snippets/summaries to help users choose among the various recommendations. It is common for summarization systems to process each document one at a time, though it would be desirable for summaries to compare and contrast the choices to help users appreciate similarities and differences between various options. One of the work items for this proposal is to produce summaries that improve over two baselines: (1) tl;dr from Semantic Scholar [1], and (2) ChatGPT [75].

Table 4 shows some tl;dr summaries from the Semantic Scholar API [1] for three candidates from Table 2. These summaries tend to be extracts from abstracts. To illustrate baseline (2), we ran ChatGPT on the text in Table 4 and it produced the following output:

> *The recent research on the dangers of large language models (LLMs) highlights the importance of considering their potential negative impacts on society. As language models become increasingly sophisticated, there is a risk of them degenerating into toxic text or perpetuating biases and harmful ideologies. Therefore, it is crucial to evaluate the risks and benefits of LLMs before developing and deploying them.*
>
> *One recommendation for addressing the risks of LLMs is to weigh the environmental and financial costs of building and training these models. LLMs are known to require large amounts of energy and computational resources, which can have significant environmental impacts. It is important to consider these costs alongside the potential benefits of using LLMs.*

Neither of these baselines address the desiderata mentioned above. That is, they fail to compare and contrast the choices. Let d_a and d_b be two documents, and $a = f(d_a)$ and $b = f(d_b)$ be two vectors for the two documents. We suggest expressing comparisons and contrasts in the vector space as in Eqn (1). We can then use f^{-1} to find nearby documents, and then report properties of those nearby documents such as: titles, abstracts, references, citations, etc.

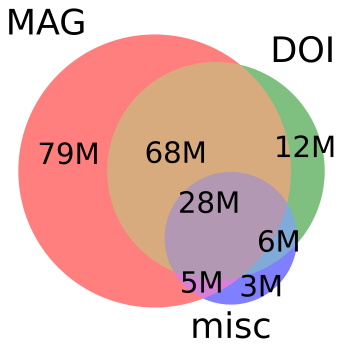
$$\begin{aligned} compare(a,b) &= ARGMIN_{c \in C}\ |c-a| + |c-b| \\ contrast(a,b) &= ARGMIN_{c \in C}\ |c-a| - |c-b| \end{aligned} \tag{1}$$

Contrasts are similar to comparisons except for a sign change. For comparisons, we plan to cast a wide net and let the set of candidates, C, be the set of 200M documents in Semantic Scholar. For contrasts, we plan to limit the set of candidates to documents that are reasonably close to a for some threshold, T. That is, $|c-a| < T$.

Data, Applications & Evaluations

For evaluations, we plan to start with benchmarks on github [84, 84] with baseline results reported in [12, 85]. In addition to these benchmarks from Semantic Scholar, there are also a number of influential benchmarks such as OGB (Open Graph Benchmark) [86, 87], which was used in KDD-cup [88].

OGB was designed to compare graph learning algorithms. While much of the work based on that benchmark is very relevant to this proposal, the rules of the benchmark prohibit the use of outside data such as data from Semantic Scholar. In addition, since the focus is on graph learning, and not on practical applications, the benchmark does not publish the mapping of paper ids to titles and abstracts, and the mapping of author identifiers to author names. While one could probably



1. 182M papers from MAG [76, 77]
2. 114M papers from DOI [78]
3. 35M papers from PubMed [79]
4. 6M papers from DBLP [80]
5. 5M papers from PubMedCentral [81]
6. 2M papers from arXiv [82]
7. 80k papers from ACL Anthology [83]

Figure 5: Seven sources in Semantic Scholar [1].

reverse engineer that mapping, doing so would violate the spirit of the benchmark. Unfortunately, without such a mapping, it is not clear how methods based on this benchmark can be used in practical applications.

The literature on graph learning classifies tasks into (a) node-level, (b) link-level and (c) graph-level. The example of (a) in [87] is closely related to tasks of interest for this proposal, unlike the examples of (b) and (c), which are closer to Knowledge Graph Completion and Chemistry. OBG uses MAG240M for (a). MAG240M is based on MAG (Microsoft Academic Graph) [77]. Unfortunately, Microsoft retired MAG at the end of 2021, though MAG has been transferred to a nonprofit, OpenAlex [89]. OpenAlex is making the data available for bulk access. More importantly, they will keep MAG up to date, which is important because the citation graph is expanding rapidly, as discussed in Figure 4.

Figure 5 shows the importance of MAG. Semantic Scholar is based on 7 sources, but the 5 smaller ones (*misc*) contribute just 3M papers that are not already covered by MAG and DOI. Our field tends to focus on ACL Anthology and arXiv, two relatively small sources.

The 200M papers cover a diverse set of fields of study (fos), though with an emphasis on STEM: Medicine (45M), Chemistry (13M), Computer Science (13M), Biology (13M), Materials Science (10M), Engineering (8M), Physics (7M), Psychology (7M), Mathematics (5M), Political Science (4M), Business (4M), Sociology (3M), Geography (3M), Economics (3M), Environmental Science (3M), Geology (3M), History (2M), Art (2M), Philosophy (1M). We have found other sources such as SSRN (Social Science Research Network) [90] to be more useful for Social Studies, Law and Journalism, even though SSRN is relatively small (1M papers). SSRN also makes use of data on downloads, which is extremely useful, especially in fields where it is less common for papers to cite one another. Unfortunately, SSRN does not make its data available for bulk download.

MAG240M is based on a citation graph from an older version of MAG. This graph has 122M papers (nodes) with 1.3B citations (edges). MAG240 also makes use of an author graph with 122M authors (nodes) that wrote 386M papers (edges). The evaluation is a prediction task: predict one of about 150 subject area labels for 1.4M arXiv papers. The data is split by publication year: train (arXiv papers published in 2018 or before), validation (arXiv papers published in 2019), test (arXiv papers published in 2020).

There are a few similar tasks in [84] such as: fos (field of study), and MeSH descriptors. These tasks are more attractive because they provide a mapping from numeric label identifiers to meaningful strings. Moreover, there are only 2M arXiv papers, and it is not clear that subject area labels for these papers will generalize well to papers from other sources. There are many more papers in Semantic Scholar with fos labels (148M), and these labels are easier to interpret. The

MeSH task is attractive since the National Library of Medicine has been labeling PubMed papers with MeSH terms for many years. PubMed (35M papers) is also considerably larger than arXiv (2M papers).

As for link prediction, there are a couple of tasks in [84] such as cite prediction that are attractive because they generalize to important practical problems. Recommender systems should be able to answer questions such as what should I read and what should I cite. The cite prediction task provides pairs of document identifiers with titles and abstracts. The pairs are labeled with 1 or 0, indicating whether the first paper of the pair cites the second member of the pair or not. The task is to predict the label.

The proposed work will improve over evaluations above in two respects. First, most of the evaluations above make use of titles and abstracts and little else. We plan to take advantage of additional features (citations, citing sentences and full text) in a way that is robust to missing/corrupted values. In addition, the evaluation should guide the community toward building effective solutions to practical tasks such as:

1. Label Prediction: Predict labels such as subject areas, fos (field of study), MeSH.
2. Link Prediction for Recommender systems: what should I read? what should I cite?
3. Search / Rank Retrieval / Relevance Feedback: given a query (text and/or documents with relevance labels): find documents in the collection that are similar to the query.
4. Paper Reviewer Matching
5. Summarize collection of documents: compare and contrast clusters (as opposed to summarizing each document in isolation)

Many of these tasks have been discussed above.

Deliverables / Work Items

1. Better access to literature
2. Resources: Many embeddings for many papers; More models to be posted on HuggingFace; Code to be posted on GitHub
3. Summarization methods to compare and contrast across small (and large) collections of documents
4. Support incremental updates to embeddings based on citation graphs
5. Evaluation: Better numbers, as well as better benchmarks
6. Establish that combinations of text and links are better together (than either by itself)
7. Establish that citing sentences are useful
8. Improve methods for assigning papers to reviewers
9. Theory: Unified framework of deep nets and Linear Algebra

References Cited

[1] Semantic-Scholar, "Semantic scholar academic graph api: Providing a reliable source of scholarly data for developers," https://www.semanticscholar.org/product/api, 2017.

[2] J. R. Firth, "A synopsis of linguistic theory, 1930-1955," *Studies in linguistic analysis*, 1957.

[3] K. W. Church and P. Hanks, "Word association norms, mutual information, and lexicography," *Computational Linguistics*, vol. 16, no. 1, pp. 22–29, 1990. [Online]. Available: https://www.aclweb.org/anthology/J90-1003

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[5] Y. Goldberg and O. Levy, "word2vec explained: deriving mikolov et al.'s negative-sampling word-embedding method," *arXiv preprint arXiv:1402.3722*, 2014.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[7] A. Popescul and L. H. Ungar, "Statistical relational learning for link prediction," 2003.

[8] L. Getoor and C. P. Diehl, "Link mining: a survey," *SIGKDD Explor.*, vol. 7, pp. 3–12, 2005.

[9] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, 2016, pp. 855–864.

[10] W. L. Hamilton, "Graph representation learning," *Synthesis Lectures on Artifical Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159, 2020.

[11] anonymous, "Graph neural networks," 2020. [Online]. Available: https://snap-stanford.github.io/cs224w-notes/machine-learning-with-networks/graph-neural-networks

[12] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld, "SPECTER: Document-level representation learning using citation-informed transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2270–2282. [Online]. Available: https://aclanthology.org/2020.acl-main.207

[13] M. Yasunaga, J. Leskovec, and P. Liang, "LinkBERT: Pretraining language models with document links," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association

for Computational Linguistics, May 2022, pp. 8003–8016. [Online]. Available: https://aclanthology.org/2022.acl-long.551

[14] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol. 18, no. 4, pp. 467–480, 1992. [Online]. Available: https://www.aclweb.org/anthology/J92-4003

[15] D. M. Blei, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, pp. 55–65, 2010.

[16] A. B. Dieng, F. J. R. Ruiz, and D. M. Blei, "Topic modeling in embedding spaces," *Transactions of the Association for Computational Linguistics*, vol. 8, pp. 439–453, 2020. [Online]. Available: https://aclanthology.org/2020.tacl-1.29

[17] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," *arXiv preprint arXiv:1207.4169*, 2012.

[18] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[19] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Information processing & management*, vol. 24, no. 5, pp. 513–523, 1988.

[20] T. K. Landauer, P. W. Foltz, and D. Laham, "An introduction to latent semantic analysis," *Discourse processes*, vol. 25, no. 2-3, pp. 259–284, 1998.

[21] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371

[22] J. Zhang, Y. Dong, Y. Wang, J. Tang, and M. Ding, "Prone: Fast and scalable network representation learning." in *IJCAI*, vol. 19, 2019, pp. 4278–4284.

[23] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.

[24] J. Qiu, Y. Dong, H. Ma, J. Li, K. Wang, and J. Tang, "Network embedding as matrix factorization: Unifying deepwalk, line, pte, and node2vec," *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, 2017.

[25] H. Cai, V. W. Zheng, and K. C.-C. Chang, "A comprehensive survey of graph embedding: Problems, techniques, and applications," *IEEE transactions on knowledge and data engineering*, vol. 30, no. 9, pp. 1616–1637, 2018.

[26] R. Kinney, "Conference Peer Review with the Semantic Scholar API," https://blog.allenai.org/conference-peer-review-with-the-semantic-scholar-api-24ab9fce2324, 2021.

[27] D. Mimno and A. McCallum, "Expertise modeling for matching papers with reviewers," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 500–509.

[28] https://mimno.infosci.cornell.edu/data/nips_reviewer_data.tar.gz, 2007.

[29] R. Kinney, "Conference Peer Review with the Semantic Scholar API," https://blog.allenai.org/conference-peer-review-with-the-semantic-scholar-api-24ab9fce2324.

[30] https://openreview.net/.

[31] https://www.softconf.com/.

[32] https://cmt3.research.microsoft.com/.

[33] K. W. Church and V. Kordoni, "Emerging trends: Sota-chasing," *Natural Language Engineering*, vol. 28, no. 2, pp. 249–269, 2022.

[34] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.

[35] O. Levy and Y. Goldberg, "Neural word embedding as implicit matrix factorization," in *NIPS*, 2014.

[36] "Learning of social representations," https://www.semanticscholar.org/paper/93b050f5bf0567a675979cd564cbe66ff9c3a78f.

[37] "Topic-aware latent models for representation learning on networks," https://www.semanticscholar.org/paper/21ee2cc0bf41c1b74efb6104edd4df73416b46c1.

[38] "Simwalk: Learning network latent representations with social relation similarity," https://www.semanticscholar.org/paper/e294339b402ce055d5a5198becc35b2dbbd20a9a.

[39] "Deep representation learning on complex graphs," https://www.semanticscholar.org/paper/bb11bec51c2e069ef0ddba4eb3117c9dbc8a4584.

[40] "node2vec: Scalable feature learning for networks," https://www.semanticscholar.org/paper/36ee2c8bd605afd48035d15fdc6b8c8842363376.

[41] "LINE: Large-scale information network embedding," https://www.semanticscholar.org/paper/0834e74304b547c9354b6d7da6fa78ef47a48fa8.

[42] "A comprehensive survey of graph embedding: Problems, techniques, and applications," https://www.semanticscholar.org/paper/006906b6bbe5c1f378cde9fd86de1ce9e6b131da.

[43] "metapath2vec: Scalable representation learning for heterogeneous networks," https://www.semanticscholar.org/paper/c0af91371f426ff92117d2ccdadb2032bec23d2c.

[44] "Deepwalk: online learning of social representations," https://www.semanticscholar.org/paper/fff114cbba4f3ba900f33da574283e3de7f26c83.

[45] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014, pp. 701–710.

[46] T. L. Scao, T. Wang, D. Hesslow, L. Saulnier, S. Bekman, S. Bari, S. R. Biderman, H. ElSahar, N. Muennighoff, J. Phang, O. Press, C. Raffel, V. Sanh, S. Shen, L. Sutawika, J. Tae, Z. X. Yong, J. Launay, and I. Beltagy, "What language model to train if you have one million gpu hours?" https://www.semanticscholar.org/paper/What-Language-Model-to-Train/bb15f3727f827a3cb88b5d3ca48415c09b40a88f.

[47] M. Mosbach, M. Andriushchenko, and D. Klakow, "On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines," https://www.semanticscholar.org/paper/On-the-Stability/8b9d77d5e52a70af37451d3db3d32781b83ea054.

[48] A. Matton and L. de Oliveira, "Emergent properties of finetuned language representation models," https://www.semanticscholar.org/paper/Emergent-Properties-of/79fdff5339017ec92b979efa4dff33d21a69b66e.

[49] S. Gehman, S. Gururangan, M. Sap, Y. Choi, and N. A. Smith, "Realtoxicityprompts: Evaluating neural toxic degeneration in language models," https://www.semanticscholar.org/paper/RealToxicityPrompts/399e7d8129c60818ee208f236c8dda17e876d21f.

[50] K. McGuffie and A. Newhouse, "The radicalization risks of gpt-3 and advanced neural language models," https://www.semanticscholar.org/paper/The-Radicalization/02fde8bfd9259a4f53316579eb0bf97213559e5c.

[51] A. Abid, M. S. Farooqi, and J. Y. Zou, "Persistent anti-muslim bias in large language models," https://www.semanticscholar.org/paper/Persistent-Anti-Muslim/4c2733d191e347753bb28afa46a1c55c65e085be.

[52] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" https://www.semanticscholar.org/paper/On-the-Dangers-of-Stochastic/6d9727f1f058614cada3fe296eeebd8ec4fc512a.

[53] ——, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[54] E. Bernhardsson, "Annoy," https://github.com/spotify/annoy, 2013.

[55] F. Research, "Faiss," https://github.com/facebookresearch/faiss, 2019.

[56] N. Peinelt, D. Nguyen, and M. Liakata, "tbert: Topic models and bert joining forces for semantic similarity detection," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020, pp. 7047–7055.

[57] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[58] "BERTopic," https://maartengr.github.io/BERTopic/index.html.

[59] "Topic modeling bert+lda," https://www.kaggle.com/code/dskswu/topic-modeling-bert-lda.

[60] J. A. Fisher, "The value of natural sounds," https://www.semanticscholar.org/paper/The-Value-of-Natural-Sounds-Fisher/f3044464aae6d74ef8b7c9a85810b5164e837378.

[61] F. V. Winnett, "Re-examining the foundations," https://www.semanticscholar.org/paper/Re-Examining-the-Foundations/52e0d57cccd3ffc9097ac9ee95c1a2214fdc1c7a.

[62] R. H. Pfeiffer, "Facts and faith in biblical history," https://www.semanticscholar.org/paper/Facts-and-Faith-in-Biblical-History-Pfeiffer/639c1e93818b157541bf5522a7cb2cf564119479.

[63] M. Liberman and A. Prince, "On stress and linguistic rhythm," https://www.semanticscholar.org/paper/On-stress/b8ec853894551c0e7a822df50dc04eccd613d46f.

[64] ——, "On stress and linguistic rhythm," *Linguistic inquiry*, vol. 8, no. 2, pp. 249–336, 1977.

[65] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, 1998, pp. 604–613.

[66] P. I. Nakov, A. S. Schwartz, M. Hearst *et al.*, "Citances: Citation sentences for semantic analysis of bioscience text," in *Proceedings of the SIGIR*, vol. 4. Citeseer, 2004, pp. 81–88.

[67] V. Qazvinian and D. Radev, "Identifying non-explicit citing sentences for citation-based summarization." in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010, pp. 555–564.

[68] A. Abu-Jbara and D. Radev, "Reference scope identification in citing sentences," in *Proceedings of the 2012 conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, 2012, pp. 80–90.

[69] N. Craswell, D. Hawking, and S. Robertson, "Effective site finding using link anchor information," in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 2001, pp. 250–257.

[70] A. M. Turing, "On computable numbers, with an application to the entscheidungsproblem," *J. of Math*, vol. 58, no. 345-363, p. 5, 1936.

[71] J. C. Gower and G. B. Dijksterhuis, *Procrustes problems*. OUP Oxford, 2004, vol. 30.

[72] "scipy.linalg.orthogonal_procrustes," https://docs.scipy.org/doc/scipy/reference/generated/scipy.linalg.orthogonal_procrustes.html.

[73] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020.

[74] "Nodevectors," https://github.com/VHRanger/nodevectors.

[75] https://chat.openai.com/.

[76] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, and A. Kanakia, "Microsoft academic graph: When experts are not enough," *Quantitative Science Studies*, vol. 1, no. 1, pp. 396–413, 2020.

[77] B.-J. Hsu, I. Shen, D. Eide, A. Chen, and R. Rogahn, "Microsoft academic graph," https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/.

[78] "DOI foundation," https://www.doi.org/.

[79] "Pubmed," https://pubmed.ncbi.nlm.nih.gov/.

[80] "DBLP," https://dblp.org/.

[81] "Pubmed central," https://www.ncbi.nlm.nih.gov/pmc/.

[82] "arXiv," https://arxiv.org/.

[83] "Acl anthology," https://aclanthology.org/.

[84] A. Singh, M. D'Arcy, A. Cohan, D. Downey, and S. Feldman, "Scirepeval: A multi-format benchmark for scientific document representations," https://github.com/allenai/scirepeval.

[85] ——, "Scirepeval: A multi-format benchmark for scientific document representations," *ArXiv*, vol. abs/2211.13308, 2022.

[86] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," *Advances in neural information processing systems*, vol. 33, pp. 22 118–22 133, 2020.

[87] W. Hu, M. Fey, H. Ren, M. Nakata, Y. Dong, and J. Leskovec, "Ogb-lsc: A large-scale challenge for machine learning on graphs," *arXiv preprint arXiv:2103.09430*, 2021.

[88] ——, "Kdd cup ogb large-scale challenge (ogb-lsc) session 1," https://www.youtube.com/watch?v=MblxCqwengI, 2021.

[89] J. Priem, H. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," *arXiv preprint arXiv:2205.01833*, 2022.

[90] "Social science research network," https://papers.ssrn.com/sol3/DisplayJournalBrowse.cfm.

[91] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '99. New York, NY, USA: Association for Computing Machinery, 1999, p. 50–57. [Online]. Available: https://doi.org/10.1145/312624.312649

[92] D. M. Blei, A. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2001.

[93] M. Grootendorst, "Bertopic," https://maartengr.github.io/BERTopic/index.html, 2022.

[94] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*, 3rd ed. USA: Cambridge University Press, 2020.

[95] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.

[96] J. Priem, H. A. Piwowar, and R. Orr, "Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts," *ArXiv*, vol. abs/2205.01833, 2022.

[97] F. R. Chung, *Spectral graph theory*. American Mathematical Soc., 1997, vol. 92.

[98] A. Cohan, W. Ammar, M. van Zuylen, and F. Cady, "Structural scaffolds for citation intent classification in scientific publications," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 3586–3596. [Online]. Available: https://aclanthology.org/N19-1361

[99] T. Kudo and J. Richardson, "SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71. [Online]. Available: https://aclanthology.org/D18-2012

[100] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[101] K. W. Church, "Emerging trends: Deep nets thrive on scale," *Natural Language Engineering*, vol. 28, pp. 673 – 682, 2022.

[102] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3645–3650. [Online]. Available: https://www.aclweb.org/anthology/P19-1355

[103] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[104] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019.

[105] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," *NeurIPS*, 2020.

[106] R. Dale, "GPT-3: What's it good for?" *Natural Language Engineering*, vol. 27, no. 1, pp. 113–118, 2021.

[107] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, and N. Fiedel, "Palm: Scaling language modeling with pathways," 2022. [Online]. Available: https://arxiv.org/abs/2204.02311

[108] S. Bubeck and M. Sellke, "A universal law of robustness via isoperimetry," *Advances in Neural Information Processing Systems*, vol. 34, pp. 28 811–28 822, 2021.

[109] Y. Li and Y. Liang, "Learning overparameterized neural networks via stochastic gradient descent on structured data," *Advances in neural information processing systems*, vol. 31, 2018.

[110] A. Brutzkus, A. Globerson, E. Malach, and S. Shalev-Shwartz, "Sgd learns over-parameterized networks that provably generalize on linearly separable data," *International Conference on Learning Representations*, 2018.

[111] Z. Allen-Zhu, Y. Li, and Y. Liang, "Learning and generalization in overparameterized neural networks, going beyond two layers," *Advances in neural information processing systems*, vol. 32, 2019.

[112] S. Oymak and M. Soltanolkotabi, "Toward moderate overparameterization: Global convergence guarantees for training shallow neural networks," *IEEE Journal on Selected Areas in Information Theory*, vol. 1, no. 1, pp. 84–105, 2020.

[113] L. Bottou, "Stochastic gradient descent tricks," *Neural Networks: Tricks of the Trade: Second Edition*, pp. 421–436, 2012.

[114] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Carnegie Mellon University, USA, Tech. Rep., 1994.

[115] K. Church, "Better together team github," https://github.com/kwchurch/JSALT_Better_Together, 2023.

[116] https://jsalt2023.univ-lemans.fr/en/better-together-text-context.html, 2023.

[117] https://huggingface.co/allenai/specter1.

[118] https://huggingface.co/allenai/specter2.

[119] https://huggingface.co/malteos/scincl.

[120] https://huggingface.co/michiyasunaga/LinkBERT-large.