

The University of Melbourne

IBM Research Australia

contact: karin.verspoor@unimelb.edu.au

We present COVID-SEE, a system for medical literature discovery based on the concept of information exploration, which builds on several distinct text analysis and natural language processing methods to structure and organise information in publications, and augments search by providing a visual overview supporting exploration of a collection to identify key articles of interest. We developed this system over COVID-19 literature to help medical professionals and researchers explore the literature evidence, and improve findability of relevant information. COVID-SEE is available at covid-see.cis.unimelb.edu.au.

The outbreak of Coronavirus disease 2019 (COVID-19) has led to a rapid and proactive response from medical and AI communities worldwide. In information retrieval and natural language processing, efforts have concentrated on building datasets and tools for efficiently managing the growing literature on COVID-19, historical coronaviruses, and other related diseases (Hutson, 2020). One of the main stimuli for the creation of new technology was the release of the COVID-19 Open Research Dataset (CORD-19) (Wang et al., 2020), a regularly updated collection of coronavirus-related publications and preprints. But while many tools have emerged for the purposes of article retrieval and question answering,¹ relatively less work has been devoted to designing systems that go beyond returning a list of (relevant) documents, and that attempt to leverage domain knowledge to organise and present information found within the literature. With our work, we aim to fill this gap by making available a web application that combines a search

In both the retrieval and visualisation components, we adopt several well-established NLP techniques. Article retrieval is powered by an existing neural search engine specifically developed for the CORD-19 dataset (Zhang et al., 2020). The results are shown as a list of document hits with metadata about the articles. Separately, the dashboard represents the current active collection with three distinct *interactive views*. The first is a **relational concept view** in which we employ Sankey diagrams to organise the medical concepts found in the articles according to broad, clinically-relevant

1



Figure 1: Sankey diagram of PICO concepts and relations for 58 articles retrieved for the query *incubation period of COVID-19*. Links between concepts can be selected to reveal papers that contain those relations. (See Fig. 2)

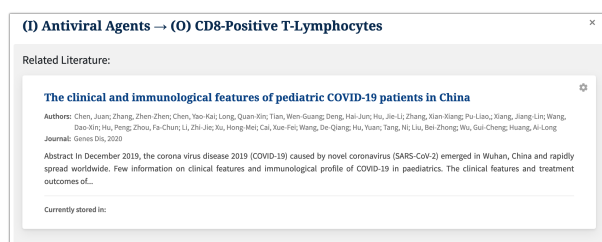


Figure 2: Selection of an Intervention-Outcome link reveals a paper containing the specific concept relations.

PICO categories (Richardson et al., 1995), such as population, intervention and outcome. In this view, more salient relations carry more weight, and once a relation is clicked, the corresponding articles are revealed. We use an example based on the query *incubation period of COVID-19* to illustrate this functionality (Fig. 1). The second, **topic view** (Fig. 3), is more thematic and shows representative topics for the current collection. For this component, we trained a global topic model on medical concepts extracted from CORD-19. Our third component is a **concept cloud view** (Fig. 4), showing the most salient concepts for each active document.

2 Data

CORD-19 is currently the most extensive coronavirus literature corpus publicly available (Wang et al., 2020). The dataset contains all COVID-

19 and coronavirus-related research (e.g. SARS, MERS, etc.) from different sources, including PubMed’s PMC open access corpus, research articles from a corpus maintained by the WHO, and bioRxiv and medRxiv pre-prints. The corpus is updated daily and consisted of more than 130,000 documents as of 6 June. Together with the release of the CORD dataset, a Kaggle challenge² took place in which we also participated.

3 System overview

3.1 Information retrieval

After submitting a query, a list of retrieved documents is shown. Each document entry can be expanded to display the entirety of the abstract as well as metadata (authors, journal, source, year, license). The user can also filter by criteria such as year and source. Articles can be selected and added to the user’s briefcase. This briefcase represents the set of documents that the user wishes to keep track of, and can be visualised, versioned, and exported.

We make use of an existing information retrieval system for the CORD-19 dataset, available at covidex.ai (Zhang et al., 2020). This provides us with a well-performing retrieval system—by our analysis, in the top 30% of submissions to the TREC-COVID shared task (Roberts et al., 2020).

² kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

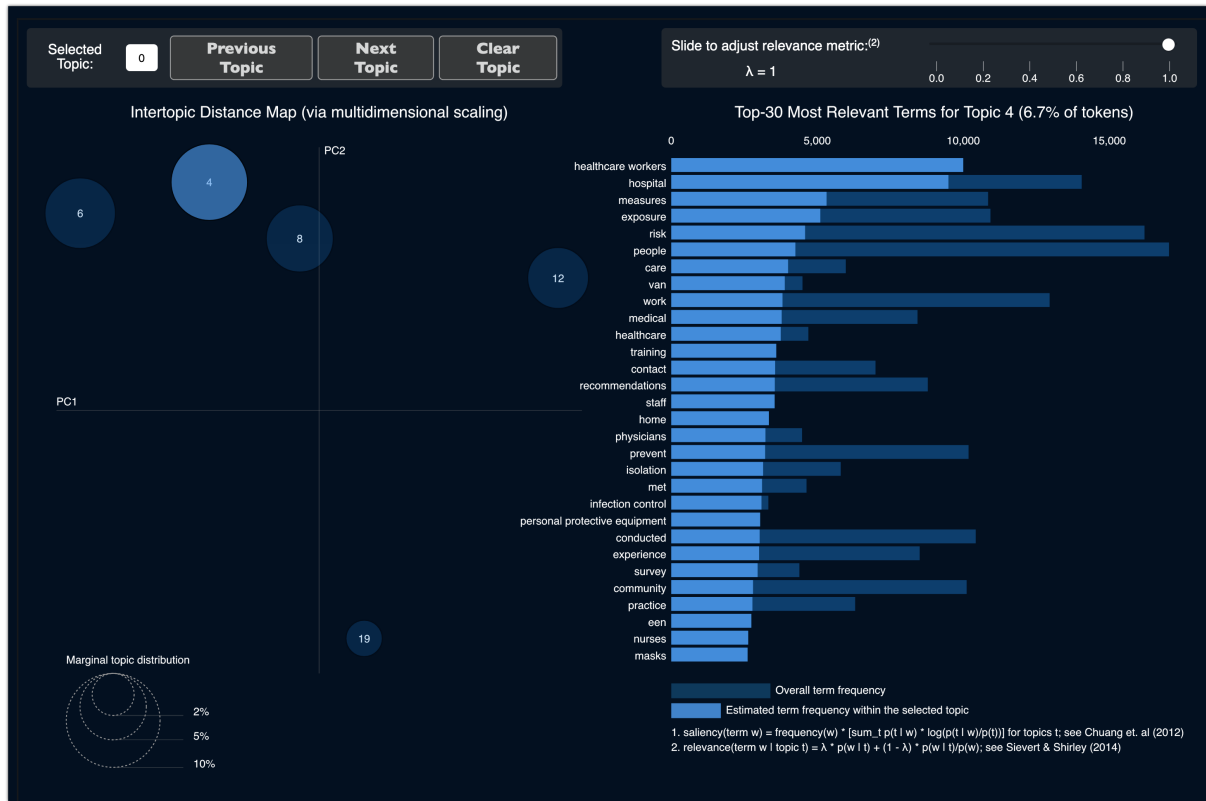


Figure 3: Topic visualisation for the 58 articles related to the query *incubation period of COVID-19*.

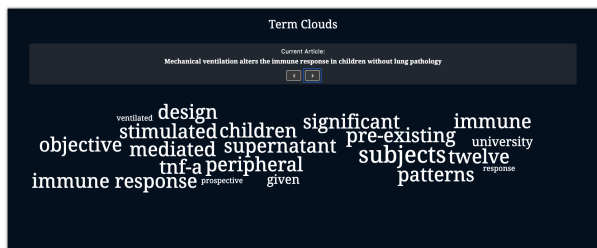


Figure 4: Concept cloud view of a single document.

3.2 Relational concept view

Given a collection of retrieved documents, the user can directly proceed to the dashboard to explore the collection. A key view over the collection is via a Sankey diagram frame, which organises the identified medical concepts into PICO categories (Population, Intervention/Comparator, Outcome), and shows how they interact. The diagram displays relations between pairs of medical concepts, where the strength of a relation corresponds to the number of supporting abstracts in which that concept pair is attested. To obtain the PICO concepts, we follow a two-step procedure, as described below.

PICO span identification We train a BiLSTM-CRF model (Lample et al., 2016) on the EBM-NLP

	P	R	F1
All	0.73	0.65	0.69
Population	0.78	0.79	0.78
Intervention	0.57	0.64	0.60
Outcome	0.80	0.58	0.67

Table 1: PICO labelling results on EBM-NLP.

dataset (Nye et al., 2018), which contains $\sim 5,000$ MEDLINE³ abstracts of reports of randomised clinical trials annotated with textual spans that describe the PICO elements. As pretrained word representations for the model, we use 200-dimensional word2vec embeddings induced on PubMed abstracts and MEDLINE articles (Hakala et al., 2016). We observe a micro-averaged F1 score of 0.69 on the EBM-NLP test set, and report the full results in Table 1; these are comparable to those reported in Nye et al. (2018). For our purposes, getting the textual span exactly correct is less important, as our final representations are not the snippets but the medical concepts that occur inside them. The PICO identification step is applied to all abstracts of the CORD-19 dataset.

³ www.nlm.nih.gov/bsd/pmresources.html

PICO snippet	PICO category	MeSH terms
<i>patients presenting with RTI</i>	Population	Patients; Respiratory Tract Infections
<i>Mass vaccination campaigns with par- enteral vaccines</i>	Intervention	Immunization Programs; Vaccines; Parenteral Nutrition; Macrophage Activation Syndrome
<i>cumulative COVID-19-related hospital- ization and death rates</i>	Outcome	Hospitalization; Mortality

Table 2: Examples of extracted PICO textual spans and MeSH terms found in them. The PICO concepts we use are the MeSH terms, typed by their respective PICO category (e.g. Vaccines+Intervention).

Recognition of PICO concepts In this step, we recognise medical concepts that occur within the identified PICO spans. This is part of a larger concept recognition procedure that we describe in more detail in Section 3.3. Here, we take the terms corresponding to Medical Subject Headings (MeSH),⁴ which is a structured vocabulary maintained by the National Library of Medicine. We use MeSH due to its transparent hierarchical structure, making it easy to control term granularity. Additionally, it is also an established way to represent keywords found in medical articles indexed by PubMed. A few examples of extracted PICO concepts are shown in Table 2.

Establishing relations The output of the previous two steps is the extracted MeSH terms for each article, typed by their respective PICO category. In the Sankey diagram, we then display pairwise relations between Population–Intervention and Intervention–Outcome concepts. Two concepts are related when they are found in the same abstract. We also include the strength of the connection, corresponding to the absolute frequency of that concept pair in the entire dataset. The user can select a concept relation link of interest, browse the relevant articles containing that relation, and save them to the briefcase for further exploration.

The rendering of the diagram is dynamic. This means that whenever a document collection selected by the user changes, the diagram is updated. In addition, the Sankey view is interactive, allowing the user to click on any relation to highlight the relevant part of the diagram and show the supporting documents. A planned future release of our tool will support *semantic search*, which will enable the use of PICO-typed MeSH terms as search criteria.

3.3 Medical concept recognition

To allow for more conceptual representation of the documents, we transform them into an unordered set of Unified Medical Language System

(UMLS: Lindberg et al. (1993)) concepts. These concepts are extracted using MetaMap⁵ from the document’s abstract, or, if it is absent, from the first two paragraphs. This choice of representation is motivated as follows: (1) it avoids splitting multi-word concepts (such as *degenerative disease of the central nervous system*) into less meaningful units (*of, the, central, etc.*), as would happen in the case of token-based representation; and (2) it maps different lexico-grammatical variations of a given term into a single concept, thus reducing noise in the data, and highlighting important keywords. For example, concept C0000731 occurs in the articles as *abdominal distension, abdominal distention, bloating, distended abdomens, swelling of abdomen, etc.*, which would not be captured by typical approaches to text normalisation such as lemmatisation, stemming, or *n*-gram overlap. We represent each MetaMap-disambiguated concept by the lexicalisation that occurs most frequently in the document collection, as distinct from the MetaMap “preferred term”, which is often a technical description rather than its lexical form (e.g. we use *colon* instead of preferred term *Colon structure (body structure)*). As the last step, we remove stopwords based on PubMed’s stopword list⁶ and the 100 most frequent tokens in the dataset. The resulting representation (Table 3) forms the foundation of the remaining analysis and visualisations.

3.4 Topic view

To enable document exploration based on common semantic themes, we provide a topic modelling view. We use latent Dirichlet allocation (LDA) (Blei et al., 2003) to learn topics over the whole dataset, and represent the topic mixtures for a given subset of articles (based on the briefcase) as a 2-dimensional map using principal component analysis (PCA) via the pyLDAvis package.⁷ LDA represents each document as a mixture of topics, and each topic as a mixture of words. We choose the

⁴ ncbi.nlm.nih.gov/mesh

⁵ metamap.nlm.nih.gov/pubmedhelp.T.stopwords/

⁶ ncbi.nlm.nih.gov/books/NBK3827/table/
⁷ pyli.org/project/pyLDAvis

Abstract	porcine deltacoronavirus (pdcov) causes severe diarrhea and vomiting in affected piglets. the aim of this study was to establish the basic, in vitro characteristics of the life cycle such as replication kinetics, cellular ultrastructure, virion morphology, and induction of autophagy of pdcov.
Concept IDs	C4038448 C1314792 C1443924 C0042963 C0392760 C2948600 C0557651 C0040363 C0443211 C1527178 C2827718 C1521970 C0023675 C0598312 C0022702 C0007634 C0041623 C0042760 C0332437 C0857127 C0004391
Concepts	porcine deltacoronavirus, causes, severe diarrhea, vomiting, affected, aim, study, to, establish, based, in vitro, characteristics, life cycle, replication, kinetics, cells, ultrastructural, virion, morphology, induction, autophagy
Final terms	porcine deltacoronavirus, causes, severe diarrhea, vomiting, affected, aim, establish, in vitro, characteristics, life cycle, replication, kinetics, ultrastructural, virion, morphology, induction, autophagy

Table 3: Text representation steps

optimal number of topics (20) based on the C_v topic coherence measure (Röder et al., 2015). After learning the topic distribution for the whole dataset, we display only those topics which are representative of the selected subset, that is, ones for which the average weight across the selected documents is over a threshold t , which we set to 0.05. Thus, the topic selection is dynamic for each set of selected documents.

3.5 Concept cloud view

For each of the documents in the user’s briefcase, we display a wordcloud⁸ containing the 20 most representative concepts. We regard each of the articles in the briefcase as a target corpus, and the remainder of the articles in the briefcase as a background corpus, and compare concept distributions using the log-likelihood test (Rayson and Garside, 2000). This highlights concepts that differentiate a particular document from others in the briefcase, even if they discuss similar topics and have the same set of frequent terms. For instance, if the briefcase contains articles regarding mechanical ventilation and we choose one of them,⁹ the top 10 concepts chosen by log-likelihood will be similar to 10 most frequent concepts, but with two visible differences (Table 4): the log-likelihood results do not contain common terms as *mechanical ventilation*, frequently used in all documents in the briefcase; on the other hand, they include more document-specific concepts as *ai (avian influenza)*. Table 4 also shows how operating at the concept level captures multi-word terms such as *mechanical ventilation* and *admitted to intensive care unit*.

⁸ github.com/chriszhou/react-wordcloud

⁹ pubmed.ncbi.nlm.nih.gov/18440440

3.6 Technical details

All data is stored in a graph database (neo4j¹⁰). The front-end of our web application interacts with the database via the Cypher language and the py2neo¹¹ library. The website was built with React¹² and Flask¹³, and topic visualisations are supported by pyLDAvis¹⁴. Core analysis was done in Python.

4 Related work

In the first months of the COVID-19 crisis, many tools were released to support literature exploration. We present here the systems that, like ours, focus on the visualisation of concepts and relations for retrieved documents, or that use concept information to guide search. We provide a summary of the existing tools and their functionalities in Table 5.

AllenAI’s SciSight Faceted Search (Hope et al., 2020)¹⁵ is a tool for exploring how authors and topics interact over time. The user can select the desired topics from different PICO-like categories, which act as a filter for the shown articles. It also provides a co-mentions view with chord diagrams, which displays associations between diseases and chemicals, or between proteins, genes and cells. This is conceptually similar to our goal of representing concept relations, but their diagram view does not support PICO.

The identification of PICO elements in SciSight is based on **DOC Search**,¹⁶ where a user constructs a query by accepting auto-suggestions based on an ontology. The selected terms can be combined with a variety of Boolean and ontological operators. The user can then use the data visualisations to further narrow down the search results. DOC Search represents the document collection with different views, but does not display topics

¹⁰ neo4j.com ¹¹ py2neo.org ¹² reactjs.org ¹³ flask.palletsprojects.com

¹⁴ github.com/bmabey/pyLDAvis ¹⁵ scisight.apps.allenai.org

¹⁶ covid-search.doctorevidence.com

	Concepts	Tokens
Fre-quent	sars-cov, airborne, mechanical ventilation, people, countries, killed, required, medical, community, threat	mechanical, airborne, ventilation, respiratory, people, countries, killed, required, intensive, care
Log-likelihood	epidemic, sars-cov, airborne, people, killed, community, threat, ai, countries, required, admitted to intensive care unit	airborne, people, killed, infectious, diseases, caused, community, ai, will, countries

Table 4: Representative concepts vs. tokens

or relations between PICO concepts, and does not support natural language queries. That said, DOC Search is a powerful tool for navigating biomedical research publications.

IBM COVID-19 Navigator¹⁷ is similar in functionality to DOC Search. The tool fully embraces the use of the UMLS knowledge graph to support semantic search, with Boolean search operators and UMLS semantic relationships. It does not provide PICO or visualisation functionality.

COVID-19 LOVE¹⁸ from the Epistemonikos project organises the literature according to study type (systematic reviews, broad syntheses, primary studies), as well as PICO categories and question types (e.g. diagnosis, prognosis), but is not based on automatic analysis or on the CORD collection, and does not provide visual exploration.

Trialstreamer (Nye et al., 2020) finds and summarises new clinical trial publications, registrations, and preprints in both COVID-19-related and broader literature. PICO snippets of text are shown, together with the key findings and an indication of the risk of bias. *Evidence maps* link interventions to outcomes, and predict a trial’s findings.

Other systems include: **NIH’s LitCovid** (Chen et al., 2020)¹⁹, which provides—based on human verification—categorisation of papers according to diagnosis, treatment, etc.; **WellAI**’s tool²⁰ for concept search, enabling retrieval of documents via highly associated concepts; **SemViz** (Tu et al., 2020), using a knowledge graph from Blender Lab²¹ to create interaction plots and word clouds for chemicals, genes and diseases; and **COVID Intelligent Search**²², which offers filters for various medical categories (e.g. drugs, indication), and suggests textual extracts as queries. **Le Bras et al. (2020)** present work on coarse-to-fine exploration of literature based on hierarchical topic models, however not visualisation of arbitrary sets of articles. **Ahamed and Samad (2020)** construct

	COVID-SEE (ours)	SciSight	DOC Search	COVID-19 Navigator	LitCOVID	SemViz	WellAI	COVID Intel. Search	Le Bras et al. 2020	COVID-19 LOVE	Trialstreamer
Search:											
NL/IR	✓	×	✓	×	✓	×	×	✓	×	×	✓
Concepts	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
PICO	✓	✓	✓	×	×	×	×	×	×	✓	✓
Visualisation:											
Concepts	✓	?	?	×	×	✓	×	×	×	×	?
Relations	✓	✓	×	×	×	✓	×	×	×	×	×
Topics	✓	×	×	×	×	×	×	?	×	×	×

Table 5: Comparison with related systems.

a graph from entity co-occurrences that enables centrality-based ranking of drugs, pathogens and biomolecules.

5 Limitations and future work

COVID-SEE has been designed to facilitate more interactive exploration of the COVID-19 literature, through integration of sub-collection thematic analysis, document-level visual concept summaries, and PICO-structured concept relations. It does not currently support semantic search based on PICO-style queries; a natural extension would be to add this functionality utilising the PICO and UMLS concept pre-processing. A recommendation system for articles which have similar topic distributions could be added. In terms of visual representations, we will experiment with expanding beyond the MeSH term vocabulary to include more specific terminology, and use of the hierarchical relationships that exist between terms. Finally, we are planning a user study with medical professionals to evaluate the potential of COVID-SEE as a knowledge discovery tool.

6 Screencast

The screencast can be viewed at <https://youtu.be/vLtXuTz-LU>.

¹⁷ covid-19-navigator.mybluemix.net

¹⁸ app.iloveevidence.com/loves/5e6fdb9669c00e4ac072701d

¹⁹ ncbi.nlm.nih.gov/research/coronavirus

²⁰ wellai.health/covid

²¹ blender.cs.illinois.edu/covid19

²² covidsearch.sinequa.com

References

- Sabber Ahamed and Manar Samad. 2020. Information Mining for COVID-19 Research From a Large Volume of Scientific Literature. *arXiv:2004.02085*.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.
- Q. Chen, A. Allot, and Z. Lu. 2020. [Keep up with the latest coronavirus research](#). *Nature*, 579(7798):193.
- Carsten Görg, Hannah Tipney, Karin Verspoor, William A Baumgartner, K Bretonnel Cohen, John Stasko, and Lawrence E Hunter. 2010. Visualization and language processing for supporting analysis across the biomedical literature. *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 420–429.
- Kai Hakala, Suwisa Kaewphan, Tapio Salakoski, and Filip Ginter. 2016. [Syntactic analyses and named entity recognition for PubMed and PubMed central — up-to-the-minute](#). In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 102–107, Berlin, Germany. Association for Computational Linguistics.
- Tom Hope, Jason Portenoy, Kishore Vasan, Jonathan Borchardt, Eric Horvitz, Daniel S. Weld, Marti A. Hearst, and Jevin D. West. 2020. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *bioRxiv*.
- Matthew Hutson. 2020. [Artificial-intelligence tools aim to tame the coronavirus literature](#). *Nature*.
- Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. [Neural architectures for named entity recognition](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.
- Pierre Le Bras, Azimeh Gharavi, David A. Robb, Ana F. Vidal, Stefano Padilla, and Mike J. Chantler. 2020. Visualising COVID-19 Research. *arXiv:2005.06380*.
- Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. 1993. The Unified Medical Language System. *Yearbook of Medical Informatics*, 2(01):41–51.
- Gary Marchionini. 2006. [Exploratory search: From finding to understanding](#). *Commun. ACM*, 49(4):41–46.
- Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. [A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.
- Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C. Wallace. 2020. [Trialstreamer: Mapping and browsing medical evidence in real-time](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 63–69, Online. Association for Computational Linguistics.
- Patrick Cheong-Iao Pang, Karin Verspoor, Shanton Chang, and Jon Pearce. 2015. [Conceptualising health information seeking behaviours and exploratory search: result of a qualitative study](#). *Health and Technology*, 5(1):45–55.
- Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. *The workshop on comparing corpora*, pages 1–6.
- W Scott Richardson, Mark C Wilson, Jim Nishikawa, Robert S Hayward, et al. 1995. The well-built clinical question: a key to evidence-based decisions. *Acp j club*, 123(3):A12–3.
- Kirk Roberts, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, Kyle Lo, Ian Soboroff, Ellen Voorhees, Lucy Lu Wang, and William R Hersh. 2020. [TREC-COVID: Rationale and Structure of an Information Retrieval Shared Task for COVID-19](#). *Journal of the American Medical Informatics Association*. Ocaa091.
- Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408.
- John Stasko, Carsten Görg, and Zhicheng Liu. 2008. [Jigsaw: Supporting investigative analysis through interactive visualization](#). *Information Visualization*, 7(2):118–132.
- Jingxuan Tu, Marc Verhagen, Brent Cochran, and James Pustejovsky. 2020. Exploration and Discovery of the COVID-19 Literature through Semantic Visualization. *arXiv:2007.01800*.
- Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. [CORD-19: The Covid-19 Open Research Dataset](#). In *Proceedings of the Workshop on NLP for COVID-19 at ACL 2020*, Online. Association for Computational Linguistics.

Edwin Zhang, Nikhil Gupta, Rodrigo Nogueira,
 Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly De-
 ploying a Neural Search Engine for the COVID-19
 Open Research Dataset: Preliminary Thoughts and
 Lessons Learned. arXiv:2004.05125.