

NLP Workbench: Efficient and Extensible Integration of State-of-the-art Text Mining Tools

Peiran Yao, Matej Kosmajac, Abeer Waheed,
Kostyantyn Guzhva, Natalie Hervieux, Denilson Barbosa

Department of Computing Science
University of Alberta
{peiran, denilson}@ualberta.ca

Abstract

NLP Workbench is a web-based platform for text mining that allows non-expert users to obtain semantic understanding of large-scale corpora using state-of-the-art text mining models. The platform is built upon latest pre-trained models and open source systems from academia that provide semantic analysis functionalities, including but not limited to entity linking, sentiment analysis, semantic parsing, and relation extraction. Its extensible design enables researchers and developers to smoothly replace an existing model or integrate a new one. To improve efficiency, we employ a microservice architecture that facilitates allocation of acceleration hardware and parallelization of computation. This paper presents the architecture of NLP Workbench and discusses the challenges we faced in designing it. We also discuss diverse use cases of NLP Workbench and the benefits of using it over other approaches. The platform is under active development, with its source code released under the MIT license¹. A website² and a short video³ demonstrating our platform are also available.

1 Introduction

Text mining, also known as *text analytics* or *text analysis*, is the process where a user interacts with machine-supported analysis tools that transform natural language text into structured data, to gain insights and new knowledge from the text (Feldman and Sanger, 2006a). For more than two decades, text mining systems have been built for applications in various domains, such as business intelligence, analytical sociology, and medical sciences (Hearst, 1999), demonstrating irreplaceable value. Analysis tools in text mining usually take the form of machine learning (ML) and natural language processing (NLP) models and span a large spectrum of

ML and NLP subfields, such as entity linking, sentiment analysis, relation extraction, and text summarization.

Nearly every subfield of NLP involved in text mining has been rapidly evolving in recent years, with records on benchmarks being continuously broken⁴. A positive practice of releasing the code and models to the public has been adopted by a growing number of researchers⁵ to address reproducibility and accessibility issues in the field⁶. Despite efforts to make new models more accessible, non-expert users such as digital humanists and business analysts still face entry barriers when trying to apply the latest models. Some salient issues include: (1) heterogeneous software stacks required to run the models; (2) non-standardized, inconsistent input and output formats; (3) the lack of user-friendly interfaces to apply the models and visualize the results; and (4) the constraints on computation and networking resources. We build NLP Workbench with the goal of addressing these issues and further bridging the gap between state-of-the-art open NLP research and the use of these models and tools in text mining applications by non-experts.

NLP Workbench is designed with two fundamental principals in mind: for developers and NLP researchers, fast and easy adaptation of off-the-shelf models and tools; and for non-expert users such as sociologists, a user-friendly interface for both document-level and corpus-level analysis. Following these principles, NLP Workbench offers the following key features:

Platform NLP Workbench unifies corpus management, text mining tools, and visualization in a single platform. It provides a growing list of models and tools that are based on state-of-the-

¹<https://github.com/U-Alberta/NLPWorkbench/>

²<https://newskg.wdmuofa.ca>

³<https://vimeo.com/801006908>

⁴<http://nlpprogress.com>

⁵<https://paperswithcode.com>

⁶<https://aclrollingreview.org/responsibleNLPresearch/>

art research, currently offering functionalities like named entity recognition, entity linking, relation extraction, semantic parsing, summarization, sentiment analysis, and social network analysis.

Interaction A web interface is included for user interactions with the ability to visualize the model results at document and corpus levels. Users could choose to interactively apply a model on a given document and have the results saved for future queries, or to apply models in batches on selected documents in a corpus.

Architecture For development, NLP Workbench adopts containerization, allowing new models to be added independent of the software stack of existing models. For deployment, its microservice architecture allows models to be deployed in a distributed way on machines that meet the computing and networking requirements of individual models, enabling horizontal scaling.

Interface Tools in NLP Workbench can be accessed in versatile ways. Besides the web interface, non-expert users could import new documents into the platform via a browser extension. For developers and researchers, NLP Workbench provides RESTful API and remote procedure call (RPC) interfaces for easy integration with other applications and pipelines.

2 Related Work

Hearst (1999) and Cunningham et al. (2002) identified three key aspects of an effective text mining system: management of text document collections (corpora), application of text processing algorithms on the collection, and visualization of results. LINDI (Hearst, 1999) is an early prototype of such a system used for gene function discovery. GATE (Cunningham et al., 2002), a framework that is still currently maintained, provides a unified architecture for all three aspects. Similarly, NLP Workbench tries to accommodate all three aspects in a single platform. Our platform uses containerized microservices instead of Java classes for each processing module, which avoids the restrictions of underlying programming frameworks for implementing NLP algorithms. Voyant Tools (Rockwell and Sinclair, 2016) is another web-based platform that provides tools for corpus analysis and the function to write user-defined scripts. Their built-in tools are mostly limited to count-based statistics and visualization, while we are integrating large

deep learning models. UIMA (Ferrucci and Lally, 2004) attempts to define a standard protocol for managing corpora and NLP algorithms. It is more developer-oriented, unlike our application which provides a complete system that can be used for analysis directly by users.

A plethora of NLP toolkits focusing on building NLP pipelines have been developed in the past few decades, including Stanford CoreNLP (Manning et al., 2014), OpenNLP⁷, NLTK (Bird et al., 2009), spaCy (Honnibal et al., 2020) and Transformers (Wolf et al., 2020). These toolkits could be used to address the text processing algorithm aspect of text mining systems, but they do not provide a full solution. The typical strategy to incorporate new models into these toolkits is to re-implement the model in the framework of the toolkit, while we try to re-use, as much as possible, the code and models released by researchers.

Some tools specialize in only the visualization aspect. To name a few, Blloshmi et al. (2021) and Cohen et al. (2021) built tools for visualizing the results of semantic parsers – a function that is also provided by our system.

Several libraries and tools are able to manage corpora or models from multiple sources. For example, Datasets library (Lhoest et al., 2021) provides an interface to access common NLP datasets, and DataLab (Xiao et al., 2022) is a platform to examine and analyze datasets. Transformers (Wolf et al., 2020) can access models and datasets from the Hugging Face Hub⁸. Beyond corpora and models, NLP Workbench also incorporates code from multiple sources.

3 Architecture

NLP Workbench is built on top of various open source software and incorporates the code and models from many research projects. We design our architecture to leverage off-the-shelf functionalities provided by these software and projects, and to minimize the effort of integrating new code from a research project.

3.1 Workflow

Figure 1 provides a high-level overview of how users interact with NLP Workbench and how the system handles the requests.

⁷<https://opennlp.apache.org>

⁸<https://huggingface.co/docs/hub/index>

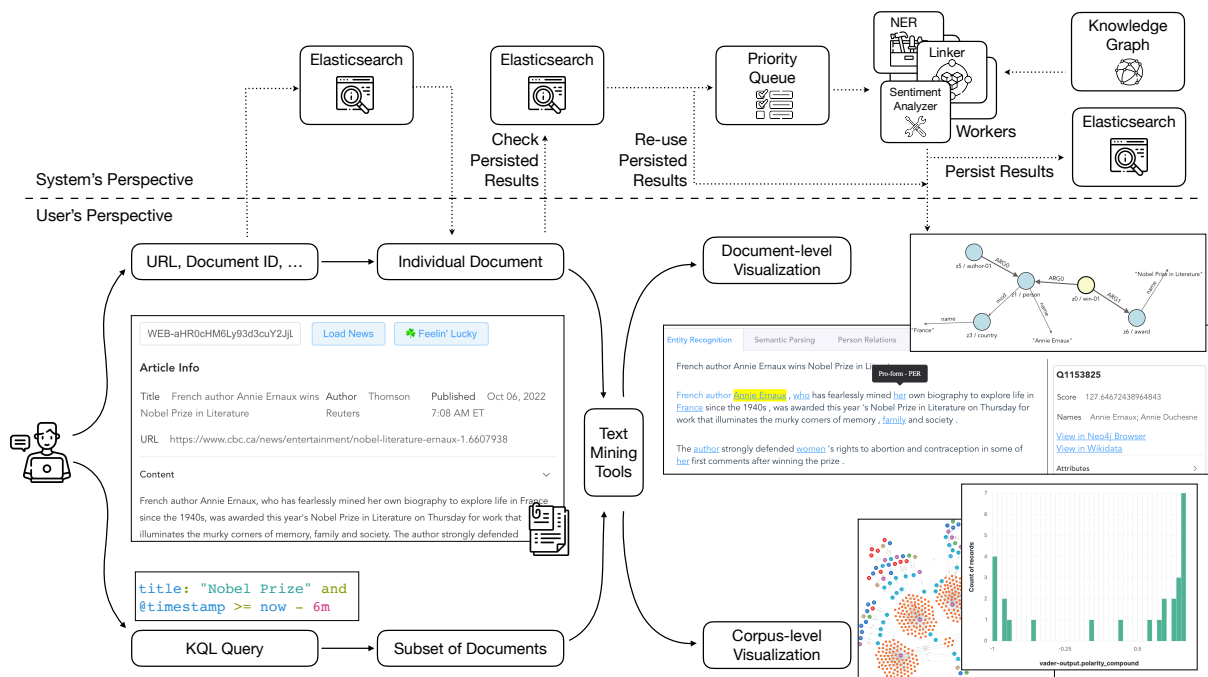


Figure 1: Workflow of NLP Workbench from the perspectives of the user and the system, as described in §3.1. For document-level visualization, we showcase the user interface for named entity recognition, coreference resolution, entity linking, and semantic parsing. For corpus-level visualization, this figure includes plots of a social network constructed from a tweet of the official Nobel Prize account, and the distribution of sentiment polarity scores of a sample of the built-in news corpus. Icons created by Freepik - Flaticon.

User Perspective From the perspective of a user, one could choose to apply text mining tools on a single document specified by its URL or ID, or apply them in batches on a set of documents specified by a query. Queries are written in the Kibana Query Language (KQL)⁹, which is a simple and intuitive text-based query language. The outputs of the tools on a single document can be visualized in the web interface, with each tool having a separate panel. Using Kibana Lens¹⁰, a user can visualize statistics calculated over the output of multiple documents, such as the distribution of sentiment polarity scores. The connected Neo4j Browser¹¹ provides an interactive web interface for exploring social networks constructed from a corpus.

System Perspective From the perspective of the system, both the corpus and the outputs of text mining tools are stored and indexed in Elasticsearch¹², a document indexing, search, and analytics engine. By storing and indexing tool outputs, we re-use previous results and avoid re-computation to im-

prove efficiency. In addition to that, Elasticsearch provides convenient tools to filter documents based on the outputs of text mining tools and visualizing statistics, which are very useful for downstream analytics.

If running a tool is indeed necessary, the task is added to a priority queue. Ad hoc and interactive requests, issued when a user is examining a single document and applying tools on it, are prioritized over batched requests that run in the background. Each tool or model has workers processing the tasks in the queue. This ensures that users performing interactive analysis experience little latency even when the number of workers is limited, which is usually the case in practice as deep learning models are often resource-intensive and it is infeasible to have multiple instances running in parallel.

3.2 Pipelining and Scheduling

Text mining tools often rely on the outputs of other tools or NLP models and are built as pipelines. For example, both entity linking and relation extraction require named entity recognition and coreference resolution. To ensure efficiency, re-computing the outputs of tools that are already available should be avoided, and tools should be run in parallel if pos-

⁹<https://www.elastic.co/guide/en/kibana/current/kuery-query.html>

¹⁰<https://www.elastic.co/kibana/kibana-lens>

¹¹<https://neo4j.com/developer/neo4j-browser/>

¹²<https://www.elastic.co/what-is/elasticsearch>

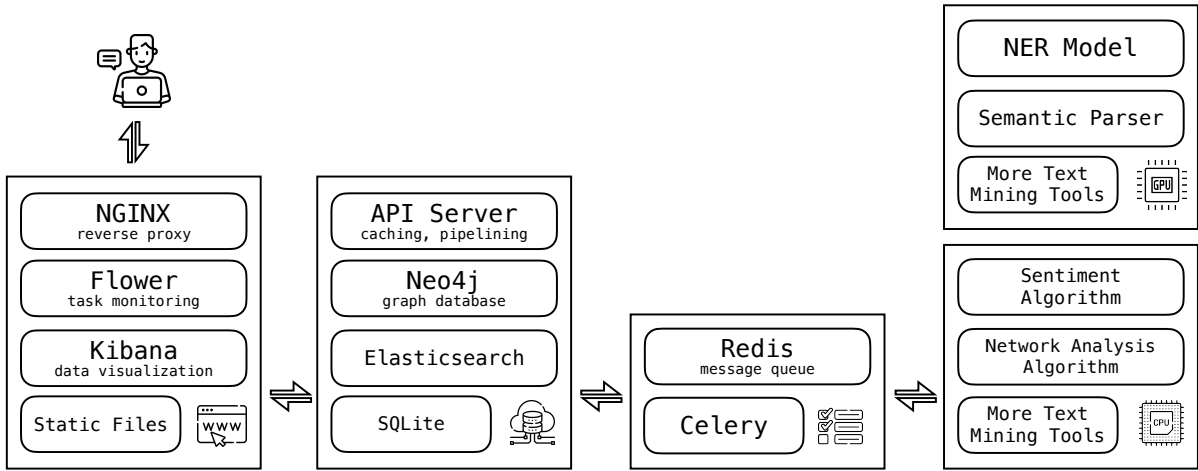


Figure 2: Microservice architecture of NLP Workbench. Each rectangle represents a physical machine, with its capability indicated by the icon at the bottom right corner. Each rounded rectangle represents a container, with the tool and function it provides indicated by the text inside. Container to physical machine allocation is for illustration purposes only and is adjusted to fit the need when the system is deployed in production.

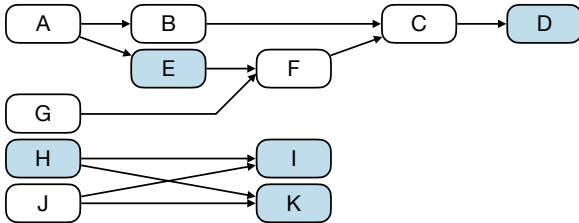


Figure 3: Example of a batched task with the directed acyclic graphs of dependencies. Shaded nodes represent tools that the user requests to run on the document, and unshaded nodes represent tools that are needed to provide the inputs to the shaded nodes.

sible. In addition to persisting and re-using outputs as discussed in Section 3.1, we design a pipelining and scheduling system that automatically detects the dependencies between tools and schedules the tasks in a way that eliminates re-computation and encourages parallelism.

The dependencies of a tool can naturally be represented as a directed acyclic graph (DAG), where inbound edges represent the dependencies. When multiple tools are requested to be run on a single document, we gather the direct and transitive dependencies of these tools in a single graph, as shown in Figure 3. The connected components of the graph are DAGs. Within each DAG, the tools are run in topological order; and disjoint DAGs are executed in parallel.

In the example illustrated in Figure 3, the user requests to run tool E, D, H, I, and K on a document. The scheduler will automatically find all

dependencies (A to K) and run two chains in parallel: A-B-E-G-F-C-D and H-J-I-K¹³.

3.3 Containerized Microservices

A major obstacle to integrating third party code is the *dependency hell* problem: it is an NP-complete problem to find a set of compatible versions of all software library dependencies (Burrows, 2005; Cox, 2016), and in reality a compatible set may not exist. This is especially true for deep learning models (Han et al., 2020; Huang et al., 2022), which often require specific versions of software libraries. In the case of popular deep learning frameworks, TensorFlow (Abadi et al., 2016) 2.0 introduces breaking API changes that are not back-compatible. Both TensorFlow and PyTorch (Paszke et al., 2019) are compiled with specific versions of CUDA (Nickolls et al., 2008) and cuDNN (Chetlur et al., 2014), and that makes different framework versions hard to coexist. Manually fixing the code to make it compatible with a specific version of a library is often tedious and error-prone (Han et al., 2020).

For deployment, a practical problem is that it is often difficult or costly to find a single physical machine that satisfies the computing and networking requirements of all the components: deep learning models require GPU for inference, database management systems consume large amounts of memory and disk space, and web servers need ac-

¹³There is more than one valid topological sorting for a DAG.

cess to the Internet. One solution to this problem is the ability to deploy components of NLP Workbench on multiple machines, which is achieved by our design.

We solve both problems discussed above at once by deploying both text mining tools and infrastructure components as *containerized microservices*. Each component is deployed as a Docker container (Merkel, 2014) that encapsulates the software and its dependencies. The containers communicate with each other via the RPC and message queue functions provided by Celery¹⁴. Figure 2 illustrates how the containerized microservices are deployed on separate machines with different capabilities. Such a microservice architecture allows us to overcome the problem of heterogeneous technologies and simplifies horizontal scaling when needed (Newman, 2015).

4 Components

NLP Workbench already includes a variety of tools and models for text mining. Most of the components come from state-of-the-art research in the respective subfields. Others are baseline implementations to demonstrate NLP Workbench’s extensibility, showing that developers can straightforwardly incorporate new tools and build pipelines from existing ones. One benefit of the flexible and modular design as described in §3 is that all built-in tools and models can easily be replaced or upgraded. Existing tools and models in NLP Workbench include:

Named Entity Recognition The task, known as NER, is to identify mentions to entities such as people, organizations, and locations. We incorporated the NER model from PURE (Zhong and Chen, 2021), which achieved good performance by simply fine-tuning BERT (Devlin et al., 2019).

Coreference Resolution To determine which entity a pronoun refers to, we adopted the heuristic algorithm by Cunningham et al. (2002) that is based on recency and type agreement.

Entity Linking Mentions to entities in the text are disambiguated and linked to Wikidata (Vrandečić and Krötzsch, 2014) entities. Candidate entities are generated by a fuzzy match on name. In addition to name similarity, the ranking of candidates utilizes the cosine similarity between the sentence

embeddings (Reimers and Gurevych, 2019) of the context and the descriptions of the candidate entity from Wikipedia and Wikidata.

Relation Extraction The user can extract structured facts in the form of knowledge triples like (*Annie Ernaux*, Country, *France*) from a text. The underlying model (Mesquita et al., 2019) combines syntax and semantic features as well as BERT embeddings to predict the relation between entities.

Semantic Parsing Semantic parsing provides a structured representation of the meaning of a sentence, allowing users to obtain information like who did what to whom, when, and where without caring about the form. NLP Workbench uses AMRBART (Bai et al., 2022), a sequence-to-sequence model based on BART (Lewis et al., 2020) and pre-trained on a large graph corpus, to parse sentences into AMR graphs (Banarescu et al., 2013).

Summarization We build an application on top of semantic parsing to create natural language summaries of events related to people in the document, partly to demonstrate the simplicity of building pipelines in NLP Workbench. For each sentence in the document, we prune its AMR graph to only contain the nodes and edges of pattern *subject-predicate-object*, where the subject or object is a person. The pruned AMR graphs are then converted to natural language using AMRBART.

Sentiment Analysis The sentiment of a document is predicted by VADER (Hutto and Gilbert, 2014), a fast and accurate rule-based algorithm optimized for social media posts. A sentiment polarity score is produced and can be used to classify the sentiment as positive, neutral, or negative.

Social Network Analysis For corpora consisting of social media posts, NLP Workbench is equipped with a tool that builds graphs of social network interactions from posts. Powered by the graph database Neo4j¹⁵, the tool can be used to visualize the network and perform analyses, such as running centrality algorithms like PageRank (Page et al., 1999) to identify influential users.

5 Use Cases

Text mining has been proven useful in a variety of domains, such as corporate finance, patent research, life sciences, and many others (Feldman

¹⁴<https://docs.celeryq.dev>

¹⁵<https://neo4j.com/>

and Sanger, 2006b). NLP Workbench, as a full-fledged text mining platform, has been or has the potential to be applied in many of these domains. Some of the use cases are described in this section.

Digital Humanities Accessible and reliable NLP tools are useful in digital humanities projects such as Linked Infrastructure for Networked Cultural Scholarship (LINCS)¹⁶. A major component of the LINCS project is to generate linked data from natural language cultural heritage texts. The key steps are NER, coreference resolution, entity linking, and relation extraction, all of which are made available in NLP Workbench. Through the simple user interface, digital humanists and their students can choose the available models that suit their data, and easily connect these steps into a custom workflow.

The Centre for Artificial Intelligence, Data, and Conflict (CAIDAC)¹⁷ houses human right scholars interested in how armed groups and extremists use social media to promote their agendas. NLP Workbench offers such researchers convenient tools for collecting tweets¹⁸ for analysis. Posts can be subjected to the NLP tools of interest, and the social network underlying the corpus can be visualized, explored, and analyzed. All of these tasks are done through an accessible interface, requiring no programming from the users, enabling them to perform more and larger studies in a fraction of the time otherwise required.

Business Analytics Business analysts ask questions like “*are recent news reports about Apple Inc. positive or negative?*”. These type of questions can easily be answered by NLP Workbench. After performing NER and entity linking on the news articles, the analyst can conduct a **semantic search** to find the articles that are related to *Apple Inc.* rather than apple the fruit. Then, the analyst can use the sentiment analysis tool and visualize the distribution of sentiment polarity scores with Kibana Lens, as shown in the bottom right screenshot in Figure 1.

NLP Research All NLP models in NLP Workbench can be accessed via RESTful API and RPC, or used directly as containers. For researchers who wish to perform inferences with the models on their own data, they could use the interfaces provided by NLP Workbench, without needing to set up the environment to run the models.

¹⁶<https://lincsproject.ca>

¹⁷<https://www.tracesofconflict.com/>

¹⁸Support for Telegram and other platforms is in progress.

6 Roadmap

NLP Workbench is still in its early stages of development, and we are actively working on improving the system. Besides usability, stability, and security updates, we plan to work on the following major features in the near future:

Human-in-the-loop NLP Adding annotation support to the web interface will allow users to provide feedback to the outputs of models. This will help researchers to collect domain-specific labelled data and improve the performance of the models in a human-in-the-loop fashion (Wang et al., 2021).

Improved Corpus Management Managing document collections is a crucial aspect of text mining (Hearst, 1999; Cunningham et al., 2002). Currently, corpora are manually imported into Elasticsearch or created by crawling social media. We hope to improve the way users access document collections. This can be done by connecting NLP Workbench to Datasets (Lhoest et al., 2021) and DataLab (Xiao et al., 2022) where popular text datasets are already available. In addition to crawling from social media, we also plan to support creating a new corpus by doing web search on a search engine.

More Text Mining Tools The extensible design of NLP Workbench allows us to keep existing tools and models up to date by replacing them when better models are released, and integrate emerging text mining tools to the system. For example, we hope to add claim extraction models to facilitate fact checking tasks (Hassan et al., 2017).

Multi-modal Analysis Social media posts often refer to or contain information in other modalities (images, video, audio) of interest. At the same time, there is growing interest in grounding NLP models and analysis on knowledge extracted from videos and other sources. While adding support for processing different media in NLP Workbench is as easy as adding more NLP tools, we are interested in integrating these models so that co-training or grounding can be automated to the extent possible.

7 Conclusion

We introduced NLP Workbench, a platform that caters to all three major aspects of text mining systems: corpus management, text mining tools, and user interface. We explained what design features make NLP Workbench efficient and extensible, and how it can be used in a variety of applications.

Limitations

We have already identified several important features that are not yet implemented in NLP Workbench, as discussed in §6: the platform needs an annotation feature for human-in-the-loop AI; it should have access to commonly used public corpora; and it should include text mining tools such as one for claim extraction. There are some intrinsic limitations that even the state-of-the-art models in NLP Workbench do not solve. For example, long tail entities may not be covered by the knowledge graph, and current entity linking models do not have a notion for out-of-knowledge-graph entities (Shen et al., 2023). This will result in long tail entities always being incorrectly linked. Beyond social network analysis, our current design does not have the user interface or models for other corpus-level analyses, such as topic modeling (Blei et al., 2003). And finally, all NLP models and algorithms in NLP Workbench are targeted at English text. Although we have been able to deal with corpora in other languages by translating them to English using Marian MT (Junczys-Dowmunt et al., 2018), it is not yet clear whether performance can be improved by directly using models trained on other languages.

Ethics Statement

By encapsulating the models, NLP Workbench lowers the entry barrier for non-experts to use state-of-the-art AI models. The microservice architecture, which allows models to be deployed on multiple servers with different capabilities rather than a single omnipotent server, also makes this text mining platform more accessible. Containerizing third-party models also helps with reproducibility and transparency. There have been attempts of using NLP Workbench to analyze datasets to help understand propaganda, misinformation, and disinformation related to war and terrorism. However, users must be warned that, as NLP Workbench uses third-party data and models without modification, outputs obtained from NLP Workbench are inevitably affected by the bias inherent in the datasets and models.

Acknowledgements

The work is supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Canadian Foundation for Innovation

(CFI), AI4Society¹⁹ and a gift from Scotiabank. Certain computing resources are provided by the Digital Research Alliance of Canada.

References

- Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. [Tensorflow: A system for large-scale machine learning](#). In *12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016, Savannah, GA, USA, November 2-4, 2016*, pages 265–283. USENIX Association.
- Xuefeng Bai, Yulong Chen, and Yue Zhang. 2022. [Graph pre-training for AMR parsing and generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6001–6015, Dublin, Ireland. Association for Computational Linguistics.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Dis-course*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O’Reilly Media Inc.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. [Latent dirichlet allocation](#). *J. Mach. Learn. Res.*, 3:993–1022.
- Rexhina Blloshmi, Michele Bevilacqua, Edoardo Fabiano, Valentina Caruso, and Roberto Navigli. 2021. [SPRING goes online: End-to-end AMR parsing and generation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 134–142, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Daniel Burrows. 2005. Modelling and resolving software dependencies.
- Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. 2014. [cuDNN: Efficient primitives for deep learning](#). *CoRR*, abs/1410.0759.
- Jaron Cohen, Roy Cohen, Edan Toledo, and Jan Buys. 2021. [RepGraph: Visualising and analysing meaning](#)

¹⁹<https://ai4society.ca/>

- representation graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–86, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Russ Cox. 2016. [Version SAT](https://research.swtch.com/version-sat). Retrieved November 29, 2022, from <https://research.swtch.com/version-sat>.
- Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. [GATE: an architecture for development of robust HLT applications](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 168–175, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Ronen Feldman and James Sanger. 2006a. [Introduction to text mining](#). In *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, page 1–18. Cambridge University Press.
- Ronen Feldman and James Sanger. 2006b. [Text mining applications](#). In *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*, page 273–314. Cambridge University Press.
- David Ferrucci and Adam Lally. 2004. [UIMA: An Architectural Approach to Unstructured Information Processing in the Corporate Research Environment](#). *Natural Language Engineering*, 10(3-4):327–348.
- Junxiao Han, Shuiguang Deng, David Lo, Chen Zhi, Jianwei Yin, and Xin Xia. 2020. [An empirical study of the dependency networks of deep learning libraries](#). In *2020 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pages 868–878.
- Naeemul Hassan, Gensheng Zhang, Fatma Arslan, Josue Caraballo, Damian Jimenez, Siddhant Gawsane, Shohedul Hasan, Minumol Joseph, Aaditya Kulkarini, Anil Kumar Nayak, Vikas Sable, Chengkai Li, and Mark Tremayne. 2017. [Claimbuster: The first-ever end-to-end fact-checking system](#). *Proc. VLDB Endow.*, 10(12):1945–1948.
- Marti A. Hearst. 1999. [Untangling text data mining](#). In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 3–10, College Park, Maryland, USA. Association for Computational Linguistics.
- Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. [spaCy: Industrial-strength natural language processing in Python](#).
- Kaifeng Huang, Bihuan Chen, Susheng Wu, Junmin Cao, Lei Ma, and Xin Peng. 2022. [Demystifying dependency bugs in deep learning stack](#). *CoRR*, abs/2207.10347.
- Clayton J. Hutto and Eric Gilbert. 2014. [VADER: A parsimonious rule-based model for sentiment analysis of social media text](#). In *Proceedings of the Eighth International Conference on Weblogs and Social Media, ICWSM 2014, Ann Arbor, Michigan, USA, June 1-4, 2014*. The AAAI Press.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in C++](#). In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. [Datasets: A community library for natural language processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. [The Stanford CoreNLP natural language processing toolkit](#). In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.
- Dirk Merkel. 2014. [Docker: Lightweight Linux containers for consistent development and deployment](#). *Linux J.*, 2014(239).

- Filipe Mesquita, Matteo Cannavicchio, Jordan Schmedek, Paramita Mirza, and Denilson Barbosa. 2019. [Knowl-edgenet: A benchmark dataset for knowledge base population](#). 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 2019, Hong Kong, China, November 3-7, 2019*, pages 749–758. Association for Computational Linguistics.
- Sam Newman. 2015. Microservices. In *Building Microservices*. O’Reilly Media, Inc.
- John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. 2008. [Scalable parallel programming with CUDA](#). *ACM Queue*, 6(2):40–53.
- Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. [The pagerank citation ranking: Bringing order to the web](#). Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.
- Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. [PyTorch: An imperative style, high-performance deep learning library](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Geoffrey Rockwell and Stéfan Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. The MIT Press.
- Wei Shen, Yuhan Li, Yinan Liu, Jiawei Han, Jianyong Wang, and Xiaojie Yuan. 2023. [Entity linking meets deep learning: Techniques and solutions](#). *IEEE Transactions on Knowledge and Data Engineering*, 35(3):2556–2578.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wiki-data: A free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Zijie J. Wang, Dongjin Choi, Shenyu Xu, and Diyi Yang. 2021. [Putting humans in the natural language processing loop: A survey](#). In *Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing*, pages 47–52, Online. Association for Computational Linguistics.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Yang Xiao, Jinlan Fu, Weizhe Yuan, Vijay Viswanathan, Zhoumianze Liu, Yixin Liu, Graham Neubig, and Pengfei Liu. 2022. [DataLab: A platform for data analysis and intervention](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 182–195, Dublin, Ireland. Association for Computational Linguistics.
- Zexuan Zhong and Danqi Chen. 2021. [A frustratingly easy approach for entity and relation extraction](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.