

# TRSRD: a database for research on risky substances in tea using natural language processing and knowledge graph-based techniques

Yongmei Wang<sup>†,\*</sup>, Peng Wang<sup>†</sup>, Yongheng Zhang<sup>†</sup>, Siyi Yao<sup>†</sup>,  
Zhipeng Xu and Youhua Zhang<sup>†</sup>

Anhui Provincial Engineering Laboratory for Beidou Precision Agriculture Information, School of Information and Computer, Anhui Agricultural University, 130 Changjiangxilu, Heifei, Anhui 230036, P.R.China

\*Corresponding author: Corresponding author: Tel: +86 13505510152; Email: [wym0152@foxmail.com](mailto:wym0152@foxmail.com)

Correspondence may also be addressed to Peng Wang. Email: [WPENGxss@gmail.com](mailto:WPENGxss@gmail.com)

<sup>†</sup>These authors contributed equally to this work.

Citation details: Wang, Y., Wang, P., Zhang, Y. *et al.* TRSRD: a database for research on risky substances in tea using natural language processing and knowledge graph-based techniques. *Database* (2023) Vol. 2023: article ID baad031; DOI: <https://doi.org/10.1093/database/baad031>

## Abstract

During the production and processing of tea, harmful substances are often introduced. However, they have never been systematically integrated, and it is impossible to understand the harmful substances that may be introduced during tea production and their related relationships when searching for papers. To address these issues, a database on tea risk substances and their research relationships was constructed. These data were correlated by knowledge mapping techniques, and a Neo4j graph database centered on tea risk substance research was constructed, containing 4189 nodes and 9400 correlations (e.g. research category-PMID, risk substance category-PMID, and risk substance-PMID). This is the first knowledge-based graph database that is specifically designed for integrating and analyzing risk substances in tea and related research, containing nine main types of tea risk substances (including a comprehensive discussion of inclusion pollutants, heavy metals, pesticides, environmental pollutants, mycotoxins, microorganisms, radioactive isotopes, plant growth regulators, and others) and six types of tea research papers (including reviews, safety evaluations/risk assessments, prevention and control measures, detection methods, residual/pollution situations, and data analysis/data measurement). It is an essential reference for exploring the causes of the formation of risk substances in tea and the safety standards of tea in the future.

Database URL: <http://trsrd.wpengxs.cn>

## Introduction

Tea is an aromatic beverage made from the processed leaves of the tea tree and is one of the most popular beverages in the world (1). Tea has unique flavors (2) and is rich in beneficial chemicals such as catechins, theaflavins, and 5-nethylglutamine (3). The polyphenols in tea can promote human health (4), and the tea has medicinal value (5). Some studies have shown that drinking tea every day has been shown to reduce weight (6, 7), prevent cancer (8), and regulate intestinal flora (9) and may also reduce the risk of cardiovascular disease (10). The food safety standards of the importing country are a crucial factor that can impact the tea trade flow (11, 12). As research into the food safety considerations of tea is conducted, more harmful substances that may be introduced during tea production are being identified. In addition to pesticide exposure in tea production (13, 14) and heavy metal risks (15), there are also residues of other chemicals (16, 17), such as 9,10-anthraquinone (18) and perchlorate (19). Nevertheless, a vast amount of information regarding

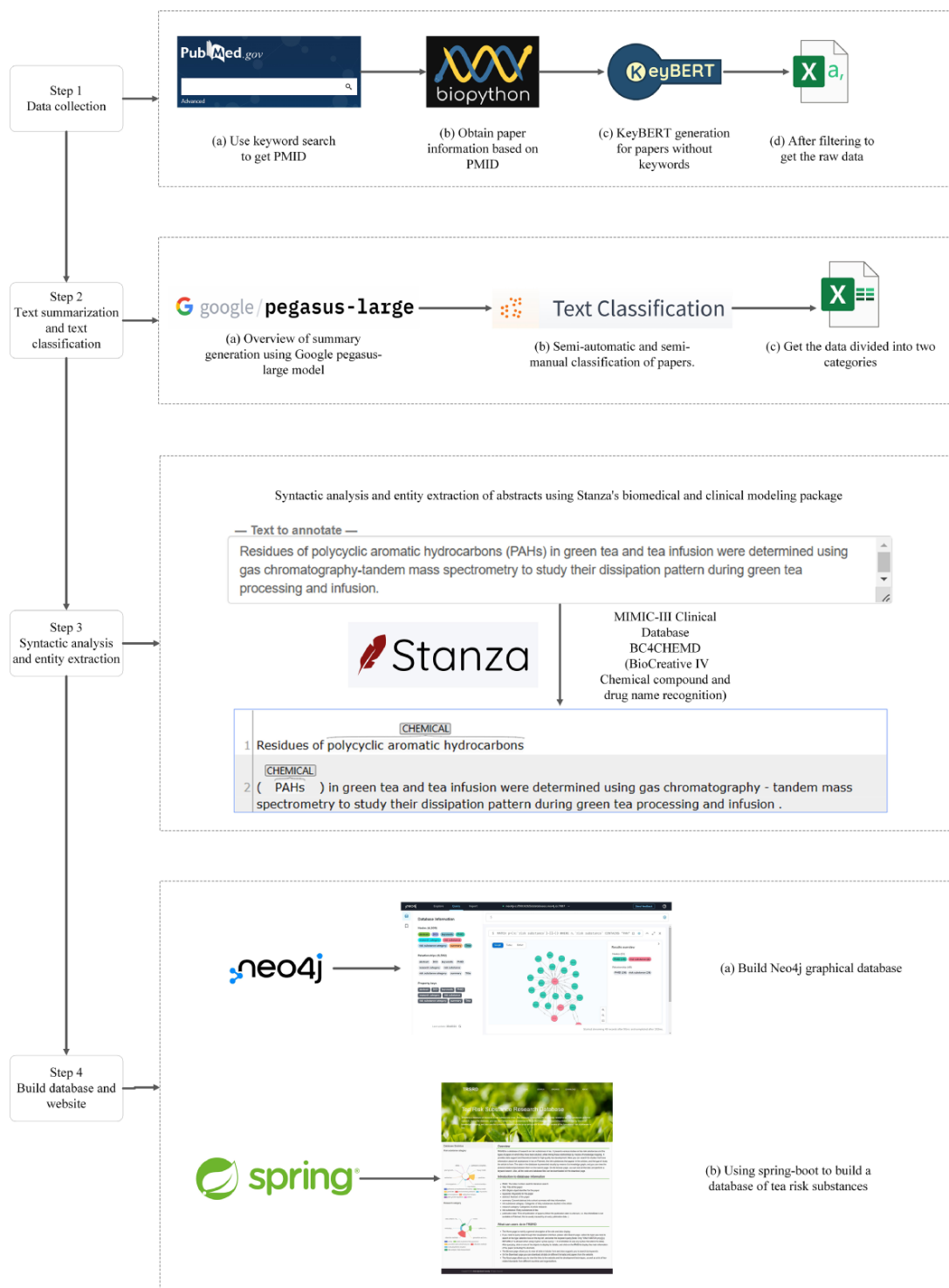
hazardous chemicals in tea is dispersed throughout various literature sources, which poses difficulties for researchers in searching for relevant papers and integrating knowledge into a systematic framework. Consequently, this situation impedes the progress of research. Therefore, it is essential to establish a complete database for research on tea risk substances.

In order to solve the above problems, we have constructed a database that catalogs tea risk substances and their corresponding research relationships. We used text mining technology (20) to collect papers that discuss harmful substances in tea. We used text abstracts and text classification to filter and classify papers. We classified all papers into nine categories of hazardous substances and six categories of research, to distinguish papers discussing various harmful substances and identify the specific research focus on such substances. This enables researchers to easily search and comprehend relevant information (21). Subsequently, the categorized data underwent syntactic analysis and chemical entity extraction to independently identify the chemical substances

discussed in each paper. These data were then linked through knowledge mapping techniques (22) correlation to create a Neo4j database containing 4189 nodes and 9400 correlations, which allows the researcher to better discover hidden relationships between studies (23).

Our main contributions to this study are as follows.

- Papers on harmful substances in tea were consolidated, summarized, and classified. The chemical substances discussed in these papers were independently identified through syntactic analysis and entity extraction. The papers have been comprehensively collected and compiled for the first time.



**Figure 1.** The Neo4j database construction process.

- A workflow has been developed from document mining to text classification to entity extraction, linking the scattered knowledge in different documents using knowledge graphs.
- This database provides a visual view of data and research data, which can be used to find relationships between studies on harmful substances in tea for reference and research citation.

## Steps and methods

This study classified the papers by extracting essential information from the literature abstracts. Each paper was labeled with its risk substance category (divided into nine categories) and its research category (divided into six categories). The text of the literature abstracts was syntactically analyzed according to the classification results, and chemical entities were separated from them, resulting in 955 chemical substances. The processed data were then used as PubMed Unique Identifier (PMID) to establish the corresponding relationships. This resulted in the construction of a Neo4j graph database of tea risk substances and their research relationships. The construction process is shown in [Figure 1](#).

## Data collection

First, all relevant literature PMIDs were obtained from PubMed using keywords such as tea, pesticide, and food safety. Subsequently, PMIDs were obtained from the PubMed database using Biopython (24) to extract the desired literature's titles, DOIs, keywords, and abstracts. At this stage, papers without keywords or abstracts were removed from the dataset. For the remaining papers without keywords but with an abstract, KeyBERT (25) was used to generate six keywords, resulting in 5177 usable results. Finally, the initial data were obtained by retrieving and filtering information from the titles, keywords, and abstracts and deleting papers irrelevant to tea. The valuable information was selected here to facilitate later overview and classification operations on this information without the effect of useless information (26).

**Table 1.** Risk substance categories and their classification criteria

Risk substance category	Classification criteria
Pollutants (comprehensive discussion)	This item is used when two or more different pollutants are included.
Heavy metals	This item is used when only heavy metals are contained.
Pesticides	Only contains pesticides or insecticides for this item.
Environmental pollutants	This item is used when only environmental pollutants are contained.
Mycotoxins	This item is used when only mycotoxins are contained.
Microorganisms	This item is used when only microorganisms are contained.
Radioactive isotopes	This item is used when only radioactive isotopes are contained.
Plant growth regulators	This item is used when only plant growth regulators are contained.
Others	This is the item if it is not one of the above categories.

## Text summarization and classification

After obtaining the initial filtered data, the long abstracts were compressed by Google's pegasus-large model (27) for paragraph comprehension, information compression (28), and language generation (29) to produce an abstract overview of only a few sentences, the purpose of this step is to compress the long abstracts into short, fluent, readable texts that retain the most salient information (30). The title, keywords, abstract, and summary are utilized collectively to classify each paper into two categories: the risk substance category (comprising of inclusion pollutants (comprehensive discussion), heavy metals, pesticides, environmental pollutants, mycotoxins, microorganisms, radioactive isotopes, plant growth regulators, and others) and the paper's research category (including review, safety evaluation/risk assessment, prevention and control measures, detection methods, residual/pollution situation, and data analysis/data measurement). If the keywords and abstracts contain the content of a risky substance category and are of a single type, they will be assigned directly to that category, otherwise, they will be classified as pollutants (comprehensive discussion). As there is currently no classification model for tea research, the remaining papers were manually classified, and those that were not manually classified were checked twice at a later stage to ensure correctness. [Table 1](#) shows the risk substance categories and their classification criteria. [Table 2](#) presents the classification criteria for the paper study categories.

## Syntactic analysis and entity extraction

After classifying the paper, in order to find out the chemical substances, we need to first analyze the syntactic structure of the sentences (31) in the abstract, and then, meaningful information is identified from the split syntactic structure by named entity recognition (32). With the research and development of natural language processing, natural language processing models trained on specific datasets are also often used in the biomedical field. Biomedical natural language processing is often used for word sense disambiguation, named entity recognition, information extraction, and relation extraction (33).

To ensure the accuracy of named entity recognition, we chose Stanza (34), a neural natural language processing

**Table 2.** Research categories of dissertations and their classification criteria

Research category	Classification criteria
Review	Summary overview of hazardous substances
Safety evaluation/risk assessment	Assess the risk of the hazardous substance or study its harm to humans
Prevention and control measures	Reduce the residue of the hazardous substance or avoid the use of the hazardous substance
Detection methods	Innovative detection methods for this hazardous substance
Residual/pollution situation	Study the residual pattern of the pollutant or investigate the pollution of a certain area by the pollutant
Data analysis/data measurement	Large-scale data collection and analysis of pollutants at a site

package customized for biomedical text processing, using the biomedical and clinical syntactic analysis and named entity recognition models provided in Stanza. For the abstracts after classification, chosen a syntactic analysis pipeline trained on the MIMIC clinical dataset (35) and a named entity recognition model pre-trained on the BC4CHEMD corpus (36). After separating the chemical entities, these substances are grouped into corresponding papers to build a database.

### Build database and website

For database selection, we chose the Neo4j graph database in order to store the connections and relationships between data in more flexible numbers with data elements. In Neo4j, the data are stored like a whiteboard, which makes Neo4j flexible compared to other graph databases. While doing the

above advantages, the Neo4j graph database also has better performance (37).

Finally, these data are aggregated, building nodes and relationships using the Neo4j graph database to create a Tea Risk Substance Research Database (TRSRD). TRSRD allows researchers to easily access and analyze the data we have collected and processed.

## Results

### Knowledge mapping for tea risk substance research

With 4189 nodes and 9400 associations, TRSRD divides all the literature into nine risk substance categories and six paper research categories and also extracts 955 different tea risk substances from the classified papers to build a knowledge graph, so that researchers can quickly visualize the different hazardous substances in tea without having to do a lot of searching and summarizing in the field. This allows researchers to quickly visualize the different studies on harmful substances in tea without having to do extensive searches in the field, which helps in the referencing and citation of research. A complete visualization of the data in the database is shown in Figure 2.

### Neo4j database management website

The developers mainly use this site to maintain and check the data, and when the data need to be updated, new data can be redeployed from this site. It is also possible to view the visualization of all the data in the database as shown in Figure 3.

### Website (TRSRD) page

The TRSRD website provides a friendly and intuitive interactive interface that allows users to browse the site's introduction, search the Neo4j database (returned as visible results), view data (in tabular form), and download all data. We built



Figure 2. Overview of all node relationships in the Neo4j database.

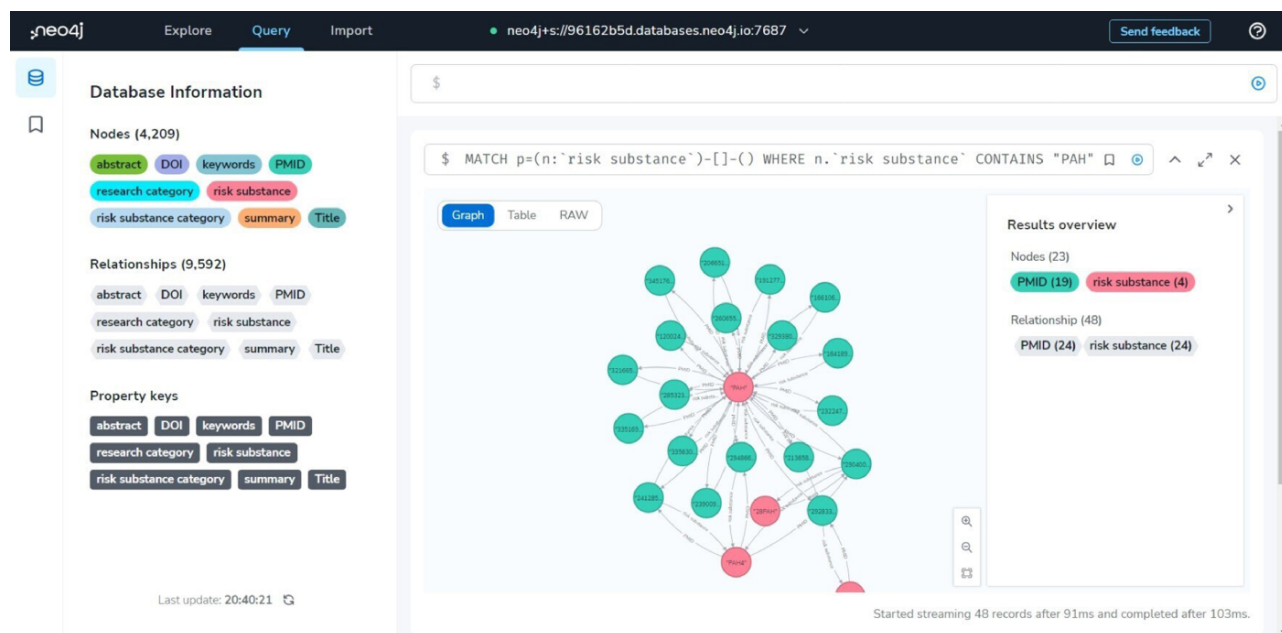


Figure 3. Neo4j database management site showcase.





The search page on the website enables users to input specific keywords and obtain visible search results, facilitating intuitive data retrieval and query. Clicking on a node allows users to view the <id> of the node and the corresponding value (Figure 4B). The browse page displays all the data in a table format and searches for the specified keywords (Figure 4C). On the browse page, you can view all temporarily included paper data (sorted by publication date), extracted risk substance data and the Maximum Residue Levels standards specified by different countries or organizations.

## Download and About page

The download page lets users download the data format they need in Neo4j database import format, CSV table format, and JavaScript Object Notation format (Figure 4D). On the About page, users can view links to research and data related to the site and links to technical web development, which can be accessed by clicking on the corresponding website. In addition, in the contact section below, users can get our email (Figure 4E).

## Conclusion

As tea becomes increasingly popular in various countries, more and more people are becoming aware of it. At the same time, research into the safety of tea has gradually increased. However, papers on tea quality and safety have never been systematically integrated. As a result, researchers cannot quickly and easily search for harmful substances in their desired field. In this paper, we have constructed a database centered on tea risk substance research by filtering and classifying existing tea research data through natural language processing and performing named entity extraction. Researchers can use TRSRD to better understand the risk substances in tea and the corresponding research, which is an essential reference for exploring the formation of risk substances in tea and future safety standards for tea.

As tea research continues to develop, we will continue to understand the harmful substances in tea and improve its safety standards. Our current work may still have shortcomings, such as the classification of chemical substances is not precise enough, not for the whole industry chain to make a more detailed way of classification. In addition, the model used on named entity identification does not achieve perfect identification and still requires manual screening of data. In the future, we will continue to monitor the latest research results in this field, incorporate more hazardous substances into the TRSRD, enrich the classification in the TRSRD, improve the classification criteria, and make the TRSRD a reference and citation platform among tea researchers.

## Data availability

All required data are contained in the database website and are available for download by all users. The database website address is <http://trsrd.wpengxs.cn>.

## Funding

Research Projects of Anhui Higher Education Institutions (Natural Science, 2022AH040122).

## Conflict of interest statement

None declared.

## References

- Khan,N. and Mukhtar,H. (2013) Tea and health: studies in humans. *Curr. Pharm. Des.*, **19**, 6141–6147.
- Zhai,X., Zhang,L., Granvogl,M. *et al.* (2022) Flavor of tea (*Camellia sinensis*): a review on odorants and analytical techniques. *Compr. Rev. Food Sci. Food Saf.*, **21**, 3867–3909.
- Graham,H.N. (1992) Green tea composition, consumption, and polyphenol chemistry. *Prev. Med.*, **21**, 334–350.
- Khan,N. and Mukhtar,H. (2019) Tea polyphenols in promotion of human health. *Nutrients*, **11**, 39.
- Saeed,M., Naveed,M., Arif,M. *et al.* (2017) Green tea (*Camellia sinensis*) and l-theanine: medicinal values and beneficial applications in humans—a comprehensive review. *Biomed. Pharmacother.*, **95**, 1260–1275.
- Yang,C.S., Zhang,J., Zhang,L. *et al.* (2016) Mechanisms of body weight reduction and metabolic syndrome alleviation by tea. *Mol. Nutr. Food Res.*, **60**, 160–174.
- Suzuki,T., Pervin,M., Goto,S. *et al.* (2016) Beneficial effects of tea and the green tea catechin epigallocatechin-3-gallate on obesity. *Molecules*, **21**, 1305.
- Yang,C.S., Wang,H., Li,G.X. *et al.* (2011) Cancer prevention by tea: evidence from laboratory studies. *Pharmacol. Res.*, **64**, 113–122.
- Zhang,Z., Zhang,Y., Li,J. *et al.* (2021) The neuroprotective effect of tea polyphenols on the regulation of intestinal flora. *Molecules*, **26**, 3692.
- Chung,M., Zhao,N., Wang,D. *et al.* (2020) Dose–response relation between tea consumption and risk of cardiovascular disease and all-cause mortality: a systematic review and meta-analysis of population-based studies. *Adv. Nutr.*, **11**, 790–814.
- Wei,G., Huang,J. and Yang,J. (2012) The impacts of food safety standards on China's tea exports. *China Econ. Rev.*, **23**, 253–264.
- Chen,J. and Liu,X.R. (2016) *Analysis of Tea Pesticide Residue Standards and Testing Methods*. Atlantis Press, Amsterdam, pp. 876–879.
- Gurusubramanian,G., Rahman,A., Sarmah,M. *et al.* (2008) Pesticide usage pattern in tea ecosystem, their retrospects and alternative measures. *J. Environ. Biol.*, **29**, 813–826.
- Lu,E.H., Huang,S.Z., Yu,T.H. *et al.* (2020) Systematic probabilistic risk assessment of pesticide residues in tea leaves. *Chemosphere*, **247**, 125692.
- Zhang,J., Yang,R., Chen,R. *et al.* (2018) Accumulation of heavy metals in tea leaves and potential health risk assessment: a case study from Puan County, Guizhou Province, China. *Int. J. Environ. Res. Public Health*, **15**, 133.
- Abd El-Aty,A.M., Choi,J.H., Rahman,M. *et al.* (2014) Residues and contaminants in tea and tea infusions: a review. *Food Addit. Contam. A*, **31**, 1794–1804.
- Cladière,M., Delaporte,G., Le Roux,E. *et al.* (2018) Multi-class analysis for simultaneous determination of pesticides, mycotoxins, process-induced toxicants and packaging contaminants in tea. *Food Chem.*, **242**, 113–121.
- Wang,X., Zhou,L., Luo,F. *et al.* (2018) 9,10-Anthraquinone deposit in tea plantation might be one of the reasons for contamination in tea. *Food Chem.*, **244**, 254–259.
- Liao,Z., Cao,D. and Gao,Z. (2022) Monitoring and risk assessment of perchlorate in tea samples produced in China. *Food Res. Int.*, **157**, 111435.
- Cohen,K.B. and Hunter,L. (2008) Getting started in text mining. *PLoS Comput. Biol.*, **4**, e20.
- Minaee,S., Kalchbrenner,N., Cambria,E. *et al.* (2021) Deep learning-based text classification: a comprehensive review. *ACM Comput. Surv.*, **54**, 62:1–62. 40.
- Auer,S., Kovtun,V., Prinz,M. *et al.* (2018) Towards a knowledge graph for science. In: *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*. WIMS '18, Association for Computing Machinery, New York, pp. 1–6.
- Xu,J., Kim,S., Song,M. *et al.* (2020) Building a PubMed knowledge graph. *Sci Data*, **7**, 205.
- Cock,P.J.A., Antao,T., Chang,J.T. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
- Grootendorst,M., Mishra,A., Matsak,A. *et al.* MaartenGr/KeyBERT: v0.7.0. <https://maartengr.github.io/KeyBERT>.

26. Ananiadou,S. and McNaught,J. (2006) *Text Mining for Biology and Biomedicine*. Artech House, Inc. <https://research.manchester.ac.uk/en/publications/text-mining-for-biology-and-biomedicine> 30 December 2022, accessed date last.
27. Zhang,J., Zhao,Y., Saleh,M. *et al.* (2020) PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In: *Proceedings of the 37th International Conference on Machine Learning* . PMLR. pp. 11328–11339.
28. Wolff,J.G. (1993) Computing, cognition and information compression. *AI Commun.*, **6**, 107–127.
29. Gatt,A. and Krahmer,E. (2018) Survey of the state of the srt in natural language generation: core tasks, applications and evaluation. *J. Artif. Intell. Res.*, **61**, 65–170.
30. Fabbri,A.R., Kryściński,W., McCann,B. *et al.* (2021) SummEval: re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguist.*, **9**, 391–409.
31. Sager,N. (1967) Syntactic analysis of natural language. In: Alt FL, Rubinoff M (eds) *Advances in Computers*. Vol. 8. Amsterdam: Elsevier, Amsterdam, pp. 153–188.
32. Goyal,A., Gupta,V. and Kumar,M. (2018) Recent named entity recognition and classification techniques: a systematic review. *Comput. Sci. Rev.*, **29**, 21–43.
33. Houssein,E.H., Mohamed,R.E. and Ali,A.A. (2021) Machine learning techniques for biomedical natural language processing: a comprehensive review. *IEEE Access*, **9**, 140628–140653.
34. Zhang,Y., Zhang,Y., Qi,P. *et al.* (2021) Biomedical and clinical English model packages for the Stanza Python NLP library. *J. Am. Med. Inform. Assoc.*, **28**, 1892–1899.
35. Johnson,A.E.W., Pollard,T.J., Shen,L. *et al.* (2016) MIMIC-III, a freely accessible critical care database. *Sci Data*, **3**, 160035.
36. Krallinger,M., Rabal,O., Leitner,F. *et al.* (2015) The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J. Cheminform*, **7**, S2.
37. Huang,H. and Dong,Z. Research on architecture and query performance based on distributed graph database Neo4j. In: *2013 3rd International Conference on Consumer Electronics, Communications and Networks*, Xianning, China. pp. 533–536.