Address
Artificial Researcher IT GmbH
i²c Innovation Incubator Center
Floragasse 7, 1040, Wien
Austria

Contact
Linda Andersson, CTO/CEO
+43 699 1782 1600
linda.andersson@artificialresearcher.com
www.artificialresearcher.com

Date
30/06/2020

Artificial Researcher's Passage Retrieval Service

A showcase demo on passage retrieval services
passageretrieval.artificialresearcher.com/

The collections available for search in this demo are the COVID-19 Open Research Dataset (Covid-19) data set, a sample set from the EP Full-Text Data for Text Analytics (Patent), and a sample of different scientific publication from the technical and medical sciences provided by CORE.uk (Science).

For access to our larger collections, provided through APIs, please contact us at demo@artificialresearcher.com

ABOUT ARTIFICIAL RESEARCHER

We are an information technology company and a start-up in the text mining industry. The company was founded by TU Wien alumni and entrepreneur Linda Andersson together with her sisters Jenny Andersson and Nina Andersson and fellow researcher Dr. Florina Piroi. In 2018 we launched the product idea 'Artificial Researcher in Science' which received the Commercial Viability Award from the Austrian Angel Investors Association. This is also the time Dr. Wolfgang Sachsenhofer joined the Artificial Researcher team by supporting the co-founders with his expertise in corporate venturing and technology commercialization.

Artificial Researcher provides industry and academia with unified platforms that increase the productivity of end-users, by exploring and implementing innovative text mining technologies based on over 15 years text mining know-how.

Artificial Researcher has received funding for two research prototype projects 'Artificial Researcher in Science', supported by Vienna Business Agency, and 'Artificial Researcher in Open Access', supported by Austria Wirtschaft Service. Within these two projects, we explore and create functional prototypes to develop novel and innovative scientific knowledge management systems, including patent data in 'Artificial Researcher in Open Access', tailored to the needs of information search professionals, researchers and students.

Artificial Researcher

USAGE

Using this demo a user has the unique opportunity to search in different scientific domains (under the heading COLLECTION in the demo), using a text segment as input, which can be combined with bibliographic meta-data filters. The user can paste a text segment, for example a sentence up to a paragraph, on a topic of interest. To this text, the user can add meta-data filters on YEAR, AUTHOR, AFFILIATION, KEYWORD and TAXONOMY.

The AUTHOR (or patent inventor) and AFFILIATION (or patent assignee) fields gives a user the possibility to write fragments of a name or the organization/institute. This data is analysed by our system using a special n-gram filter.

The TAXONOMY field can be applied to the COLLECTION field: for the patent data we provide IPC and CPC taxonomies – or subject classifications – to the sub-class level; for the medical collection, Covid-19, we support the MeSH subject classification; and for the Science text collections (CORE.uk) we provide a baseline subject classification by an automatic text classifier that assigns IPC sub classes to the scientific publication.

In the KEYWORD field the user can apply two types of filtering to her search: for the Patent and Science data, the user can use the WIPOcatchWord Index[1] as a filtering mechanism; for Covid-19 the user has the opportunity to filter the search using PICO [2]elements by writing PICO "patient with heart conditions" as well as an extended set of medical keywords extracted from MeSH and other domain-specific text resources.

Once the search request is submitted, the system retrieves the relevant paragraphs of text (i.e. passages), and displays automatically assigned keywords, bibliographic data together with a link to the full-text document where the paragraph occurs. The user can select the paragraphs of special interest and download them for further processing and analysis. Due to copyright reasons, not all paragraphs retrieved can be displayed. In these cases, we provide a link to the document publisher's websites.

COLLECTIONS

We give a short description of the collections that are available to the users of our demo.

COVID-19 Open Research Dataset (Covid-19)

This dataset is provided as a resource to encourage text-mining researchers to generate new insights and support the fight against this infectious disease (Wang et al 2020). The corpus is updated on regular basis with new publications from peer-reviewed journals and archival services like bioRxiv, medRxiv[3].

[1] https://www.wipo.int/meetings/en/doc_details.jsp?doc_id=892

[2] PICO is an acronym for annotating medical abstracts with the following classes: Population, Intervention, Comparison, and Outcome.

[3] https://www.semanticscholar.org/cord19/download

EP Full-Text Data for Text Analytics (Patent)

The European Patent Office has made a Full-Text collection available especially for data analysis and text mining. This data set contains basic bibliographic information and includes full-text for each European patent publication[4].

Scientific Open Access Publications (Science)

The CORE.uk data set are composed of a sample set of technical and medical scientific publication published in English. The paragraph segmentation provided by CORE.uk[5] is still experimental which causes the quality of the search results (retrieved paragraphs) to vary significantly.

TECHNOLOGY

The technology used in our demo is based upon Linda Andersson PhD research at TU Wien. Concretely, this demo deployment uses the advanced baseline in Linda Andersson's research result[6] during 2016 and 2017.

The PICO functionality uses different training models, one re-implementing the work presented in (Beltagy et al 2019), and another based upon the work and data set presented in (Zlabinger et al 2018), which have been extend in on going master thesis project[7](Schweikert 2020).

DATA QUALITY

The data are automatically processed from PDF format to JSON and XML format, the quality varies especially for paragraphs containing chemical compound descriptions, formulas, and tables. We appreciate you taking the time to send us feedback that help us improve our automatic extraction process. Should you find any inconsistencies let us know by copying the file id and sending it to demo@artificialresearcher.com.

[4] https://www.epo.org/searching-for-patents/data/bulk-data-sets/text-analytics.html
[5] https://core.ac.uk/
[6] http://www.ifs.tuwien.ac.at/~andersson/
[7] https://picoweb.ifs.tuwien.ac.at/pico-prediction-service/pubmed/?pubmed_id=16855426

Artificial Researcher

ACKNOWLEDGMENT

The 'Artificial Researcher's Passage Retrieval Service' is a part of the 'Artificial Researcher in Open Access' project that has been developed with the support of the Austria Wirtschaft Service.

REFERENCES

Andersson L, Rekabsaz N., Hanbury A. (2017) Automatic query expansion for patent passage retrieval using paradigmatic and syntagmatic information The first WiNLP Workshop co-located with with the Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver

Andersson L, Lupu M, Palotti J, Hanbury A, Rauber A. (2016) When is the time Ripe for Natural Language Processing for Patent Passage Retrieval? In Proceedings of the 25th ACM International Conference on Conference on Information and Knowledge Management (CIKM 2016)

Zlabinger M., Andersson L., Hanbury A., Andersson M., Quasnik V., Brassey J. (2018) Medical Entity Corpus with PICO Elements and Sentiment Analysis Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan

Beltagy Iz, Kyle Lo, and Arman Cohan. (2019) "SciBERT: A pretrained language model for scientific text." arXiv preprint arXiv:1903.10676. https://www.aclweb.org/anthology/D19-1371.pdf

Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Eide, D., Funk, K., Kinney, R.M., Liu, Z., Merrill, W., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O., & Kohlmeier, S. (2020). CORD-19: The Covid-19 Open Research Dataset. ArXiv.

Florian Schweikert (2020). "An eLearning tool to identify population, intervention, comparison and outcome (sentiment) for medical students." Master Thesis, to be submitted September 2020.