EPFL

Blue Brain Project

Blue Brain toolbox for searching and mining scientific literature

Blue Brain Search is a Python library for performing semantic searches, mining, and structured information extraction from large databases of scientific literature.

Every year, hundreds of thousands of scientific papers are published around the world to add to the vast quantity already available. An essential element of scientific research is access to data and information. But, no human can possibly read all these papers. Blue Brain Search was built to solve this problem and works across any scientific database.

An example of this, is in 2020, when the Blue Brain Project answered the call to action for the world's artificial intelligence (AI) experts to develop text and data mining tools that can help the medical community develop answers to high priority scientific questions in particular in relation to COVID-19.

In response to the COVID-19 pandemic, the COVID-19 Open Research Dataset (CORD-19) of over 500,000 scholarly articles was made open access, including over 200,000 with full text papers related to COVID-19, SARS-CoV-2, and other coronaviruses. The CORD-19 dataset is the most extensive coronavirus literature collection available for data mining to date and the coalition behind it has challenged AI experts to apply their skills in natural language processing and other machine learning techniques in order to generate new insights that may help in the ongoing fight against COVID-19.*

Accordingly, Blue Brain built an open source two-component framework for Knowledge Graph guided literature review using its Machine Learning and Data and Knowledge Engineering expertise. Blue Brain Search mines and extracts structured information from text sources. Blue Graph then generates a Knowledge Graph from text concepts extracted and performs graph analytics. Blue Brain Search and Blue Graph enabled Blue Brain scientists to read, analyze and synthesize the knowledge contained in over 240'000 open access scientific papers available in the CORD-19 dataset v47.

The Knowledge Graph guided review of the COVID-19-related literature performed using Blue Brain Search and Blue Graph enabled the BBP scientists to reveal how glucose helps the SARS-CoV-2 virus. The resulting study - A machine-generated view of the role of Blood Glucose Levels in the severity of COVID-19 was published in Frontiers in Public Health.

* - https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge

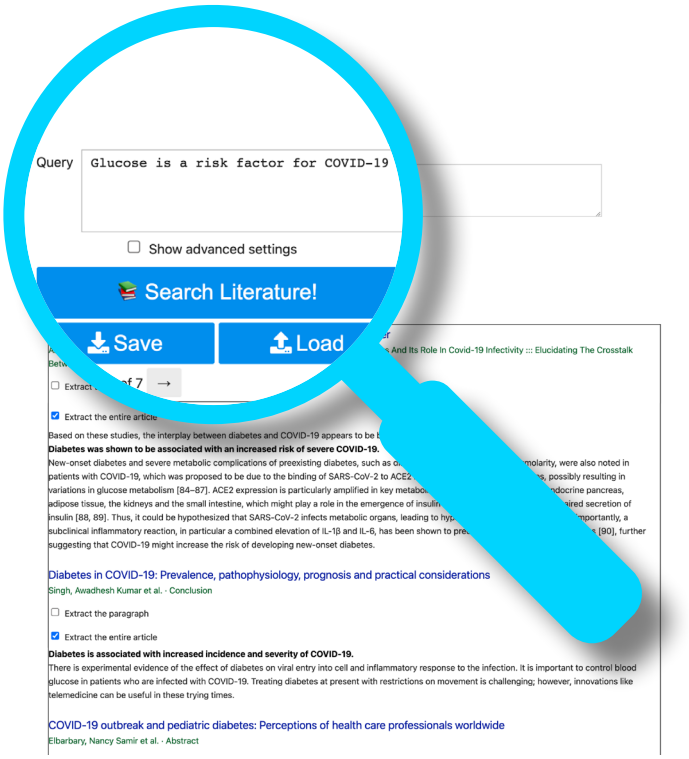
Using **Blue Brain Search**

Blue Brain Search facilitates the set-up of a server for search and mining. These functionalities are then displayed through a user interface consisting of two widgets that users can employ in their Jupyter Notebooks – Search and Mining.

First, the **Search Widget** allows the user to perform various kinds of searches of the literature dependent on their needs. These include metadata filtering or exact term matching, and also those based on semantic search.

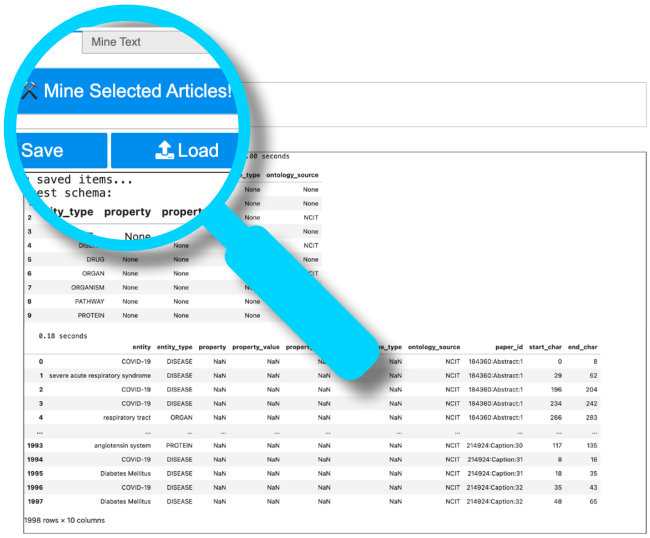
A semantic search is a data searching technique where a search query is not just finding keywords but also determining the intent and contextual meaning of the words used in the search. This then produces more meaningful search results by evaluating and understanding the search phrase and finding the most relevant results in a website, database or any other data repository.

A Machine Learning (ML) model for 'sentence embedding' facilitates the semantic search. The ML model computes numerical features representing the semantics of a sentence in a vector of numbers and these can then be compared with other vectors to find sentences with similar content



'The first step in Blue Brain's framework for Knowledge Graph guided literature review employs machine learning tools for performing semantic search of relevant papers given a user-specified search phrase (or query), for example, 'Glucose as a risk factor in COVID-19'

Francesco Casalegno, Machine Learning Section Manager, Blue Brain Project



Second, the **Mining Widget** allows the user to extract structured information from the literature related to their query of interest found using the Search Widget. This information is collected in a table with rows representing entries and columns representing data type or relations or metadata.

The Mining Widget therefore takes as inputs

1. The output of the Search Widget
2. A 'schema' that details which information the user wants to extract (i.e. I want to extract all the information possible about 'organs', but I don't need information about 'cells').

It then generates a table with the mined structured information.

Currently the Mining Widget supports the extraction of entities using Machine Learning models for 'named entity recognition', i.e. a model that detects and classifies spans in the text that refer to entity types of interest (e.g. 'cell', 'organ', 'chemical').

The Mining Widget also produces metadata, which explains where the information was extracted from.

"Techniques in research in biology are constantly progressing, becoming more and more advanced and powerful, and consequently generating more and more data in record time. As humans, not only can we not keep up to date with all the information generated every day, but we also can't simply read them, process them or obtain outputs without the computational performance of machines.

When the CORD-19 database was released, the Search tool extracted and ranked more than 400'000 single entities, and then easily identified the term "glucose" as a frequently mentioned entity among other very generic terms (virus, coronavirus, lung, blood, pulmonary disease ...); a task that a human couldn't have done. Even more sophisticated, the search tool selected the most relevant papers to a specific user query 'glucose as a risk factor for COVID-19', that allowed us to discover the findings that are presented in the paper - A Machine-Generated View of the Role of Blood Glucose Levels in the Severity of COVID-19, illustrating the immense value of Blue Brain Search".

Dr. Emmanuelle Logette, Molecular Biologist, Blue Brain Project
Lead Author of *'A Machine-Generated View of the Role of Blood Glucose Levels in the Severity of COVID-19'.*

Entity Linking

Entry Id	Type
Diabetes mellitus (NCIT_C298	DISEASE
Glucose (NCIT_C2831)	CHEMICAL
Leukocyte (NCIT_C12529)	CELL TYPE
COVID-19 infection (NCIT_C171133)	DISEASE

Build into a Knowledge Graph with Blue Graph

About EPFL's **Blue Brain Project**

The aim of the EPFL Blue Brain Project, a Swiss brain research initiative founded and directed by Professor Henry Markram, is to establish simulation neuroscience as a complementary approach alongside experimental, theoretical and clinical neuroscience to understanding the brain, by building the world's first biologically detailed digital reconstructions and simulations of the mouse brain.

Blue Brain Search is open source and you can find the code on GitHub: https://github.com/BlueBrain/Search

License - LGPL-3.0 License.

Installing this Python package is very easy as we have released all versions on PyPI: https://pypi.org/project/bluesearch/.

Simply run `pip install bluesearch` in your terminal.

Software Authors

Francesco Casalegno, Emilie Delattre, Pierre-Alexandre Fonta, Jan Krepl, Stanislav Schmidt and Anil Tuncel

Funding & Acknowledgements

This project was supported by funding to the Blue Brain Project, a research center of the Ecole polytechnique fédérale de Lausanne, from the Swiss government's ETH Board of the Swiss Federal Institutes of Technology.

Feedback

Sharing the tools we are using at the Blue Brain is part of our open science policy and we welcome contributions and ideas from external users.

Contact details

Francesco Casalegno
Machine Learning Section Manager
Blue Brain Project
francesco.casalegno@epfl.ch