Corpus creation: Open Access repository retrieval and analysis

Presented by: Dr. Renu Kumari

Project Scientist, NIPGR

Program manager, #semanticClimate

October 24, 2025

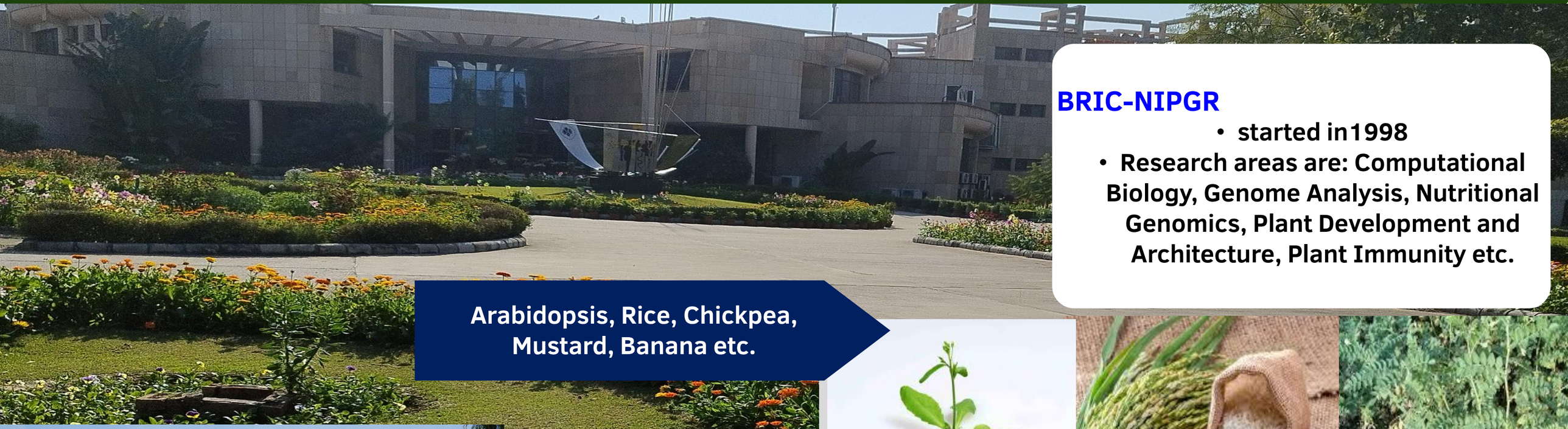
Open Access Week 2025

#semanticClimate
Transforming information into actionable knowledge

Learning by Doing!

Current Team

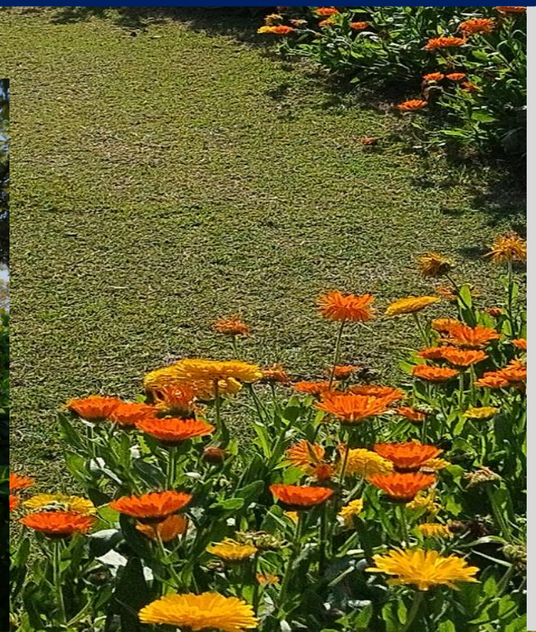
BRIC-National Institute of Plant Genome Research (BRIC-NIPGR), New Delhi, India



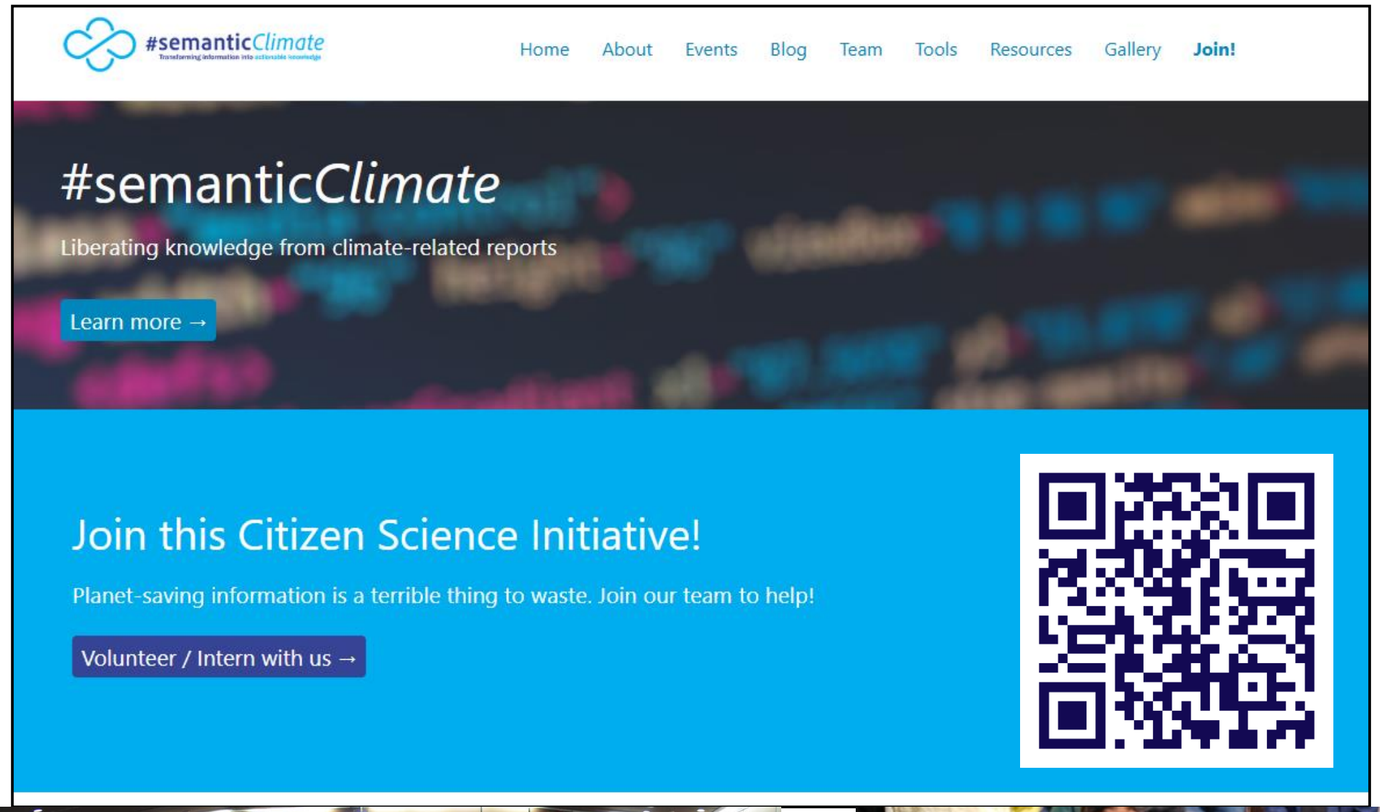
BRIC-NIPGR

- started in1998
- Research areas are: Computational Biology, Genome Analysis, Nutritional Genomics, Plant Development and Architecture, Plant Immunity etc.

Arabidopsis, Rice, Chickpea, Mustard, Banana etc.



#semanticClimate



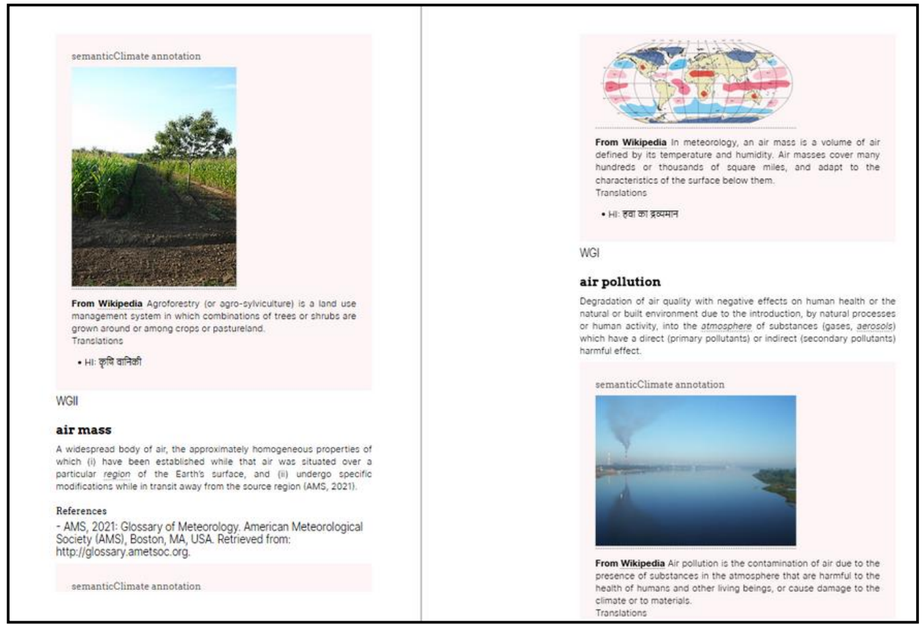
Internship, Workshop and Outreach

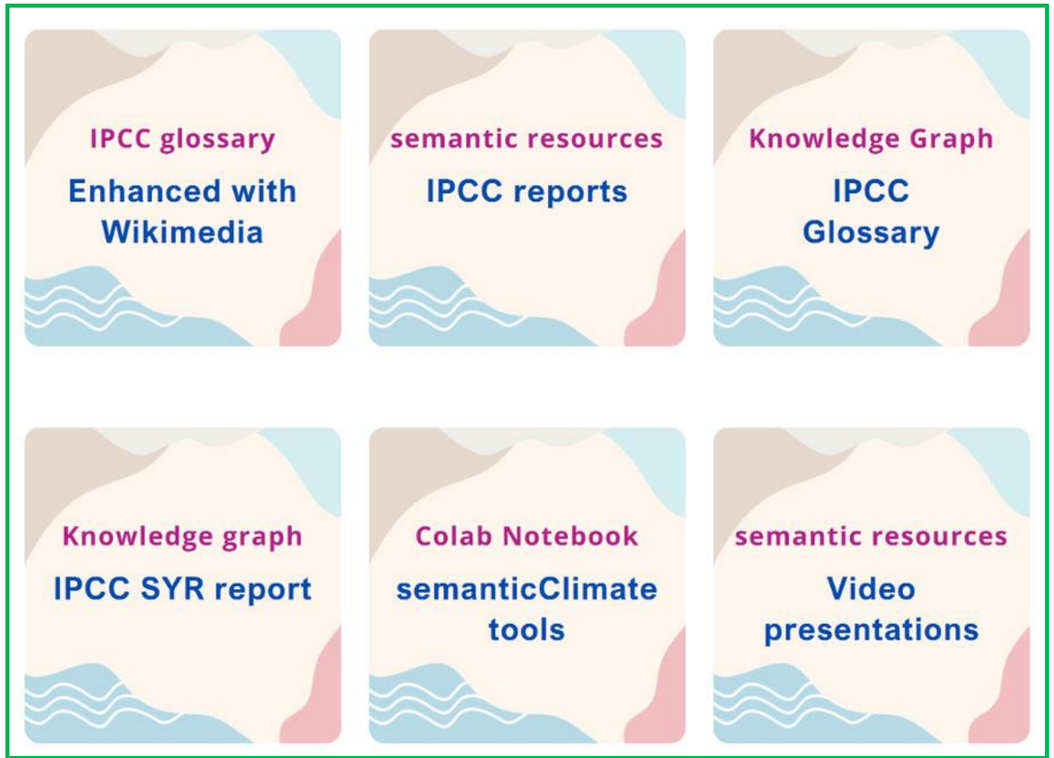
https://semanticclimate.github.io/p/en/

#semanticClimate resources

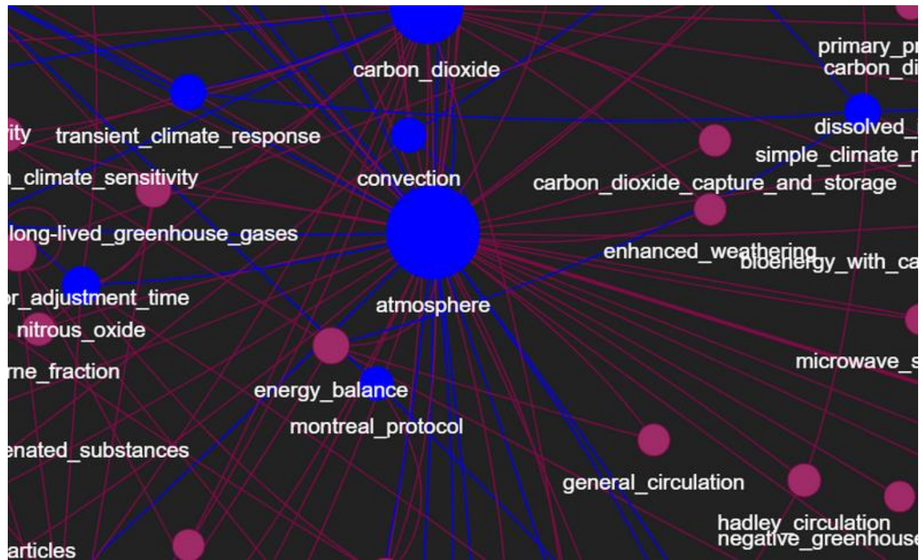
https://semanticclimate.github.io/p/en/posts/resources/

IPCC Glossary enhanced with Wikipedia

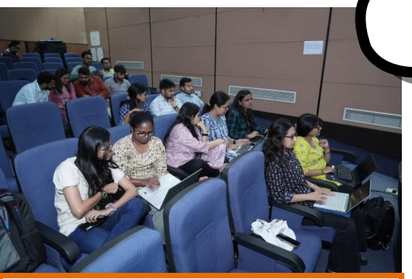




IPCC Glossary Knowledge Graph



Context of the presentation



Researchers, Academician policy makers

for writing thesis, report, projects

Need updated research articles on specific query in a single line step

#semanticClimate
Transforming information into actionable knowledge

pygetpapers

creates semantic corpus of scientific literatures in machine readable format

Content of the presentation

1. Introduction about corpus and their applications **semantic corpus**
2. About the tool **pygetpapers**
3. Workflow of the tool **pygetpapers**
4. About the tools
   - **txt2phrases** to extract keywords/keyphrases
   - **amilib** **Encyclopedia**
5. Introduction of colab notebook platfom to run the tool

What is Corpus?

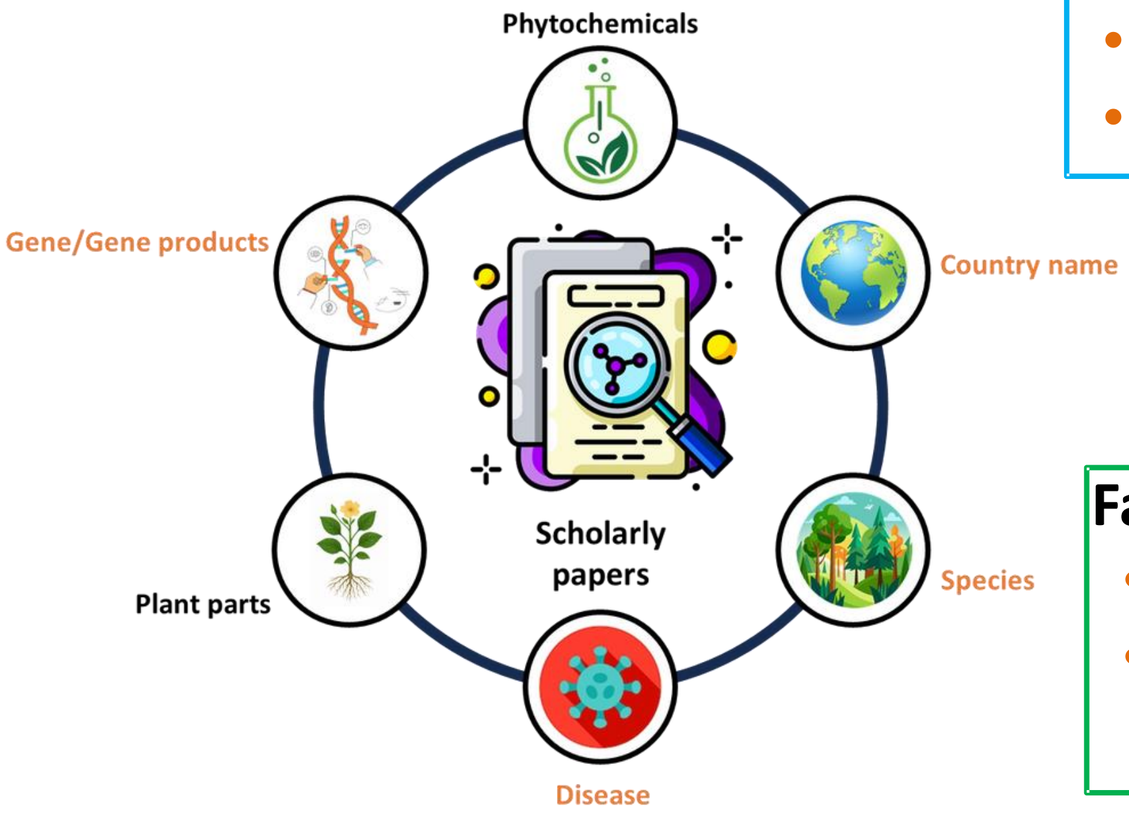
It is a structured collection of scholarly articles and research papers that can be used for further analysis.

Challenges in current scenario

- Exponential growth of publications
- Difficult to keep up with the latest developments
- Literature exists in various formats: mainly PDFs
- Not machine-readable or structured formats
- Limited Access to the repositories and journals
- No single platform for getting access to all research outputs
- Bulk downloading is often restricted
- Technical Barriers to automate article retrieval for people with no coding

Applications of Scientific Literature corpus?

Named Entity Recognition



To train **Natural Language Processing (NLP)** models for the following:

- **Named Entity Recognition (NER)**
- **Automated summarization**

Facilitate literature reviews:

- **identifying gaps**
- **formulating hypotheses**

So, we need a tool which can create curated corpus in a machine-readable

pygetpapers (corpus builder)

#semanticClimate tool used to create corpus

pygetpapers

- **pygetpapers** is a tool to assist text miners.

Developed by: Ayush Garg and Peter Murray-Rust

GitHub repo: https://github.com/petermr/pygetpapers

Workflow

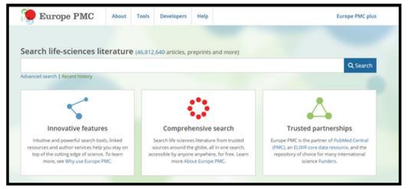
Add query term
("Ocean Heat")

Input

pygetpapers

Open access repository
(ex: Europe PMC)

EuropePMC



bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

medRχiv
THE PREPRINT SERVER FOR HEALTH SCIENCES

Crossref

OpenAlex

Creates datatables for the corpus

download PDF files

Give papers in XML/HTML format

Output

Features of pygetpapers

Advanced Features

1 Date Range Queries

Search papers published between dates

--startdate 2023-01-01 --enddate 2023-12-31

2 Term-based Queries

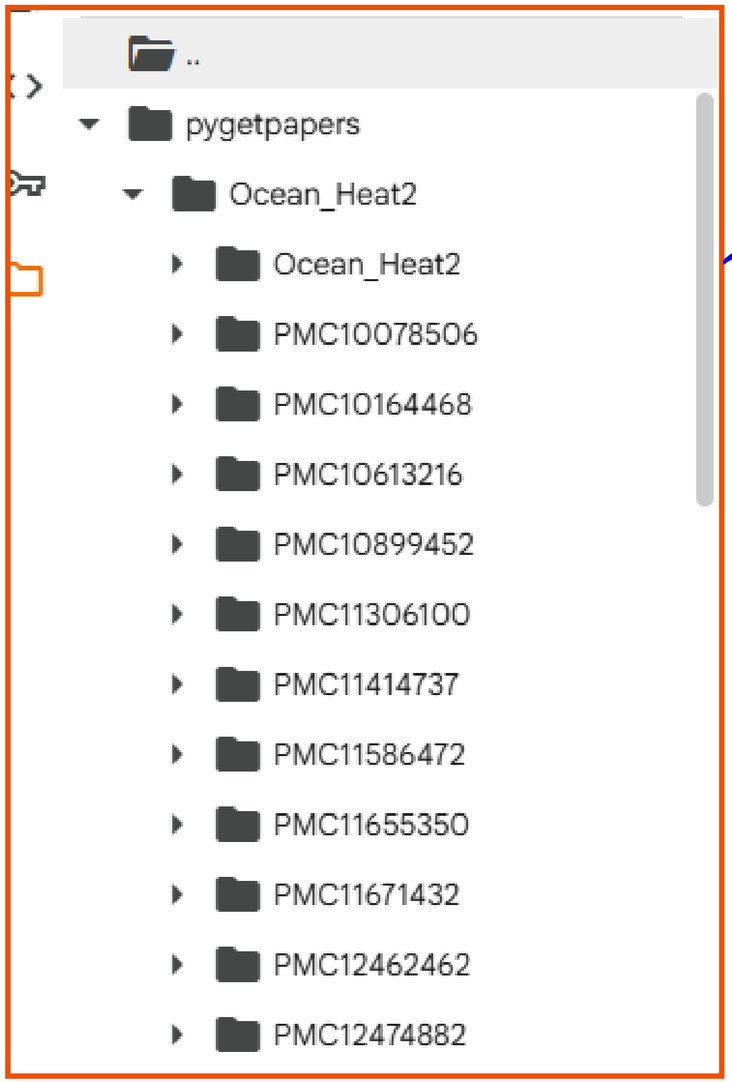
terms.txt with comma-separated terms

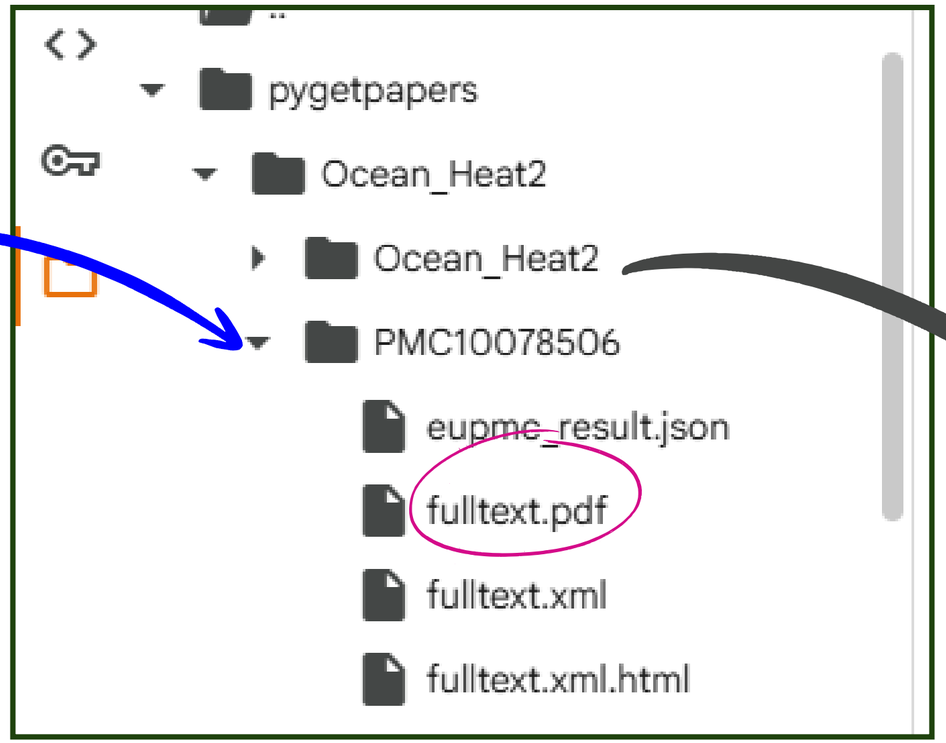
machine learning, artificial intelligence, deep learning

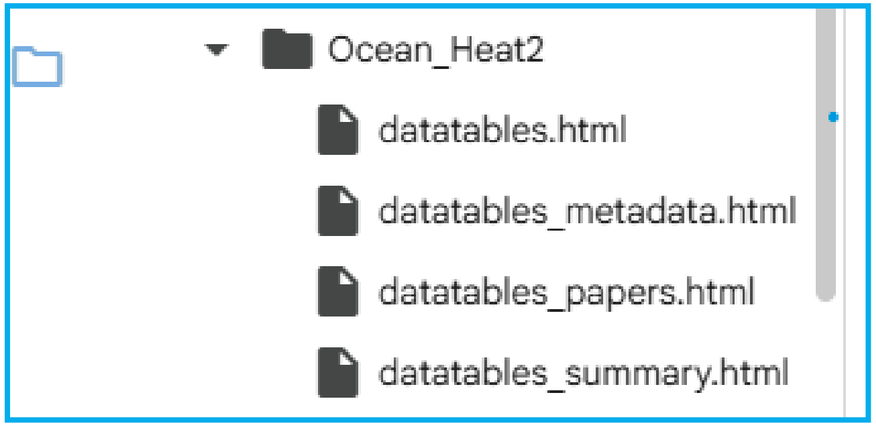
Output Formats

1. JSON Metadata
2. CSV Output
3. HTML Tables
4. Datatables
5. Full-text Downloads

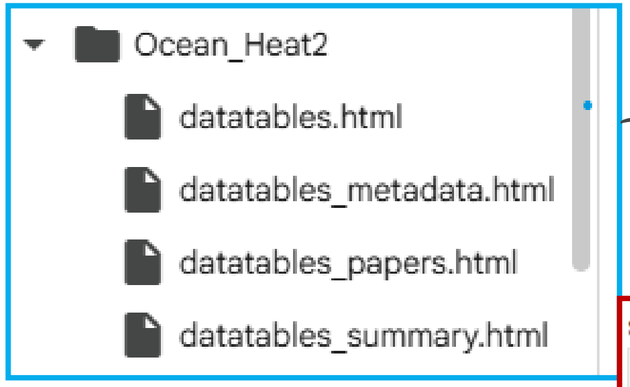
Result: Structure of the corpus







Summary table for the corpus



- PMCID (for Europe PMC)
- Title,
- Authors,
- DOI,
- link to PDF, XML, HTML

Show 25 entries per page Search:

Select	ID	Title	Authors	Abstract	Journal	DOI	PMID	PMCID	Date	XML	PDF	Suppl	HTML
☐	PMC5653740	The Subpolar North Atlantic Ocean Heat Content Variability and its Decomposition.	Zhang W, Yan XH.	The Subpolar North Atlantic (SPNA) is one of the most important areas to global climate because its ocean heat content (OHC) is highly correlated with...	Scientific reports	10.1038/s41598-017-14158-6	29062083	PMC5653740	2017-10-23	XML	PDF	✖	HTML
☐	PMC7991649	Author Correction: The causality from solar irradiation to ocean heat content detected via multi-sca...	Wang G, Zhao C, Zhang M, Zhang Y, Lin M, Qiao F.	No abstract available	Scientific reports	10.1038/s41598-021-86723-z	33762654	PMC7991649	2021-03-24	XML	PDF	✖	HTML
☐	PMC6347704	Global reconstruction of historical ocean heat storage and transport.	Zanna L, Khatiwala S, Gregory JM, Ison J, Heimbach...	Most of the excess energy stored in the climate system due to anthropogenic greenhouse gas emissions has been taken up by the oceans, leading to therm...	Proceedings of the National Academy of Sciences of the United States of America	10.1073/pnas.1808838115	30617081	PMC6347704	2019-01-07	XML	PDF	✖	HTML

Application of the semantic corpus

keyphrases to encyclopedia

query

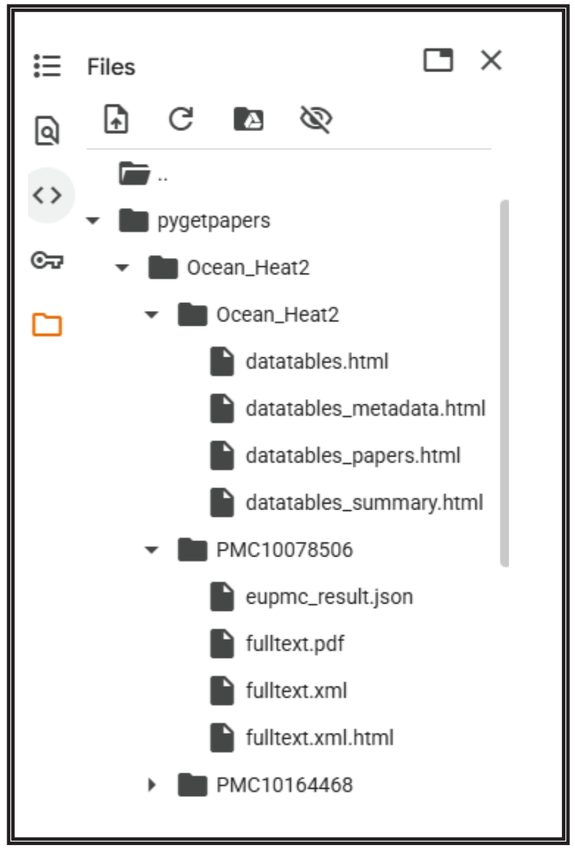
Ocean Heat

search research papers

pygetpapers

1

semantic corpus



poleward expansion of the zonal mean **Hadley cell** in the Southern Hemisphere
bserved poleward expansion of the zonal mean **Hadley cell** in the Northern
causes of the observed strengthening of the Pacific Walker circulation since the
hening trend is outside the range of trends simulated in the coupled models
general characteristics of the tropospheric large-scale circulation (*high confidence*),

Text from climate report

Hadley cell, monsoons, solar dimming

keywords/keyphrases

2

txt2phrases 0.2.0

```
pip install txt2phrases
```

amilib

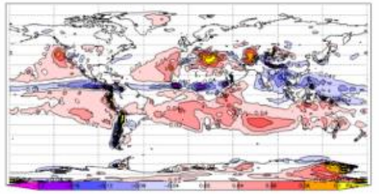
3

- HTML to TXT conversion
- Keyword extraction: Use Hugging Face Transformers
- Per-chapter TF-IDF-based keyword classification

Encyclopedia

search term: Hadley cells Wikipedia Page

The **Hadley cell**, also known as the **Hadley circulation**, is a global-scale tropical atmospheric circulation that features air rising near the equator, flowing poleward near the tropopause at a height of 12–15 km (7.5–9.3 mi) above the Earth's surface, cooling and descending in the subtropics at around 25 degrees latitude, and then returning equatorward near the surface. It is a thermally direct circulation within the troposphere that emerges due to differences in insolation and heating between the tropics and the subtropics. On a yearly average, the circulation is characterized by a circulation cell on each side of the equator. The Southern Hemisphere Hadley cell is slightly stronger on average than its northern counterpart, extending slightly beyond the equator into the Northern Hemisphere. During the summer and winter months, the Hadley circulation is dominated by a single, cross-equatorial cell with air rising in the summer hemisphere and sinking in the winter hemisphere. Analogous circulations may occur in extraterrestrial atmospheres, such as on Venus and Mars.



Average vertical velocity (in pascals per second) at the 500 hPa pressure height in July from 1979–2001. Ascent (negative values) is concentrated close to the solar equator while descent (positive values) is more diffuse; their distribution is an imprint of the ascending and descending branches of the Hadley circulation.

#semanticClimate
Transforming information into actionable knowledge

Platform to run the tool

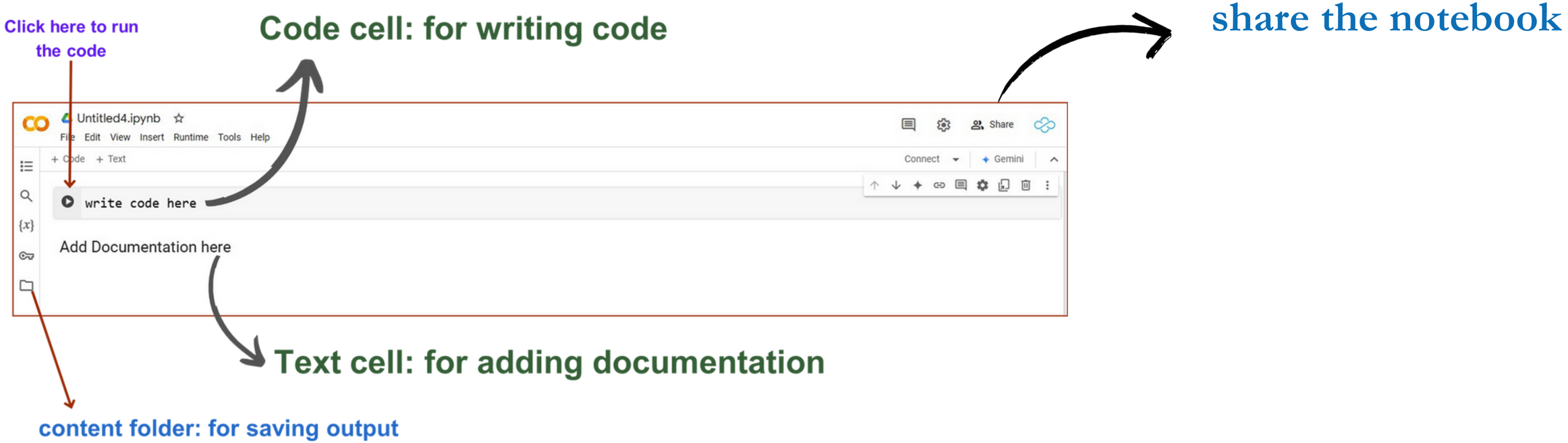
Google Colab Notebook

Importance of Google Colab Notebook

- Open Jupyter E-Notebook environment
- No pain with setups, versions
- Human Machine friendly
- Supports interactive programming
- Easy learn and explore new tools

Google Colab (Collaboratory)

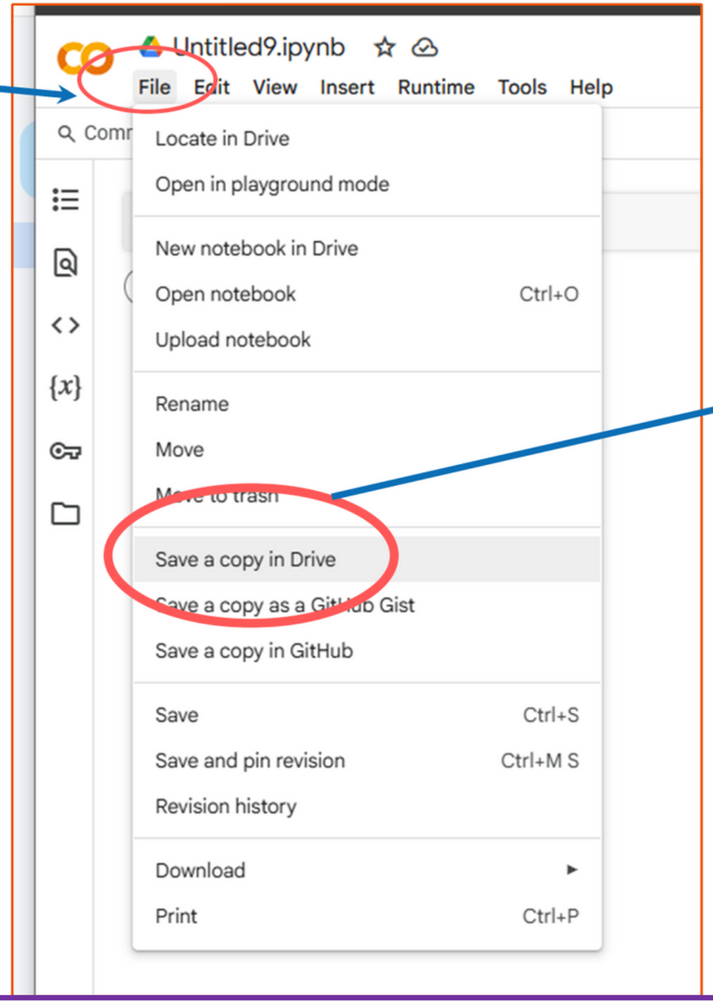
free, cloud-based platform to write and execute python-based codes



Need Google account to get started!

Google Colab (Collaboratory)

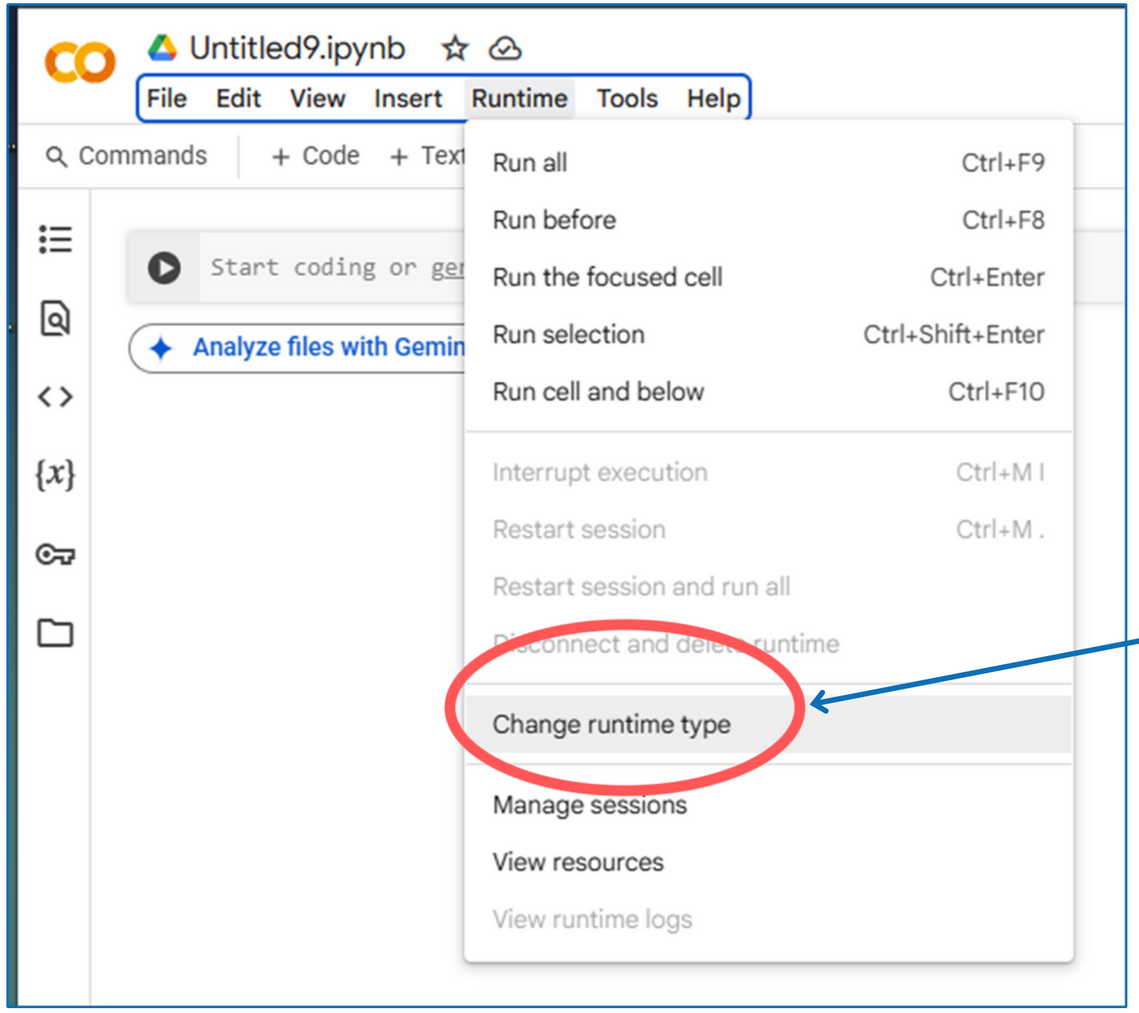
Click to File



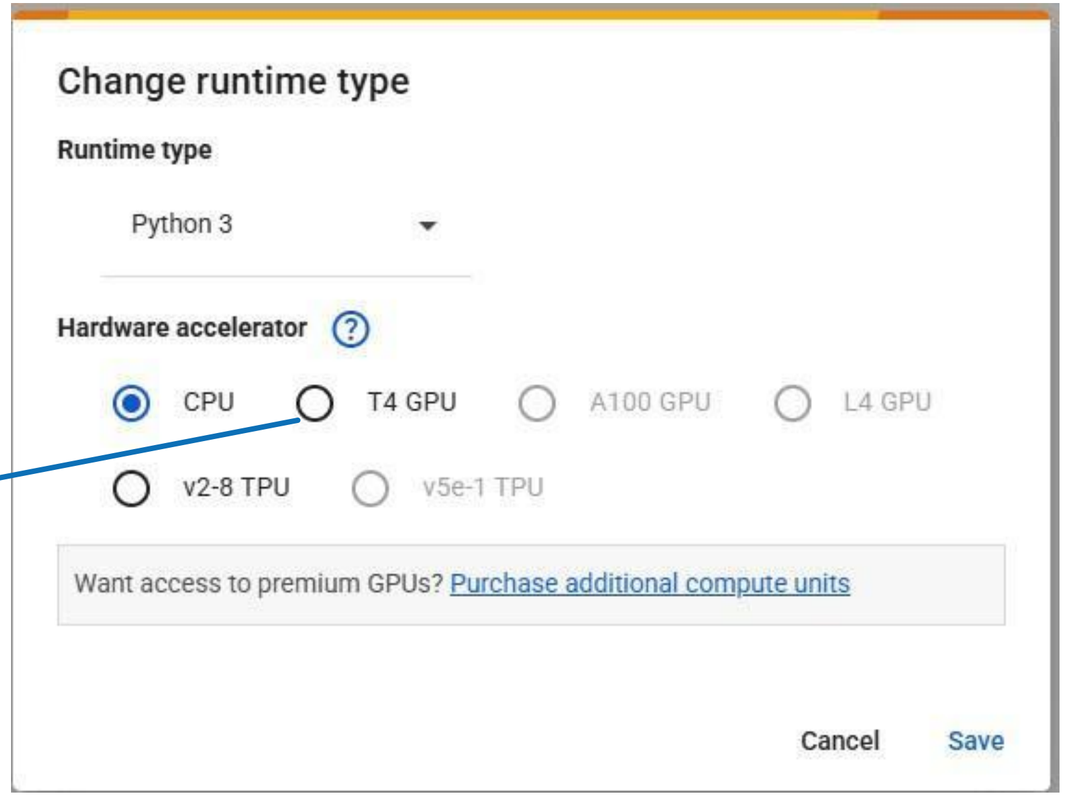
Save a copy in Drive

Before using this Colab, Save a copy to your own Google Drive:
Click on "File" > "Save a copy in Drive"

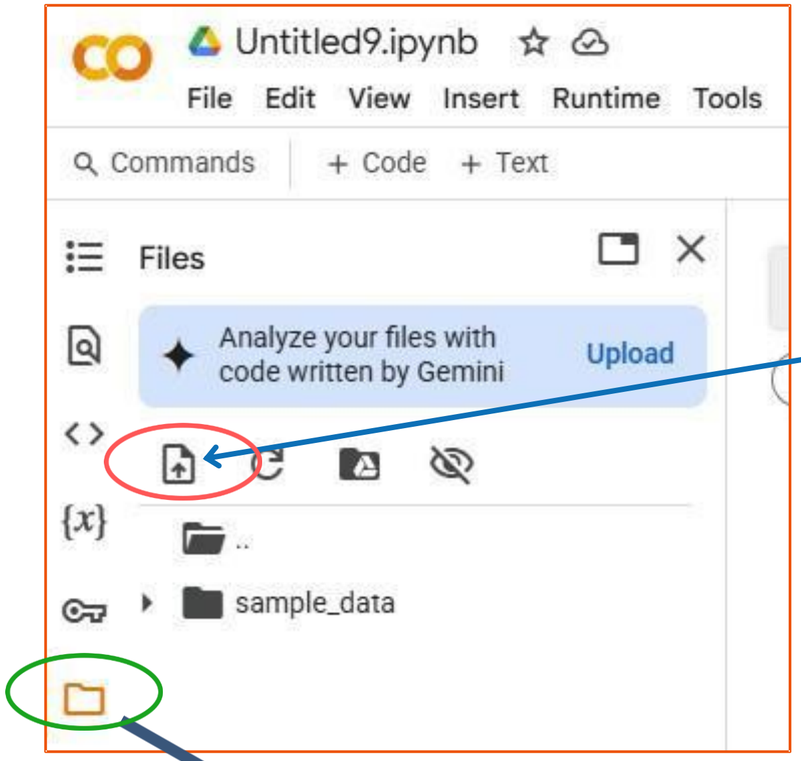
Change Runtime for more memory requirement



T4 GPU



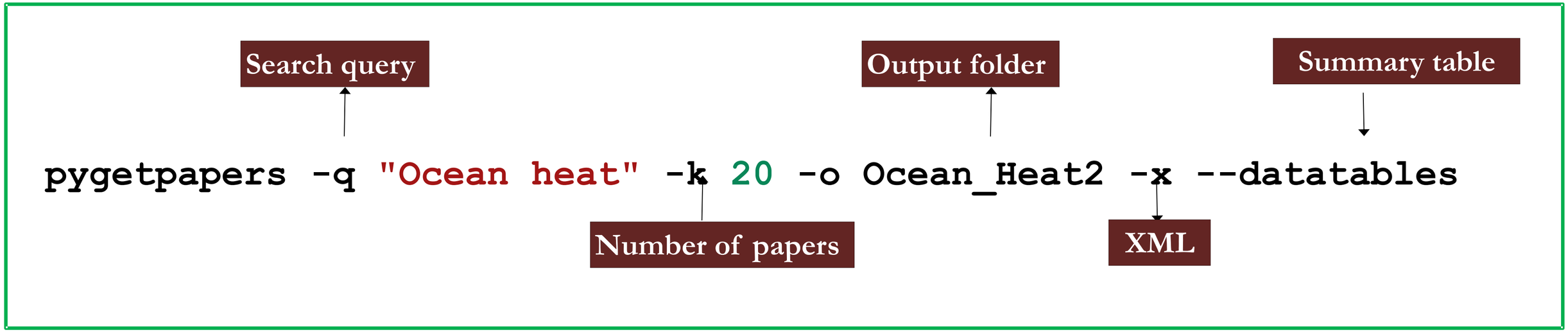
Input and Output files



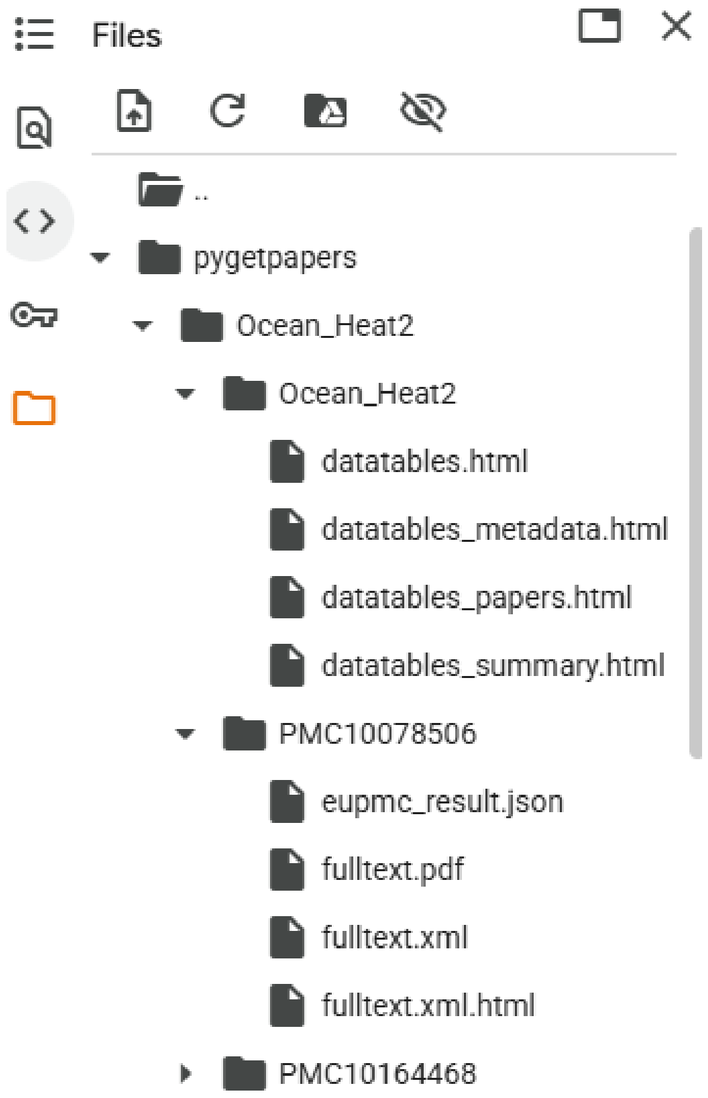
Upload the files

For all the results and input files

Querying pygetpapers in colab notebook



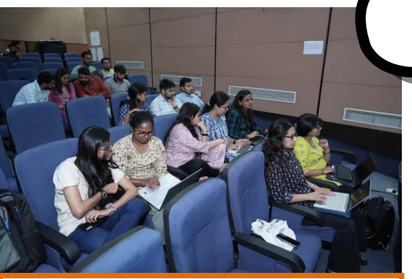
Result: Corpus created for the query



Show 25 entries per page Search:

Select	ID	Title	Authors	Abstract	Journal	DOI	PMID	PMCID	Date	XML	PDF	Suppl	HTML
☐	PMC5653740	The Subpolar North Atlantic Ocean Heat Content Variability and its Decomposition.	Zhang W, Yan XH.	The Subpolar North Atlantic (SPNA) is one of the most important areas to global climate because its ocean heat content (OHC) is highly correlated with...	Scientific reports	10.1038/s41598-017-14158-6	29062083	PMC5653740	2017-10-23	XML	PDF	✖	HTML
☐	PMC7991649	Author Correction: The causality from solar irradiation to ocean heat content detected via multi-sca...	Wang G, Zhao C, Zhang M, Zhang Y, Lin M, Qiao F.	No abstract available	Scientific reports	10.1038/s41598-021-86723-z	33762654	PMC7991649	2021-03-24	XML	PDF	✖	HTML
☐	PMC6347704	Global reconstruction of historical ocean heat storage and transport.	Zanna L, Khatiwala S, Gregory JM, Ison J, Heimbach...	Most of the excess energy stored in the climate system due to anthropogenic greenhouse gas emissions has been taken up by the oceans, leading to therm...	Proceedings of the National Academy of Sciences of the United States of America	10.1073/pnas.1808838115	30617081	PMC6347704	2019-01-07	XML	PDF	✖	HTML

Summary



Researchers, Academician policy makers

for writing thesis, report, projects

query

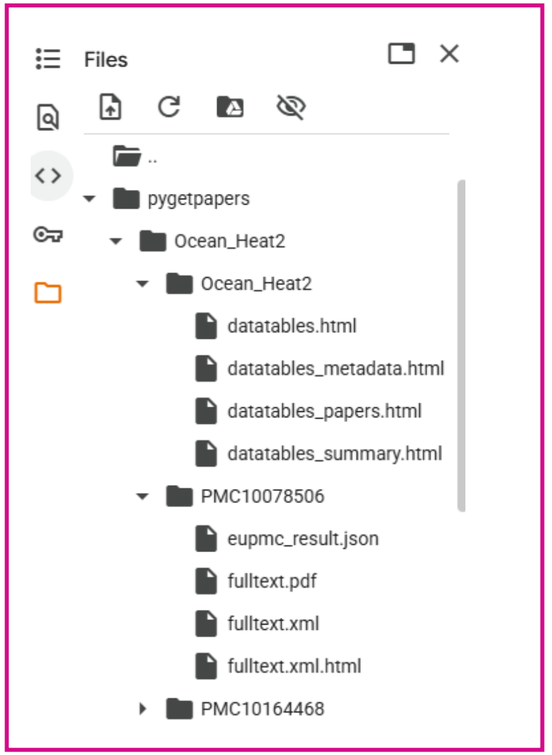
Ocean Heat

updated research articles on specific query in a single line step

pygetpapers

#semanticClimate
Transforming information into actionable knowledge

semantic corpus



Summary

query

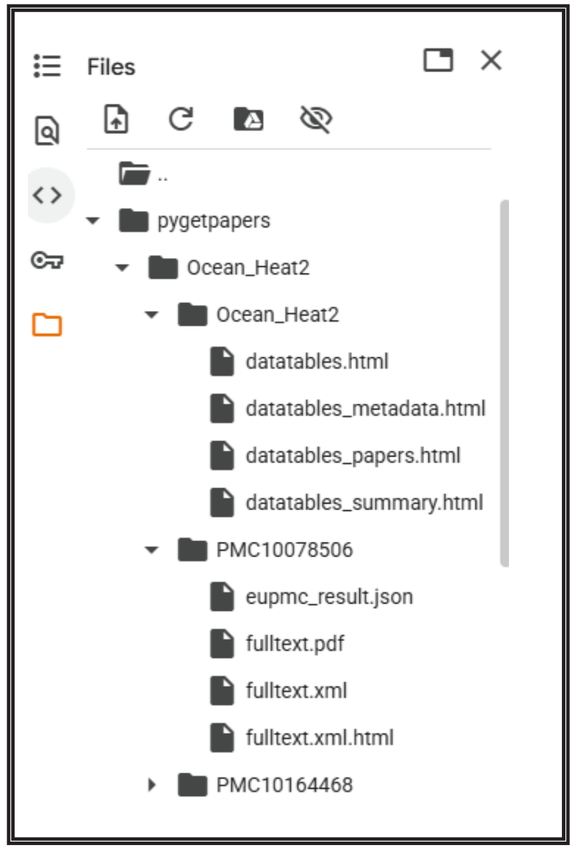
Ocean Heat

search research papers

pygetpapers

1

semantic corpus



poleward expansion of the zonal mean Hadley cell in the Southern Hemisphere
bserved poleward expansion of the zonal mean Hadley cell in the Northern
causes of the observed strengthening of the Pacific Walker circulation since the
hening trend is outside the range of trends simulated in the coupled models
general characteristics of the tropospheric large-scale circulation (*high confidence*),

Text from climate report

2

txt2phrases

keywords/keyphrases

Hadley cell, monsoons, solar dimming

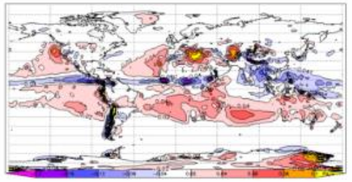
amilib

3

Encyclopedia

search term: Hadley cells Wikipedia Page

The **Hadley cell**, also known as the **Hadley circulation**, is a global-scale tropical atmospheric circulation that features air rising near the equator, flowing poleward near the tropopause at a height of 12–15 km (7.5–9.3 mi) above the Earth's surface, cooling and descending in the subtropics at around 25 degrees latitude, and then returning equatorward near the surface. It is a thermally direct circulation within the troposphere that emerges due to differences in insolation and heating between the tropics and the subtropics. On a yearly average, the circulation is characterized by a circulation cell on each side of the equator. The Southern Hemisphere Hadley cell is slightly stronger on average than its northern counterpart, extending slightly beyond the equator into the Northern Hemisphere. During the summer and winter months, the Hadley circulation is dominated by a single, cross-equatorial cell with air rising in the summer hemisphere and sinking in the winter hemisphere. Analogous circulations may occur in extraterrestrial atmospheres, such as on Venus and Mars.



Average vertical velocity (in pascals per second) at the 500 hPa pressure height in July from 1979–2001. Ascent (negative values) is concentrated close to the solar equator while descent (positive values) is more diffuse; their distribution is an imprint of the ascending and descending branches of the Hadley circulation.

#semanticClimate
Transforming information into actionable knowledge

BRIC
a DBT Organization

NIPGR

THANK YOU

NEW YORK INSTITUTE OF TECHNOLOGY
Library

International Open Access Week

SPARC

FOLLOW US

E-mail semanticclimate@gmail.com

Website https://semanticclimate.github.io/p/en/

GitHub https://github.com/semanticClimate

X (Formerly Twitter) #semanticClimate

Encyclopedia

#semanticClimate

GitHub repo: https://github.com/petermr/pygetpapers

Link to demonstration notebook

URL https://colab.research.google.com/drive/16DPjzO3gR8dc2M_don4tSQSjjcjeuSUJ?usp=sharing

Demonstration by: Udita Agarwal