Indian Academy of Sciences, Bengaluru
Indian National Science Academy, New Delhi
The National Academy of Sciences India, Prayagraj

SUMMER RESEARCH FELLOWSHIPS — 2025

Format for the final Report*,^

Name of the candidate : Haarthi Vallabhaneni

Application Registration no. : ENGS3376

Date of joining : 26-05-2025

Date of completion : 21-07-2025

Total no. of days worked : 56 days(8weeks)

Name of the guide : Dr. Gitanjali Yadav, Staff Scientist VI

Guide's institution : National Institute of Plant Genome Research, New Delhi

Project title : Semantic Analysis of IPCC AR6 WGII Chapter 4: Water

Address with pin code to which the certificate could be sent:

Dno:40-25/1-29/1, Maszid Street Patamatalanka, Near
Benzcircle Ntr District Vijayawada 520010 Andhra Pradesh

E-mail ID: haarthivallabhaneni13@gmail.com

Phone No: 7416831263

TA Form attached with final report : YES _______ [✓ NO]

If, NO, Please specify reason : worked from home

V. Haarthi

Signature of the candidate
Date: 24-07-2025

Signature of the guide
Date: 24-07-2025

IMPORTANT NOTES:

*** This format should be the first page of the report and should be stapled with the main report. The final report could be anywhere between 20 and 25 pages including tables, figures etc.**

^ The final report must reach the Academy office within 10 days of completion. If delayed fellowship amount will not be disbursed.

(For office use only; do not fill/tear)

Candidate's name:	Fellowship amount:
Student: Teacher:	Deduction:
Guide's name:	TA fare:
KVPY Fellow: INSPIRE Fellow:	Amount to be paid:
PFMS Unique Code:	A/c holder's name:
Others	

REPORT

Semantic Analysis of IPCC AR6 WGII Chapter 4: Water

Submitted by:

Haarthi Vallabhaneni
ENGS3376

in partial fulfilment of the requirements of
IASc-INSA-NASI SRFP 2025

Under the Supervision of

Dr. Gitanjali Yadav
NIPGR, New Delhi

Dr. Peter Murray-Rust
University of Cambridge, UK

National Institute of Plant Genome Research New Delhi 110067

Contents

1. Introduction

Climate change is no longer a distant threat, its impacts are visible across ecosystems, economies, and communities worldwide. Recognizing the urgency, the global scientific community established the Intergovernmental Panel on Climate Change (IPCC) under the United Nations and the World Meteorological Organization to provide comprehensive assessments of climate science, impacts, and response options.

The Sixth Assessment Report (AR6) is the most recent and detailed effort by the IPCC. It consists of contributions from three Working Groups and culminates in the 2023 Synthesis Report, described by UN Secretary-General António Guterres as “the survival guide for humanity.” The Synthesis Report distills the findings of the three major reports:

- Working Group I (WGI) – The Physical Science Basis
- Working Group II (WGII) – Impacts, Adaptation, and Vulnerability
- Working Group III (WGIII) – Mitigation of Climate Change

These reports are deeply interconnected, and together they form a foundation for science-based policy action. The Synthesis Report brings their key messages together into a unified narrative, offering a holistic overview of the climate crisis and the urgent steps needed.

Among the various chapters available across the IPCC AR6 reports, I chose to work on Chapter 4 of Working Group II – Water[1].

Summary of IPCC AR6 WGII Chapter 4: Water

This chapter presents a comprehensive assessment of how climate change is affecting the global water cycle. It explores observed and projected changes in precipitation, snow and ice melt, droughts, floods, and water availability, along with their consequences for human societies and ecosystems.

Key highlights from the chapter include:

- **Intensification of the water cycle**: Global warming is leading to more intense rainfall events and longer dry spells, disrupting agricultural practices and water supply systems.
- **Widespread water-related risks**: At 2°C of global warming, between 0.9 to 3.9 billion people are projected to experience increased water stress. Vulnerabilities are especially high in regions with limited adaptive capacity.
- **Impacts on multiple sectors**: Water-related climate risks directly affect agriculture, sanitation, urban planning, public health, energy production, and biodiversity.
- **Adaptation and responses**: Over 60% of documented climate adaptation strategies globally focus on water-related issues, demonstrating the central role of water in climate resilience planning.

By working on this chapter, I aimed to extract meaningful keywords, build a domain-specific dictionary, and create an interactive semantic structure that could help scientists, educators, and the public explore the content more effectively.

2. Aims and Objectives

My work under the #semanticClimate initiative was guided by the broader goal of transforming climate-related documents, especially the IPCC's Sixth Assessment Reports into semantic, structured, and accessible formats. The intention is to make these critical scientific materials easier to explore, search, and understand by both humans and machines. This is essential for increasing engagement, promoting open science, and enabling global action on climate change.

During the internship, my objectives were focused on the following key areas:

- **Understanding the Project and Tools:**

Before contributing meaningfully to the project, it was important to first understand the software ecosystem and the vision behind #semanticClimate. The tools developed by the team are designed to process large scientific documents and convert them into machine-readable, semantically enriched content. I spent time exploring the core libraries, attending team discussions, reviewing documentation, and setting up the development environment. This helped me align my work with the project's values of open knowledge, transparency, and collaborative science.

- **Keyword Extraction:**

Identifying and extracting relevant keywords from climate reports is essential because these terms represent the scientific concepts, phenomena, and concerns discussed in the chapter. They form the basis for glossaries, semantic tagging, and information retrieval. My task involved analyzing the chapter's content to generate a curated list of domain-specific keywords. These terms capture the essence of the chapter and support downstream tasks like dictionary building and knowledge graph construction.

- **Dictionary Development:**

A dictionary serves as a bridge between technical content and public understanding. It provides clear, contextual definitions that help readers grasp complex scientific terminology. After generating the keyword list, I created a structured dictionary that linked each term with accessible descriptions. The goal was to enhance comprehension for non-expert audiences and support multilingual translation and educational use. This dictionary also contributes to semantic interoperability by integrating linked open data.

- **Knowledge Graph Design:**

Climate reports often contain dense and complex structures, making them difficult to navigate. Creating a knowledge graph helps visualize the hierarchy and relationships between sections, enabling intuitive exploration of the document. I constructed a graph that reflected the structure of Chapter 4, highlighting its major themes and subtopics. This serves as a visual map for readers and can also be embedded into interactive platforms for enhanced digital engagement.

- **Alpha testing tools:**

As part of ensuring the reliability and usability of the tools developed under the #semanticClimate initiative, I participated in alpha testing key software components. This involved running the tools on sample chapters, reporting bugs, and suggesting improvements based on user experience. The focus was on validating outputs such as keyword lists, dictionaries, and semantic markup. I contributed feedback to enhance the robustness and accessibility of the tools for future users. Through this process, I gained experience in software quality assurance and helped ensure that the tools meet the standards of accuracy, usability, and open collaboration.

Each of these objectives contributes to #semanticClimate's larger mission of building an open, semantic infrastructure for climate science—empowering users to interact with critical data in meaningful and accessible ways.

3. Activities & Accomplishments

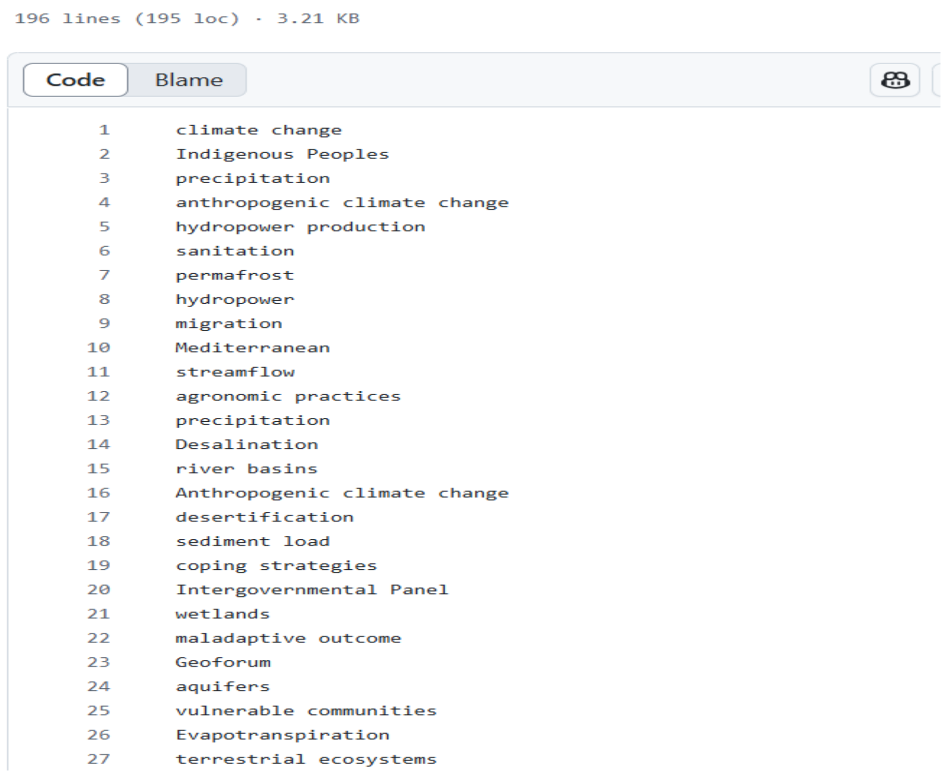
The core of my internship work focused on semantically enriching Chapter 4 (“Water”) of the IPCC AR6 Working Group II report. This involved a series of technical and research-driven tasks aimed at making the chapter's content more accessible, machine-readable, and interoperable. By extracting key scientific terms, creating structured dictionaries, and visualizing the chapter's structure as a knowledge graph, I contributed to building the foundational components necessary for advanced search, retrieval, and educational applications. These processes were aligned with #semanticClimate’s mission to enable open, transparent, and reusable climate knowledge.

3.1 Wordlist Extraction for IPCC AR6 WG2 Chapter 4

The first step in enriching the content semantically was the extraction of domain-specific keywords from Chapter 4 ("Water") of the IPCC AR6 WGII report. This chapter discusses critical water-related climate issues, and identifying key terms within it is essential for tasks such as tagging, glossary generation, and knowledge mapping.

- **Downloaded HTML File:** Acquired the cleaned and preprocessed HTML version of Chapter 4 from the IPCC content repository. This version was specifically formatted to support machine reading and downstream processing.
- **Environment Setup:** Created a virtual environment using Anaconda Navigator to ensure isolation and compatibility. Installed necessary Python dependencies.
- **Keyword Extraction:** Utilized a custom script (keyword_extraction.py) to preprocess and analyze the chapter text. This included tokenization, stopword removal, and frequency analysis. The output was a curated wordlist capturing important scientific and technical terms related to water and climate.

-
  ```
  python keyword_extraction.py -t text_file_path.txt -s ./
  ```



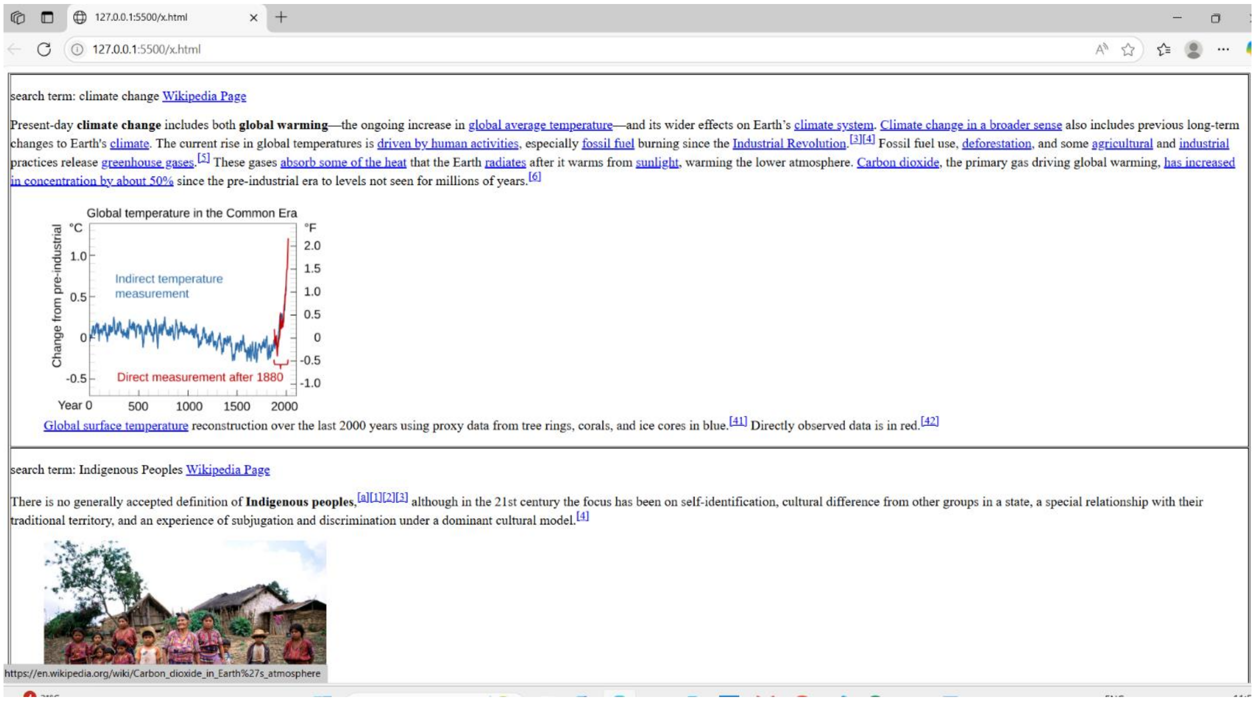
```
climate change
Indigenous Peoples
precipitation
anthropogenic climate change
hydropower production
sanitation
permafrost
hydropower
migration
Mediterranean
streamflow
agronomic practices
precipitation
Desalination
river basins
Anthropogenic climate change
desertification
sediment load
coping strategies
Intergovernmental Panel
wetlands
maladaptive outcome
Geoforum
aquifers
vulnerable communities
Evapotranspiration
terrestrial ecosystems
```

Extracted wordlist

3.2 Dictionary Creation for IPCC AR6 WG2 Chapter 4

The second major task was transforming the extracted keywords into a structured semantic dictionary. This dictionary aims to bridge the gap between complex scientific language and public comprehension by linking each term to an accessible definition.

- **Input Preparation:** Processed the generated keyword list and ensured the terms were contextually relevant and distinct.
- **Dictionary Generation:** Used the amilib framework to link each keyword to its corresponding description sourced from Wikipedia. The dictionary was produced in HTML format for easy integration into browsers and educational platforms.
- **Optional Features:** Included semantic figures where available.
- amilib DICT --words your_wordlist_path.txt wordlist.txt --description wikipedia --dict output_dict_path.html --figures --operation create


127.0.0.1:5500/x.html
127.0.0.1:5500/x.html
Global temperature in the Common Era
°C
°F
Change from pre-industrial
1.0
0.5
0
-0.5
2.0
1.5
1.0
0.5
0
-0.5
-1.0
Indirect temperature measurement
Direct measurement after 1880
Year 0
500
1000
1500
2000
https://en.wikipedia.org/wiki/Carbon_dioxide_in_Earth%27s_atmosphere

search term: climate change Wikipedia Page

Present-day **climate change** includes both **global warming**—the ongoing increase in global average temperature—and its wider effects on Earth's climate system. Climate change in a broader sense also includes previous long-term changes to Earth's climate. The current rise in global temperatures is driven by human activities, especially fossil fuel burning since the Industrial Revolution.[3][4] Fossil fuel use, deforestation, and some agricultural and industrial practices release greenhouse gases.[5] These gases absorb some of the heat that the Earth radiates after it warms from sunlight, warming the lower atmosphere. Carbon dioxide, the primary gas driving global warming, has increased in concentration by about 50% since the pre-industrial era to levels not seen for millions of years.[6]

Global surface temperature reconstruction over the last 2000 years using proxy data from tree rings, corals, and ice cores in blue.[41] Directly observed data is in red.[42]

search term: Indigenous Peoples Wikipedia Page

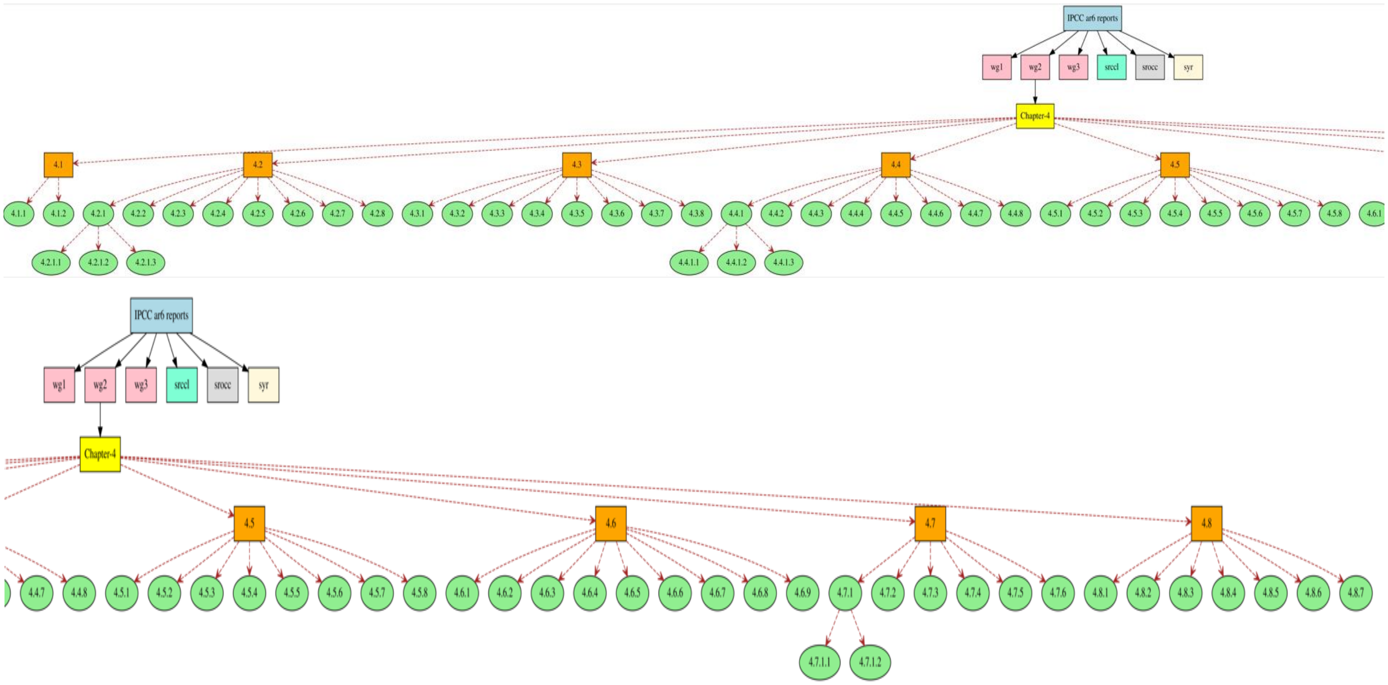
There is no generally accepted definition of **Indigenous peoples**,[a][1][2][3] although in the 21st century the focus has been on self-identification, cultural difference from other groups in a state, a special relationship with their traditional territory, and an experience of subjugation and discrimination under a dominant cultural model.[4]

Dictionary Created

3.3 Knowledge Graph Creation

This graph presents a hierarchical and thematic overview of the chapter's contents, helping both technical and non-technical users navigate complex material intuitively.

- **Table of Contents Extraction:** Parsed the section headers and subheaders from the chapter to understand the internal structure and thematic flow.
- **Graph Design:** Designed a semantic knowledge graph using the Graphviz library. Each node represents a section or subsection, while the edges indicate logical or thematic relationships. Dotted red edges were used to signify implicit or loosely connected topics. Tooltips were added to provide additional context when hovering over nodes.
- **Output Generation:** The final graph was rendered as an SVG (ipcc_ch04_graph.svg) for embedding into Jupyter Notebooks, technical documents, or web interfaces.

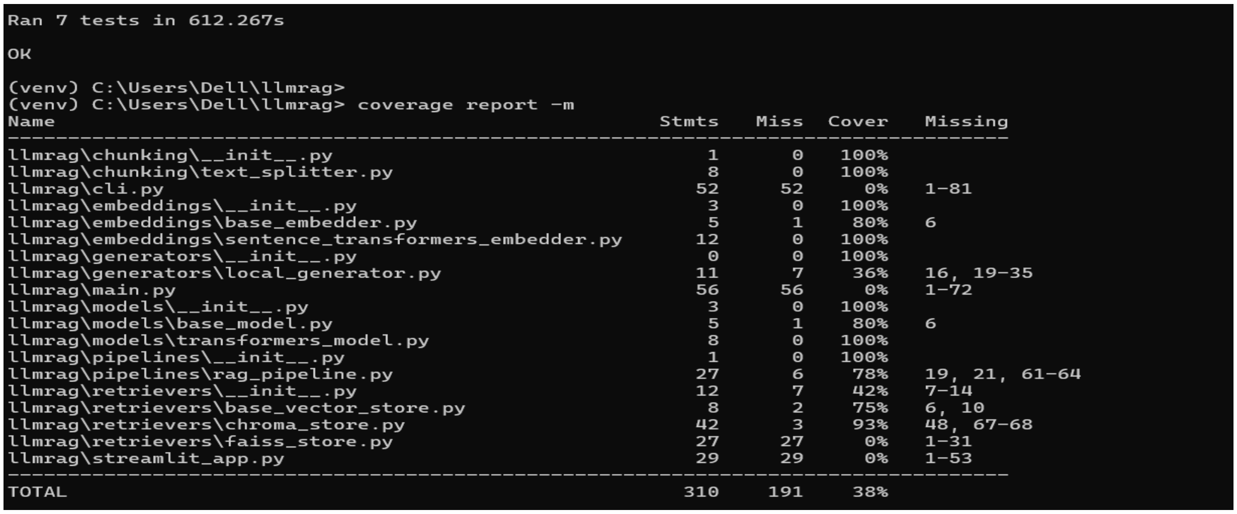


Knowledge Graph created using Graphviz

3.4 Alpha testing of tool

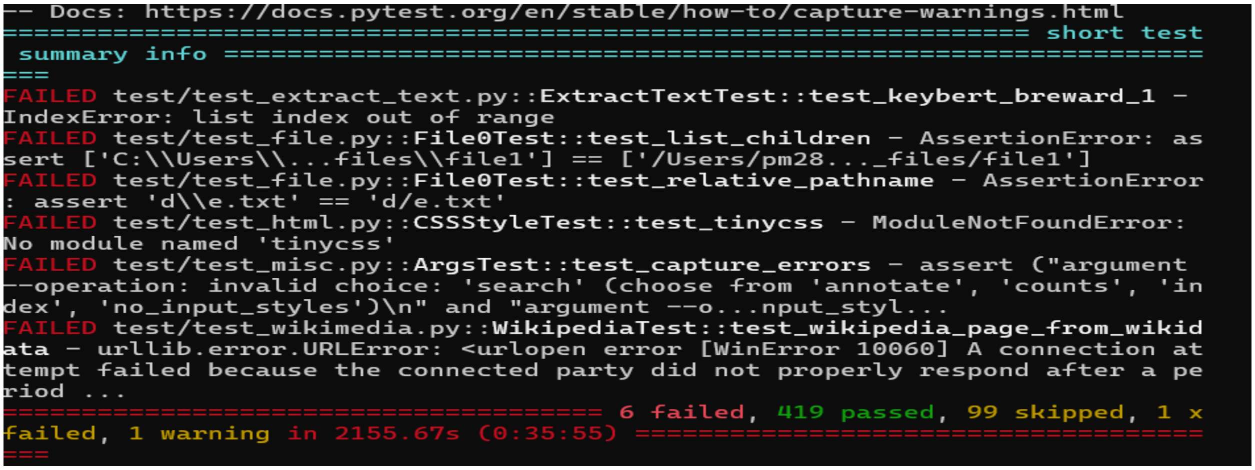
Contributed to improving tool robustness by participating in early-stage testing (alpha phase) of key Semantic Climate tools.

- **Tested Tools:**
  - Pygetpapers (for scientific paper retrieval)
  - amilib (for semantic enrichment and dictionary creation)
  - LLMRAG (for AI-assisted document retrieval and analysis)
- Reported bugs, usability issues, and provided enhancement suggestions during internal testing phases.
- Verified outputs against expected results and helped validate tool behavior under different input scenarios.



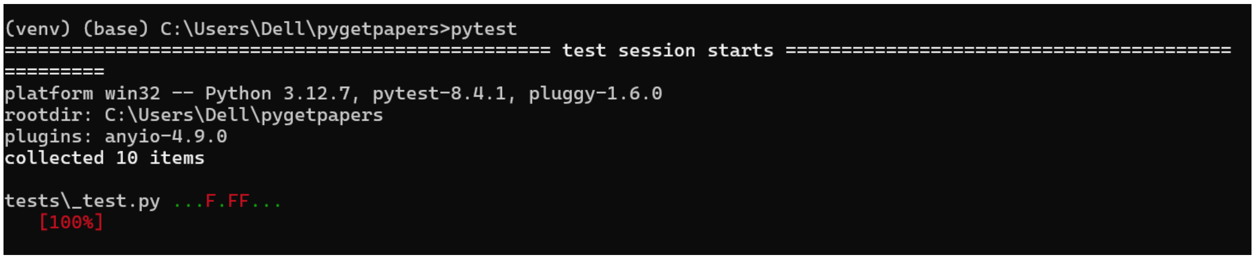
```
Ran 7 tests in 612.267s

OK

(venv) C:\Users\Dell\llmrag>
(venv) C:\Users\Dell\llmrag> coverage report -m
Name                                                 Stmts   Miss  Cover   Missing
----------------------------------------------------------------------------------
llmrag\chunking\__init__.py                              1      0   100%
llmrag\chunking\text_splitter.py                         8      0   100%
llmrag\cli.py                                           52     52     0%   1-81
llmrag\embeddings\__init__.py                            3      0   100%
llmrag\embeddings\base_embedder.py                       5      1    80%   6
llmrag\embeddings\sentence_transformers_embedder.py     12      0   100%
llmrag\generators\__init__.py                            0      0   100%
llmrag\generators\local_generator.py                    11      7    36%   16, 19-35
llmrag\main.py                                          56     56     0%   1-72
llmrag\models\__init__.py                                3      0   100%
llmrag\models\base_model.py                              5      1    80%   6
llmrag\models\transformers_model.py                      8      0   100%
llmrag\pipelines\__init__.py                             1      0   100%
llmrag\pipelines\rag_pipeline.py                        27      6    78%   19, 21, 61-64
llmrag\retrievers\__init__.py                           12      7    42%   7-14
llmrag\retrievers\base_vector_store.py                   8      2    75%   6, 10
llmrag\retrievers\chroma_store.py                       42      3    93%   48, 67-68
llmrag\retrievers\faiss_store.py                        27     27     0%   1-31
llmrag\streamlit_app.py                                 29     29     0%   1-53
----------------------------------------------------------------------------------
TOTAL                                                  310    191    38%
```

output obtained when run unittest on llmrag



```
-- Docs: https://docs.pytest.org/en/stable/how-to/capture-warnings.html
==================================================================== short test
 summary info ==================================================================
===
FAILED test/test_extract_text.py::ExtractTextTest::test_keybert_breward_1 -
IndexError: list index out of range
FAILED test/test_file.py::File0Test::test_list_children - AssertionError: as
sert ['C:\\Users\\...files\\file1'] == ['/Users/pm28..._files/file1']
FAILED test/test_file.py::File0Test::test_relative_pathname - AssertionError
: assert 'd\\e.txt' == 'd/e.txt'
FAILED test/test_html.py::CSSStyleTest::test_tinycss - ModuleNotFoundError:
No module named 'tinycss'
FAILED test/test_misc.py::ArgsTest::test_capture_errors - assert ("argument
--operation: invalid choice: 'search' (choose from 'annotate', 'counts', 'in
dex', 'no_input_styles')\n" and "argument --o...nput_styl...
FAILED test/test_wikimedia.py::WikipediaTest::test_wikipedia_page_from_wikid
ata - urllib.error.URLError: <urlopen error [WinError 10060] A connection at
tempt failed because the connected party did not properly respond after a pe
riod ...
====================================== 6 failed, 419 passed, 99 skipped, 1 x
failed, 1 warning in 2155.67s (0:35:55) ======================================
===
```

output obtained when run pytest on amilib



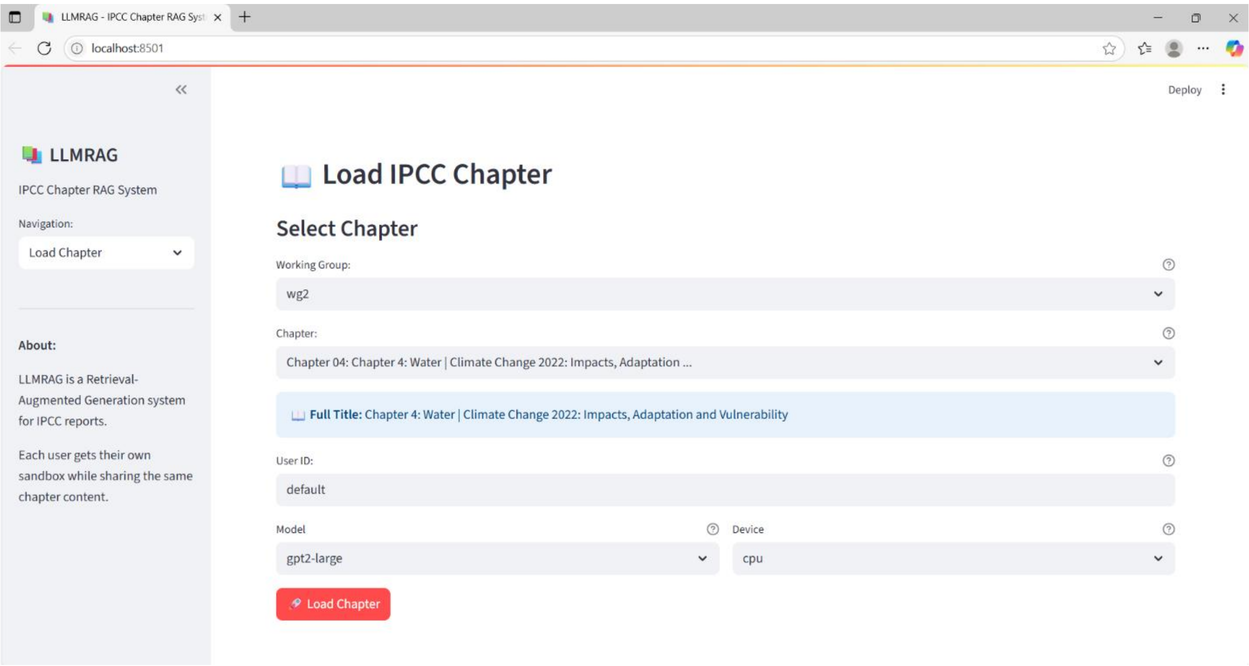
```
(venv) (base) C:\Users\Dell\pygetpapers>pytest
================================================= test session starts ========================================
=========
platform win32 -- Python 3.12.7, pytest-8.4.1, pluggy-1.6.0
rootdir: C:\Users\Dell\pygetpapers
plugins: anyio-4.9.0
collected 10 items

tests\_test.py ...F.FF...
   [100%]
```

output obtained when run pytest on pygetpapers

3.5 Streamlit chatbot debugging and collaboration

Collaborated with team to ensure that the Semantic Climate chatbot interface (built with Streamlit) worked seamlessly on Windows platforms.

- Identified platform-specific issues related to file paths and library compatibility.
- Tested code revisions locally on Windows and suggested fixes (e.g., path handling, dependency adjustments).
- Validated chatbot functionality, ensuring that users on Windows could query climate documents through the Streamlit interface.



LLMRAG

IPCC Chapter RAG System

Navigation: Load Chapter

About:

LLMRAG is a Retrieval-Augmented Generation system for IPCC reports.

Each user gets their own sandbox while sharing the same chapter content.

Load IPCC Chapter

Select Chapter

Working Group: wg2

Chapter: Chapter 04: Chapter 4: Water | Climate Change 2022: Impacts, Adaptation ...

Full Title: Chapter 4: Water | Climate Change 2022: Impacts, Adaptation and Vulnerability

User ID: default

Model: gpt2-large

Device: cpu

Semantic Climate chatbot interface

3.6 Applying Tools to Chapter 4

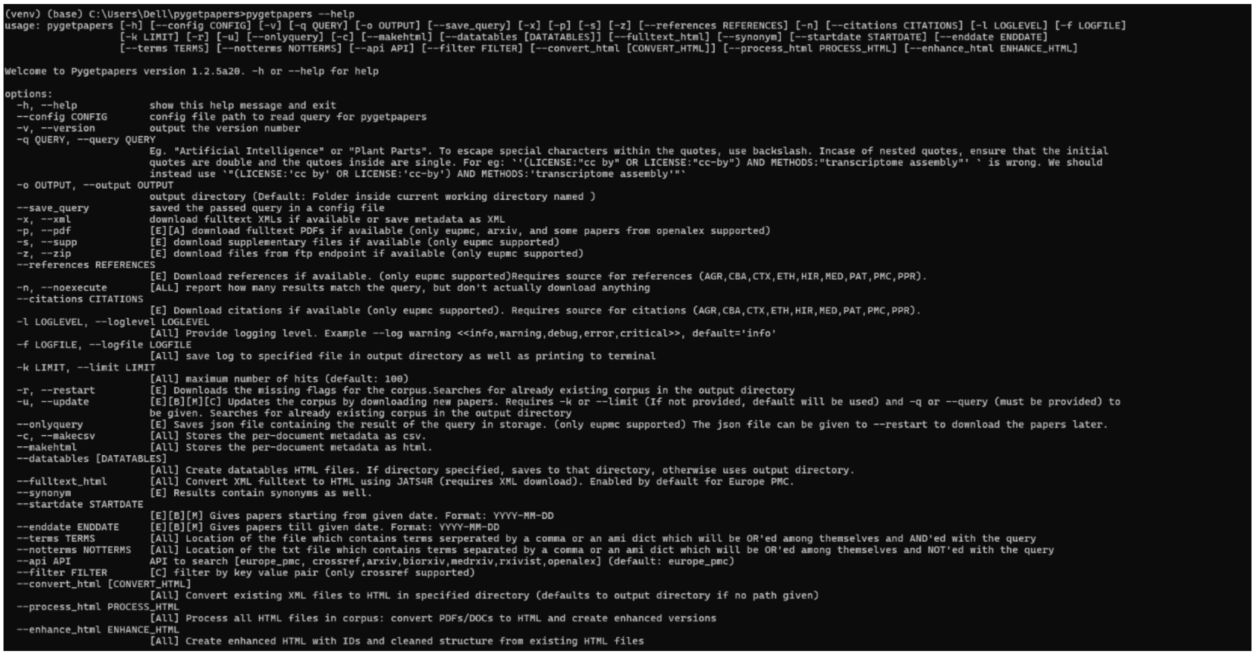
In this phase of the internship, I explored how emerging tools from the #semanticClimate ecosystem could be applied to IPCC WGII Chapter 4 ("Water") to support advanced document retrieval and interactive knowledge access. This involved using the latest versions of **Pygetpapers:** I worked with the **latest version of Pygetpapers**, a command-line tool designed for retrieving research articles from open repositories such as Europe PMC. The updated version now includes:

- Integration with DataTables: Making the results easier to navigate, filter, and analyze.
- Automatic conversion of XML to HTML: Enhancing human readability and facilitating semantic parsing.
- Improved metadata parsing: For cleaner extraction of abstracts, author details, and publication info.

Using **keywords extracted from Chapter 4**, I ran automated queries to collect relevant supplementary literature related to water, climate change, hydrology, and adaptation. This curated set of documents supports both dictionary enrichment and downstream AI analysis tasks.

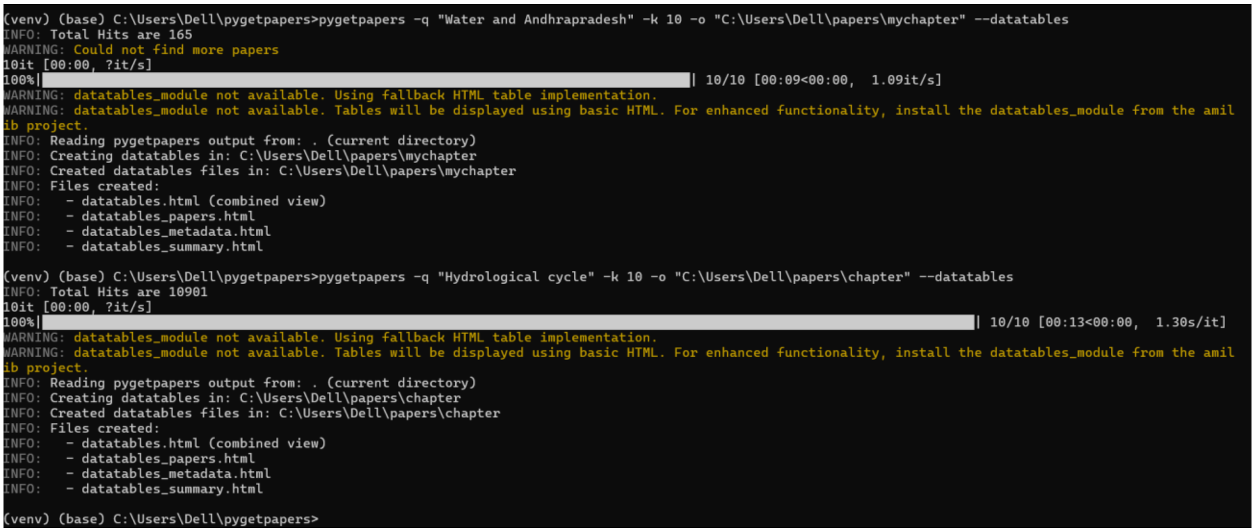
LLMRAG:

I also contributed to testing and evaluating LLMRAG, a prototype tool that combines language models with retrieval-based methods to answer user queries about climate reports and scientific articles. By applying it to Chapter 4 and, I assessed how effectively it could identify relevant passages, return accurate answers, and support context-aware dialogue.



```
(venv) (base) C:\Users\Dell\pygetpapers>pygetpapers --help
usage: pygetpapers [-h] [--config CONFIG] [-v] [-q QUERY] [-o OUTPUT] [--save_query] [-x] [-p] [-s] [-z] [--references REFERENCES] [-n] [--citations CITATIONS] [-l LOGLEVEL] [-f LOGFILE]
                   [-k LIMIT] [-r] [-u] [--onlyquery] [-c] [--makehtml] [--datatables [DATATABLES]] [--fulltext_html] [--synonym] [--startdate STARTDATE] [--enddate ENDDATE]
                   [--terms TERMS] [--notterms NOTTERMS] [--api API] [--filter FILTER] [--convert_html [CONVERT_HTML]] [--process_html PROCESS_HTML] [--enhance_html ENHANCE_HTML]

Welcome to Pygetpapers version 1.2.5a20. -h or --help for help

options:
  -h, --help            show this help message and exit
  --config CONFIG       config file path to read query for pygetpapers
  -v, --version         output the version number
  -q QUERY, --query QUERY
                        Eg. "Artificial Intelligence" or "Plant Parts". To escape special characters within the quotes, use backslash. Incase of nested quotes, ensure that the initial
                        quotes are double and the qutoes inside are single. For eg: `'(LICENSE:"cc by" OR LICENSE:"cc-by") AND METHODS:"transcriptome assembly"' ` is wrong. We should
                        instead use `"(LICENSE:'cc by' OR LICENSE:'cc-by') AND METHODS:'transcriptome assembly'"`
  -o OUTPUT, --output OUTPUT
                        output directory (Default: Folder inside current working directory named )
  --save_query          saved the passed query in a config file
  -x, --xml             download fulltext XMLs if available or save metadata as XML
  -p, --pdf             [E][A] download fulltext PDFs if available (only eupmc, arxiv, and some papers from openalex supported)
  -s, --supp            [E] download supplementary files if available (only eupmc supported)
  -z, --zip             [E] download files from ftp endpoint if available (only eupmc supported)
  --references REFERENCES
                        [E] Download references if available. (only eupmc supported)Requires source for references (AGR,CBA,CTX,ETH,HIR,MED,PAT,PMC,PPR).
  -n, --noexecute       [ALL] report how many results match the query, but don't actually download anything
  --citations CITATIONS
                        [E] Download citations if available (only eupmc supported). Requires source for citations (AGR,CBA,CTX,ETH,HIR,MED,PAT,PMC,PPR).
  -l LOGLEVEL, --loglevel LOGLEVEL
                        [All] Provide logging level. Example --log warning <<info,warning,debug,error,critical>>, default='info'
  -f LOGFILE, --logfile LOGFILE
                        [All] save log to specified file in output directory as well as printing to terminal
  -k LIMIT, --limit LIMIT
                        [All] maximum number of hits (default: 100)
  -r, --restart         [E] Downloads the missing flags for the corpus.Searches for already existing corpus in the output directory
  -u, --update          [E][B][M][C] Updates the corpus by downloading new papers. Requires -k or --limit (If not provided, default will be used) and -q or --query (must be provided) to
                        be given. Searches for already existing corpus in the output directory
  --onlyquery           [E] Saves json file containing the result of the query in storage. (only eupmc supported) The json file can be given to --restart to download the papers later.
  -c, --makecsv         [All] Stores the per-document metadata as csv.
  --makehtml            [All] Stores the per-document metadata as html.
  --datatables [DATATABLES]
                        [All] Create datatables HTML files. If directory specified, saves to that directory, otherwise uses output directory.
  --fulltext_html       [All] Convert XML fulltext to HTML using JATS4R (requires XML download). Enabled by default for Europe PMC.
  --synonym             [E] Results contain synonyms as well.
  --startdate STARTDATE
                        [E][B][M] Gives papers starting from given date. Format: YYYY-MM-DD
  --enddate ENDDATE     [E][B][M] Gives papers till given date. Format: YYYY-MM-DD
  --terms TERMS         [All] Location of the file which contains terms serperated by a comma or an ami dict which will be OR'ed among themselves and AND'ed with the query
  --notterms NOTTERMS   [All] Location of the txt file which contains terms separated by a comma or an ami dict which will be OR'ed among themselves and NOT'ed with the query
  --api API             API to search [europe_pmc, crossref,arxiv,biorxiv,medrxiv,rxivist,openalex] (default: europe_pmc)
  --filter FILTER       [C] filter by key value pair (only crossref supported)
  --convert_html [CONVERT_HTML]
                        [All] Convert existing XML files to HTML in specified directory (defaults to output directory if no path given)
  --process_html PROCESS_HTML
                        [All] Process all HTML files in corpus: convert PDFs/DOCs to HTML and create enhanced versions
  --enhance_html ENHANCE_HTML
                        [All] Create enhanced HTML with IDs and cleaned structure from existing HTML files
```

Setting up pygetpapers using command line



```
(venv) (base) C:\Users\Dell\pygetpapers>pygetpapers -q "Water and Andhrapradesh" -k 10 -o "C:\Users\Dell\papers\mychapter" --datatables
INFO: Total Hits are 165
WARNING: Could not find more papers
10it [00:00, ?it/s]
100%|██████████████████████████████████████████████████████████| 10/10 [00:09<00:00,  1.09it/s]
WARNING: datatables_module not available. Using fallback HTML table implementation.
WARNING: datatables_module not available. Tables will be displayed using basic HTML. For enhanced functionality, install the datatables_module from the amil
ib project.
INFO: Reading pygetpapers output from: . (current directory)
INFO: Creating datatables in: C:\Users\Dell\papers\mychapter
INFO: Created datatables files in: C:\Users\Dell\papers\mychapter
INFO: Files created:
INFO:   - datatables.html (combined view)
INFO:   - datatables_papers.html
INFO:   - datatables_metadata.html
INFO:   - datatables_summary.html

(venv) (base) C:\Users\Dell\pygetpapers>pygetpapers -q "Hydrological cycle" -k 10 -o "C:\Users\Dell\papers\chapter" --datatables
INFO: Total Hits are 10901
10it [00:00, ?it/s]
100%|██████████████████████████████████████████████████████████████████████████████████████| 10/10 [00:13<00:00,  1.30s/it]
WARNING: datatables_module not available. Using fallback HTML table implementation.
WARNING: datatables_module not available. Tables will be displayed using basic HTML. For enhanced functionality, install the datatables_module from the amil
ib project.
INFO: Reading pygetpapers output from: . (current directory)
INFO: Creating datatables in: C:\Users\Dell\papers\chapter
INFO: Created datatables files in: C:\Users\Dell\papers\chapter
INFO: Files created:
INFO:   - datatables.html (combined view)
INFO:   - datatables_papers.html
INFO:   - datatables_metadata.html
INFO:   - datatables_summary.html

(venv) (base) C:\Users\Dell\pygetpapers>
```

Applying pygetpapers to chapter4

3.7 Contribution to FSCI 2025 Activities

As part of Semantic Climate's involvement in the FORCE11 Scholarly Communication Institute (FSCI), I contributed to participant support and educational resources.

- Prepared general instructions for FSCI participants to engage with Semantic Climate tools.
- Summarized the Cochrane AI RAISE guidance in a simplified PDF format for broader accessibility.
- Created a video presentation summarizing the content and structure of IPCC WGII Chapter 4, intended as an educational asset for newcomers to climate policy and research.

3.8 Documentation of workflow on github

To promote transparency, reproducibility, and collaborative learning, I documented my entire workflow on GitHub, in line with Semantic Climate's commitment to open science and community knowledge sharing.

- Created detailed Markdown guides for setting up the environment, executing keyword extraction and dictionary creation scripts, and interpreting outputs.

- Included code snippets, command-line examples, and troubleshooting notes to ensure that others could replicate and learn from the process.
- Maintained a daily progress log, capturing all tasks completed during the internship, including technical milestones, testing updates, and collaborative contributions.
- Ensured that the documentation was consistent with Semantic Climate's values of openness, accessibility, and transparent communication.

This repository serves as both a technical guide and a personal learning journal, reflecting my end-to-end contributions to the project.

4. Key Deliverables

All code, scripts, visualizations, and documentation produced during the internship are available in the Semantic Climate repository:

4.1 Contents in Github Repository:

Repository Path: semanticClimate/internship_sC (branch: Haarthi)

semanticClimate/internship_sC at Haarthi

The repository includes a well-organized collection of folders and files representing the full scope of the work carried out during the internship. Key contents include:

- Files related to extraction of wordlist in the keyword_extraction folder
- Final wordlist of the chapter which contains more than 150 words (IPCC-Ch04-Wordlist)
- Python code used to extract the keywords(keyword_extraction.py) HTML file used to create the dictionary of the chapter(wg2chap04_dict.html)
- Graphviz pipeline and obtained svg file in the graphviz folder.
- Setup and instructions for the keyword extraction and dictionary creation process are mentioned in the readme file.
- General instructions made for FSCI participants are in the FSCI_instructions.md
- My daily progress throughout the internship is mentioned in the daily_progress.md

4.2 Links to the outputs:

Link to the wordlist extracted:

internship_sC/Keyword_Extraction/IPCC-Ch04-Wordlist at Haarthi · semanticClimate/internship_sC

Link to the dictionary created in HTML:

internship_sC/wg2chap04_dict.html at Haarthi · semanticClimate/internship_sC

Link to knowledge graph created using graphviz:

internship_sC/graphviz at Haarthi · semanticClimate/internship_sC

Link to the readme file containing the keyword extraction and dictionary creation process:

internship_sC/Keyword_Extraction_Process.md at Haarthi · semanticClimate/internship_sC

Link to the markdowns of general instructions and my daily progress:

internship_sC/FSCI_Instructions.md at Haarthi · semanticClimate/internship_sC

internship_sC/Daily_Progress.md at Haarthi · semanticClimate/internship_sC

Link to the video presentation:

https://github.com/semanticClimate/assisted-literature
review/blob/main/presentations/video_presentation.md

5.Tools & Technologies

To achieve the project's goals of transforming static IPCC reports into dynamic, structured, and accessible resources, a range of tools and technologies were utilized. These tools supported tasks such as keyword extraction, dictionary generation, document conversion, and testing.

Tools used to accomplish the aforementioned objectives include, but are not limited to:

- ***pygetpapers:*** A valuable text mining tool that interacts with open access scientific repositories, systematically acquiring relevant articles. It comes with the packages pygetpapers and download tools which provide various functions to download, process and save research papers and their metadata. [2].
- ***docanalysis:*** is a command-line tool that processes document collections and performs text analysis. It uses custom code along with Python tools like NLTK, and it can use spaCy or scispaCy for extracting and annotating entities. The tool creates summary data and word lists as output. [3].
- ***amilib:*** amilib has tools for finding, cleaning, converting, searching, republishing legacy documents. It is a Python library designed for document processing,and dictionary creation. We can create dictionaries using amilib from existing set of words. The library simplifies data extraction and manipulation, offering a user-friendly interface for processing data formats like HTML and XML. It ensures that complex operations like term marking and dictionary building can be performed with minimal coding effort.[4].

A non-exhaustive list of software and technologies critical to our functioning:

- Git: A distributed version control system that tracks versions of files.
- GitHub: A developer platform used extensively for hosting our code. It is accessible to the public and easy to collaborate on.
- Python: A powerful high-level programming language that is ideal for scripting and rapid application development.
- VS Code or PyCharm: A highly efficient Integrated Development Environment (IDE) that allows coding, testing and debugging for software development.
- Google Colab: A hosted Jupyter Notebook service that requires no setup to use and provides free of charge access to computing resources, including GPUs and TPUs.
- Slack: A cloud-based team platform for seamless communication and coordination

Importance of semantic toolkit: The semantic toolkit developed by #semanticClimate plays a central role in enhancing the accessibility and utility of climate science reports. IPCC documents are often published in dense, PDF formats that are difficult to search, analyze, or link contextually. The semantic toolkit addresses this by:

- Structuring content semantically (e.g., converting PDF to HTML),
- Extracting high-value keywords and terms,
- Linking those terms to open knowledge bases (such as Wikidata or Wikipedia),
- Generating visual navigational aids like knowledge graphs,
- Creating contextual dictionaries to aid in interpretation.

Ultimately, the semantic toolkit empowers researchers, educators, and the public to interact with climate science in more meaningful and accessible ways[5].

6. Conclusion

The Eight weeks of my internship with #SemanticClimate have been an enriching and insightful journey, blending technical skill-building with meaningful contributions to climate science. I gained hands-on experience with open science tools such as Pygetpapers, amilib and worked on the semantic processing of Chapter 4 from the IPCC Working Group II report.

My contributions included creating a structured keyword wordlist, developing a contextual dictionary, and designing a knowledge graph to support semantic search and accessibility. I also documented the extraction and dictionary-building process in Markdown for GitHub, and applied document retrieval tools to Chapter 4 using Pygetpapers and LLMRAG. Additionally, I participated in alpha testing of core tools, collaborated on interface testing for a chatbot prototype, and created participant-facing resources and a video presentation for FSCI 2025.

This internship has significantly strengthened my technical skills and deepened my understanding of how open, structured scientific knowledge can support climate action. It has provided me with valuable exposure to real-world research workflows, collaborative open-source practices, and the intersection of technology and environmental science.

7. References

[1] IPCC Sixth Assessment Report (WGII – Chapter 4). https://www.ipcc.ch/report/ar6/wg2/

[2] pygetpapers – GitHub Repository. https://github.com/petermr/pygetpapers

[3] docanalysis – GitHub Repository. https://github.com/petermr/docanalysis

[4] amilib – GitHub Repository. https://github.com/petermr/amilib

[5] semanticClimate . https://semanticclimate.github.io/p/en/