

REPORT

Extracting Knowledge from Scientific Literature and Climate report

Submitted by:

Saurav Mishra

SUMMER RESEARCH FELLOW

FLFS47



Under the Supervision of

Dr. Gitanjali Yadav
Staff Scientist VI
NIPGR, New Delhi

Dr. Suneel Kateriya
Professor SBT
JNU, New Delhi

Table of Content

1. Introduction	3
(A)Extracting Data and Knowledge from Climate report	3
(B)Automated Literature Review(ALR)	3
(C)System Biology Analysis of CO2 Fixation, Sequestration and Concentration	3
2. Objective	4
(A) Extracting Knowledge from Climate report	4
Summary (Chapter 6 of the IPCC AR6 WGIII: “Energy Systems”)	4
(B) Automated Literature Review (ALR)	6
(C)System Biology Analysis of CO2 Fixation, Sequestration and Concentration	9
3. Tools and Technologies Used	10
4. Skills Acquired	11
5. References	11
6. Plan for the Next Month	11
7. Conclusion	12
8. Acknowledgements	12

1. Introduction

(A)Extracting Data and Knowledge from Climate report

This study project is a collaborative initiative focused on extracting structured climate knowledge from large textual sources, particularly the Intergovernmental Panel on Climate Change (IPCC) reports. These reports contain critical scientific assessments on climate change, but they are written in dense, unstructured natural language, making it difficult for machines and researchers to quickly extract and query relevant knowledge.

In this project we apply tools from Natural Language Processing (NLP), machine learning, and ontology engineering to convert textual climate data into structured, machine-readable formats such as knowledge graphs or dictionaries. This enables improved searchability, automated reasoning, and data reuse in climate research and policymaking.

(B)Automated Literature Review(ALR)

Automated Literature Review (ALR) is a computational approach to systematically collect, process, and analyze large volumes of scientific literature using tools such as natural language processing (NLP), keyword extraction, and structuring. It reduces manual effort by identifying relevant information, extracting key terms, and organizing knowledge into structured formats like knowledge graphs. In this project, ALR was applied to query terms to build a dict that supports further analysis and understanding of climate change terminology.

(C)System Biology Analysis of CO₂ Fixation, Sequestration and Concentration

Systems biology is an interdisciplinary field that integrates biology, computation, and mathematics to study complex interactions within biological systems as a whole, rather than focusing on individual components. In the context of CO₂ fixation, sequestration, and concentration, systems biology enables a holistic understanding of how organisms such as plants, microalgae, and cyanobacteria capture, convert, and store atmospheric carbon dioxide.

Using high-throughput data (genomics, proteomics, transcriptomics) and computational tools such as Protein–Protein Interaction (PPI) networks, metabolic modeling, and gene regulatory networks, researchers can analyze the key proteins, enzymes, pathways, and environmental factors involved in carbon assimilation.

2. Objective

(A) Extracting Knowledge from Climate report

During the first month of the FAST-SF fellowship, my primary focus was on understanding the objectives of the project and beginning hands-on work with tools for knowledge extraction. My tasks and contributions included:

Method Used

To extract structured knowledge from the IPCC climate report (Chapter 6: *Energy Systems*), an Automated Literature Review (ALR) approach was employed. The methodology combined text extraction, natural language processing (NLP), and structuring. Specifically, the pipeline involved:

- Using Python scripts for keyword extraction and preprocessing,
- Organizing the extracted terms into a reusable and machine-readable format for further analysis.

This approach ensured efficient handling of large-scale technical documents and reduced manual bias in keyword selection.

Workflow

- Document Preparation
- Keyword Extraction
- Structuring
- Dictionary Output

As part of this project, I am working on:

- Extracting key concepts and entities from Chapter 6 of the IPCC AR6 WGIII report.
- Building a domain-specific dictionary([Fig.i](#)) using tools such as Pyget papers and Amilib.
- Exploring how these extracted terms can support the development of a climate encyclopedia

In this topic I worked on **Chapter 6 of the IPCC AR6 WGIII**

Summary (Chapter 6 of the IPCC AR6 WGIII: “Energy Systems”)

This chapter discusses how global energy systems contribute to greenhouse gas emissions and explores pathways for transforming energy production, distribution, and consumption to reduce emissions. It covers topics such as:

- Renewable energy integration (solar, wind, hydro, etc.),
- Electrification of end-use sectors,
- Energy efficiency improvements,
- Fossil fuel phase-out strategies,
- Role of hydrogen, bioenergy, and carbon capture.

It highlights that the energy sector is the largest contributor to global greenhouse gas emissions, mainly due to the combustion of fossil fuels for electricity, heat, and transportation. The chapter explores mitigation strategies such as a rapid shift to renewable energy sources, electrification of end-use sectors (like transport and industry), improvements in energy efficiency, and the deployment of emerging technologies like green hydrogen, carbon capture and storage (CCS), and advanced nuclear

Result

WordList extracted from **Chapter 6 of the IPCC AR6 WGIII**

- agri voltaics
- green hydrogen
- Photobioreactor
- carbon sequestration
- Microgrid
- intermittent energy
- stratospheric cooling
- resilience building
- green ammonia
- thermal cracking

Mini climate encyclopedia created([Fig:1](#))



<p>search term: agrivoltaics Wikipedia Page</p> <p>Agrivoltaics (agrophotovoltaics, agrisolar, or dual-use solar) is the dual use of land for solar energy and agriculture ^{[2][3][4]} The technique was first conceived by Adolf Goltzberger and Armin Zastrow in 1981. ^[5]</p>  <p>Vertical solar panels, east to west orientation, with bifacial modules near Donaueschingen, Germany ^[1]</p>
<p>search term: appliances Wikipedia Page</p> <p>Appliance may refer to:</p>
<p>search term: battery storage Wikipedia Page</p> <p>A battery energy storage system (BESS), battery storage power station, battery energy grid storage (BEGS) or battery grid storage is a type of energy storage technology that uses a group of batteries in the grid to store electrical energy. Battery storage is the fastest responding dispatchable source of power on electric grids, and it is used to stabilise those grids, as battery storage can transition from standby to full power in under a second to deal with grid contingencies ^[1]</p>  <p>Tehachan Energy Storage Project, Tehachan, California</p>

Fig. i Mini climate Encyclopedia

Wordlist:

https://github.com/semanticClimate/internship_sC/blob/Saurav/Saurav_Work/Wordlist_chapter_06.txt

Mini climate Encyclopedia:

https://github.com/semanticClimate/internship_sC/blob/Saurav/Saurav_Work/ch6_keywordup.html

(B) Automated Literature Review (ALR)

Automated Literature Review (ALR) is a computational approach to systematically collect, process, and analyze large volumes of scientific literature using tools such as natural language processing (NLP), keyword extraction, and structuring. It reduces manual effort by identifying relevant information, extracting key terms, and organizing knowledge into structured formats like dictionaries or knowledge graphs. In this project, ALR was applied to climate-related texts that support further analysis and understanding of climate change . We will get wordcloud for Countries ([Fig:ii](#)), wordcloud for Drugs ([Fig:iii](#)), wordcloud for Plants ([Fig:iv](#))

Workflow

Tools Used	Search Query	Code Used
Pygetpapers	<p>Query : Bio Fuel and Carbon Fixation</p> <p>No.of article on: Bio Fuel and Carbon Fixation</p>	<pre>!pygetpapers --query “Bio fuel” AND “Carbon Fixation” - -xml -n --output chapter6 -- save_query</pre> <pre>!pygetpapers --query "'Carbon fixation" AND "Bio fuel"' --xml -n --startdate "2010-01-01" --enddate "2025-04-01" --output medplant --save_query</pre>
Docanalysis	<p>List of countries Specific to INT(introduction),RES(result),CONS(conclusion) and DIS(discussion)</p> <p>Displaying Datatables</p> <p>List of DRUG</p>	<pre>[!python -m docanalysis.docanalysis \ --project_name /content/Pyrolysis \--make_section \ --search_section INT, RES, CON, DIS \--dictionary COUNTRY --output pyrolysis_Country.csv]</pre> <pre>from IPython.core.display import display, HTML # Path to the HTML file html_file_path = '/content/ch6.html' # Read the HTML file with open(html_file_path, 'r', encoding='utf-8') as file: html_content = file.read() # Display the HTML content display(HTML(html_content))</pre> <pre>python -m docanalysis.docanalysis ^ --project_name biofuel_result ^ --make_section ^ --search_section INT,RES,CON,DIS ^ --dictionary DRUG ^ --output biofuel_drugT.csv</pre>

Result

Total no.of article found: 19

List of Countries:

- India
- Canada
- Ukraine
- Cambodia
- Turkey
- Brazil
- Poland
- Argentina
- Romania
- Nigeria

Wordcloud of countries:



Fig:ii

CSV file:

https://github.com/semanticClimate/internship_sC/blob/Saurav/Saurav_Work/biofuel_country.csv

List of DRUG:

- Acid
- Ethanol
- Succinic
- Eicosapentaenoic

Wordcloud For Drugs:



Fig:iii- Figure showing extracted Drugs from literature

CSV file:

https://github.com/semanticClimate/internship_sC/blob/Saurav/Saurav_Work/biofuel_DRUG.csv

(C)System Biology Analysis of CO₂ Fixation, Sequestration and Concentration

As part of the assigned project under the guidance of Dr. Katariya, I initiated a comprehensive exploration of the systems biology behind carbon dioxide fixation and its regulation at the molecular level. My work included the following:

- Studied the Calvin Cycle in depth, understanding each biochemical reaction, key enzymes like RuBisCO, PEP carboxylase, and cofactors (ATP, NADPH), and how these contribute to CO₂ fixation in plants and algae.
- Explored Protein–Protein Interaction (PPI) networks relevant to CO₂ fixation using tools like STRING ([Fig.V](#)) Focused particularly on the model microalga *Chlamydomonas reinhardtii*, analyzing interactions of carbon-fixing proteins such as rbcL and carbonic anhydrase.
- Studied recent advancements in AI-generated proteins, understanding how computational tools are used to design functional proteins that can potentially improve enzyme efficiency or CO₂-binding properties.

Result

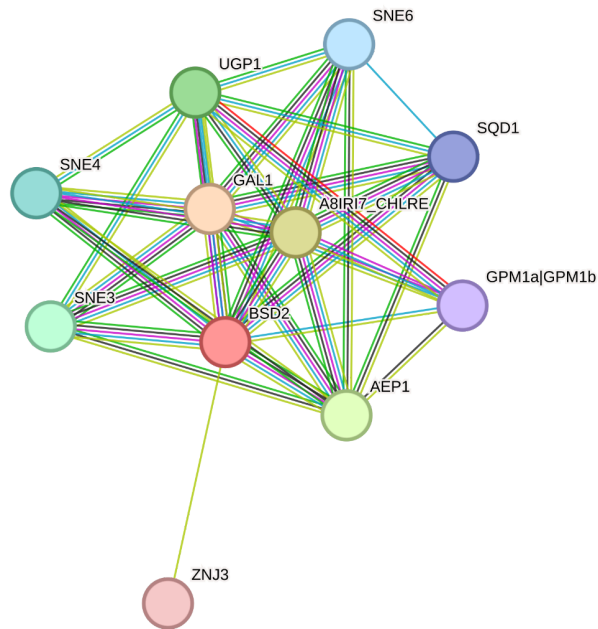


Fig.V: Protein–Protein Interaction (PPI) Network Centered on RuBisCO in *Chlamydomonas reinhardtii*.

3. Tools and Technologies Used

- Python : for scripting, text processing, and keyword extraction.
- Amilib :for extraction and dictionary generation.

- GitHub : for code version control and collaboration.
- Streamlit : basic understanding of web interface for data exploration.
- pygetpapers: A valuable text mining tool that interacts with open access scientific repositories, systematically acquiring relevant articles.
- docanalysis: An integrated suite of open-source Command Line tools that enables users to download scientific literature from europepmc.org based on specific query criteria
- VS Code or PyCharm: A highly efficient Integrated Development Environment (IDE) that allows coding, testing and debugging for software development.
- Slack: A cloud-based team platform for seamless communication and coordination.
- Git: A distributed version control system that tracks versions of files.
- stringDb: A software or database which shows the interaction between proteins or enzymes

4. Skills Acquired

Over the course of the month, I developed new skills in the following areas:

- Working with scientific documentation formats (HTML, Markdown, LaTeX).
- Understanding and applying basic Natural Language Processing techniques.
- Installing and using domain-specific tools (pygetpapers, amilib).
- Contributing to open-science repositories.
- Collaborating in a research environment.
- Initial exposure to ontology-based modeling and web technologies.
- Reading scientific literature and identifying key terminologies.

5. References

StringDB: <https://string-db.org/>

Google Collab ALR:

<https://colab.research.google.com/drive/1JI4efY1rBnZxRM1qSqVZdPRwQo0OICDH?usp=sharing>

GIT HUB: https://github.com/semanticClimate/internship_sC/tree/Saurav/Saurav_Work

IPCC : <https://www.ipcc.ch/report/ar6/wg3/>

6. Plan for the Next Month

In the upcoming weeks, I plan to:

- Begin integration with algal CO₂ fixation concepts into the pipeline.

- Work on a prototype system for querying climate data using the dictionary.
- Continue contributing to collaborative open-science projects under Dr. Gitanjali Yadav's guidance.
- Explore protein-protein interaction networks using model organisms like *Chlamydomonas reinhardtii* under Dr. Suneel Kateriya JNU

7. Conclusion

During the internship, I extensively used Git as a version control system. This helped me get skilled in managing code changes, collaborating with team members, and maintaining a clean and efficient codebase. I was actively involved in daily testing and debugging sessions, which included running and creating unit tests for our open-source tool, amilib. Overall, this internship was a great learning experience. It provided me with practical skills in version control, testing, documentation and open-source technologies, and Python programming. The integration of Automated Literature Review (ALR) and provides a powerful framework for systematically analyzing and organizing climate-related scientific information. ALR enables the efficient extraction of key terms and concepts from large volumes of climate literature, reducing manual effort while ensuring comprehensive coverage. This approach builds upon this by structuring the extracted terms into a meaningful and navigable dictionary, facilitating knowledge discovery, interoperability, and reuse. Together, these tools enhance our understanding of complex climate systems and support data-driven research, policy analysis, and innovation in climate mitigation and adaptation strategies.

8. Acknowledgements

I sincerely thank Dr. Gitanjali Yadav and Dr. Suneel Kateriya for their mentorship, continuous guidance, and encouragement. I also acknowledge the Indian Academy of Sciences for awarding me the FAST-SF fellowship and giving me the opportunity to work at NIPGR, New Delhi. Special thanks to Dr. Peter Murray-Rust and Dr. Renu Kumari for their mentorship and the support throughout the internship