**SEMANTIFYING IPCC REPORTS**

IPCC reports have data about climate change that is required to save this world.

All this information is locked away in pdf’s with more than 1000 pages and numerous images and tables that are neither machine friendly nor human friendly.

But, a machine can read all these pdf’s in less time than a human and extract the data with proper tools.

We are creating tools to hack this large data and extract useful and meaning full structured knowledge from these reports and make them available for all.

* variety of technologies
* aimed at a wide variety of sectors
* simple (simpler than IPCC report) - content and the distribution
* comparison/ exploration of alternative tools (PDF2HTML)
* dealing with moving target (draft and final versions)

**deliverables**

* researcher report book (example, template)

**IPCC message**

* financial implications

**what I learnt**

* coding skills
* working in groups

**Chapter analysis**

ipcc reports are divided into chapters, each chapter focusing on an aspect that is related to climate change.

Chapter15: “finance and investment”

Deals with financial implications in reducing the climate change.

Chapter link (<https://github.com/petermr/semanticClimate/blob/27e0e069729a75947d30dd8941b7d1ad4685ecaa/ipcc/ar6/wg3/Chapter15/fulltext.pdf>)

**Chapter analysis** flow

* **Converting the chapter15 pdf into html**

We are developing a tool called pyami for this conversion.

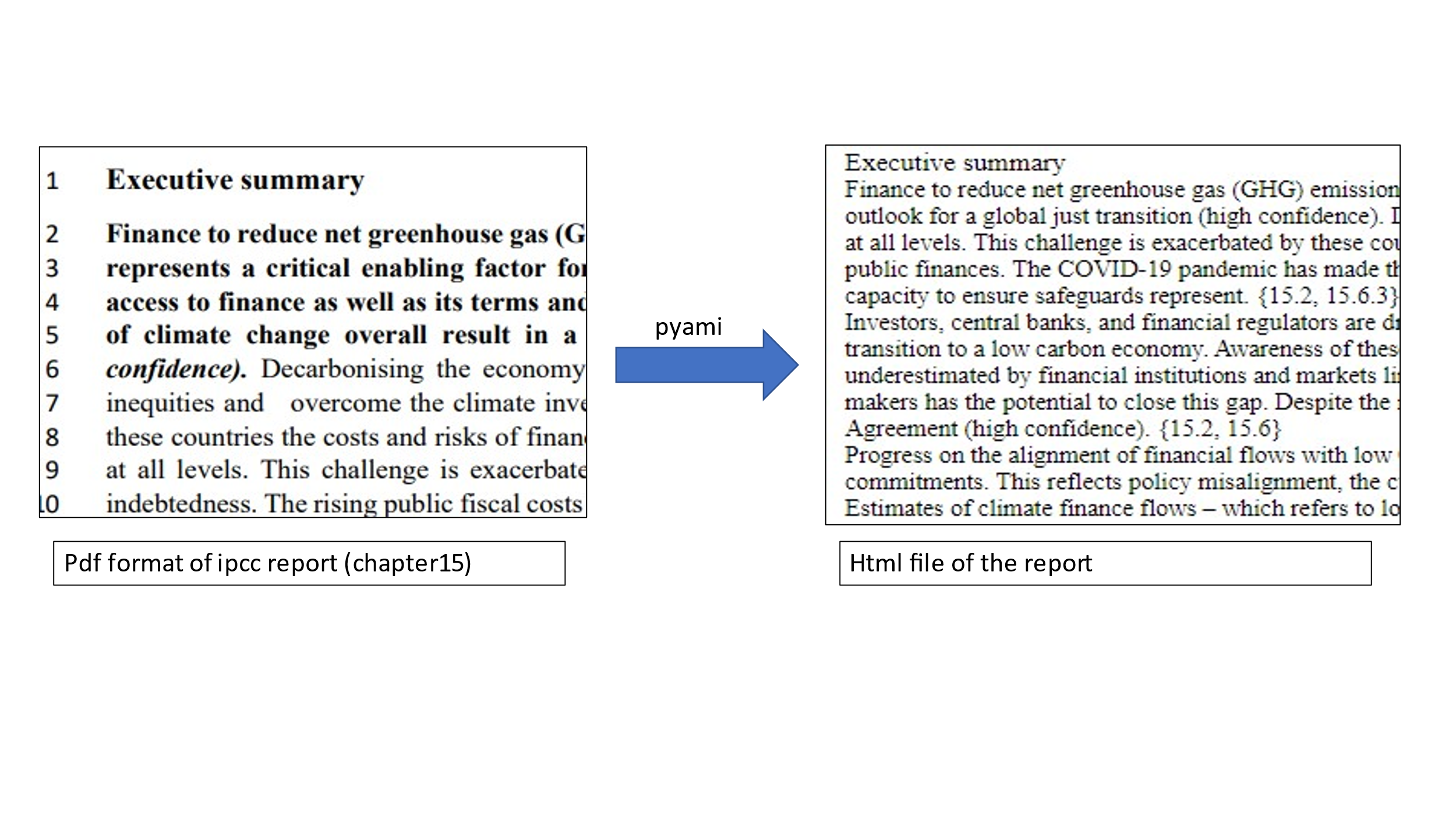
Language used for the tool – python

Initially the tests are written for a mac environment

Goal- to optimise the tool for all the systems (windows, mac, unix, linux )

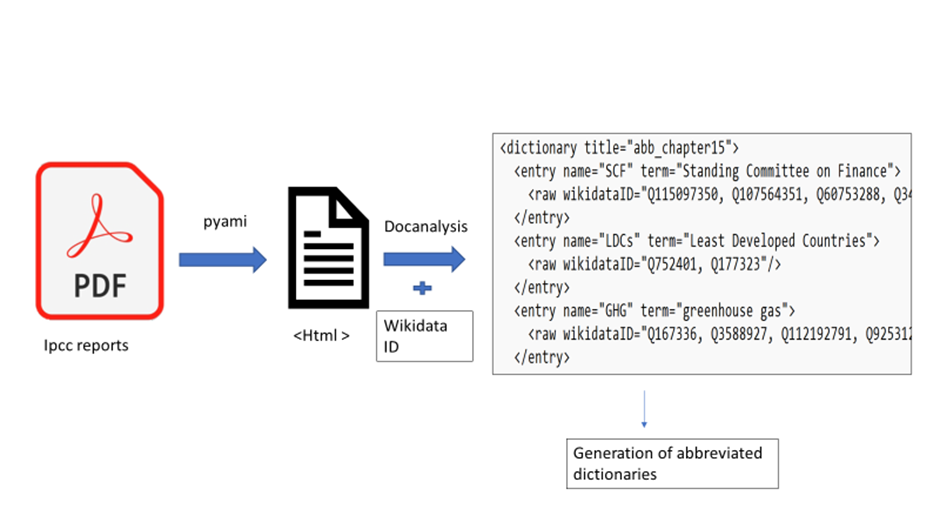
**Pyami –** this is a semantic reader for scientific literature. It reads the documents in bulk ,converts and helps us analysing it.

Html file of chapter15: (<https://github.com/petermr/semanticClimate/blob/27e0e069729a75947d30dd8941b7d1ad4685ecaa/ipcc/ar6/wg3/Chapter15/fulltext.flow.html>)



(We use dictionaries and keywords to annotate the html file of the chapter so that they should be able to represent the available metadata of the keywords,abbrevations)

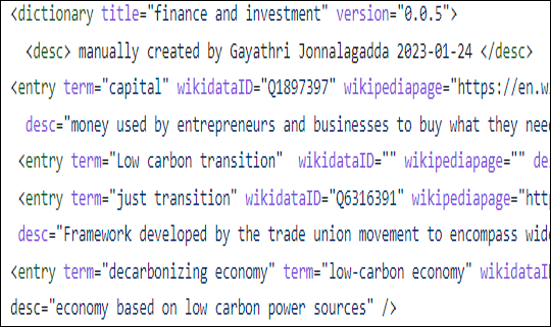
* **Docanalysis**

This command line tool helps in text mining and generating abbreviated dictionaries from the HTML file of the chapter.Abbreviation dictionary of chapter15: (<https://github.com/petermr/semanticClimate/blob/27e0e069729a75947d30dd8941b7d1ad4685ecaa/ipcc/ar6/wg3/Chapter15/dict/ipcc_chapter15_abb_dict.xml>)

* **Manual dictionary (chapter/climate related terms)**

Terms/words that are felt important or difficult and picked from the chapter by manually reading the chapter and created an xml dictionary file with entries like

* Term
* WikidataId
* Definition
* Wikidatapage (link)

Manual dictionary of chapter 15 : (<https://github.com/petermr/semanticClimate/blob/27e0e069729a75947d30dd8941b7d1ad4685ecaa/ipcc/ar6/wg3/Chapter15/dict/ipcc_chapter15_man_dict.xml>)

* **Keyword extraction**

Extracting keywords/keyphrases from the HTML file using methods

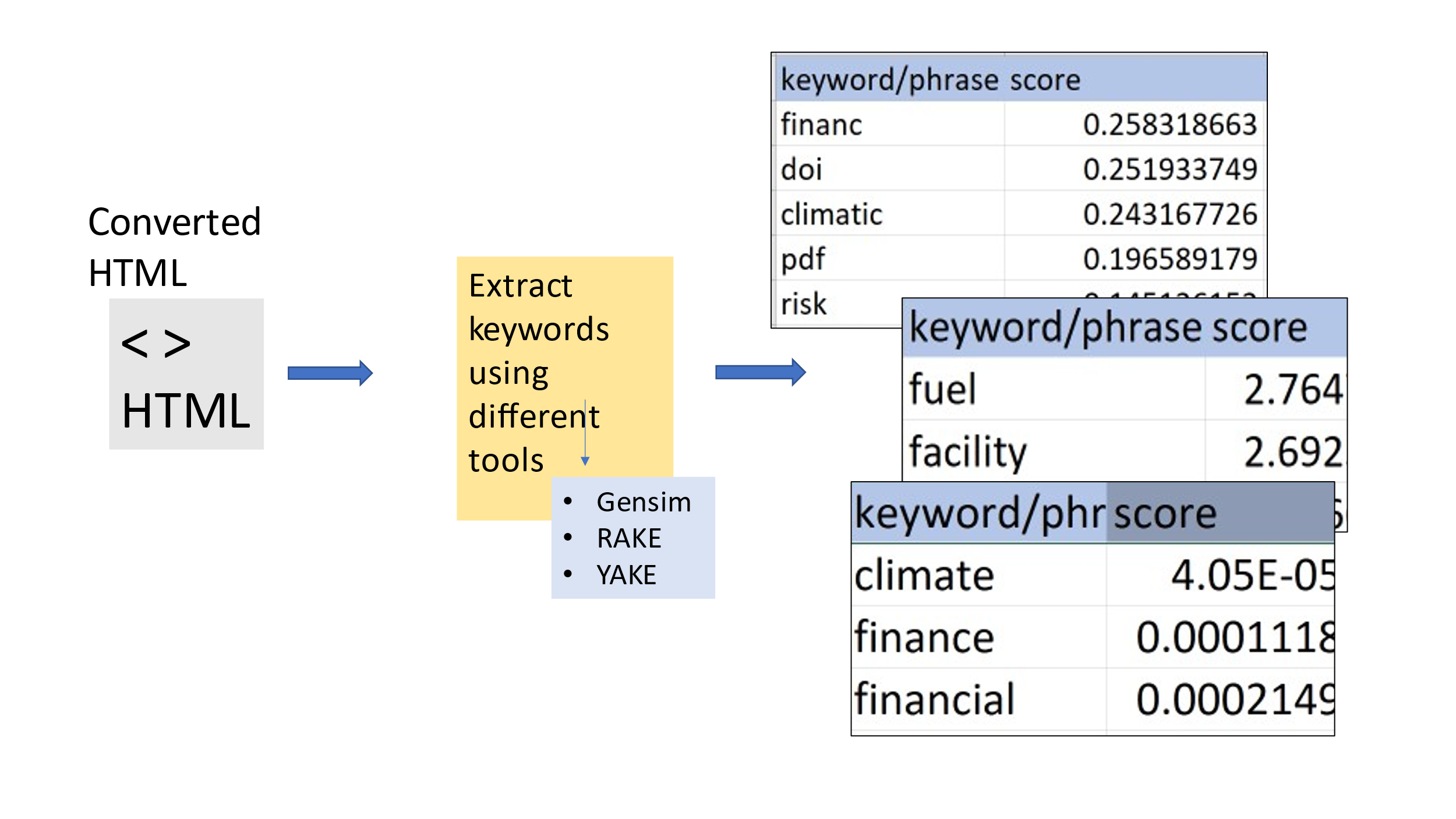
* **Genism** : is an open-source library for natural language processing, using modern statistical machine learning

(<https://github.com/petermr/semanticClimate/blob/27e0e069729a75947d30dd8941b7d1ad4685ecaa/ipcc/ar6/wg3/Chapter15/raw/Rake_keywords.csv>)

* **Rake** : short for Rapid Automatic Keyword Extraction algorithm, is a domain independent keyword extraction algorithm which tries to determine key phrases in a body of text by analyzing the frequency of word appearance and its co-occurance with other words in the text.

(<https://github.com/petermr/semanticClimate/blob/27e0e069729a75947d30dd8941b7d1ad4685ecaa/ipcc/ar6/wg3/Chapter15/raw/Rake_keywords.csv>)

* **Yake** : Yet another keyword extractor, light-weight unsupervised automatic keyword extraction method which rests on text statistical features extracted from single documents to select the most important keywords of a text.

(<https://github.com/petermr/semanticClimate/blob/27e0e069729a75947d30dd8941b7d1ad4685ecaa/ipcc/ar6/wg3/Chapter15/raw/yake_keywords.csv>)

Readme files for different sections of chapter are also created

* Introduction
* Executive summary
* Table of contents
* Images
* Chapter FAQs

Further I am going to create word cloud (using the keywords that are extracted) which is a group of words that are used repeatedly in the chapter.

Also aiming to develop **knowledge graphs** that connects different sections of the chapter and also the sections that can be related to/ mentioned in other chapters which gives a overview of the total reports.

* Semantifying data is important as it saves time and energy.
* Tools that are being developed will be made available online that can be used from anywhere in the world.
* It helps young researchers to grasp knowledge fast and learn/implement it
* It gives a new perspective for educationalists towards large scientific literature.
* Expands the learning community.

**Knowledge graphs**

Knowledge graphs are a way to integrate information that is extracted from various data sources. Knowledge graphs have started to play a central role in representing the information extracted using natural language processing and computer vision.

Knowledge graphs can be an input for machine learning models for better predictions.

Knowledge graphs have nodes and edges, where every node represents some structured data and every edge represents its connection to that node(s).

A knowledge graph may look something like this (example only)

