University of California Berkeley

Climate Synapse: Using Natural Language Processing and Web Development to connect climate investors to researchers and inventors

Martin Liu

Table of Contents

1. Executive Summary

As countries start pledging to achieve net zero emissions by 2050, the need for investment in these clean energy technologies increases. The International Energy Agency (IEA) has listed 23 technology groups and over 500 specific technologies that have the potential to help achieve this.[1] However, this is not enough. Knowledge about such technologies is often spread through word-of-mouth rather than a generalized search system. The goal of our group is to match the descriptions of said technology with the databases of scientific papers and patents using Natural Language Processing. Thus, creating a search system for finding clean technologies. This search system was then developed into a demo website using Streamlit that can be used to find related patents and papers of the IEA listed technologies. Our team successfully built a website that searches for related patents and papers. However, improvements to the NLP model as well as backend website deployment could be improved.

2. Problem statement

For this project, we are focusing on three main players in the field of climate technology. First, the Investors: Venture capitalists, governments, companies, philanthropists, players who want to invest in climate technology, be it for profit or to meet national goals. For instance, the World Fund is a European Climate Tech VC fund investing in startups that contribute to solving the climate crisis. It focuses on the key emitting sectors Energy, Food & Agriculture, Manufacturing, Buildings, and Transport. [2]

Second, Inventors: established firms, research labs and science competitions, players who are working on a marketable product or technology that can both be turned into a business while also helping achieve climate goals. For instance, Climeworks AG is a startup focusing on Direct Air Capture, MIT Climate & Energy Prize is a competition for university students to potentially launch their idea into a company. [3]

Lastly, researchers: companies, governments and universities, players who are working on the theoretical aspect of climate technology. The figure below is a visual representation of the three players in climate technology working together.

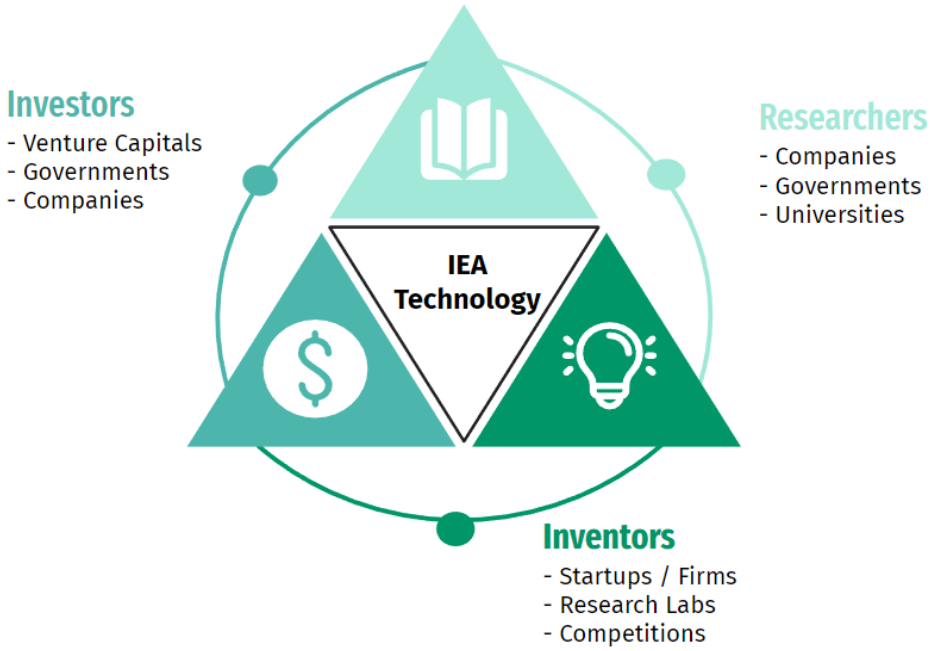

Figure 1: Investors, Inventors and Researchers working together on climate related technology as listed by IEA

Technology advances through combinations of innovations. [4] Currently there is no platform connecting these players together and knowledge is usually spread through word-of-mouth. The proposal sent by RMI hypothesizes that by creating greater visibility of technological innovation this can accelerate the speed of such technologies being integrated into the economy. A platform similar to this is the employment website “climatebase.com” where employers and employees from the climate industry connect to facilitate the sharing of knowledge and employment.

3. Technical Section: Connecting scientific texts using NLP

3.1. Connecting Data Sources: IEA’s ETP Clean Energy Technology Guide

The ETP Clean Energy Technology Guide is an interactive framework that contains information for over 500 individual technology designs and components across the whole energy system that contribute to achieving the goal of net-zero emissions. A web-scrapping bot was used to extract information regarding all the technologies. An example of the technology used for Electric Vehicles; in particular Lithium-Ion batteries can be found below.

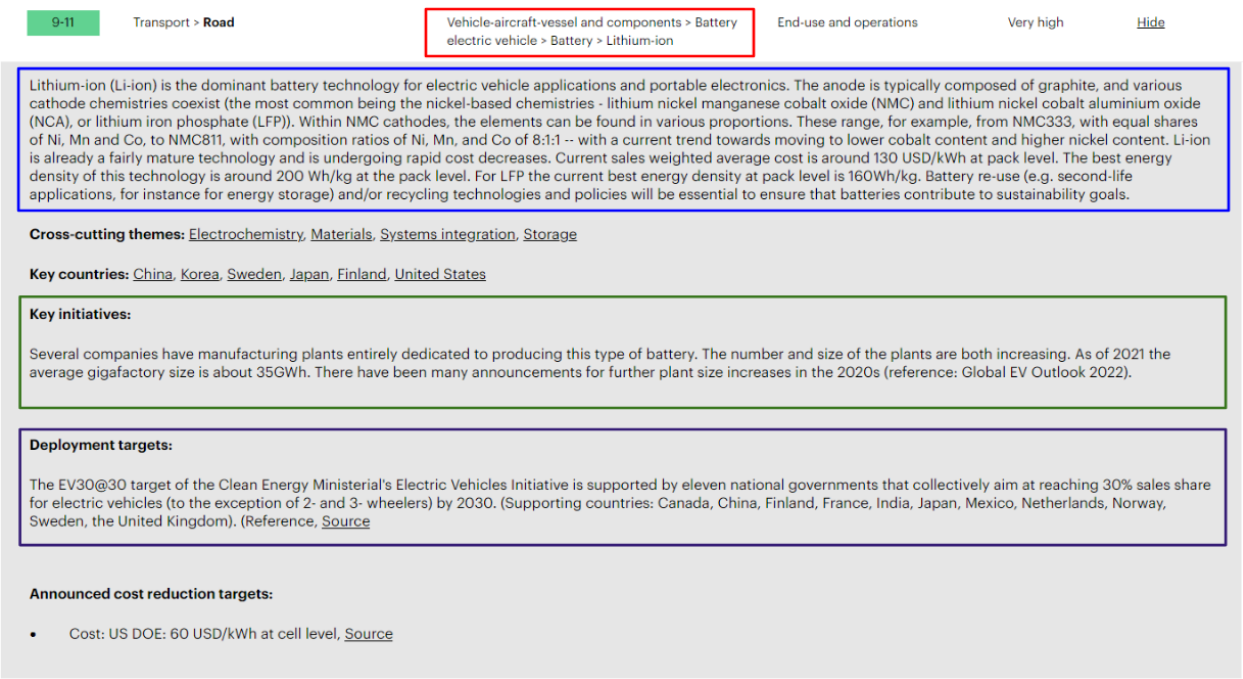

Figure 2: Information on climate technologies from IEA website

All technologies follow a standard format as shown above in Figure 2. The information boxed in red represents the name of technology used; boxed in blue describes how the technology works; boxed in green elaborate on which countries have adopted this technology and their work on it thus far; boxed in purple details what some of these countries hope to achieve by when.

Looking at this we knew that the description of the Lithium-Ion battery would most likely line up with an abstract of a paper or patent. It also includes specific descriptions of the technology such as “best energy density of this technology is around 200 Wh/kg” and details on the material the battery uses. The name of the technology could be used to query for patents or papers, and the other details to fine-tune our search system results.

3.2 Connecting Data Sources: Patents and Papers

Next, we looked at the database for patents and papers. The database used for patents is PatentsView and for papers is OpenAlex. [5] In order to retrieve information, we accessed their APIs. (Application Programming Interface) The figure below shows how keywords are extracted from IEA and queried into PatentsView along with the endpoints that can be queried upon and the data that can be retrieved from the API. Endpoints refer to the different categories of data that can be queried from the API.

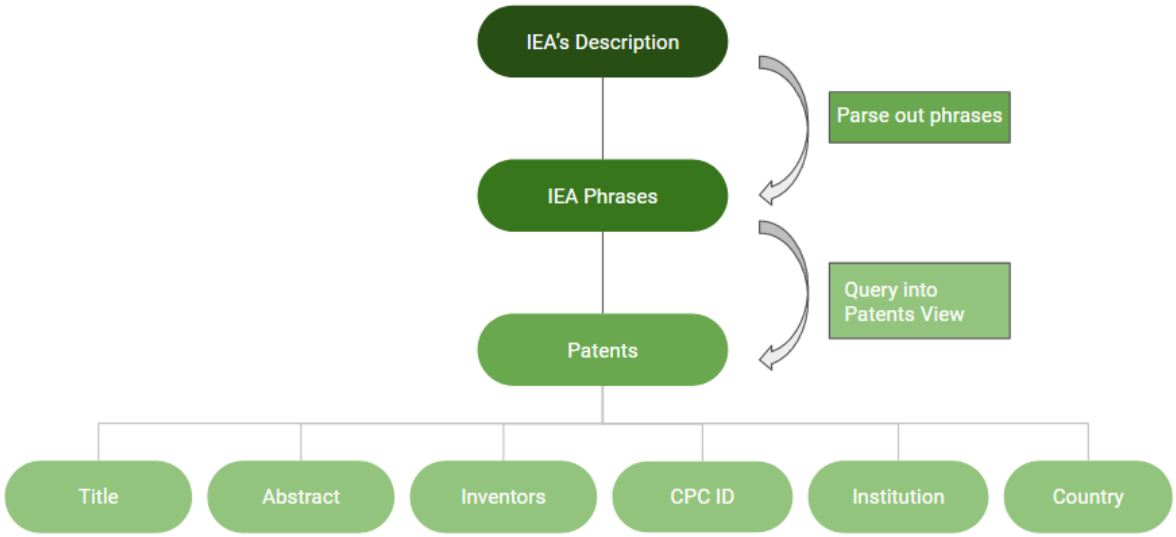

Figure 3: Flowchart of how data is extracted from IEA and queried into patents view and the possible endpoints

The inventors refer to the people who worked on the patent, country refers to where the patent was filed. For the purposes of this project, we only looked at patents filed in the United States. Lastly, we can also categorize our patents based off what field they are in using the Cooperative Patent Classification (CPC) ID, which is an extension of the International Patent Classification (IPC).[6] It is divided into nine sections, A-H and Y, which in turn are sub-divided into classes, sub-classes, groups, and sub-groups. For the purposes of this project, we look at patents with the code Y02 as those fall under climate related patents.

Using Figure 3 as a reference, the process of getting Lithium-Ion battery related patents is as follows. First, we would parse out the data from IEA technology and only select the phrases that summarize the technology, in this case it would be, "Vehicle-aircraft-vessel and components", "Battery electric vehicle", "Battery", "Lithium-ion". Next, we use the data we extracted earlier from IEA's description of their technology to query into Patents View. What we intend to find are patents with abstracts containing these specific phrases. From there we take these patents and extract its Title, Abstract, Inventors, Institutions, Country, and CPC ID.

The methodology is followed when finding Research papers with Open Alex, the key difference is that the CPC ID is now replaced with "concepts". These concepts can range from being energy or biology related. For the purposes of this project, we focus on concepts that are "carbon emission" or "climate" related, this way it would be analogous with the CPC that patents are labeled with.

3.3 Using NLP to match IEA text to abstracts of Patents and Papers

The issue with selecting patents and papers this way is that we might get unrelated patents/papers that coincidentally have those keywords. For instance, consider the phrase “Battery electric vehicle”, ideally the patent would be describing some type of battery for electric vehicles that is able to achieve more efficient energy usage, or stronger battery life. However, a patent with those keywords could also be one related to “Optimized charging and discharging of a plug-in electric vehicle”, which is exactly what happened during our query. While somewhat related, this is not exactly what we were looking for.

To get a closer connection between the technology and the patents/paper we compare the abstracts of the text, which naturally contain more information, and use Natural Language Processing to compare how similar these texts are. After which, we rank the similarity of the patent/paper to the technology description.

3.3.1 Brief Introduction into Natural Language Processing

For us to assess how similar the abstracts are to the IEA descriptions we must first encode these texts, essentially converting them from words to fixed-dimensional vectors. This encoding process requires a pre-trained language model. For our purposes we used ‘all-MiniLM-L6-v2’, a transformer-based language model that uses a neural network architecture to learn contextual representations of words and sentences.[7] This is because it was important for the model to be able to capture the context that these technologies had. If we looked back at our example on Lithium-ion batteries, the model had to capture the context that the batteries were being discussed in, which was improving the materials and energy density of the batteries.

3.3.2 Comparing embedded texts

Using the ‘all-MiniLM-L6-v2’ model, we were able to embed the description of IEA technology and the abstracts. The similarity between those two texts were determined using cosine similarity, as described in the equation below.

$$cosine\_similarity(abstract, IEA\_tech) = \frac{abstract\ .IEA\_tech}{\|abstract\|.\|IEA\_tech\|}$$

‘abstract’ and ‘IEA_tech’ refer to the respective encoded vectors. The higher the cosine similarity score, the more similar the encoded vectors are which imply that the texts are also more similar. This process is done for all the other patents and papers that were queried, with the IEA_tech vector remaining the same in all comparisons.

3.4 Building a website for deployment

Since the goal of this project was to demonstrate the possibility that such a platform could be possible, we decided that simplicity and ease of use would be our priority when finding an appropriate program or software. The team ultimately decided on Streamlit, an open-source Python library used to create web applications for machine learning projects.

3.4.1 Streamlit website demo

Using Streamlit we were able to make a functioning demo website that can be used to find patents and papers related to the IEA technology. A screenshot of the website can be seen below in Figure 4. Using our example of Battery Electric vehicles, we can see how a user would navigate the website to perform that search.

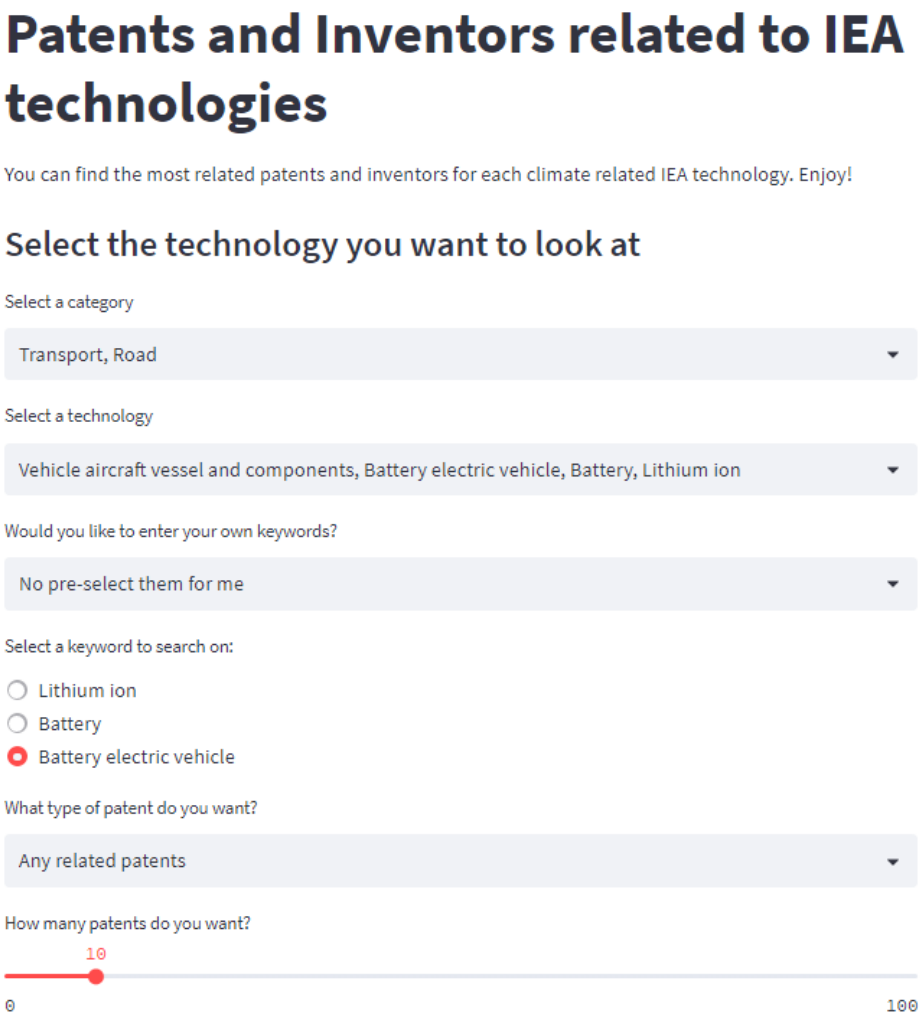

Figure 4: Screenshot of Streamlit website to search for patents and inventors related to IEA technology

In the first dropdown box, the user can pick from 23 categories and further select the technology of the category. The next dropdown box allows users to select a technology based on the prior category chosen. From there, they can have the option of typing their own keywords or choosing the pre-selected ones. The number and type of patents can then be chosen, namely climate or non-climate related.

After the user selects the options, they can find the closest related patents. This is displayed as a dataframe in Figure 5.

	title comparision	abstract comparison	title	citations	date
US-10020494	0.5877	0.638	Anode containing active material-coated graphene sheets and lithium-ion batteries c	9	2018
US-10003068	0.6956	0.6314	High capacity anode materials for lithium ion batteries	2	2018
US-10020491	0.5132	0.5978	Silicon-based active materials for lithium ion batteries and synthesis with solution pr	1	2018
US-10014522	0.6251	0.5868	Cathode material for lithium-ion secondary battery	1	2018
US-10014126	0.4705	0.5537	Lithium-ion supercapacitor using graphene-CNT composite electrode and method fo	1	2018
US-10002718	0.529	0.5219	Lithium ion capacitor	1	2018
US-10008715	0.5864	0.5166	Cathode material for lithium-ion secondary battery, method for manufacturing same	1	2018
US-10002717	0.4646	0.5061	High performance lithium-ion capacitor laminate cells	1	2018
US-10014552	0.6092	0.5039	Lithium ion rechargeable battery	8	2018
US-10014553	0.5296	0.4798	Electrolyte formulations for lithium ion batteries	1	2018

Figure 5: Sample dataframe ranking the closest related patents based off user selection

The title and abstract comparison scores are calculated using the cosine similarity equation mentioned in Section 3.3.2. The subsequent columns are different endpoints that we extracted from the API when querying for information, for instance, title, citations, and date.

Users can also specify to display the names of inventors and rank them based on the number of times they have a patent related to the technology. Figure 6 shows a sample data frame of this.

Inventor's name	PatentsView inventor's id	Number of occurence	Number of patents	Number of US patents citations
Charan Masarapu	fl:ch_ln:masarapu-1	2	14	47
Yongbong Han	fl:yo_ln:han-210	2	14	42
Haixia Deng	fl:ha_ln:deng-8	2	15	68
Yogesh Kumar Anguchamy	fl:yo_ln:anguchamy-1	2	13	45
Herman A. Lopez	fl:he_ln:lopez-4	2	42	319
Subramanian Venkatachalam	fl:su_ln:venkatachalam-3	2	27	166
Wanjun Cao	fl:wa_ln:cao-4	1	15	5
Harry Chen	fl:ha_ln:chen-233	1	10	285
Yuima Kimura	fl:yu_ln:kimura-9	1	7	2
Nobuhiro Okada	fl:no_ln:okada-8	1	39	32

Figure 6: Sample data frame ranking the closest related patents based off user selection

The inventor Chara Masarapu would have 2 occurrences of patents that are climate related and have Lithium-ion in their patent abstract. In addition, other statistics of interest such as the number of patents and citations the user has are also displayed.

Lastly, a map can be generated showing where these patents originated from, as seen in Figure 7.

Figure 7: Map of where the inventors are located in for the sample technology

3.4.2 Additional Website features

In addition to performing these functions for patents, this can also be done for research papers with OpenAlex. Another feature added is that users can also input their own description of technology rather than use the ones from IEA. A screenshot of the how that can be done is seen in Figure 8.

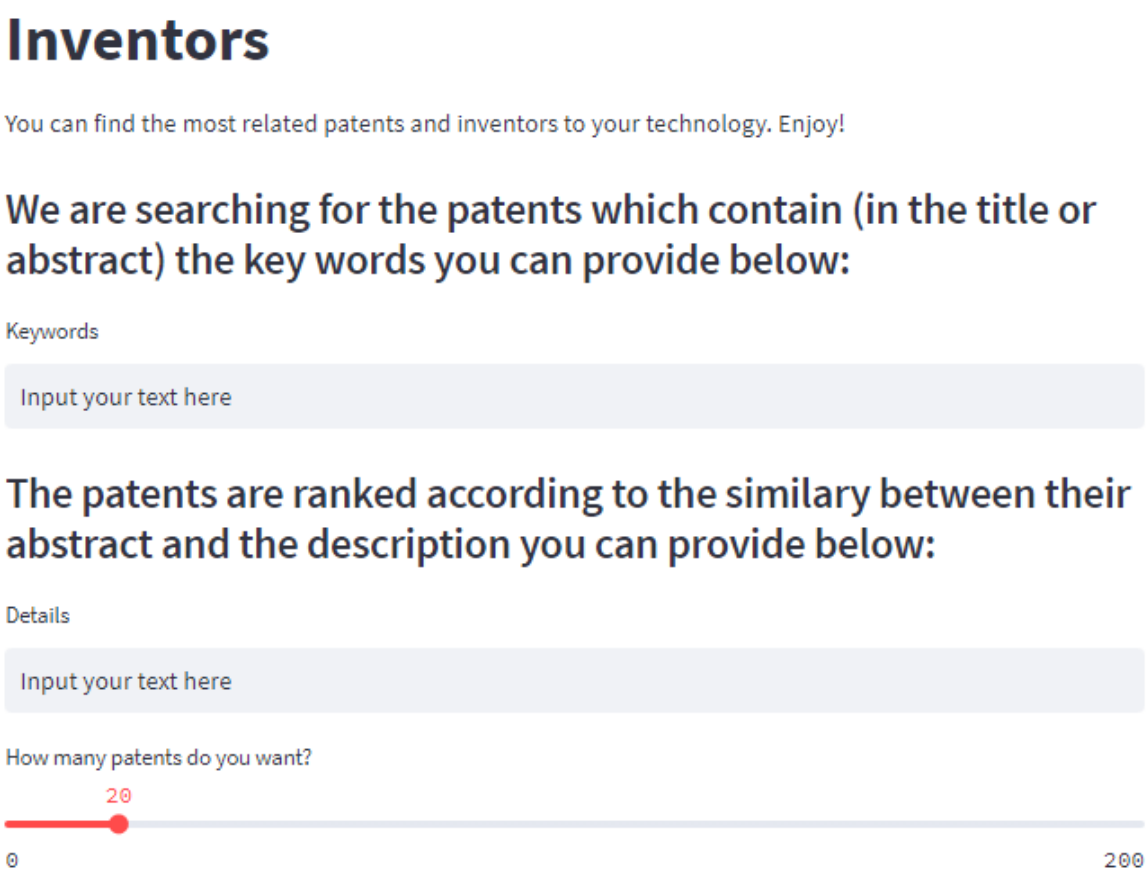

Figure 8: Screenshot of website for finding technology specific to user's input

The first input box is for users to type in keywords that they would like to search on, and the second input box is the details of the technology they are interested in. Here the user will have full control over what keywords they wish to search for within OpenAlex or PatentsView and what kind of information they are trying to match.

The website can be accessed here: https://mmaliu97-cl-synapse-project-emmaclimate-siteintroduction-2vg58n.streamlit.app/

4. Discussion and Future Work

While the model performs well in terms of comparing IEA technology text with Patent and Paper abstracts, there is still room for improvement as the current NLP model used is not trained on any specific kind of text. An NLP model trained on climate technology texts from Patents and Papers might be better suited when doing such text comparisons.

Streamlit served as a useful tool to create a demo website. One additional functionality it could add is to create a user account functionality for other investors, researchers, inventors, or even the general public interested in climate technology to communicate and network with one another.

Lastly, there are other types of climate-related information that can be investigated, for instance companies working in climate tech or climate-related competitions hosted by universities and companies. By finding out more climate-related information, we can better connect investors to the people working in climate technology.

Citations

[1] IEA. “ETP Clean Energy Technology Guide – Data Tools.” *IEA*, 21 Sept. 2022, www.iea.org/data-and-statistics/data-tools/etp-clean-energy-technology-guide.

[2] Fleitmann, Maximilian. “Top 9 Climate Tech vc Investors | Startup Investors.” *Www.basetemplates.com*, 30 Sept. 2022, www.basetemplates.com/investors/top-9-climate-tech-vc-investors#:~:text=1. Accessed 1 May 2023.

[3] MIT Climate & Energy Prize. [Online]. Available: https://cep.mit.edu/. [Accessed: 01-May-2023].

[4] W. B. Arthur, *The nature of technology: What it is and how it evolves*. New York: Free Press, 2011.

[5] Priem, J et al, OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, *Proceedings of the 26thInternational Conference on Science and Technology Indicators*, 2022

[6] E. P. Office, “Cooperative patent classification (CPC),” *EPO*. [Online]. Available: https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/classification/cpc.html#:~:text=The%20Cooperative%20Patent%20Classification%20(CPC,%2C%20groups%20and%20sub%2Dgroups. [Accessed: 30-Apr-2023].

[7] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using Siamese Bert-Networks,” *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

Acknowledgement

Map data copyrighted OpenStreetMap contributors and available from https://www.openstreetmap.org

I would like to thank Emma Scharfmann and Professor Lee Fleming for their guidance and help on this project. I would also like to thank RMI for giving me the opportunity to work on this project. Lastly, I would like to thank my capstone project team, Julien Raffy, Zhuo Fan Li, and Yi Zhe Zhao for their support and help throughout the school year.