Project Paper

Group Members

John Christenson - jchristenson

Manuel Mendez - manuelm

Pablo Montenegro Helfer - pablomonhel

Santiago Satizabal – ssatizabal

Abstract

This project analyses the implementation of energy policy legislation in Pennsylvania and Texas, two states with high energy production[1], without a person reading a single legislative bill.

The project consists of four stages. First, we collect energy-related data, including the keywords, energy index input, and official variables collected from the U.S. Energy Information Administration. Second, we scrape the text of legislative bills using the Open States API. Third, we process the data and analyze it. Finally, we create a visualization of graphs and tables in an interactive dashboard.

The Energy Policy Index ranges from 0 to 100 and tells each bill's relative implementation of energy policy. A bill with no energy keywords will have an index of 0, and the bill with most keywords will have an index of 100.

This analysis enables the comparison of bills inside and between states using the energy policy index, the analysis of the most frequent subjects written in overall bills, and the comparison of energy-related official data between states.

[1] per the U.S. Energy Information Administration.

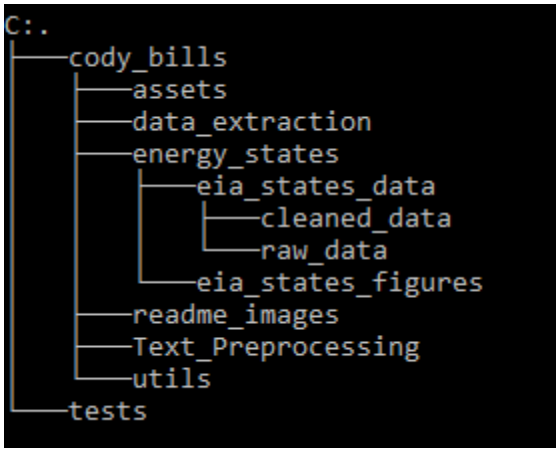
Overall Structure of the Project

The project is structured into four main parts:

1. The data, legislative bills, and extraction from the openstates.org webpage. This stage is presented in the "data_extraction" folder, inside the "cody_bills" folder, in the project repository on GitHub.
2. The text preprocessing, keyword search, index creation, and text analysis. This stage is presented in the "Text_Preprocessing" folder, inside the "cody_bills" folder, in the project repository on GitHub.
3. The official energy policy data extraction, processing, and visualization. This stage is presented in the "energy_states" folder, inside the "cody_bills" folder, in the project repository on GitHub.
4. The dashboard design and execution. This stage is presented in the "cody_bills" folder, in the project repository on GitHub, and uses the "app.py", "__init__" and "__main__" files for execution, and the "assets" and the "utils" folders for input storage and helper function usage, respectively.

The composition of the repository's folders is shown in the following image:

Figure 1. Project Repository Composition - Folders



Responsibilities of each member

Name	Module	Tasks	Files
Manuel	Research	Understanding the Open States API, researching methods to scrape text/html files and pdf files and convert them into strings	
	API extraction	- Make the requests to the Open States API for the bills that works for any state and any date of creation - Do an implementation that makes the extraction work for apikeys that have limited requests per day	- cody_bills - data_extraction - scraper.py
	Data scraping	- Scraping html and pdf links into python objects and subsequently into strings - Creating dictionaries for each bill with metadata and the text for the bill - Saving the dictionaries as json files	- cody_bills - data_extraction - scraper.py - bills_Pennsylvania.json - bills_Texas.json - Text_Preprocessing - bills_pennsylvania.json - bills_texas.json
	Document writing	- Introduction for the README.md - Wrote the abstract and worked on the other parts of the paper	- README.md - proj-paper.pdf
John	Research	Research and define the keywords (words and bigrams) that define energy policy	
	Data pull	- Downloaded U.S. Energy Information Administration csv data, and brought it into raw_data folder	- cody_bills - energy_states - eia_states_data - raw_data
	Data Cleaning	- Created multiple functions clean, format, and prepare the raw data using pandas. The now cleaned data was placed into the cleaned_data folder as a txt file holding csv data - Created a pytest to test the cleaning process	- cody_bills - energy_states - eia_states_data - cleaned_data - raw_data - eia_clean.py - states_dict.py - tests - test_eia_clean.py
	Visualization	- Created four bar graphs from the cleaned data using multiple functions - The graphs are optimized to both be called directly to the dash live and separately saved as a png in the eia_states_figures folder - Created a pytest to test the data visualizations creation	- cody_bills - energy_states - eia_states_data - cleaned_data - energy_dataviz.py - tests - test_energy_dataviz.py
Santiago	Research	Researched about text analysis, count words and sliding window algorithms	
	Data Cleaning	- Created a function clean_and_tokenize() that takes the text of each bill and cleans punctuation, digits and tokenizes the texts using regex and other tools.	- cody_bills - Text_Preprocessing - text_analysis.py
	Data Processing	-Several functions that generate counts of words or ngrams used for wordclouds -Function for sliding window algorithm that computes an index that grows with the number of times a	- cody_bills - Text_Preprocessing - text_analysis.py - assets

		keyword/keyngram appears in a text -Function to normalize the index using the data for the two states and using a feature scaling I = (x - min(x))/(max(x)-min(x)) -Function to open the scraped bills clean them, generate the word clouds and save them into the assets folder to be used in the dashboard -Function to open the scraped bills clean them, generate the index, normalize it, and separate the dataframe for each state to be displayed as the metadata in the dashboard.	- table_pennsylvania.json - table_texas.json
	Visualization	-Four wordclouds (for each state, unigrams and bigrams -Tables for each state with the normalized index to be displayed as metadata in dashboard -get_histogram() function to display the distribution of the Energy Policy Index for the two states with and without zeros (in team with Pablo)	- cody_bills - Text_Preprocessing - text_analysis.py - assets - words_pennsylvania.png - words_texas.png - bigrams_pennsylvania.png - bigrams_texas.png - utils - helper_functions.py
	Project folder design	-Wrote and designed the README.md in the parent folder	- README.md
Pablo	Research	Energy Index creation, from preprocessing to normalization	
	Visualization	- Dashboard design and creation with Python module called Dash - Input manipulation and display in Dash, tables, graphs and images	- cody_bills - app.py - __init__.py - __main__.py - utils - helper_functions.py
	Project folder design	Project folder creation, design, update	
	Testing	Tested app display, module installation, colleagues' code execution	
	Coworker discussion	Discussion sessions with coworkers to study code implementation, algorithms, graph design,folder design, testing, and input output connection	
	Document writing	- Wrote section on README.md - Wrote chapters on "proj-paper.pdf"	- README.md - proj-paper.pdf

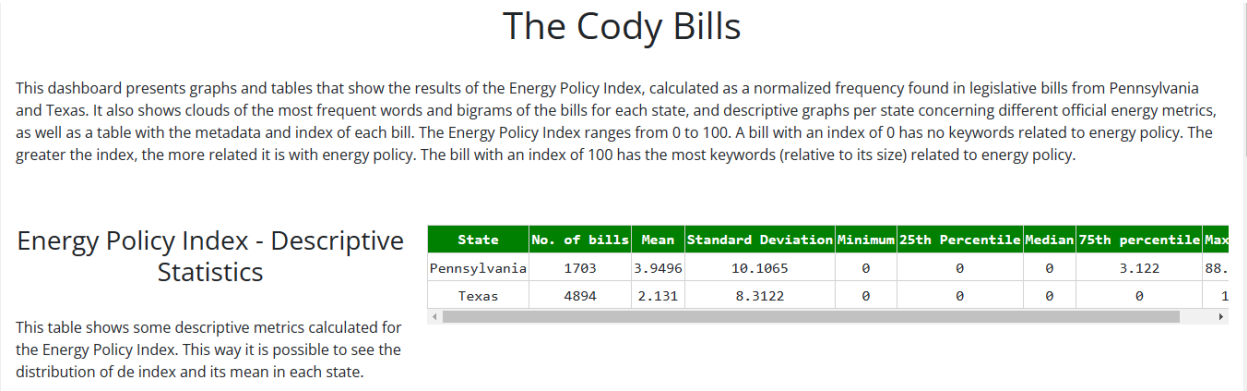
Application Guide

It is important to briefly mention how the index was created and how to interpret it. As mentioned above, the index ranges from 0 to 100 and shows each bill's relative implementation of energy policy. It was created using a sliding window algorithm on each bill, where each keyword (a word or a bigram) was searched inside each window, and if found, a counter was added by 1. In the case of the bigrams, a 1 was added to the counter only if both the words composing it were found in the

window. Subsequently, the counter was divided by the number of words in the bill to compare longer and shorter bills with greater ease. Finally, when the counter for each bill was computed, both in Pennsylvania and Texas, all of the counters were normalized with a min-max function, where the bills with no keywords in them would have a value of 0, and the one with most keywords would have a value of 100.

The dashboard has 5 main panels that the user can interact with. These are presented below.

1) Energy Policy Index - Descriptive Statistics: This table shows some descriptive metrics calculated for the Energy Policy Index. This way, it is possible to see the distribution of the index and its mean in each state.



The Cody Bills

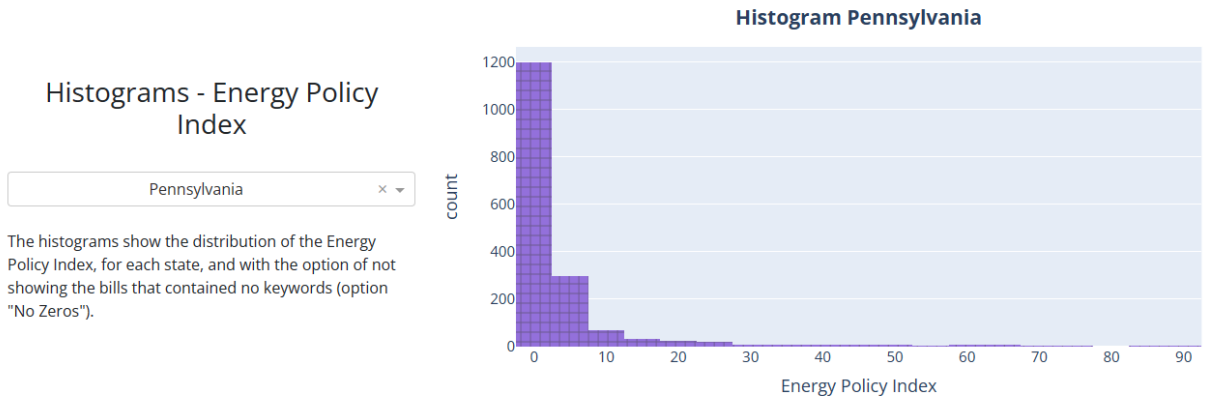
This dashboard presents graphs and tables that show the results of the Energy Policy Index, calculated as a normalized frequency found in legislative bills from Pennsylvania and Texas. It also shows clouds of the most frequent words and bigrams of the bills for each state, and descriptive graphs per state concerning different official energy metrics, as well as a table with the metadata and index of each bill. The Energy Policy Index ranges from 0 to 100. A bill with an index of 0 has no keywords related to energy policy. The greater the index, the more related it is with energy policy. The bill with an index of 100 has the most keywords (relative to its size) related to energy policy.

Energy Policy Index - Descriptive Statistics

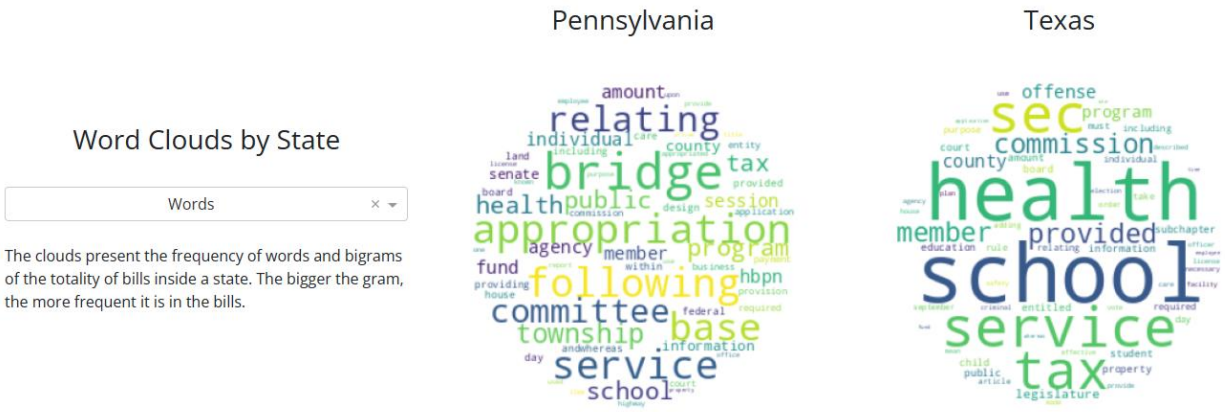
This table shows some descriptive metrics calculated for the Energy Policy Index. This way it is possible to see the distribution de index and its mean in each state.

State	No. of bills	Mean	Standard Deviation	Minimum	25th Percentile	Median	75th percentile	Max
Pennsylvania	1703	3.9496	10.1065	0	0	0	3.122	88.
Texas	4894	2.131	8.3122	0	0	0	0	1

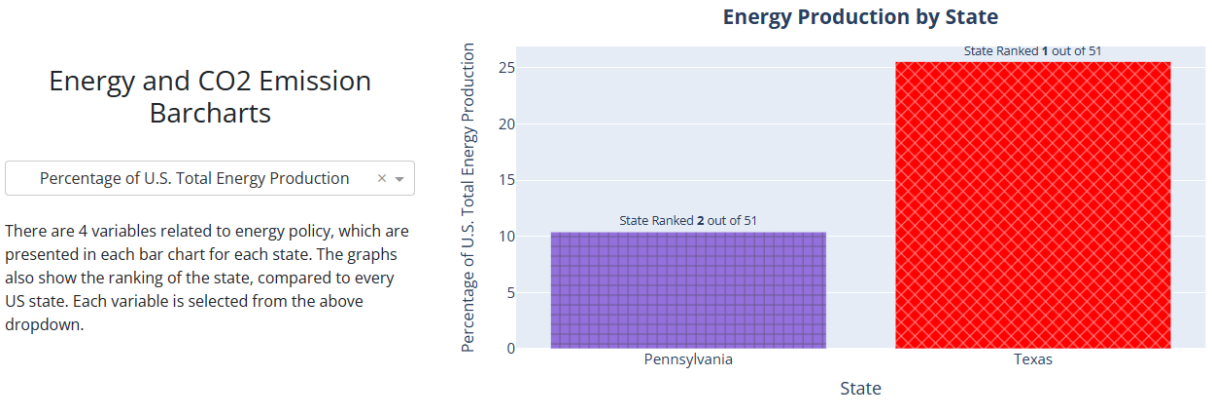
2) Histograms - Energy Policy Index: The histograms show the distribution of the Energy Policy Index, for each state, with the option of not showing the bills that contained no keywords (option "No Zeros"). The user can choose or filter between 4 options: "Pennsylvania" or "Texas" to see the distribution of all the bills for each of these states, but also can choose "Pennsylvania - No Zeros" and "Texas - No Zeros" to see the distributions excluding the bills for which we didn't find any key ngram related to the Energy Policy. Hovering over any bar of the histogram will show the interval of the normalized Energy Policy index and the number of bills within that interval (frequency).



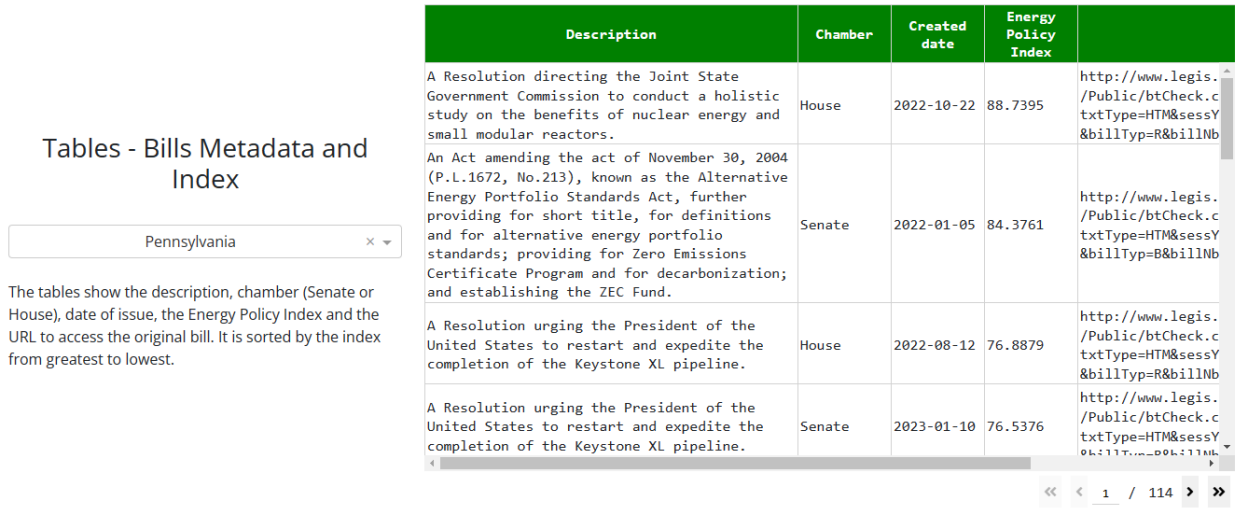
3) Word Clouds by State: The clouds present the frequency of words and bigrams of the totality of bills inside a state. The larger the size of the ngram within the cloud, the more frequent it is in the bills. The user can choose between unigrams or bigrams from the dropdown, and the word (or bigram) - clouds of each state will be displayed side by side for comparisons.



4) Energy and CO2 Emission Barcharts: This panel presents 4 variables (Percentage of U.S. Total Energy production, Percentage of U.S. total Carbon Dioxide Emissions, Energy Expenditure per capita, and Energy Consumed per capita) related to energy policy the user can select from the above dropdown. Once the variable is selected, a bar chart for each state will be displayed, showing the level for the given variable and the state's ranking compared to all U.S. states.



5) Tables - Bills Metadata and Index: The tables show the description, chamber (Senate or House), date of issue, the Energy Policy Index, and the URL to access the original bill. It is sorted by the index from greatest to lowest.



Conclusions

With this project, we were able to display an interesting comparison between two states – Pennsylvania and Texas – regarding how often they use important energy-related words in their legislation. Considering that the states are the two leading producers of energy in the country and have some of the largest carbon dioxide emissions (Texas ranked 1st and Pennsylvania 4th), it was striking to find that, in our sample, more than 50% of the bills in Pennsylvania and more than 75% of the bills in Texas don't have a single word related to energy policy (as defined in our keywords). The distribution of the Energy policy index is especially skewed to the right.

We could state a plethora of plausible explanations for this phenomenon. For instance, stakeholders may want to hold back new energy legislation, or energy policy, to maintain a beneficial status quo. Nonetheless, we need further information to assess whether the absence of these words in the bills is to be expected.

Against this background, one clear step forward in our project would be to include the indicators and the bills of other states (hopefully the 50 states) to see patterns or trends on a larger scale. To observe more robust results, we could also add more years to our analysis so that our indicators (Energy consumption, carbon dioxide emissions, etc.) and the bills don't have a recency bias or a biased by any seasonal shock. All of this, it is worth noting, is contingent upon having a greater computational capacity and an API key that allows us to scrape more bills.

Finally, an advantage of our approach is that it can be applied to energy policy and other public policy-related topics. We could exploit the code we already have to expand the scope of our project and find or compute the frequency index for any other list of keywords/key-ngrams regardless of whether this list contains words related to higher education, healthcare, security, or any other relevant topic. In this case, the indicators would need to change as well. Arranging a single U.S. states data frame (data panel) containing diverse socioeconomic indicators could help in this regard.