Overview:

This project analyzes global carbon dioxide CO2 emissions data by modeling countries and their emission trends as a graph. Using the Our World in Data CO2 dataset, the program compares countries based on the similarity of their per capita emissions over time, constructs a graph where edges connect highly similar countries, and applies graph algorithms to identify clusters of countries with similar environmental behavior. It then calculates and ranks countries by their centrality in the similarity graph, identifying the most influential nations regarding emission trends. The program outputs the overall graph statistics, clusters of similar countries, and saves all discovered clusters into a text file.

CSV File: https://github.com/owid/co2-data/blob/master/owid-co2-data.csv

Functionality:

The project's key functionality is divided into several important functions. The **cosine_similarity function** compares two countries' emission vectors by calculating the cosine of the angle between them and producing a score between -1 and 1. This measures how directionally similar their CO2 emission trends are over time.

The **build_similarity_graph** function constructs a graph where each node represents a country or region. Edges are created between countries whose emission trends are highly similar based on a defined threshold of 0.975. **print_graph_stats** outputs basic properties of the constructed graph, such as the number of countries, the number of strong emission pattern connections, and the number of identified clusters.

The **print_degree_centrality** function analyzes and ranks countries based on how many strong emission pattern connections they have, which helps identify influential countries. To further structure the results, the print_clusters function groups countries into clusters of highly similar emission profiles using a union find structure and displays these clusters neatly. Finally, **save_clusters_to_fil**e takes all detected clusters and writes them into a .txt file, making the output easy to review or process further. These functions together support a modular and clean analysis of the global CO2 emission landscape using graph theory.

Workflow:

The main workflow of the program starts by using the csv_loader module to read and structure the raw OWID CO2 dataset into a usable form. The data is then passed into the utils module, where cleaning steps are applied: missing data is filled, time series are smoothed to reduce noise, and vectors are normalized to enable fair comparisons. Once the data is prepared,

the graph module constructs a graph based on emission trend similarities using cosine similarity measures. It then applies graph algorithms to analyze the graph structure, identify emission clusters, and calculate centralities. A shortened version of the final results is printed to the console, and the full cluster list is saved to a .txt file for full analysis. The main.rs file manages the orchestration of this entire process, sequentially executing these steps to ensure a smooth and modular flow of data from raw input to final analysis.

Usage:

The defined similarity threshold is set to 0.975. The program automatically loads the dataset file (owid-co2-data.csv) from the working directory, processes the data, builds the graph, and outputs all results. The project is designed to run efficiently, typically completing in just a few seconds(with release flag) even for the full dataset, thanks to careful data structure choices and optimized Rust performance. The output includes console prints showing overall graph properties, the most central countries by emission pattern similarity, and a shortened list of clustered countries. A text file named clusters_output.txt is saved, containing the full list of cluster groupings.

The tests.rs file contains five unit tests that verify the correctness of key mathematical and processing functions. It includes tests for the cosine_similarity function to ensure that identical emission vectors return a perfect similarity score of 1. It also tests the smooth function to confirm that it correctly applies a moving average to an input vector without altering its length.

Results and Interpretation:

The program I implemented successfully analyzed global CO2 per capita emission trends and grouped countries into clusters based on cosine similarity. The output text file, clusters_output.txt, shows the results of this graph-based clustering, revealing meaningful global patterns in emission behaviors. In total, the model identified 44 clusters, with the largest cluster containing 112 countries. This largest group primarily consists of low income and developing countries, such as Burundi, Bhutan, Fiji, and Yemen. Many of these are small island nations or nations with historically low industrial activity. These countries probably share flat or consistently low emission trends due to limited industrialization and slower energy transitions which explains their strong similarity.

Other major clusters show regional and economic alignment. For example, Cluster 2 groups many post-Soviet states and Eastern European countries, such as Russia, Ukraine, and Kazakhstan, along with some South American nations. These countries likely share moderately industrialized emission patterns shaped by similar geopolitical and economic histories. Cluster 3 features large southern European countries and rapidly industrializing nations like Brazil, India, and Turkey, which are economies with varied but comparable development timelines. Meanwhile, Cluster 4 includes high income countries such as the United States, Germany,

Canada, and Australia. These nations share peak emission periods followed by policy driven reductions and transitions toward cleaner energy sources, leading to consistent emission trends.

Notably, there are over 30 single country clusters with nations whose emission trends were too unique to cluster with others. Countries like Mexico, Algeria, Nigeria, Saudi Arabia, and the United Kingdom fall into this category, probably due to distinctive factors such as heavy reliance on oil, differing policies or unique histories. These outliers illustrate the effectiveness of the model in detecting non conforming behavior. Overall, the results demonstrate that the graph based clustering algorithm effectively made global groupings based on structural emission similarity. These insights could support regional climate policy coordination and future global climate modeling efforts.

Console Output:

```
[kavin@kavins-laptop co2 % cargo test                                          ]
   Compiling co2 v0.1.0 (/Users/kavin/Desktop/Rust/Final/co2)
    Finished `test` profile [unoptimized + debuginfo] target(s) in 0.93s
     Running unittests src/main.rs (target/debug/deps/co2-6775d2651021b279)

running 5 tests
test utils::tests::test_cosine_similarity_identical ... ok
test utils::tests::test_clean_data_simple ... ok
test utils::tests::test_cosine_similarity_orthogonal ... ok
test utils::tests::test_graph_building_simple ... ok
test utils::tests::test_smooth_basic ... ok

test result: ok. 5 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; fini
shed in 0.00s
```

```
kavin@kavins-laptop co2 % cargo run --release
    Finished `release` profile [optimized] target(s) in 0.07s
     Running `target/release/co2`

📄 Loading and cleaning data from owid-co2-data.csv...
✅ Successfully parsed 231 countries and regions.
🔗 Building similarity graph (threshold = 0.975)...
📊 Graph Statistics:
- Total Countries (Nodes): 231
- Total Strong Similarity Connections (Edges): 1352
- Number of Connected Components (Clusters): 44
🌟 Top Countries by Degree Centrality:
- Bermuda: 41 strong connections
- Belize: 38 strong connections
- Fiji: 38 strong connections
- Anguilla: 35 strong connections
- French Polynesia: 34 strong connections
- Andorra: 33 strong connections
- Greenland: 32 strong connections
- Faroe Islands: 31 strong connections
- Guinea: 31 strong connections
- Ghana: 30 strong connections

🗺️Emission Behavior Clusters:
- Cluster 1 (112 countries): ["Micronesia (country)", "Macao", "Saint Vincent and the Grenadines", "Burundi", "Lesotho"]
- Cluster 2 (22 countries): ["Kazakhstan", "Belarus", "Denmark", "Romania", "Latvia"]
- Cluster 3 (17 countries): ["Brazil", "Italy", "Turkey", "Portugal", "Lower-middle-income countries"]
- Cluster 4 (16 countries): ["European Union (27)", "France", "Europe (excl. EU-27)", "Oceania", "Europe"]
- Cluster 5 (9 countries): ["Africa", "Ecuador", "Serbia", "Bosnia and Herzegovina", "North Macedonia"]
- Cluster 6 (6 countries): ["Morocco", "Thailand", "Lebanon", "Syria", "Israel"]
- Cluster 7 (4 countries): ["Iran", "South Korea", "China", "Egypt"]
- Cluster 8 (3 countries): ["Tajikistan", "Kyrgyzstan", "Moldova"]
- Cluster 9 (3 countries): ["Indonesia", "Malaysia", "Paraguay"]
- Cluster 10 (2 countries): ["Philippines", "Panama"]
- Cluster 11 (2 countries): ["North Korea", "Low-income countries"]
- Cluster 12 (2 countries): ["Bahamas", "Liberia"]
- Cluster 13 (2 countries): ["Asia", "Taiwan"]
- Cluster 14 (1 countries): ["Bulgaria"]
- Cluster 15 (1 countries): ["Zimbabwe"]
- Cluster 16 (1 countries): ["Uruguay"]
- Cluster 17 (1 countries): ["Belgium"]
- Cluster 18 (1 countries): ["Europe (excl. EU-28)"]
- Cluster 19 (1 countries): ["Venezuela"]
- Cluster 20 (1 countries): ["Democratic Republic of Congo"]
- Cluster 21 (1 countries): ["Sint Maarten (Dutch part)"]
- Cluster 22 (1 countries): ["Bolivia"]
- Cluster 23 (1 countries): ["Mexico"]
- Cluster 24 (1 countries): ["United Kingdom"]
- Cluster 25 (1 countries): ["Saudi Arabia"]
- Cluster 26 (1 countries): ["Kuwait"]
- Cluster 27 (1 countries): ["Mozambique"]
- Cluster 28 (1 countries): ["Brunei"]
- Cluster 29 (1 countries): ["Asia (excl. China and India)"]
- Cluster 30 (1 countries): ["Austria"]
- Cluster 31 (1 countries): ["United Arab Emirates"]
- Cluster 32 (1 countries): ["Peru"]
- Cluster 33 (1 countries): ["Hungary"]
- Cluster 34 (1 countries): ["Ireland"]
- Cluster 35 (1 countries): ["Spain"]
- Cluster 36 (1 countries): ["World"]
- Cluster 37 (1 countries): ["Algeria"]
- Cluster 38 (1 countries): ["Aruba"]
- Cluster 39 (1 countries): ["Nigeria"]
- Cluster 40 (1 countries): ["Cuba"]
- Cluster 41 (1 countries): ["Chile"]
- Cluster 42 (1 countries): ["Albania"]
- Cluster 43 (1 countries): ["Vietnam"]
- Cluster 44 (1 countries): ["Trinidad and Tobago"]
✅  Clusters saved to 'clusters_output.txt'
The file contains a list of all clusters.
```