Isaac Weissman

Professor Barsky

Machine Learning

22 April 20222

Carving Countries

We've taken the world and split it into 10 clusters:

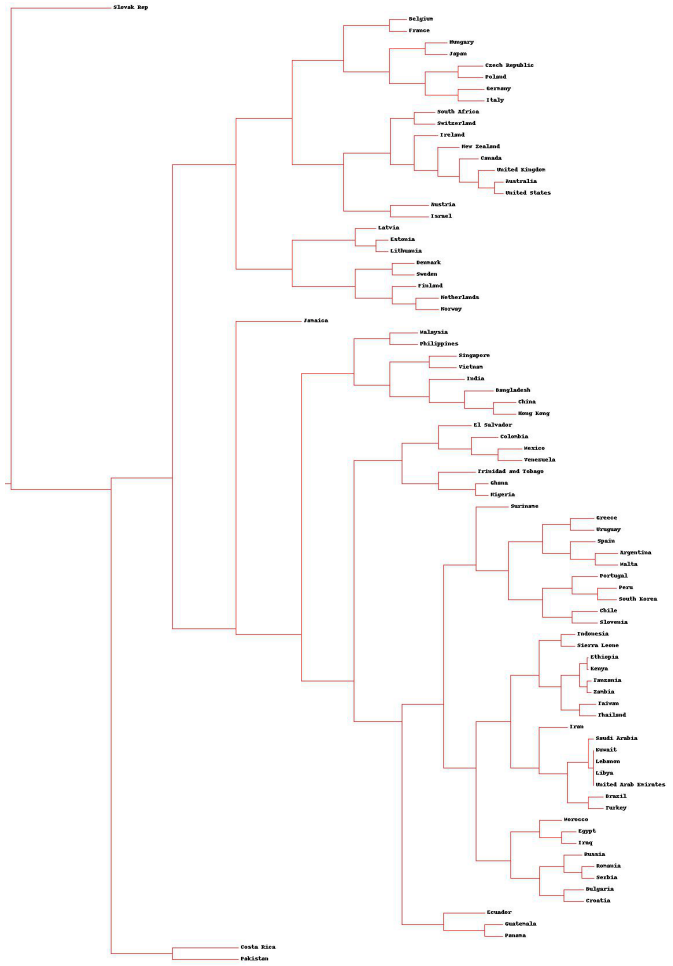
- Cluster 0 : ['China', 'Hong Kong', 'India', 'Jamaica', 'Malaysia', 'Philippines', 'Singapore', 'Slovak Rep', 'Vietnam']
- Cluster 1 : ['Chile', 'Costa Rica', 'Greece', 'Guatemala', 'Panama', 'Peru', 'Portugal', 'Slovenia', 'South Korea', 'Suriname', 'Turkey', 'Uruguay']
- Cluster 2 : ['Denmark', 'Finland', 'Netherlands', 'Norway', 'Sweden']
- Cluster 3 : ['Austria', 'Germany', 'Israel', 'South Africa', 'Switzerland']
- Cluster 4 : ['Brazil', 'Ecuador', 'Ethiopia', 'Indonesia', 'Iran', 'Kenya', 'Kuwait', 'Lebanon', 'Libya', 'Saudi Arabia', 'Sierra Leone', 'Taiwan', 'Tanzania', 'Thailand', 'United Arab Emirates', 'Zambia']
- Cluster 5 : ['Argentina', 'Belgium', 'Czech Republic', 'France', 'Hungary', 'Italy', 'Japan', 'Malta', 'Poland', 'Spain']
- Cluster 6 : ['Bangladesh', 'Bulgaria', 'Croatia', 'Egypt', 'Iraq', 'Morocco', 'Pakistan', 'Romania', 'Russia', 'Serbia']
- Cluster 7 : ['Estonia', 'Latvia', 'Lithuania']
- Cluster 8 : ['Colombia', 'El Salvador', 'Ghana', 'Mexico', 'Nigeria', 'Trinidad and Tobago', 'Venezuela']
- Cluster 9 : ['Australia', 'Canada', 'Ireland', 'New Zealand', 'United Kingdom', 'United States']

The clusters are actually fairly sensible, as is the number. We arrived at 10 clusters after graphing the sum of squared residuals when we conduct the KNN algorithm on all k from 0 to 30. 10 provided a low SSE, while not using too many clusters. It was slightly after the inflection point where diminishing returns kicked in. Certain clusters are particularly strong, like Cluster 7 being all the Baltic States, and Cluster 2 being largely Scandinavian countries. There's also a British Empire clustering in Cluster 9, which are all the major countries to have an intimate relationship with Britain. Cluster 8 is largely South American/Hispanic countries, but includes two African countries with likely similar characteristics.

Cluster 4 is perhaps the strangest, as it spans almost the whole globe and has a diverse cohort of countries. Many of those countries skew towards a centralized rule and lack true open democracy, so perhaps they are grouped for that reason. One of the suggestions to perhaps

improve the clusters was to drop the IVR variable, but I thought some of the clusters here made more sense that what I found when I dropped it--so I kept with this version.

After this, we compared these clusters with those yielded by a hierarchical clustering algorithm (using Euclidean distance--as one always should). That yielded the following Dendrogram (zoom in for details):

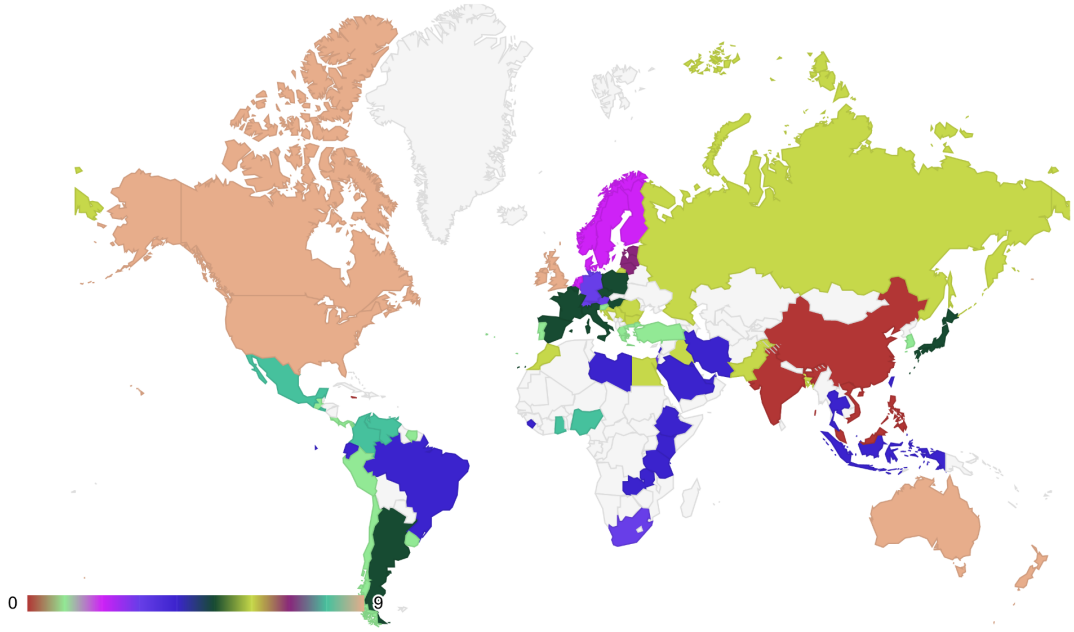


This yielded some slightly stranger clusterings, with the Slovak Republic being noticeably different from every other country.

We then took the clusters from the KNN results, and graphed them on a plot of the world. This exercise was particularly enjoyable for me as I had a lot of fun setting up the coloring for each country. A while back I had been working on a random color generator for JavaScript, but I never finished it. This project gave me a good opportunity to do that, and I made the 10 colors assigned to each cluster be randomly determined on each re-load of the webpage. Out of concern for two colors being too similar, I enforced a minimum “distance” on the colors, based on the Euclidean difference between each primary color (the RGB values). This was completely

unrelated to the assignment, but was perhaps my favorite thing to do with it--since I think the colors it gives are cool.

Here is a copy of an image of the map based on my clusters:



I don't know how to make a website, but I included the html file which produces the above image in my repository, the color code is in that file as well. I only know how to host things in VSCode's Live Server.

Word Clouds:

We also produced word clouds based on the attributes that link all of the countries within their cluster.

Cluster 0:

Cluster 1:

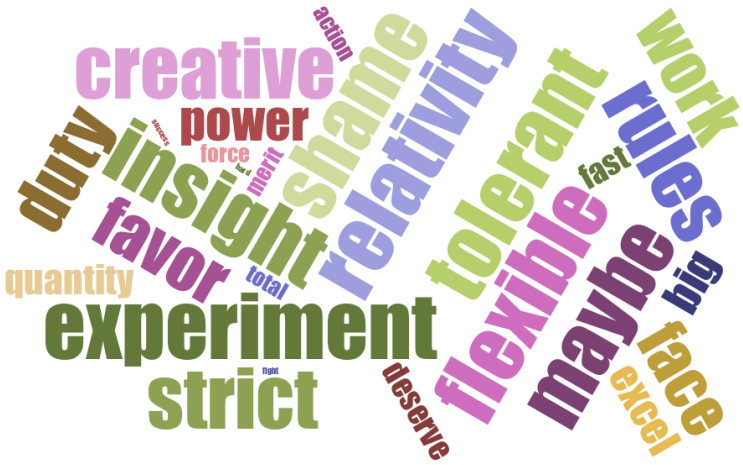
Cluster 3:

Cluster 4:

Cluster 5:

Cluster 6:

Cluster 7:



Cluster 8:

Cluster 9: