Text mining

Now that we have articulated characteristics of our constitutions using region and type, we are able to apply a text mining technique called topic modeling to the constitutions. More specifically, we are now able to address questions such as "How are the constitutions rooted in common law different from the constitutions rooted in civil law?" This missive outlines what I learned so far.

Topic modeling

Topic modeling is an unsupervised machine learning process used to enumerate latent themes in a corpus. Given an integer (N), a topic modeler will create N clusters of words where all the words in a given cluster are in high physical proximity to each other. The resulting clusters can often connote themes ("topics").

Topics

For simplicity's sake, I used 8 as the value of N, and my topic modeler came up with the following topics and elaborations. In other words, I predict the constitutions discuss law, president, state, house, etc. And the word "law" is further characterized with words like "state", "constitution", "general", "rights", etc:

topics	weights	features
law	0.24627	law state constitution general rights presiden...
president	0.23410	president parliament constitution commission a...
state	0.20192	law state president right assembly constitutio...
house	0.19123	law house members council government provision...
assembly	0.18715	law president assembly council state court con...
person	0.15343	person office court law constitution member pa...
court	0.07143	court person order provision commission electo...
constitution	0.03773	state law court constitution council president...

This table can be visualized as a pie chart, and thus making it easier to see that each of the topics carry a similar weight; few of the topics predominate. See Figure 1.

Types

Since we have additional (meta-) data describing each constitution, we can supplement our model with the additional values, pivot the result, and chart things accordingly. We can then address the question, "How does the topic of

"law" manifest itself in constitutions rooted in civil law?" Figure 2 illustrates the answer. More specifically, the topic of "law" is bright yellow. Notice how constitutions rooted in civl law emphasis "law" compared to the others. Similarly, notice how "person" is the predominate topic in the constitutions rooted in common law. Based on what I learned yesterday regarding the differences between civil law and common law, it looks to me as if the topic modeling process is working.

Regions

We can do the same thing for regions. "Depending on where a country is located, how are the topics distributed?" See Figure 3. Countries in Africa advocate "assembly". Europe and the Middle East advocate "house". Countries in South America advocate "law".

Summary

These results are preliminary, and I did not spend a whole lot of time refining the results nor making everything look pretty. That said, I believe I am able to see patterns between the constitutions I had not seen previously. Given our metadata file, we are able to do interesting comparisons. Once we articulate human values and measure then in the constitutions and blockchain forums, then we will be able to visualize our results in manner similar to the manners outlined above.

We're doing data science with words.

Eric Lease Morgan
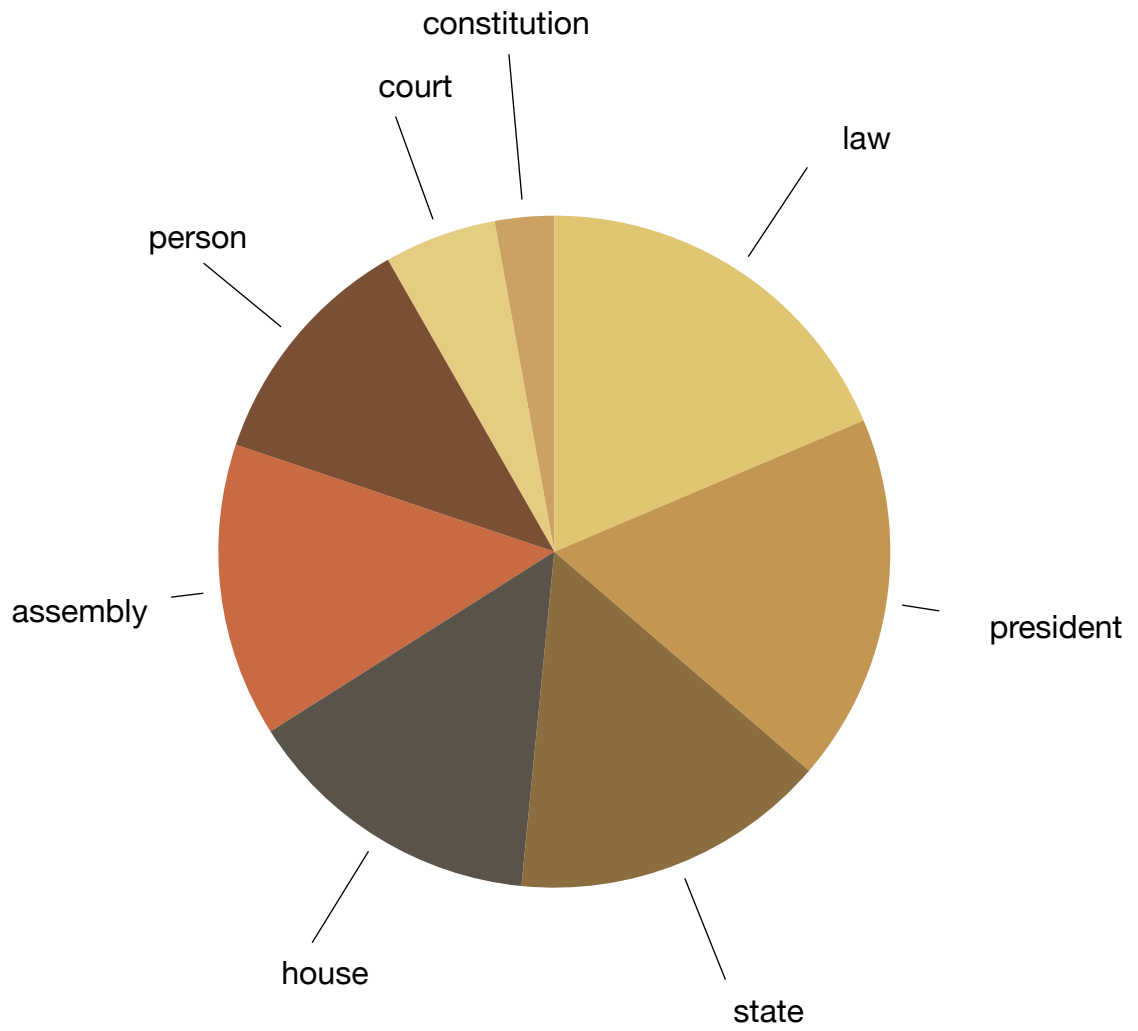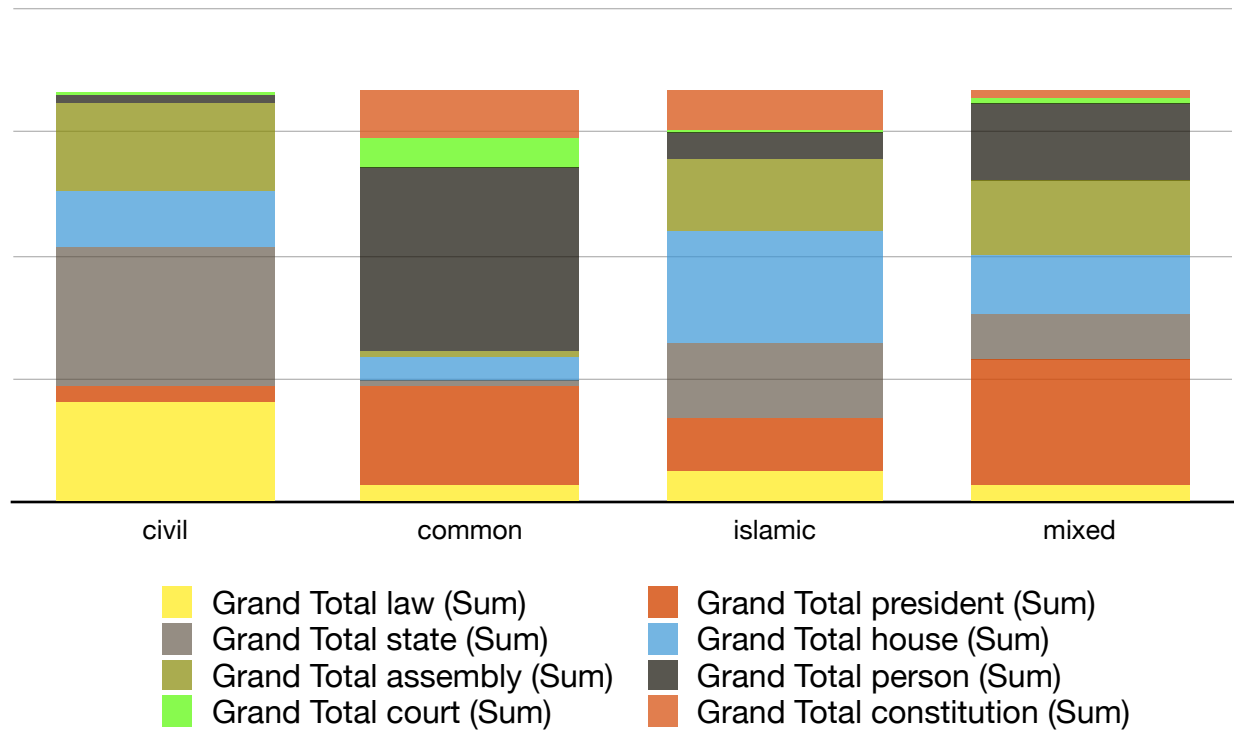Navari Family Center for Digital Scholarship

December 5, 2023

Figure 1: Topics

Figure 2: Topics by type

Figure 3: Topics by region

Grand Total law (Sum)
Grand Total president (Sum)
Grand Total state (Sum)
Grand Total house (Sum)
Grand Total assembly (Sum)
Grand Total person (Sum)
Grand Total court (Sum)
Grand Total constitution (Sum)