QuantGov

Occupation Documentation

Purpose:

Similar to the industry data in RegData, Occupation Data estimates the relevance of SOC occupation codes to regulations in the US Federal *Code of Federal Regulations* and state regulatory codes. These data can be used by researchers to quantify the amount of regulation relevant to particular occupations, which in turn could be used to estimate the effect of regulation on the economic performance of those occupations and other interesting insights.

Content:

Occupation Data contains two different datasets. The first dataset classifies occupation at the 2-digit SOC level. The estimator used for this dataset was trained on a variety of regulatory documents. The probability values in this dataset could be used to quantify occupation-relevant words and restrictions, similar to industry-relevant metrics in RegData.

The second dataset classifies occupation at the 3- and 4-digit SOC level. The estimator used for this dataset was trained only on occupational licensing regulations. There are no probability estimates for this dataset.

A big difference between the occupation data and the industry data is the assumption that a particular document is only relevant to one occupation, so the data only includes a single prediction for each document (with the probability of that prediction for the 2-digit data).

Technical Notes:

- The 2-digit estimator was built using a logit regression model, so the included probability values can be used to quantify occupation-relevant metrics across regulatory codes.
- The 3- and 4-digit estimator was built using an SVM, which while more accurate than a logit model, does not produce probability values. Since this estimator was trained on occupational licensing regulations, it is not recommended to use this dataset to analyze non-OL regulations

QuantGov

Citation:

If you use this data, please cite:

McLaughlin, Patrick A., Jonathan Nelson, Stephen Strosko, and Walter Stover. Occupation Data (dataset). QuantGov, Mercatus Center at George Mason University, Arlington, VA, 2020.

Current Version: 1.0
Release Date: 10-14-2020

Variable Descriptions:

Dataset	Variable	Description
2-Digit SOC	date_collected	Date the text was gathered
	name	A constructed name for the given document based off of metadata objects
	soc_code	2-digit SOC code to which the document is relevant
	probability	Probability the document is relevant to the classified SOC code
3 and 4-Digit SOC	date_collected	Date the text was gathered
	name	A constructed name for the given document based off of metadata objects
	soc_code	3 or 4-digit SOC code to which the document is relevant