Federal Healthcare RegData 2.0 User's Guide

Kofi Ampaabeng and Stephen Strosko

January 20, 2021

1 Purpose

Federal Healthcare RegData 2.0 builds on the Mercatus Center at George Mason University's RegData project to catalog federal and state regulations. The Federal Healthcare project uses the QuantGov platform to analyze regulations that specifically affect the provision of healthcare. This project focuses on regulations that exist in the Code of Federal Register (CFR) for the years 1970 to 2019.

2 Overview

Federal Healthcare RegData identifies the number of regulatory restrictions within a unit of regulatory text that are related to healthcare. For the 2.0 iteration of Federal Healthcare RegData, a machine learning algorithm was developed to directly identify regulatory text related to healthcare. This is a significant improvement over the 1.0 version where researchers identified all sections of the CFR that contained a healthcare regulation by hand.

Two additional changes differentiate the 2.0 version from the 1.0 version. First, the 2.0 version expands on the single year of data in the 1.0 version and now covers every year found in the RegData 3.2 release, 1970 through 2019. Second, the 2.0 version analyzes text at the CFR part level instead of the section level. The CFR is organized into 50 titles, each covering a broad topic. For example, title 42 covers health care regulations while title 45 covers public welfare. Each title is further divided into parts and parts.

The reason for choosing the part as the unit of analyses in the 2.0 is due to both algorithm accuracy and ease of use. After testing the algorithm at both the part and section level, part level results were slighly more accurate due to the larger amount of text that is in each document. RegData 3.2 also uses the part level for analysis - allowing for easy analysis between the two projects. The downfall of using the part level instead of the section level is that some parts may cover more than one topic. This is partially solved by using a probability to describe each part of regualtory text.

3 Classification

The main data series for Federal Healthcare 2.0 is a probability for each title and part combination is the CFR. This probability describes the likelihood that a piece of regulatory text is related to the topic healthcare. More specifically, we define healthcare language as language that regulates the following: the provision of healthcare, of health professionals, drugs for human use, medical products, and healthcare research. Healthcare RegData is intended to exclude regulations that pertain to other areas such as occupational health and safety, environment, social assistance, and other such regulations, which are meant to affect health outcomes, but not the provision of healthcare services. However, the distinction between these two closely related buckets of topics is still a work in progress. Many regulations that cover topics in the second bucket will still be classified with a high probability in the version 2.0 data series.

Additionally, RegData 3.2 classification results are able to be used with the Federal Healthcare 2.0 results. RegData 3.2 classifies regulatory text at the part level in the CFR as having probabilistic associations with industries found in the North American Industry Classification System (NAICS). This data series can easily be merged by using the document id for each piece of regulatory text or the unique year-title-part document index.

4 Algorithm Specifics

The algorithm used to classify regulatory text for the Federal Healthcare 2.0 project is a logit algorithm. This algorithm was chosen after testing, via 5-fold cross-validation, a selection of SVM, random forest, logit, and deep learning algorithms. While the deep learning algorithms tested also performed well, and will most likely be used in future iterations of the project, the selected logit algorithm performed just as well and ascribes to Occam's Razor.

Extensive preprocessing of the regulatory text took place before algorithm testing and training to increase algorithm accuracy. Preprocessing steps included, stemming and lemmatization, the removal of stop words, and threshold based tokenization (at the 2-word level). In addition, the title of each regulatory document was multiplied an additional ten times to provide a sort of "weight" to the algorithm's consideration of the title relative to the rest of the text in the document.

5 Additional Metadata

While the Federal Healthcare 2.0 project does not add any additional data series outside of healthcare probability, its results can be merged with any data series from the RegData 3.2 project. This includes data series that assign complexity metrics to each title and part combination, as well as data series that provide agency information for each title and part combination. These data series can easily be merged by using the document id for each piece of regulatory text or the unique year-title-part document index.

6 Technical Notes

1. Federal Healthcare RegData 2.0 (slated for future release) uses machine learning to identify healthcare regulations, resulting in a reduction of the potential errors over the 1.0 version.

2. Federal Healthcare RegData 2.0 classifies documents at the part level. Users should keep in mind that parts in the CFR can cover multiple topics and may not exclusively discuss healthcare.

3. Federal Healthcare RegData 2.0 is intended to exclude regulations that pertain to other areas such as occupational health and safety, environment, social assistance, and other such regulations, which are meant to affect health outcomes but not the provision of healthcare services. However, the distinction between these two closely related buckets of topics is still a work in progress.

Table 1: Variable Descriptions

Variable	Description Definition
document reference	The reference for the document used as a source for the prediction algorithm.
document title	The title of the document if available.
restrictions	The total number of restrictions, comprises the sum of “shall”, “must”, “may not”, “required”, “prohibited”.
shall	Occurrences of the word “shall” in the unit.
must	Occurrences of the word “must” in the unit.
may not	Occurrences of the words “may not” in the unit.
required	Occurrences of the word “required” in the unit.
prohibited	Occurrences of the word “prohibited” in the unit.
words	The total number of words in the unit.
probability (as related to healthcare)	The probability that the regulatory unit pertains to healthcare topics.
industry code	Industry classification code (NAICS).
industry probability	The probability that the regulatory unit pertains to industry identified by “industry” variable.

Table 2: Change Log

Version	Release Date	Release Type	Notes
1.0	March 2020	Major	Documents were hand selected - future iterations will use a custom algorithm.
2.0	January 2021	Major	Documents were classified via a logit algorithm.