Leveraging Natural Language Processing for Legislative Research

Federal Committee on Statistical Methodology (FCSM)
Pavani Samala – Data Scientist (CTR)
Amber Hennessy – Survey Statistician
Economic Statistical Methods Division (ESMD)

ACKNOWLEDGEMENTS

U.S. CENSUS - ESMD TEAM
ANDREANA ABLE, ROBYN HARRIS,
AMBER HENNESSY, JOY PIERSON,
& MERCERA SILVA

REVEAL TEAM
JAYRAM ATHIMOOLAM, PAVANI SAMALA
& TAYLOR WILSON

United States® Census Bureau

Outline

1. Background
2. Problem
3. Solution
    - 3.1 Web Scraping
    - 3.2 Creating a Dataset
    - 3.3 Classification Algorithm
4. Puerto Rico
5. Overall Workflow
6. Streamlit Application
7. Burden Reduction Results
8. Challenges

United States® Census Bureau

Background

- The Census of Governments (CoG) is conducted every five years. This census collects and publishes data on government organization, employment, and finance from all state and local governments in the U.S.

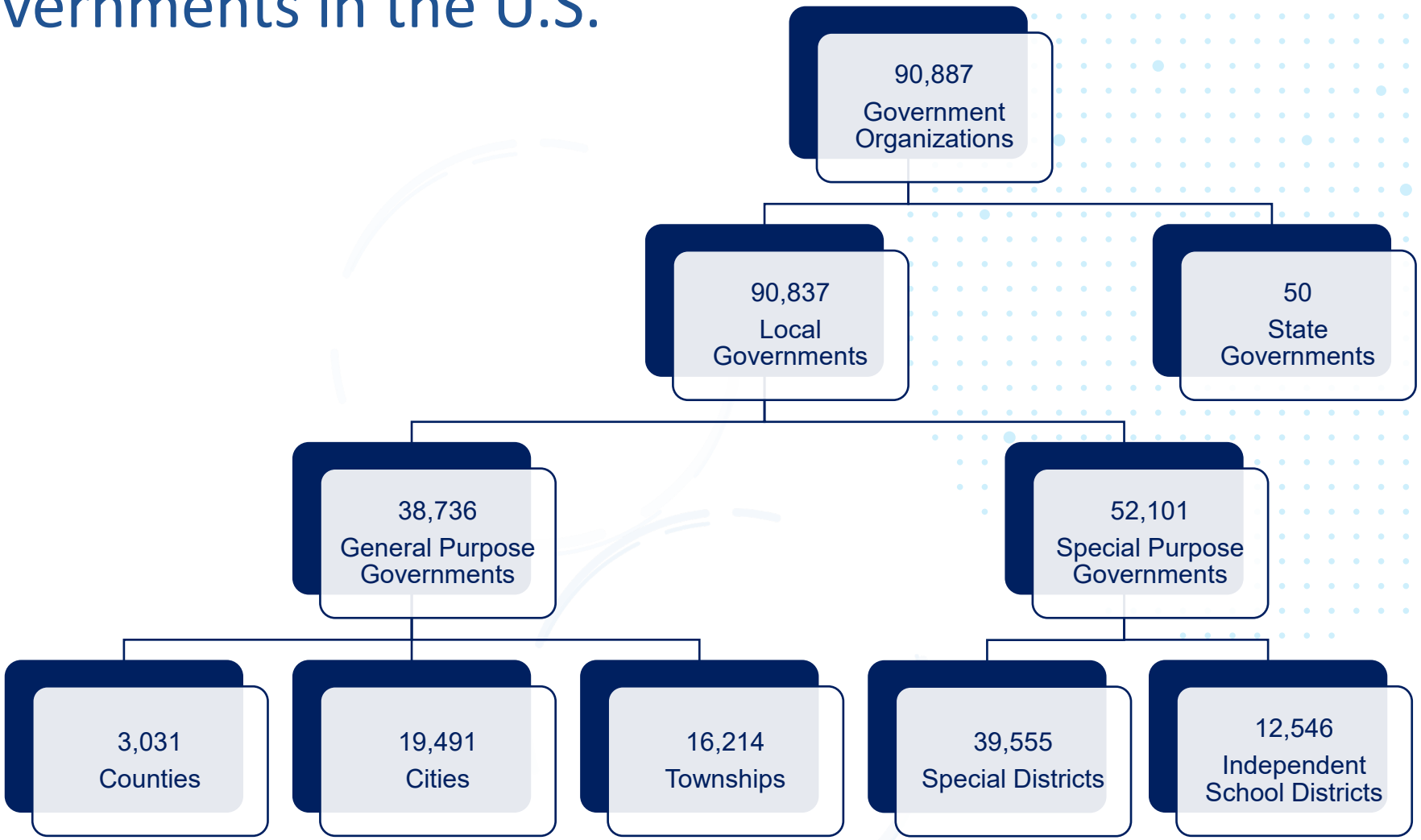
- U.S. Census Bureau definition of a government:

  A government is an organized entity having governmental character and sufficient discretion in the management of its own affairs to distinguish it as separate from the administrative structure of any other governmental unit within that state.

  An entity must possess three attributes to be counted as a government:

  - Existence as an organized entity.
  - Governmental character.
  - Substantial autonomy.

Types of Governments in the U.S.



Source: U.S. Census Bureau, 2022 Census of Governments: Organization

Legislative Research

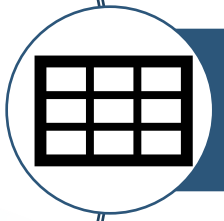
- The presence of certain language in a state's laws is an important indicator of whether the change criteria is or is not met.
  - Creation of a government (**Birth**)
  - Dissolution of a government (**Death**)
  - Consolidation of two or more governments into a new government (**Merger**)
  - Name change of a government agency (**Name Change**)
  - Identify hierarchal relationships (**Parent/Child association**)
- Why is Legislative Research Important?
  - Frame maintenance (Government Master Address File) for the CoG, annual public sector surveys, and Individual State Descriptions report

Problem

- Public Sector Frame and Classification Branch (PSFCB) is responsible for reviewing thousands of bills passed annually for all 50 states and the District of Columbia to identify legislative changes that lead to births, deaths, and mergers.
- Current process is very time consuming. Requires analysts to search state legislature websites, find, read, and assess each bill for relevance.
- Analysts identify a relatively small number of changes among the hundreds of bills reviewed, because impactful changes are infrequent.
- ***New Challenge***: Integrate Puerto Rico into the legislative research workflow

Automation Based Solution

Use web scraping to expedite the search process for gathering legislative bills for all 50 states, the District of Columbia, and Puerto Rico



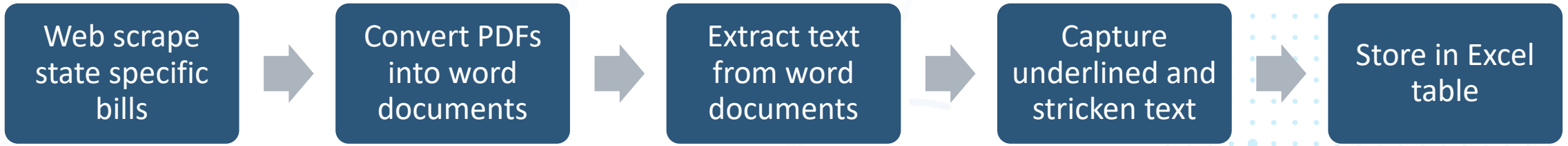
Create a structured dataset that can be inputted into a model

Use Natural Language Processing (NLP) to score bills based on relevant key words and phrases for analyst review

Web Scraping

- Each state has rules and regulations for web scraping
  - Robot.txt
  - Legiscan
- Each state has specialized code to web scrape legislative bills
- Python code needs to include error handling capabilities
  - Alert the user when a specific state's website has changed, while proceeding to scrape other states

Creating a Dataset



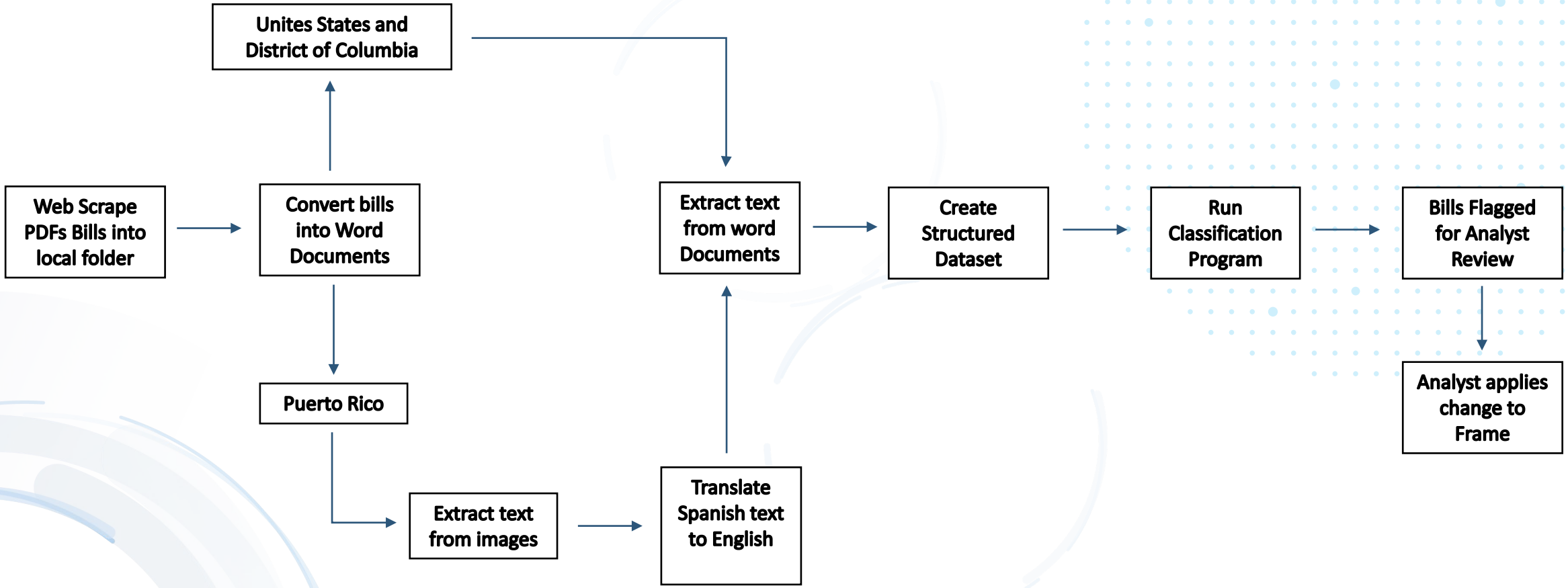
Classification Program

- Birth, death, and merger keywords
  - Unique to each state
- Text cleaning
  - Remove “decorative characters”, bill history, and stopwords
- Cosine similarity
  - Length of text does not skew similarity measure
- Positional weighting
  - Birth keywords in the beginning of the text are weighted higher
- Regular expressions
  - Search for a government

Puerto Rico

- Puerto Rico bills are not consistently stored and need additional processing
  - Bills include images of paper legislation
  - Bills are published in Spanish
- Use Object Character Recognition (OCR) to extract text from images
- Use an LLM to translate Spanish text to English text
  - The Spanish translation quality was assessed by native Spanish speakers with Puerto Rican ancestry
  - Companion research is being conducted to quantify the LLM translation quality

Overall Workflow



Streamlit Application

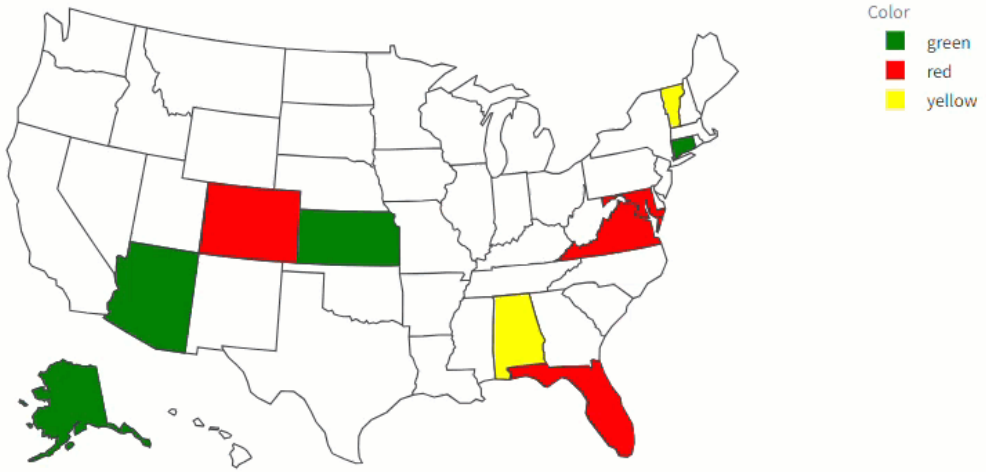
United States Census Bureau

Legal Research Automation Tool beta

About | Data Scraper | Classification | Results | Tracker

Data Scraper

Response time for various states



This tab can be used to collect all the bills passed at the state level. The **Individual** mode can scrape and

Burden Reduction Results: 2020 Legislation

State Legislature	Total Enacted Bills	Number of Bills to Review: Possible Birth, Death, Merger, or Name Change	Total Burden Reduction (%)
Arizona	90	9	90%
Minnesota	118	18	84.75%
Illinois	673	113	83.21%
Texas	1,373	249	81.86%
New York	1,349	283	79.02%
Florida	182	44	75.82%
All U.S. States Total	21,893	4,196	80.83%

Sorted by percentage of reduced analyst review by state

United States® Census Bureau

Ongoing Challenges

- Some states require permission for web scraping (API keys)
- Very rarely, some states do not share their bills to the public
- States alter their website infrastructure occasionally
- States may use inconsistent language or bill structure that may cause relevant changes to be missed
- Accessing all Python packages behind the Census Firewall and determining a long-term database solution

Thank You

pavani.j.samala@census.gov

esmd.gus.psfcb@census.gov

U.S. Census Bureau, 2022 Census of Governments, *Individual State Descriptions: 2022*
https://www.census.gov/library/publications/2024/econ/2022isd.html

Government Organization & Structure
https://www.census.gov/topics/public-sector/government-organization.html

United States® Census Bureau