Tools for Scraping and Classifying Data from Government Websites

Brian Dumbacher (brian.dumbacher@census.gov)

Background

- The U.S. Census Bureau's public sector surveys collect data from state and local governments.
- The Quarterly Summary of State and Local Government Tax Revenue (QTax) collects data on tax revenue collections.
- Respondents sometimes direct QTax analysts to their websites.
- Going directly to websites to obtain the data could reduce burden on respondents and QTax analysts and increase the timeliness of data products.

Goal

Develop tools for scraping and classifying tax revenue data from state and local government websites. Toolkit is called:

SABLE – Scraping Assisted By LEarning

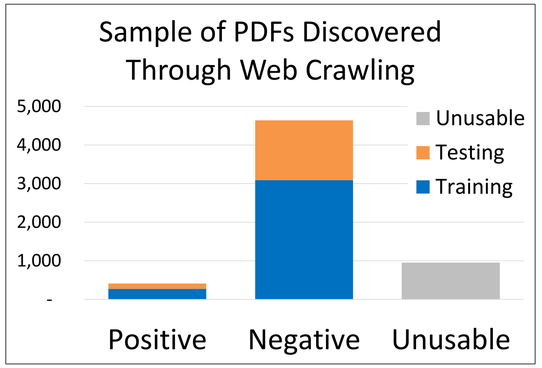
Challenges

- There are thousands of government websites with little standardization.
- Most data and publications are in Portable Document Format (PDF).

SABLE Methodology

Crawl Relevant Websites and Create a Corpus of PDFs

- Compile a list of government websites related to revenue, taxation, and finance
- Use Apache Nutch to crawl these websites and discover PDFs
- Manually classify the PDFs as positive (contains relevant data) or negative



Use Machine Learning to Classify PDFs as Positive or Negative

- Work is done using Python
- Convert PDFs to text format
- Use text analytics techniques to remove common words and create *n*-grams, which are sequences of *n* words

n-Gram Example

Snippet of PDF text after conversion:
" … total general revenue … "

1-grams: (**total**), (**general**), and (**revenue**)
2-grams: (**total**, **general**) and (**general**, **revenue**)
3-grams: (**total**, **general**, **revenue**)

- Construct 0/1 indicators that indicate the presence of *n*-grams in the PDF
- Use these indicators as model features
- Apply machine learning methods to the training set
  - Support Vector Classifier (SVC)
  - Naïve Bayes
  - Decision Tree
- Evaluate models using the test set

Results for SVC Using 1-grams and 2-grams

Confusion Matrix

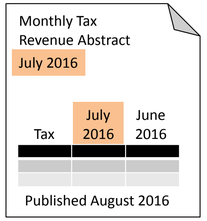
True Class	Predicted Class	
	Positive	Negative
Positive	115	22
Negative	7	1,540

F1 Score
0.888

Accuracy
0.983

Apply Models to Scrape Data and Contextual Information

- Scrape year and month
- Model based on frequency and location of year and month in PDF



Current Work

- Map non-standardized terminology in government PDFs to the Census Bureau's tax codes definitions
  - Training set is being developed
  - Same machine learning methods such as SVC can be applied
- Use SABLE to create an experimental data product such as a Monthly Summary of State Government Tax Revenue

References

- The Apache Software Foundation. (2014). Apache Nutch. *<http://nutch.apache.org>*
- Dumbacher, B. and Capps, C. (2016). Big Data Methods for Scraping Government Tax Revenue From the Web. *2016 Proceedings of the American Statistical Association, Section on Statistical Learning and Data Science*, 2940-2954.
- U.S. Census Bureau. (2017). Quarterly Summary of State & Local Taxes. *<http://www.census.gov/govs/qtax>*