

# SABLE: Tools for Web Crawling, Web Scraping, and Text Classification

Federal Committee on Statistical Methodology  
Research Conference

March 7, 2018

Brian Dumbacher  
Lisa Kaili Diamond  
U.S. Census Bureau

# Outline

- Motivation
- SABLE Overview
- Applications
- Moving to a Production Environment
- Quality
- Future Work

# Motivation

- For many economic surveys, respondent data or equivalent-quality data are available online
- Respondents sometimes direct Census Bureau analysts to their websites to obtain data
- Going directly to online sources and collecting data passively could reduce respondent and analyst burden
- For the most part, current data collection efforts along these lines are manual

# Goal

Automate the process of finding, scraping, and organizing data from online sources

- Challenges
  - Websites and the documents on them lack standardization
  - Data are often in formats not amenable to analysis right away such as Portable Document Format (PDF)

# What is SABLE?

- Scraping Assisted by Learning
- Collection of tools for
  - Crawling websites
  - Scraping documents and data
  - Classifying text
- Models based on text analysis and machine learning
- Implemented using free, open-source software
  - Apache Nutch
  - Python

# Three Main Tasks

Crawl



Scrape



Classify

Given a website,

- Scan website
- Find documents and extract text
- Apply classification model to predict whether document contains useful data

Given a document classified as useful,

- Apply model to learn the location of useful data
- Extract numerical values and corresponding text

Given scraped data,

- Preprocess data
- Apply classification model to map text to Census Bureau definitions and classification codes

# Machine Learning

- In some applications, machine learning is used to classify text
- Examples of text classes
  - Document is “Useful” or “Not Useful”
  - Census Bureau classification codes
- Machine learning models pick up on associations between word sequences and classes
- Building a training set on which to fit models is manually intensive and usually time-consuming

# Training Set

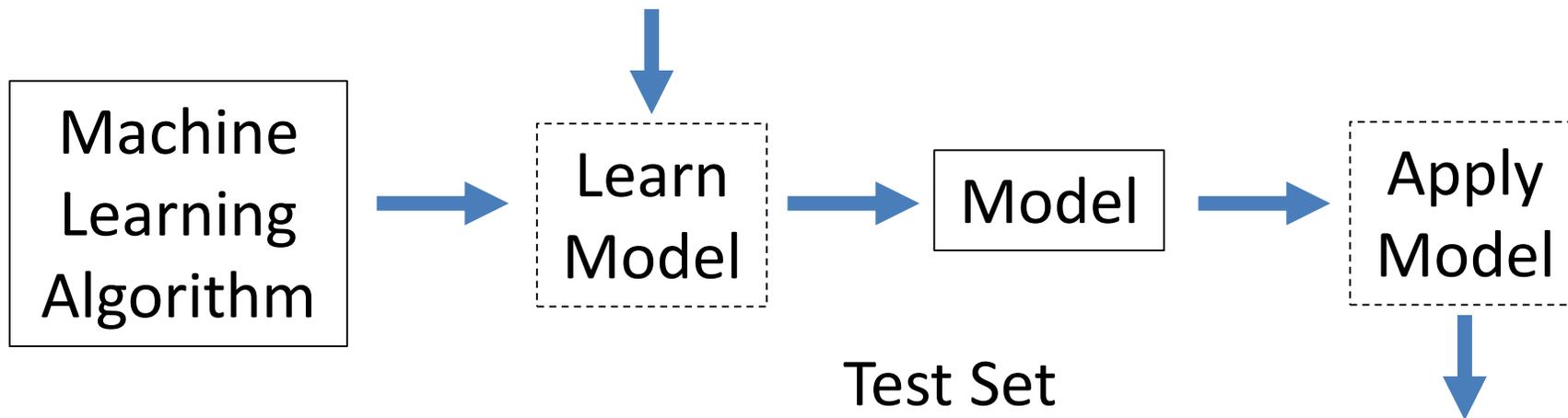
Document Text	Feature for "tax"	Feature for "revenue"	...	Feature for "tax revenue"	True Class
tax revenue for local governments	1	1		1	Useful
state government tax collections	1	0		0	Useful
instructions for filling out tax form	1	0		0	Not Useful
retirement announcement	0	0		0	Not Useful
tax revenue statistical abstract	1	1		1	Useful
upcoming road closures	0	0		0	Not Useful

# Test Set

Document Text	Feature for "tax"	Feature for "revenue"	...	Feature for "tax revenue"	True Class	Predicted Class
parks and recreation guide	0	0		0	Not Useful	?
local government tax collections	1	0		0	Useful	?

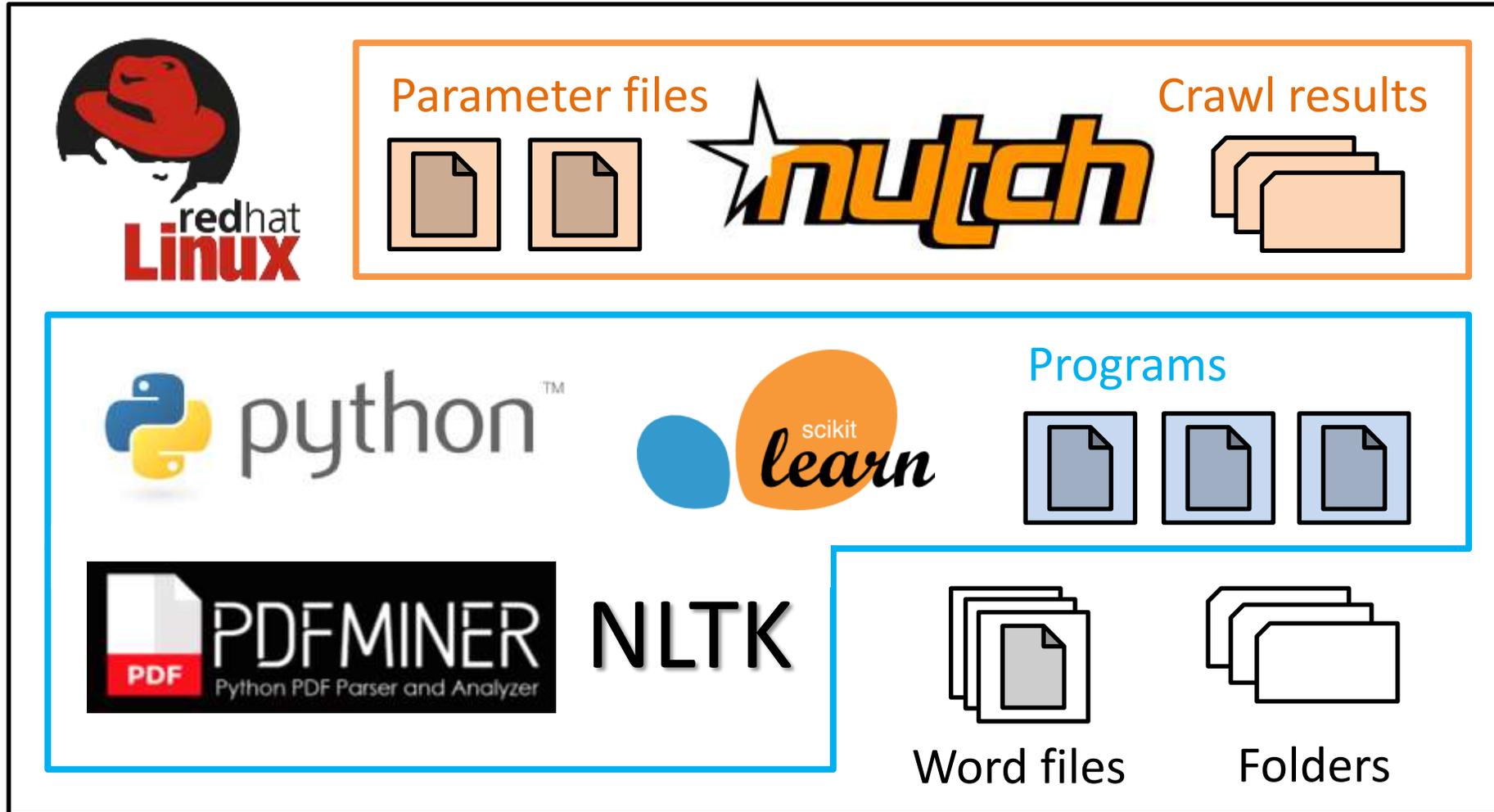
# Training Set

Document Text	Feature for "tax"	Feature for "revenue"	...	Feature for "tax revenue"	True Class
tax revenue for local governments	1	1		1	Useful
state government tax collections	1	0		0	Useful
instructions for filling out tax form	1	0		0	Not Useful
retirement announcement	0	0		0	Not Useful
tax revenue statistical abstract	1	1		1	Useful
upcoming road closures	0	0		0	Not Useful



Document Text	Feature for "tax"	Feature for "revenue"	...	Feature for "tax revenue"	True Class	Predicted Class
parks and recreation guide	0	0		0	Not Useful	?
local government tax collections	1	0		0	Useful	?

# Architecture Design



Firewall

External public website

# Applications and Research Areas

- Public sector projects
  - Find new sources of tax revenue collection statistics on government websites
  - Scrape pension statistics from specific Comprehensive Annual Financial Reports (CAFRs)
- Autocoding and write-ins
  - Assign North American Industry Classification System (NAICS) codes to establishments based on business descriptions for the Economic Census

# Tax Revenue Statistics

- Data on state government tax revenue collections can be found online in CAFRs and other publications
- Used SABLE to find additional online sources
  - Crawled websites of state governments
  - Discovered approximately 60,000 PDFs
  - Manually classified a random sample of 6,000 PDFs as “Useful” or “Not Useful”
  - Applied machine learning to build text classification models based on occurrences of word sequences

# Pension Statistics

- Likewise, data on public pension funds can be found online and in CAFRs
- Examine feasibility of scraping data
- Pension statistics
  - Service cost
  - Interest
- Two-stage approach
  - Identify tables using occurrences of word sequences
  - Apply scraping algorithm based on table structure

# REQUIRED SUPPLEMENTARY INFORMATION – PENSION

## CHANGES IN NET PENSION LIABILITY

	Fiscal Year Ended		
	2016	2015	2014
<b>Total pension liability</b>			
Service Cost (MOY)	\$ 71,218,683	\$ 70,056,133	\$ 66,696,324
Interest (includes interest on service cost)	241,733,937	231,804,221	220,238,560
Differences between expected & actual experience	(31,199,454)	(27,900,755)	-
Benefit payments, including refunds of member contributions	(146,657,716)	(137,771,219)	(131,100,585)
Net change in total pension liability	135,095,450	136,188,380	155,834,299
Total pension liability - beginning	3,260,156,781	3,123,968,401	2,968,134,102
<b>Total pension liability - ending</b>	3,395,252,231	3,260,156,781	3,123,968,401

# Autocoding and Write-ins

- The Census Bureau classifies business establishments according to NAICS
- Information for classification comes from various sources such as write-in responses to the Economic Census
- Disadvantages of assigning NAICS codes manually
  - Expensive
  - Time-consuming
  - Introduce systematic errors
- Use text classification models developed in SABLE to automate assignment of NAICS codes

# NAICS Classification Example

**Business Description**

Paintball Field, Supplies, & Games

**Standardized Text:**

paintball field supplies games

**1-Word Sequences:**

“paintball”, “field”,  
“supplies”, “games”

**2-Word Sequences:**

“paintball field”, “field  
supplies”, “supplies games”

Sporting Goods Stores 45111026

All Other Amusement  
and Recreation Industries 71399080

# Moving to a Production Environment

- Approval to use Apache Nutch 1.13
- Two Linux servers
  - Development
  - Production
- Authority to Operate (ATO)
- SABLE repository on the Census Bureau's GitHub account
  - <https://www.github.com/uscensusbureau/SABLE>
  - Programs, supplementary files, and documentation

# Quality

- Integrate quality into SABLE early on
- Establish procedures for assessing quality on a regular basis
- Crawling and scraping
  - Manual checks
  - Comparisons with respondent data
- Machine learning
  - Recruit subject matter experts to help create training sets
  - Assess quality of predictions and identify and quantify different misclassification costs

# Future Work

- Obtain Authority to Operate
- Update SABLE GitHub repository periodically
- Create a data product based on scraped data
- Research how to assign North American Product Classification System (NAPCS) codes based on product descriptions for the Economic Census

# Contact Information

- [Brian.Dumbacher@census.gov](mailto:Brian.Dumbacher@census.gov)
- [Lisa.Kaili.Diamond@census.gov](mailto:Lisa.Kaili.Diamond@census.gov)