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

Brian Dumbacher[1], Lisa Kaili Diamond[1]
Brian.Dumbacher@census.gov, Lisa.Kaili.Diamond@census.gov
[1]U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Proceedings of the 2018 Federal Committee on Statistical Methodology (FCSM) Research Conference

Abstract
For many economic surveys conducted by the U.S. Census Bureau, respondent data or equivalent-quality data can sometimes be found online such as on respondent websites and government agency websites. An automated process for finding useful data sources and then scraping and organizing the data is ideal but challenging to develop. Websites and the documents on them have various formats, structures, and content, so a long-term solution needs to be able to deal with different situations. To this end, Census Bureau researchers are developing a collection of tools for web crawling and web scraping known as SABLE, which stands for Scraping Assisted by Learning (as in machine learning). Elements of SABLE involve machine learning to perform text classification and autocoding. SABLE is based on two key pieces of open-source software: Apache Nutch, which is a Java-based web crawler, and Python. This paper gives an overview of SABLE and describes research to date, potential applications to economic surveys, efforts in moving to a production environment, and future work.

Key Words: U.S. Census Bureau, economic statistics, web crawling, web scraping, text classification

1. Introduction

1.1 Background

For many economic surveys conducted by the U.S. Census Bureau, respondent data, equivalent-quality data, and relevant administrative records can sometimes be found online. For example, the Census Bureau conducts public sector surveys of state and local governments to collect data on public employment and finance (U.S. Census Bureau, 2017a). Much of this data is publicly available on respondent websites in Comprehensive Annual Financial Reports (CAFRs) and other publications. Another example of an online data source is the Securities and Exchange Commission (SEC) EDGAR database. The EDGAR (Electronic Data Gathering Analysis and Retrieval) database contains financial filing information for publicly traded companies and is used often by Census Bureau analysts to impute missing values and validate responses for many economic surveys. Going directly to online sources such as these and collecting data passively has a lot of potential to reduce respondent and analyst burden (Dumbacher and Hanna, 2017). For the most part, the Census Bureau's processes for collecting economic data from online sources are manually intensive. Efficiency can be improved greatly by using automated methods such as web scraping (Mitchell, 2015).

1.2 Challenge

An automated process for finding useful data sources and then scraping and organizing the data is ideal but challenging to develop. Websites and the documents on them have various formats, structures, and content, so a long-term solution needs to be able to deal with different situations. To this end, Census Bureau researchers are developing tools for web crawling and web scraping that are assisted by machine learning. This collection of tools is known as SABLE, which stands for Scraping Assisted by Learning. Elements of SABLE involve machine learning to perform text classification [for a discussion of text analytics topics, see Hurwitz *et al.* (2013, chap. 13)] and autocoding (Snijkers *et al.*, 2013, p. 478). Text classification models are used for different reasons, such as predicting whether a document contains useful data or mapping scraped data to Census Bureau terminology and classification codes.

Disclaimer: Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

1.3 Outline

The rest of the paper is organized as follows. Section 2 gives an overview of SABLE, its machine learning methodology, underlying software, and architecture design. Section 3 covers potential applications and ongoing areas of research such as public sector surveys, SEC metadata, and text classification problems for assigning codes to survey write-in responses. SABLE is currently being moved from a research environment to a production environment, and Section 4 describes this effort. Lastly, Section 5 describes future work, particularly ideas for quality assurance.

2. SABLE Overview

2.1 Main Tasks

SABLE performs three main tasks: web crawling, web scraping, and text classification. Web crawling is the automated process of systematically visiting and reading web pages. Web crawlers, also known as spiders or bots, are typically used to build search engines and keep website indices up to date. For SABLE, web crawling is used to discover potential new data sources on external public websites and to compile training sets of documents for building classification models.

Web scraping involves finding and extracting data and contextual information from web pages and documents. This is an automated process and an example of passive data collection, whereby the respondent has little awareness of the data collection effort or does not need to take any explicit actions. In order to scrape data from some documents, they might have to be converted to a format more amenable to analysis. This is especially true for documents in Portable Document Format (PDF). Models based on the frequencies and locations of important word sequences can be employed to find useful data in documents.

Text classification is the task of assigning text to a category, or class, based on its content and important word sequences. SABLE uses machine learning to classify text. Text classification models can be used to predict whether a document contains useful data or to map scraped data to the Census Bureau's terminology and classification codes. The models developed for this task have also found applications beyond web scraping to the automation of classifying survey write-in responses.

Table 1, which is adapted from Dumbacher and Hanna (2017), summarizes the tasks performed by SABLE. Not all three tasks may be relevant to a given application. For example, data sources may already be determined, so it may not be necessary to perform web crawling. In this case, the problem would consist of just scraping and classifying data from known websites and documents.

Table 1. Three Main Tasks Performed by SABLE

Web Crawling
• Scan websites • Discover documents • Compile a training set of documents for building classification models
Web Scraping
• Find the useful data in a document using the frequencies and locations of important word sequences • Extract numerical values and contextual information such as data labels
Text Classification
• Predict whether a document contains useful data • Map scraped data to the Census Bureau's terminology and classification codes using data labels associated with the scraped data • Classify survey write-in responses