# Big Data Methods for Scraping Government Tax Revenue from the Web

Brian Dumbacher
U.S. Census Bureau

Cavan Capps
U.S. Census Bureau

United States Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

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

# Outline

- Introduction
- Background
  - Big Data
  - Portable Document Format
- Machine Learning
  - Training and Test Sets
  - Features and Classifiers
- Results
- Future Research

# Introduction

- Quarterly Summary of State and Local Government Tax Revenue (QTax)
- Some respondents direct QTax analysts to websites to obtain data
- Going directly to websites to obtain data on tax revenue collections could
  - Reduce respondent burden
  - Aid data review, imputation, and verification

# Introduction (cont.)

- Automated process for scraping data from government websites ideal but challenging
- Large majority of government publications are in Portable Document Format (PDF)
- Goal is to use a web crawler and predict whether a new PDF contains relevant data
  - Unstructured data
  - Text analytics
  - Classification

# Big Data

- Methods belong to realm of Big Data
- Big Data are “found” or “organic” data
- Important to consider how representative the data are of the target population
- This research is part of the Census Bureau’s effort to use Big Data to enhance its economic programs

# Portable Document Format (PDF)

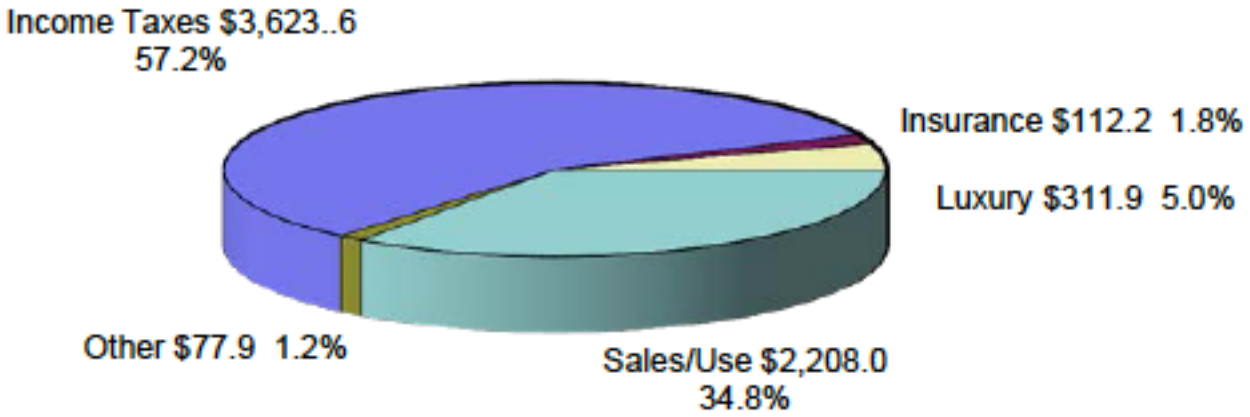
- File format for presenting documents in a way that does not depend on operating system
- Text described in terms of hierarchy
  - pages
  - textboxes
  - textlines
  - characters
- Elements assigned identification numbers and x- and y-coordinates

# PDF Conversion Algorithm

- Extract text and convert it to TXT format for text analysis and machine learning
  - Use Python module PDFMiner to extract text and output to XML format
  - Use regular expressions to parse XML file and obtain character-level information
  - Construct words character by character
  - Check each word against an English dictionary

# Conversion Example

## STATE OF ARKANSAS
## FISCAL YEAR 2015
## ESTIMATED
## Gross General Revenue
## $6,333.6 Million



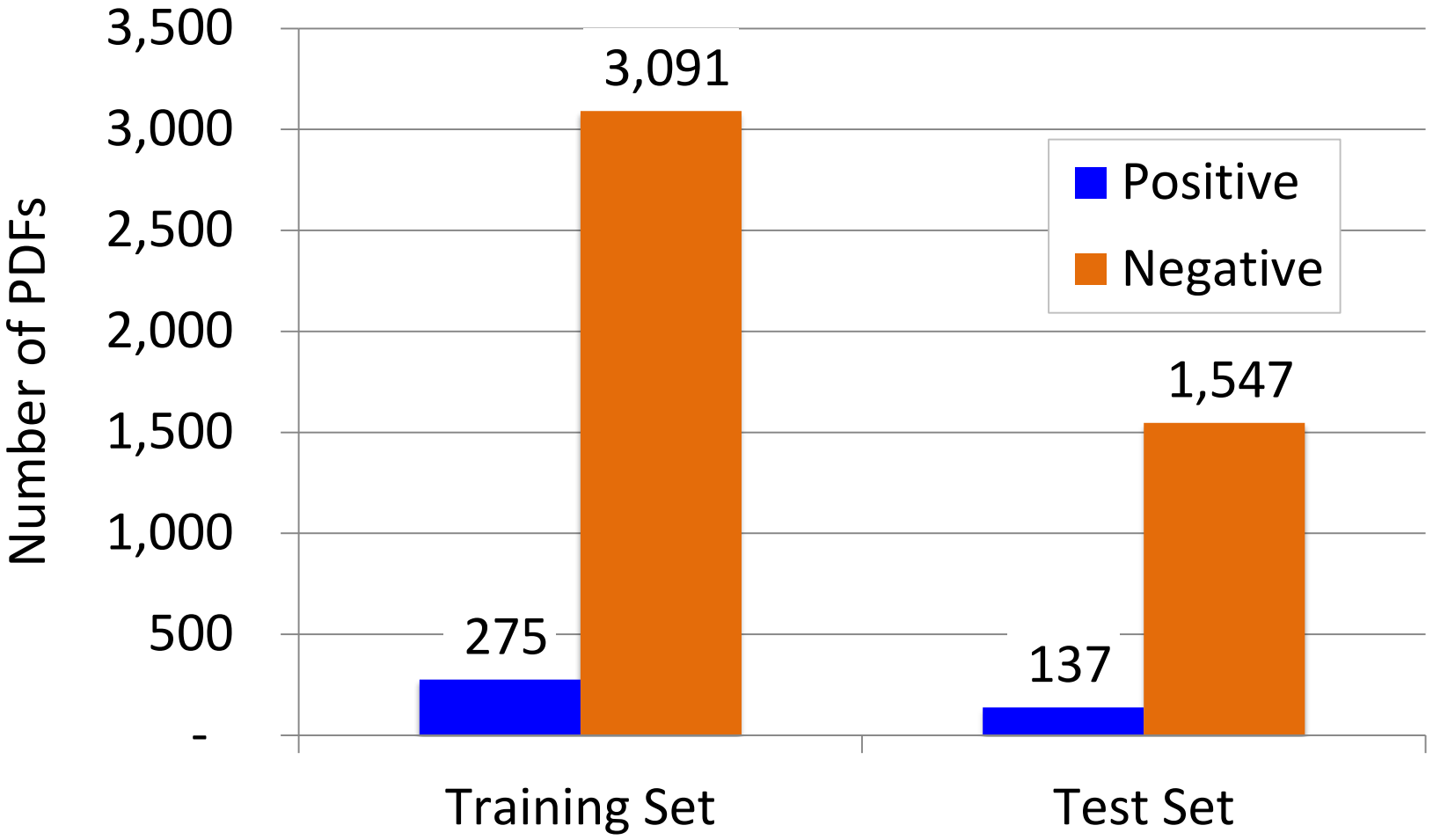
# Conversion Example (cont.)

```
state arkansas fiscal estimated gross general
revenue million income taxes insurance luxury
sales less central services educational
adequacy taxes services college savings bonds
debt service economic development incentive
educational excellence city county tourist
desegregation refunds water sewer claims
after rainy set aside total general revenue
available distribution million public schools
health human services general government
general ed inst higher ed general revenue
forecast
```

# PDFs for Machine Learning

- Collection of PDFs with class labels already assigned is required for model fitting
- Compiled list of state government websites related to finance
- Used Apache Nutch to crawl these websites and discover PDFs, 59,578 in total
- Selected a simple random sample of 6,000 PDFs and manually classified them
- 5,050 PDFs could be used in the analysis

# Training and Test Sets



# Features and *n*-grams

- For each PDF in final TXT format, a vector of features is required
- *n*-grams
  - Sequences of *n* words
  - *1*-grams are words, *2*-grams are pairs of words, and *3*-grams are sequences of three words
- Features are 0/1 indicators of *n*-grams
- Many other features are possible

# Classifiers

- Two classification methods
  - Support vector classifiers (SVC)
  - Naïve Bayes (NB)
- Seven sets of features based on combinations of *1*-grams, *2*-grams, and *3*-grams
- Python modules
  - Natural Language Toolkit (NLTK)
  - Scikit-learn

# Evaluation

- Fit classifiers on training set and apply them to the test set
- Performance measures
  - Accuracy
  - Precision
  - Recall
  - $F_1$ score

Confusion Matrix

| True Class | Predicted Class | |
|---|---|---|
| | Positive | Negative |
| Positive | TP | FN |
| Negative | FP | TN |

# Results: Important *n*-grams

| *1*-gram | *2*-gram | *3*-gram |
|---|---|---|
| (constr) | (collections, month) | (jefferson, county, county) |
| (devil) | (fuel, refunds) | (tax, tobacco, tax) |
| (curr) | (refunds, net) | (enterprises, administrative, support) |
| (riverboat) | (mining, utilities) | (companies, enterprises, administrative) |
| (addiction) | (enterprises, administrative) | (mining, utilities, construction) |
| (depict) | (statistical, report) | (sales, tax, income) |
| (boot) | (oil, severance) | (remediation, services, educational) |
| (betting) | (add, due) | (severance, tax, collections) |
| (dobson) | (activities, funds) | (county, miami, county) |
| (defraying) | (title, fee) | (county, lake, county) |

# Results: NB Method

| Features | Accuracy | $F_1$ |
|---|---|---|
| (1) | 0.857 | 0.297 |
| (2) | 0.916 | 0.262 |
| (3) | 0.919 | 0.081 |
| (1,2) | 0.914 | 0.271 |
| (1,3) | 0.920 | 0.118 |
| (2,3) | 0.918 | 0.127 |
| (1,2,3) | 0.920 | 0.172 |

# Results: SVC Method

| Features | Accuracy | $F_1$ |
|---|---|---|
| (1) | 0.984 | 0.900 |
| (2) | 0.979 | 0.858 |
| (3) | 0.972 | 0.802 |
| (1,2) | 0.983 | 0.888 |
| (1,3) | 0.979 | 0.859 |
| (2,3) | 0.979 | 0.856 |
| (1,2,3) | 0.979 | 0.857 |

# Future Research

- Tool is a combination of web crawler and classifier
- Integrate two components into an automated process
- Apply tool to the Census Bureau website and see if QTax publications are discovered and classified correctly
- Identify data in PDFs and put them in a normalized data structure

# Contact Information


- Brian.Dumbacher@census.gov
- Cavan.Paul.Capps@census.gov