NTPS Web Scraping

Center for Optimization and Data Science (CODS), US Census Bureau

Louis Avenilla

October 27, 2022, FCSM

Project Motivation

- The National Teacher and Principal Survey (NTPS) gathers information on the United States teaching staff through initial teacher listing forms and follow-up surveys

- Alternate sources of data could augment, validate, and update survey generated information
  - Vendor Supplied Data
  - **Data scraped from the Web**

Project Motivation

- Data acquisition from the Web may offer several advantages when combined or even compared with vendor supplied data

1. Control over timing
2. Transparency
3. Customizability
4. Enhanced coverage

Method Overview

1. **Query**: find websites via addresses from Google Places API
2. **Crawl** (information retrieval): explore landing page links to identify staff roster pages
3. **Extract** (information extraction): extract teacher names, positions, etc. from roster pages

United States® Census Bureau

Query

- Google Places API to acquire school websites.
  - Can ostensibly be done for any school in the country

- Some data quality concerns
  - District websites
  - Broken links
  - Incorrect websites

- Google Places API returned a URL for 90% of submitted public school addresses

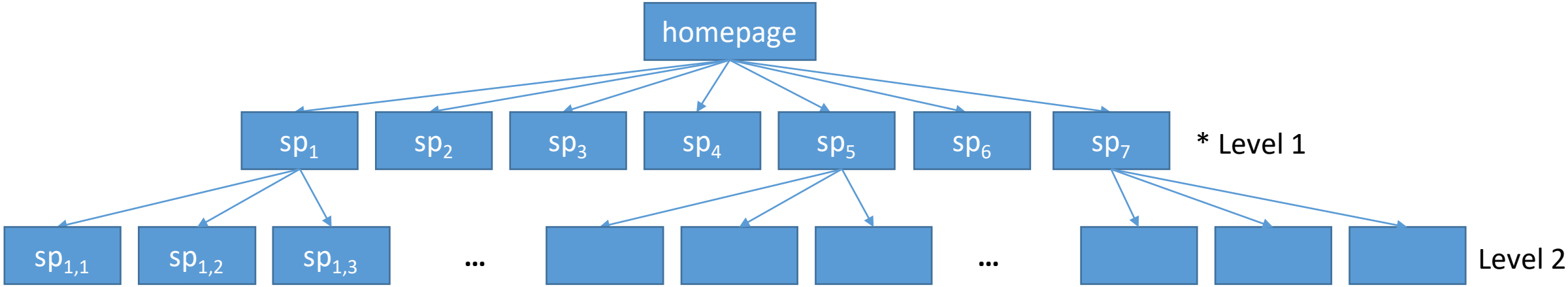
- For public schools with a returned URL, we assess the URLs from Google Places API queries to be 92% relevant[1]

[1]Doesn't account for school district URLs

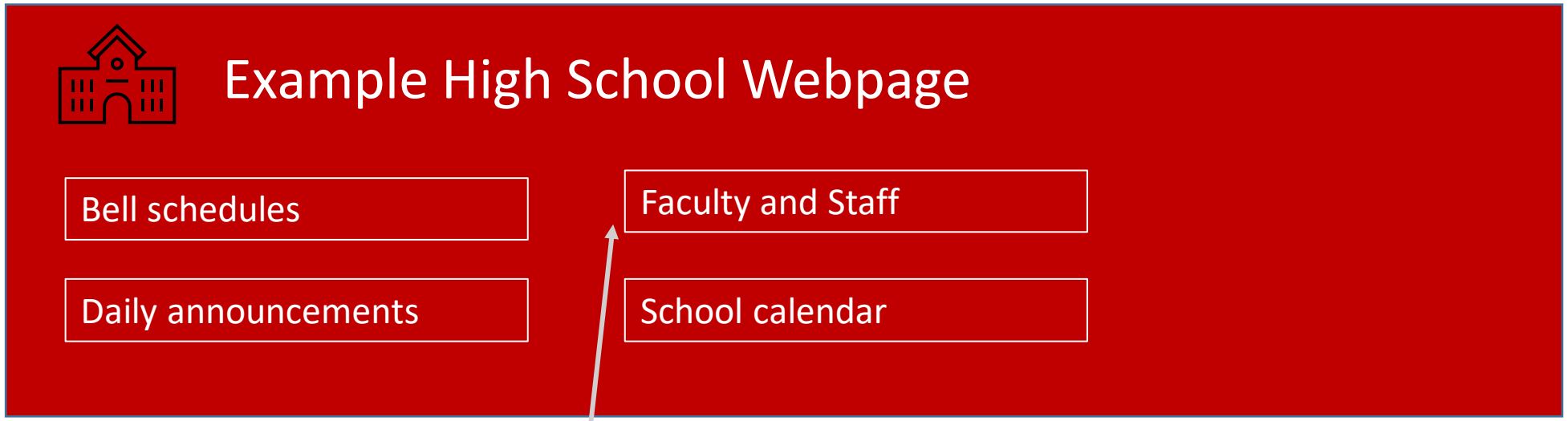
United States® Census Bureau

Crawl

- Google Places API queries provide the starting points for our initial Crawler
- The Crawler gathers level 1* of the hyperlink hierarchy for each school URL:



sp = subpage



Crawl

- Identifying - amongst the level 1 subpages - the page that lists school faculty:
  - How do we know that a page lists school faculty before acquiring that page?

Crawl

- We curated a list of expressions for faculty directories
- Then, we measured the frequency with which these expressions were used
- Finally, we construct a function that uses both the known expressions and their frequencies to estimate the likelihood that a page contains faculty directories

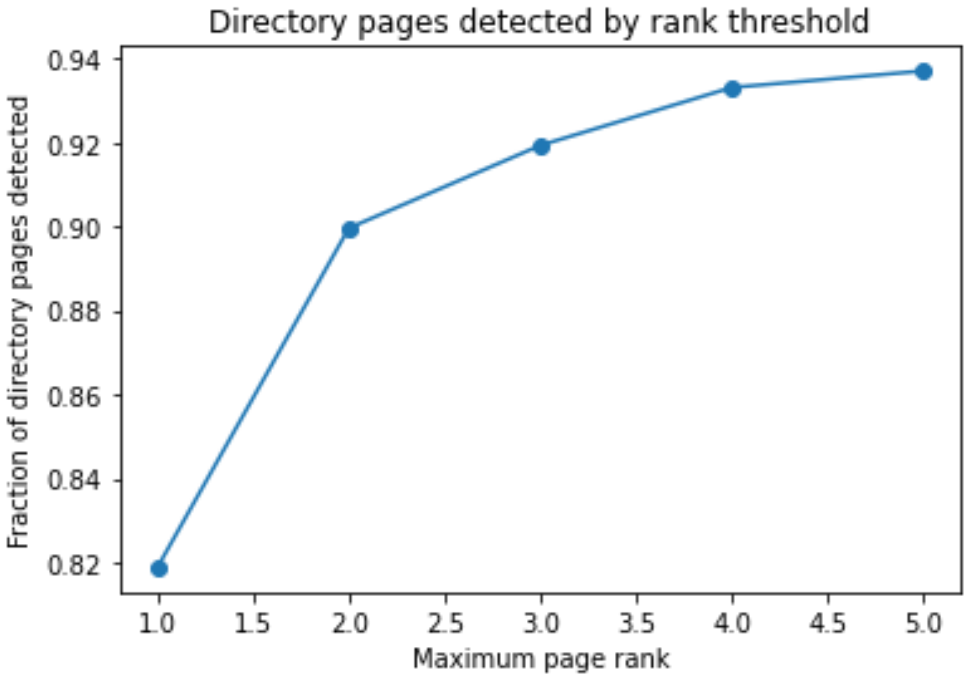
Expression	Frequency
Staff Directory	150
Staff	60
Faculty & Staff	20
Faculty and Staff	<15
Our Staff	<15
Teachers	<15
Faculty Directory	<15
Faculty	<15
Faculty & Staff Directory	<15
Teachers & Staff	<15
School Staff	<15

Table 1: curated list of expressions describing faculty directory pages

United States® Census Bureau

Crawl – How well do we detect these pages?

- For any given school, we use our function to rank its level 1 pages/links.
- We set aside the top pages for further processing (parsing)
- We manually curated approximately 3,100 pages
  - Using the top ranked page, we capture **82%** of directory pages
  - Using the top 3 ranked pages, we capture **92%** of directory pages



United States® Census Bureau

Extract - Directory Pages Sampler

First grade

Teacher name

Teacher name

Second grade

Teacher name

Teacher name

Teacher name

Third grade

Teacher name

Teacher name

School Staff

1 2 3 4 … > showing 1-4 of 65 staff

Teacher name
Titles: Second grade teacher
Emails: teacher@school.edu
Phone number: 000-000-0000

Teacher name
Titles: Second grade teacher
Emails: teacher@school.edu
Phone number: 000-000-0000

Teacher name
Titles: Second grade teacher
Emails: teacher@school.edu
Phone number: 000-000-0000

Teacher name
Titles: Second grade teacher
Emails: teacher@school.edu
Phone number: 000-000-0000

Position	Title
Teacher 1	Title A
Teacher 2	Title B
Teacher 3	Title C
Teacher 4	Title D
Teacher 5	Title E
Teacher 6	Title F
Teacher 7	Title G
Teacher 8	Title H
Teacher 9	Title I

United States® Census Bureau

Extract – Key Components

Component	Description
Parser	Extracts names, titles, and emails from webpages
Relation Extractor	Given lists of parsed names, titles, and emails, group values by person
District Detector	Determines if a page represents an individual school or a school district
District Crawler	Crawls and scrapes district information (e.g. links or teacher data) from a district page

Extract – Parser NER

- Named Entity Recognition (NER) on text peeled away from HTML elements to provide hints for people, email, and title locations

HTML

```
<div class="name" data-v-0924f08a> John Doe </div> <div class="title" data-v-0924f08a> English Teacher </div>
```

↓

Text

. John Doe . English Teacher .

↓

NER

Entity type	Value
PERSON	John Doe
TITLE	English Teacher

United States® Census Bureau

Extract – Parser Profile

- Use NER results to identify a repeating HTML profile around likely people, titles, and emails

Webpage view

Teacher name 1
Teacher title A
Example High School
xxx-xxx-xxxx

Teacher name 2
Teacher title B
Example Middle School
xxx-xxx-xxxx

Teacher name 3
Teacher title C
Example Middle School
xxx-xxx-xxxx

HTML

```
<div class="staff-info">
 <div class="name">
  Teacher name 1
 </div>
 <div class="title">
  Teacher title A
 </div>
 <div class="department">
  Example High School
 </div>
 <div class="phone">
  xxx-xxx-xxxx
 </div>
<div class="staff-info">
 <div class="name">
  Teacher name 2
 </div>
 <div class="title">
  Teacher title B
 </div>
 <div class="department">
  Example Middle School
 </div>
 <div class="phone">
  xxx-xxx-xxxx
 </div>
```

United States® Census Bureau

Extract – Parser Progress

- We ran the pipeline on a sample of approximately 80 staff directory pages

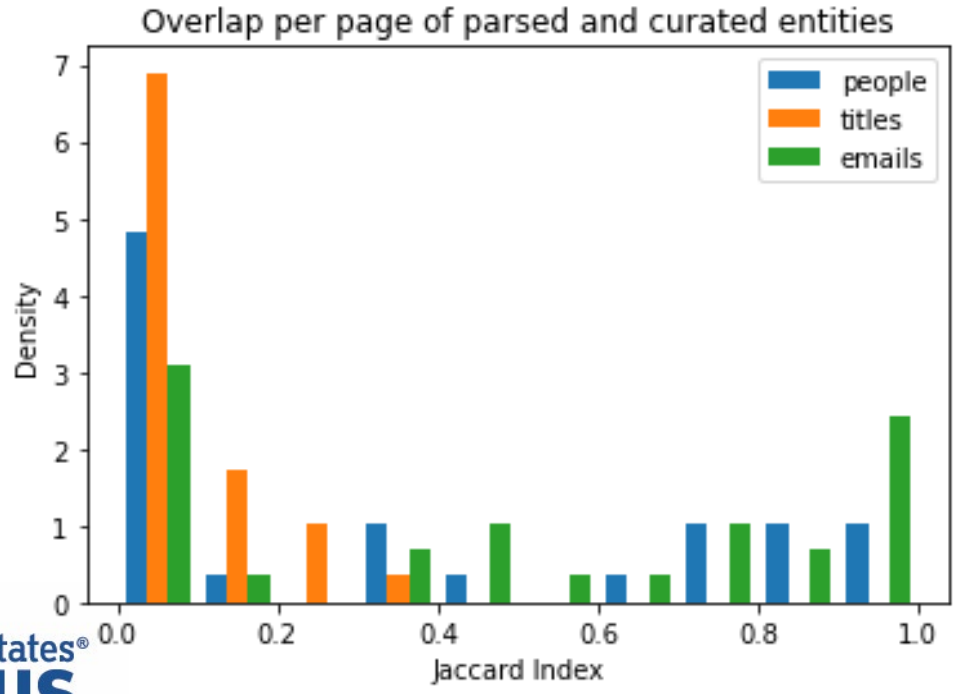
	Person	Title	Email
Total count	6,100	2,600	1,900
Percent of pages with at least one person, title, or email	90%	100%	60%
Average count per page	80	30	20

Table 2: metrics describing parser payload and coverage for a sample of staff directory pages

United States® Census Bureau

Extract – Parser Performance

- We manually curated approximately 30 pages and assessed the overlap of values per page between the parsed data and curated data



Overlap per page assessed with Jaccard Index:

$$\frac{\text{intersection}}{\text{union}}$$

United States® Census Bureau

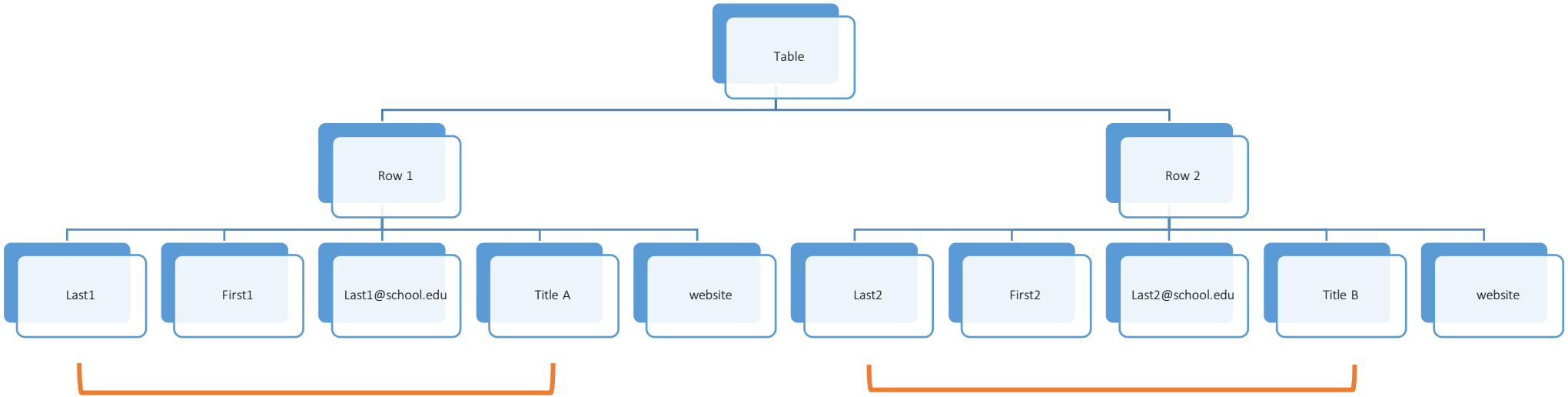
Extract – Relationship Extraction

- Traverse the HTML element hierarchy of a page to find the global minimum distance between names and titles or other elements on a page

Webpage view

Last Name	First Name	Email Address	Job Title	Website
Last1	First1	last1@school.edu	Title A	website
Last2	First2	last2@school.edu	Title B	website

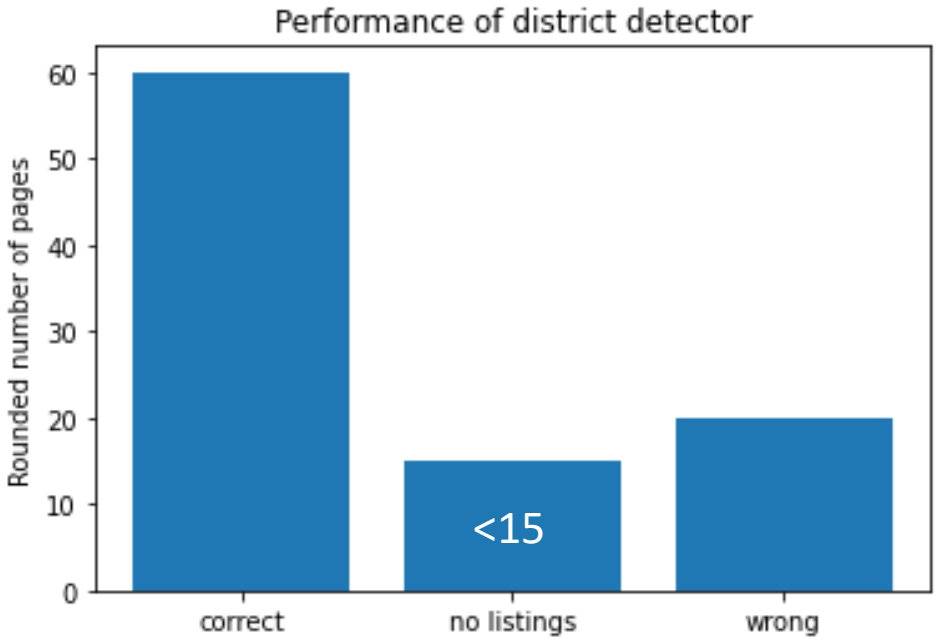
HTML hierarchy



United States® Census Bureau

Extract – District Detector

- Determines if a page represents an individual school or a school district
- Uses distribution of school names on the page



Curated ~80 schools

Recap of current status and next steps

	Query	Crawl	Extract
Achieved	1. Acquired school URLs	1. Link harvesting 2. Directory link detection 3. Dynamic scraping	1. Parsing 2. Post-processing 3. Relationship extraction 4. District detector
In progress			1. School links from district sites 2. School to staff relationship

United States® Census Bureau

Acknowledgements

Core Team

- Sara Alaoui | sara.alaoui@census.gov
- Louis Avenilla | Louis.R.Avenilla@census.gov
- Ugo Etudo | ugochukwu.o.etudo@census.gov
- Haley Hunter-Zinck | haley.s.hunter-zinck@census.gov

Special Thanks

- Patrick Campanello | patrick.campanello@census.gov
- Yathish Kolli | yathish.b.kolli@census.gov
- Anup Mathur | anup.mathur@census.gov
- Kayla Varela | kayla.m.varela@census.gov
- Allison Zotti | Allison.Zotti@census.gov

United States® Census Bureau

Thank you! Questions?