A Generic and Automated Staff Scraping Tool for School Webpages

Sara Alaoui and Haley Hunter-Zinck

Center for Optimization and Data Science (CODS)
U.S. Census Bureau

FCSM 2023

Any opinions and conclusions expressed herein are those of the author and do not reflect the views of the U.S. Census Bureau or the National Center for Education Statistics (NCES). The NCES Disclosure Review Board has reviewed this data product for unauthorized disclosure of confidential information and has approved the disclosure avoidance practices applied to this release

United States® Census Bureau

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

Project Motivation

- The National Teacher and Principal Survey (NTPS), sponsored by the National Center for Education Statistics (NCES), surveys K-12 schools and their staff
  - In order to sample teachers, the NTPS collects teacher rosters from sampled schools
  - Currently, schools submit a list of their school staff, verify a list of teachers obtained from commercial sources, or (if necessary) teachers are sampled from commercial sources
  - Sampled teachers are asked to complete a Teacher Questionnaire

- Alternate sources of data could augment, validate, and update school-submitted rosters
  - Vendor supplied data
  - **Data scraped from the Web**

The NTPS web scraper consists of three major steps to generate the final payload

Step 1: Query

Find school websites **via Google Places Application Programming** Interface (API)

Step 2: Crawl

Explore school webpage **links to identify staff** directory **pages** and download

Step 3: Extract

Extract **teacher names,** titles, and emails **from** downloaded directory pages

Final payload

Find school websites via Google Places API to provide a starting website for each school

- Query with school name and address
  - Return most relevant Google Place
  - Request associated websites, address, and name annotations

- Some data quality concerns
  - No website
  - Broken links
  - Incorrect websites
  - District websites

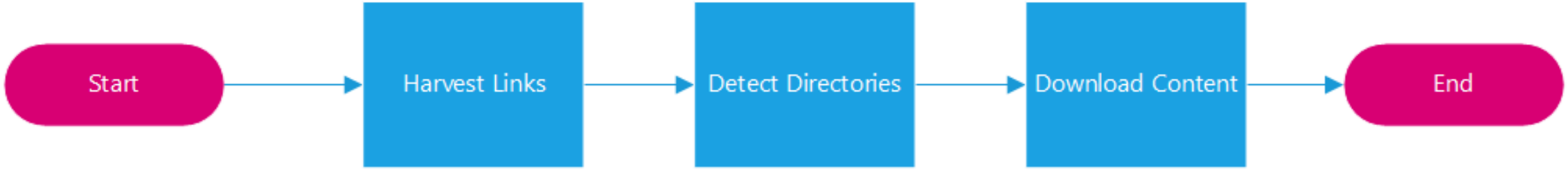
Sampled private schools are more findable than public schools on the web via Google places API

Metric	Public	Private
Number of schools	10,000	3,100
Schools in sample with a returned website	90.4%	85.3%
Website accessible and relevant*	85.0%	96.7%
Estimated return rate	76.9%	82.5%

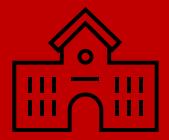
United States® Census Bureau

* determined via manual curation

Step 2: crawl | Use the returned websites to locate and download directory pages



Staff directory pages are usually linked from the school's homepage with an intuitive label



Example High School Webpage

- Bell schedules
- Faculty and Staff
- Daily announcements
- School calendar

We identify potential directory pages using a string similarity score

We curated a list of expressions for faculty directories and their prevalence

Finally, we construct a function that uses both the known expressions and their frequencies to estimate the likelihood that a page contains faculty directories

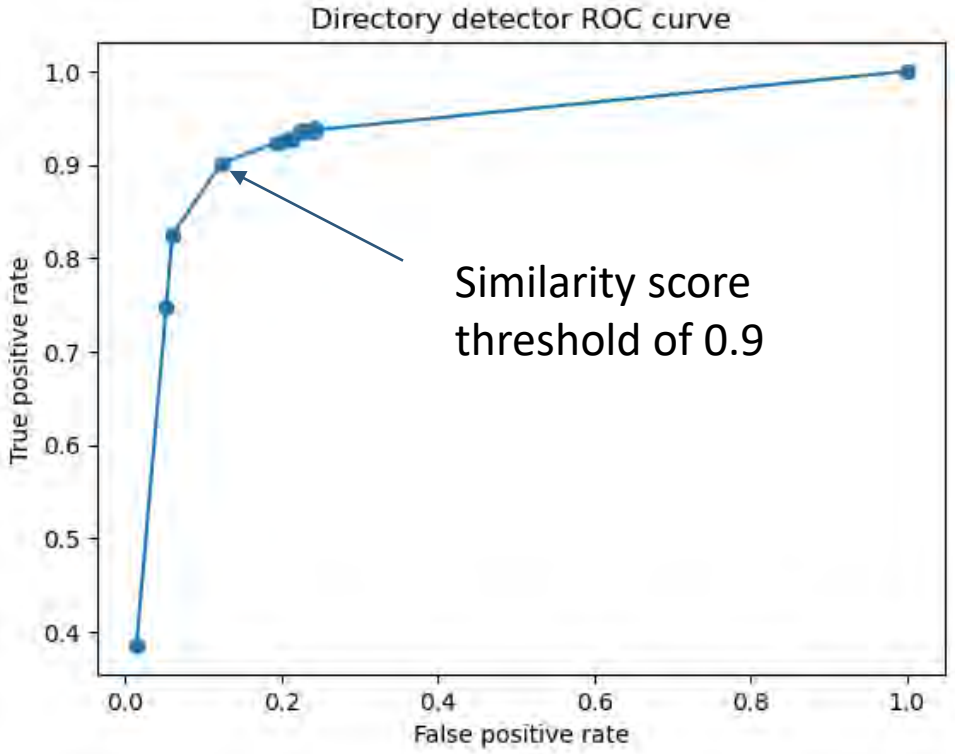
Expression	Frequency
Staff Directory	150
Staff	60
Faculty & Staff	20
Faculty and Staff	<15
Our Staff	<15
Teachers	<15
Faculty Directory	<15
Faculty	<15
Faculty & Staff Directory	<15
Teachers & Staff	<15
School Staff	<15

Curated list of expressions describing faculty directory pages for public schools

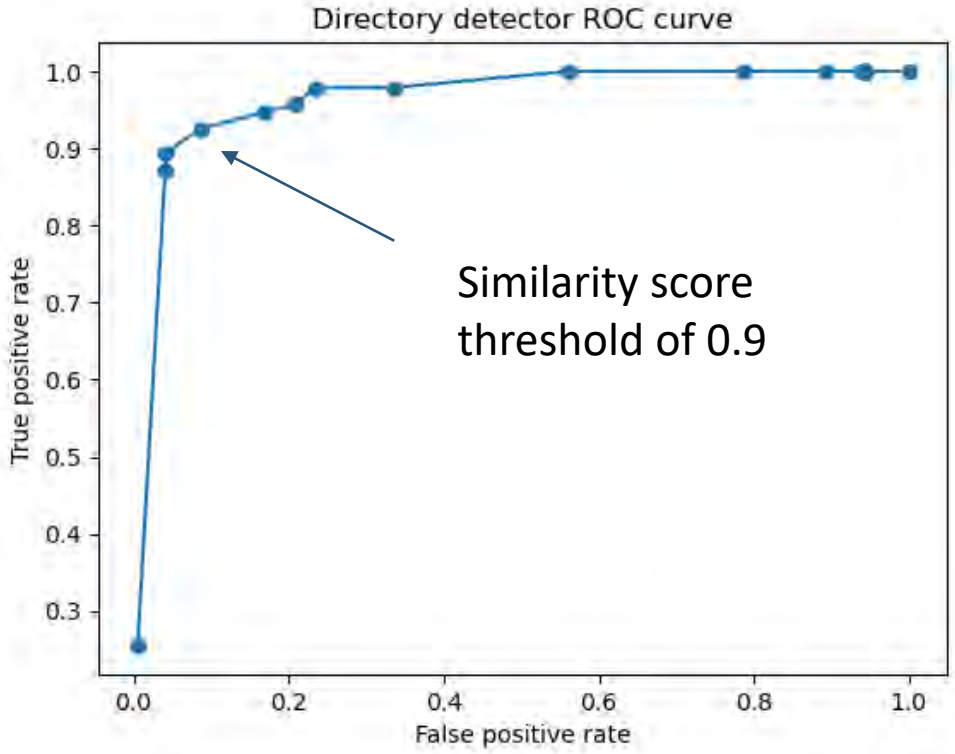
How well do we detect directory pages?

Directory detection: capture > 90% of directory pages with ~10% false positive rate at a string similarity threshold of ~ 0.9 for public and private schools

Public



Private



United States Census Bureau

Downloading webpages - dynamic content

Before page loading:

School Staff

Dynamic content rendered **upon user interaction**

After page loading:

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

Downloading webpages - pagination

Detecting pagination

- Look for typical patterns
  - **A B C … Z**
  - **1 2 3 …**
- Extract links or xpaths from paginator elements
- Add to the download queue

Example of a paginated page

School Staff

1 2 3 4 … > showing 1-4 of 65 staff

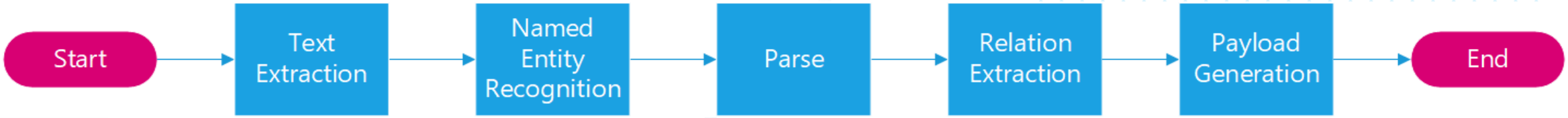
Teacher name
Titles: Second grade teacher
Emails: teacher@school.edu
Phone number: 000-000-0000

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

Step 3: extract | Text from the HTML of potential directory pages is processed to retrieve school staff information



Staff directory pages have different content and formats

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

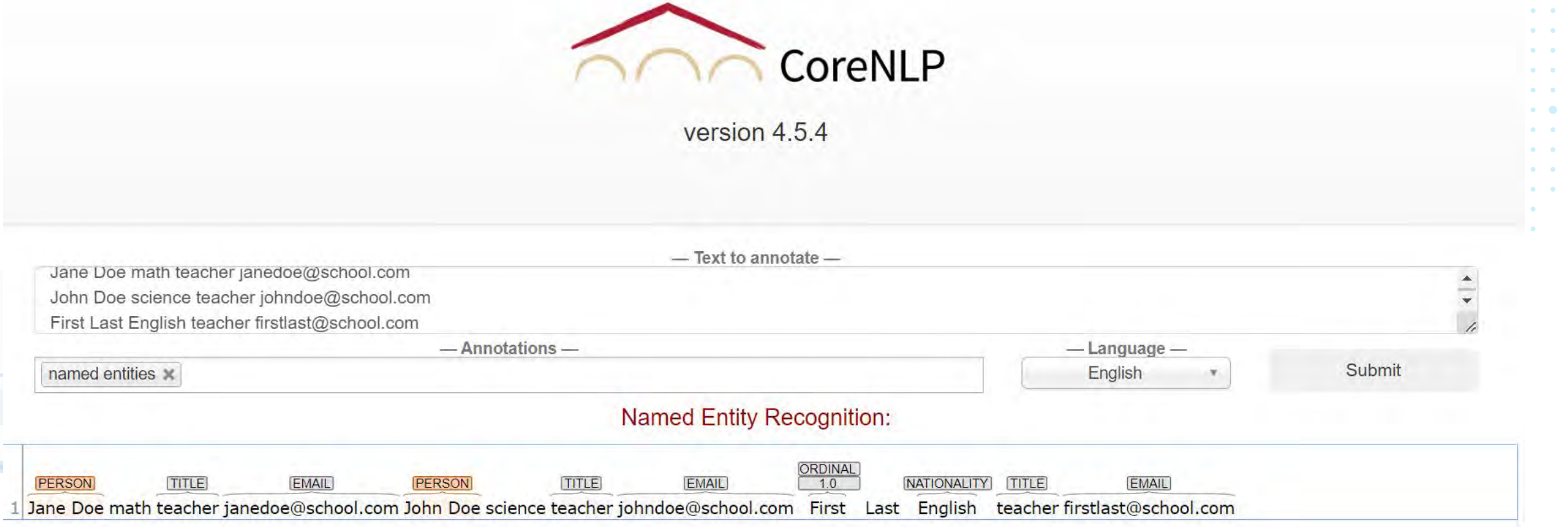
Teacher name
Titles: Second grade teacher
Emails: teacher@school.edu
Phone number: 000-000-0000

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

Extract – named entity recognition (NER)

We apply pretrained named entity recognition (NER) models in the CoreNLP package on text extracted from a page's HTML to highlight potential people, email, and title values



https://corenlp.run/

Extract – custom teacher title tagger

Motivation

- Pretrained NER models for TITLE entities from packages such as CoreNLP miss many teacher titles
- We manually curated teacher titles from school roster webpages and trained a custom teacher title tagger
- We represented each candidate text element with a features derived from the text and HTML element structure

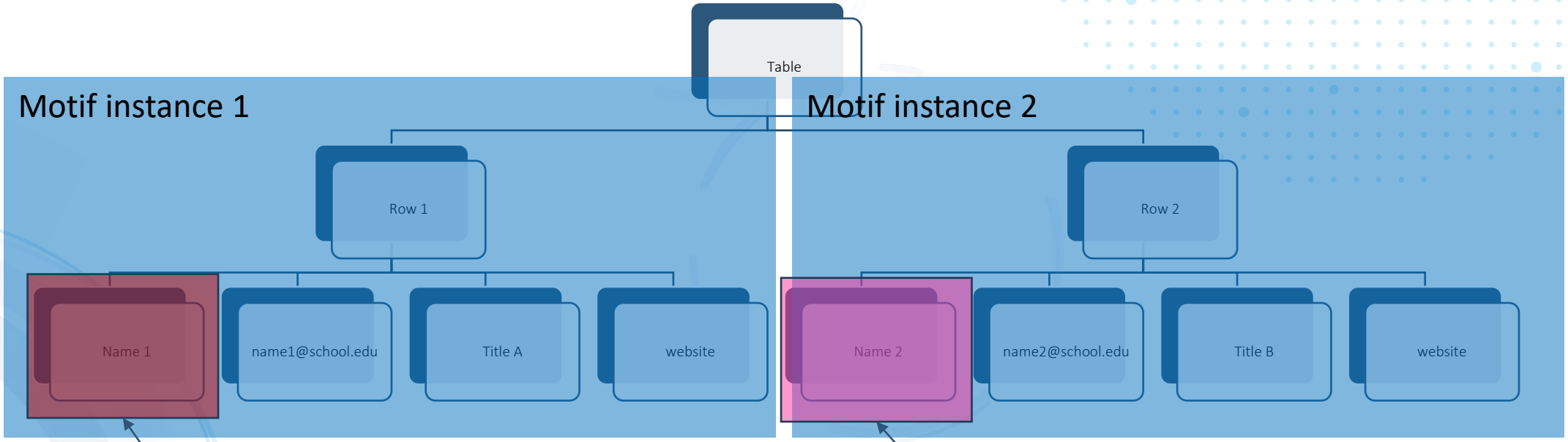
Performance

Precision	Recall	F1
0.88	0.87	0.87

Extract – identify target data element motifs

Teacher rosters frequently appear in a semi-structured format on webpages

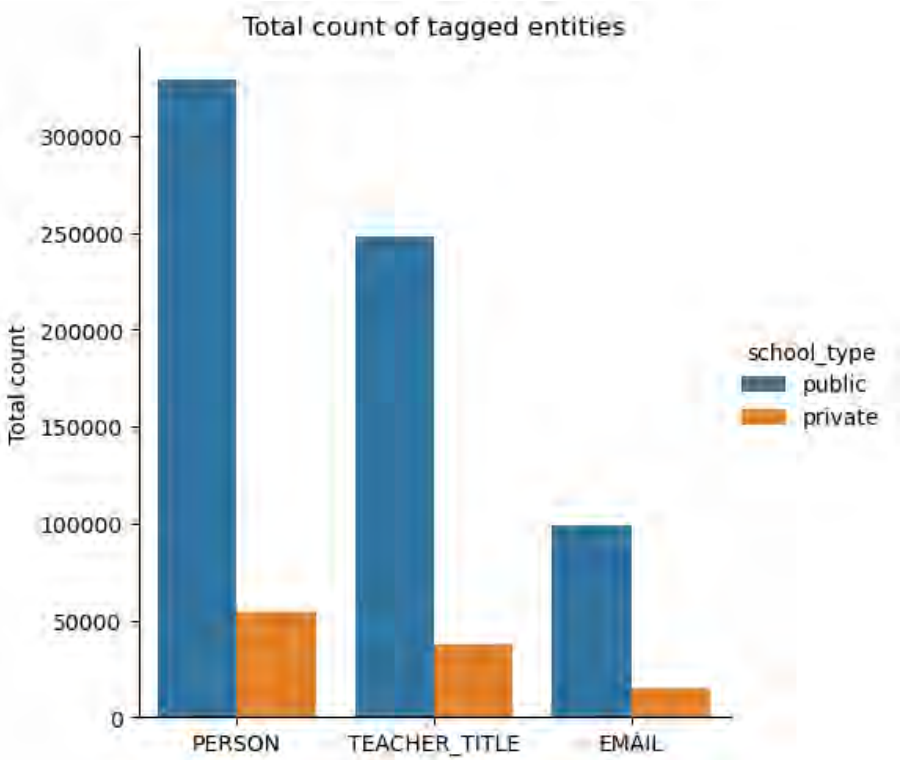
We harness this structure by identifying the most common HTML element patterns, or “motifs”, around NER values



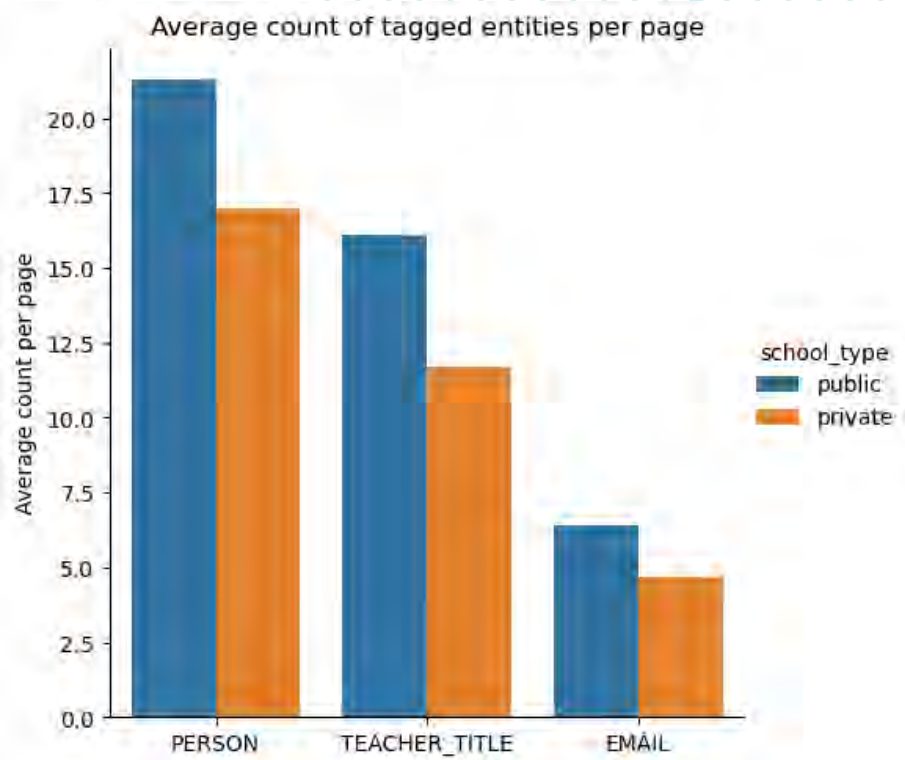
Extract – Parser Progress

Parsed payload for identified staff directory pages reveal smaller counts for private schools

Total count of parsed entities



Average count per page



United States® Census Bureau

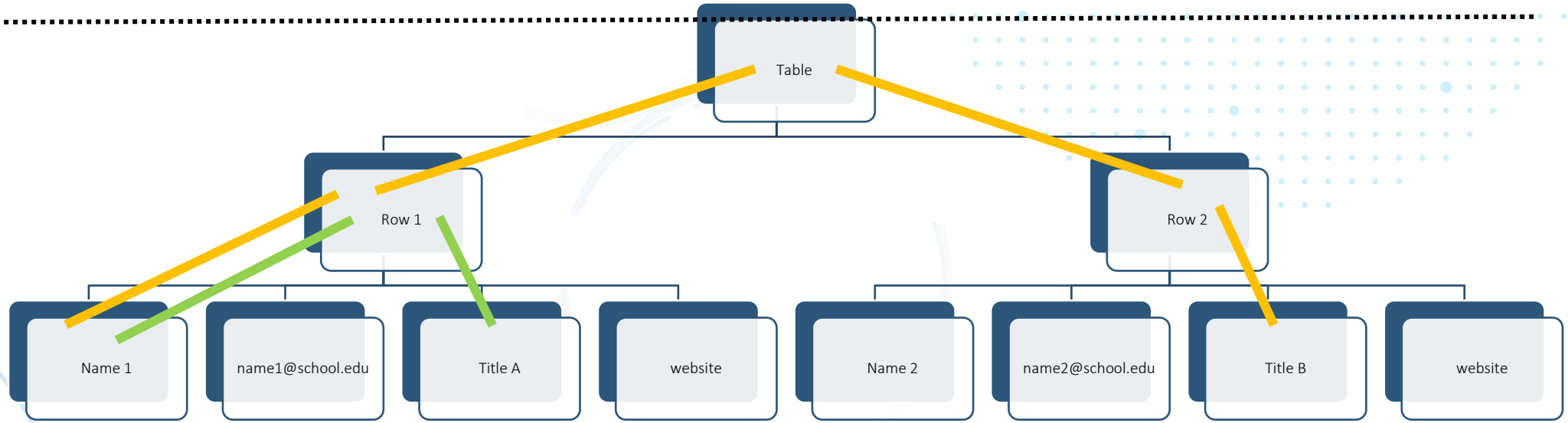
Extract – Relationship Extraction

Represent the HTML content as a graph and traverse the HTML elements to find the shortest path between names and titles or other elements on a page.

Webpage view

Name	Email Address	Job Title	Website
Name 1	name1@school.edu	Title A	website
Name 2	name2@school.edu	Title B	website

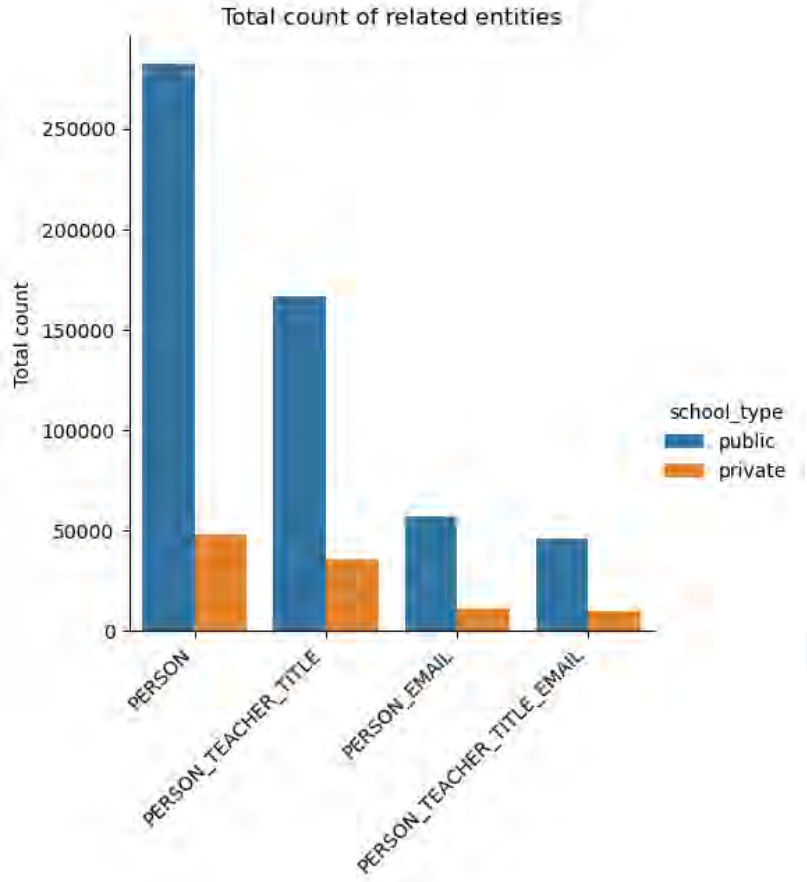
HTML hierarchy

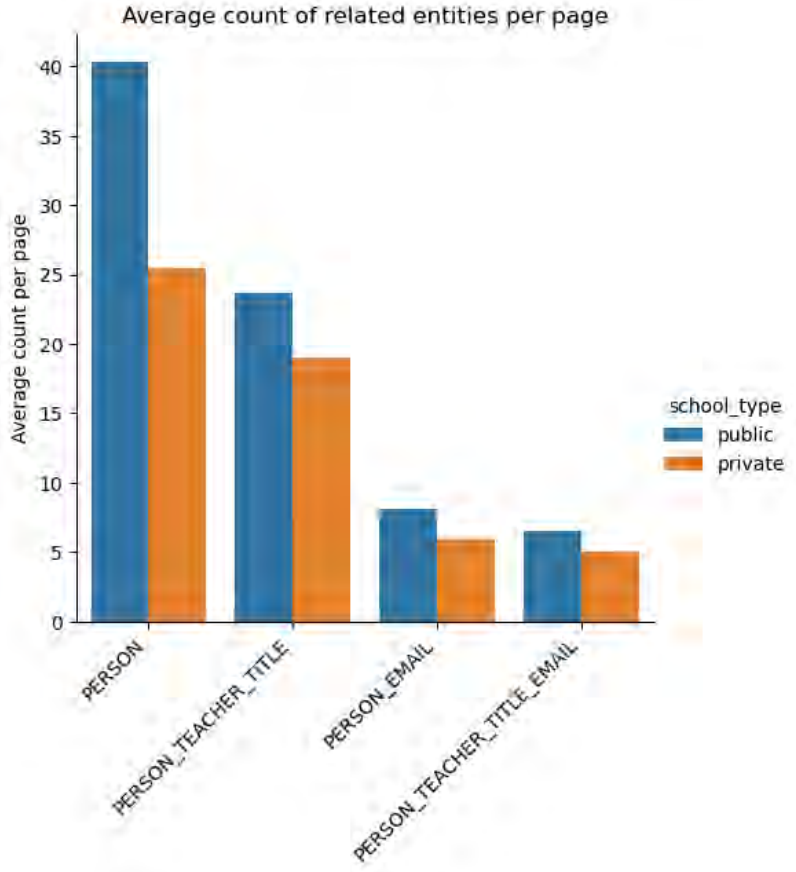


The path between **Name 1** and **Title A** has length 2 while the path to **Title B** has length 4.

Final Payload

In our final payloads, we extract more teacher names, titles, and emails from public than private school websites but capture more complete entries for private schools.





Conclusions

Limitations

- Scraping payloads limited by the ability to find a school website
- Pre-trained NER models built for full document text rather than webpage text
- Identifying pages with staff from multiple schools is difficult

Highlights

- We have developed an end-to-end web scraping and extraction pipeline for public and private school rosters
- The web scraping pipeline generalizes to many website formats
- Web scraped data can provide an alternative data source to augment traditional survey data collection methods

Acknowledgements

U.S. Census

- **Louis Avenilla**
- Patrick Campanello
- Shawna Cox
- **Ugo Etudo**
- Walter Holmes
- Yathish Kolli
- Anup Mathur
- Kayla Varela
- Allison Zotti

NTPS

- Maura Spiegelman

Questions?

sara.alaoui@census.gov

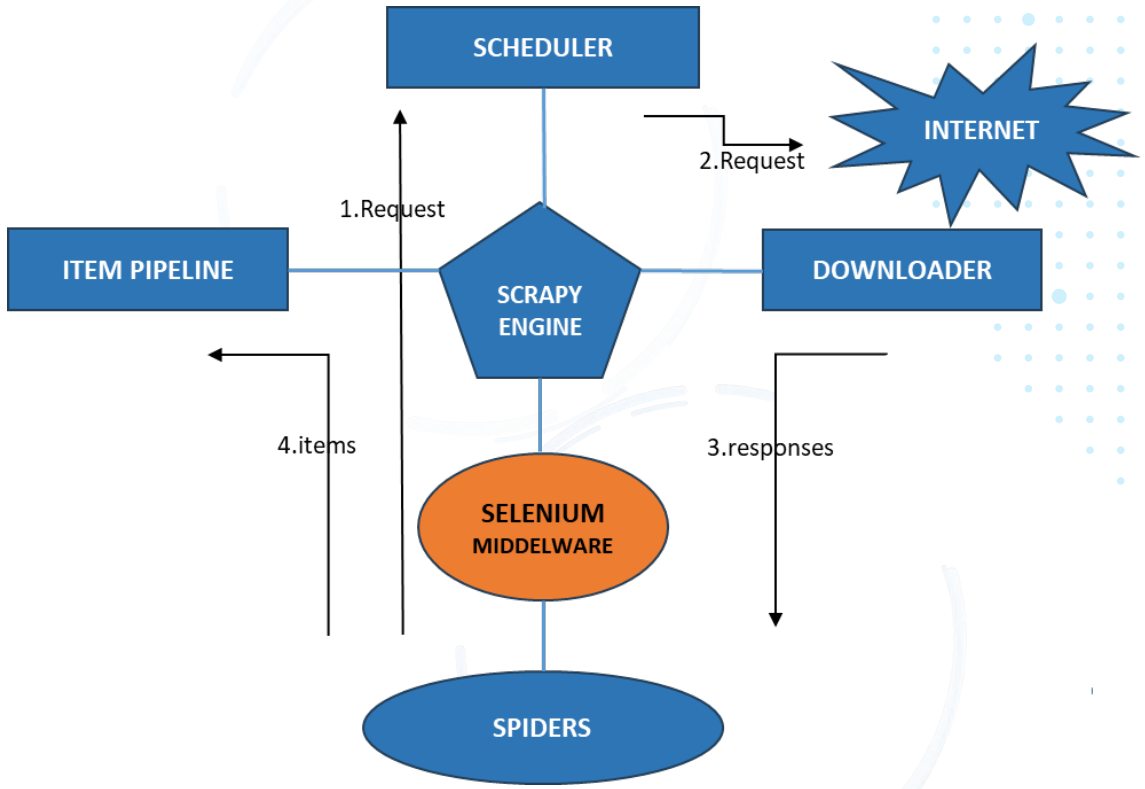
haley.s.hunter-zinck@census.gov

Supplemental Slides

United States® Census Bureau

Downloading webpages - dynamic content

Render dynamic content before download with Selenium middleware



Extract – custom TEACHER_TITLE tagger

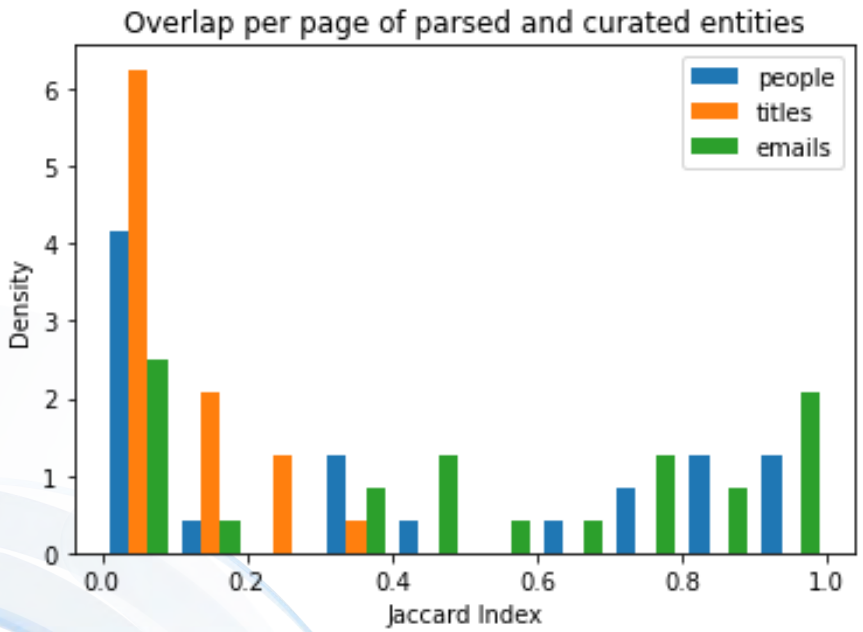
Methodology

- 3 discriminative features:
  - Structural information
  - Phrase embedding
  - Semantic distances
- Element-level predictions
- Hand + Machine-generated training data
- XGBoost model (based on cross validated model selection)
- ~ 2000 training examples ~ 400 websites

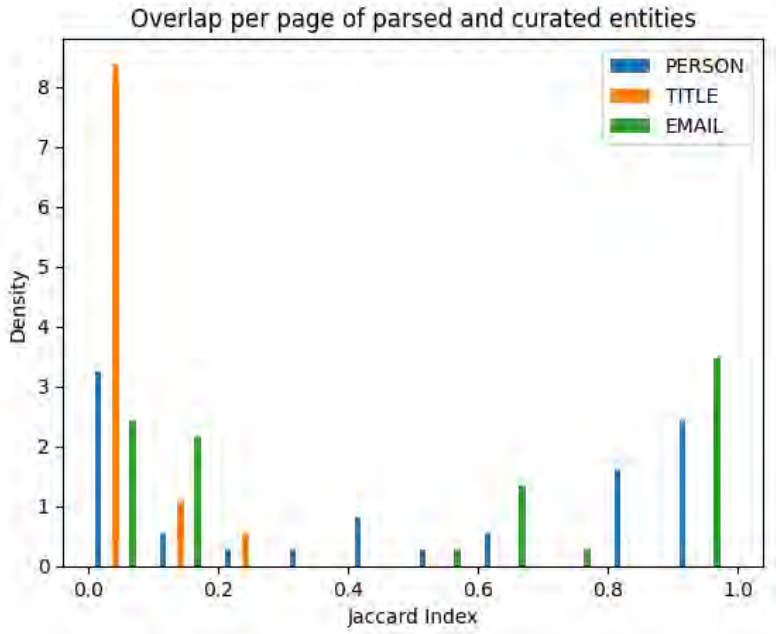
Extract – parser performance

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

Public schools

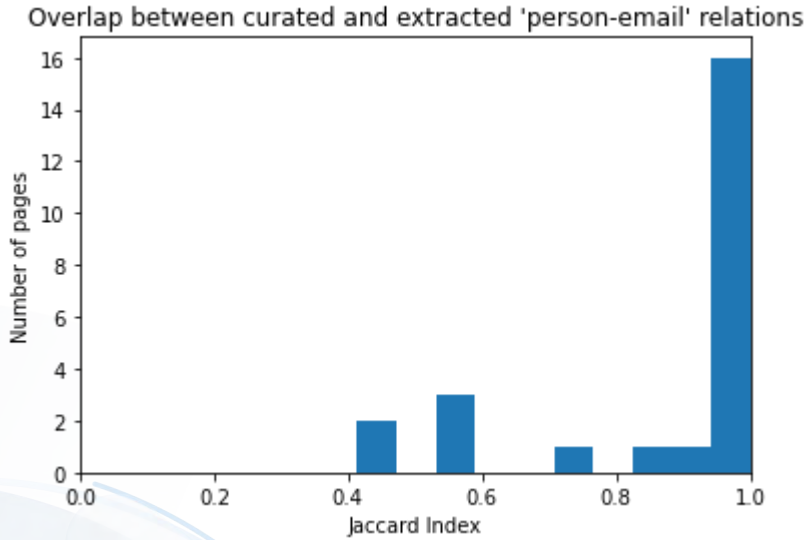


Private schools

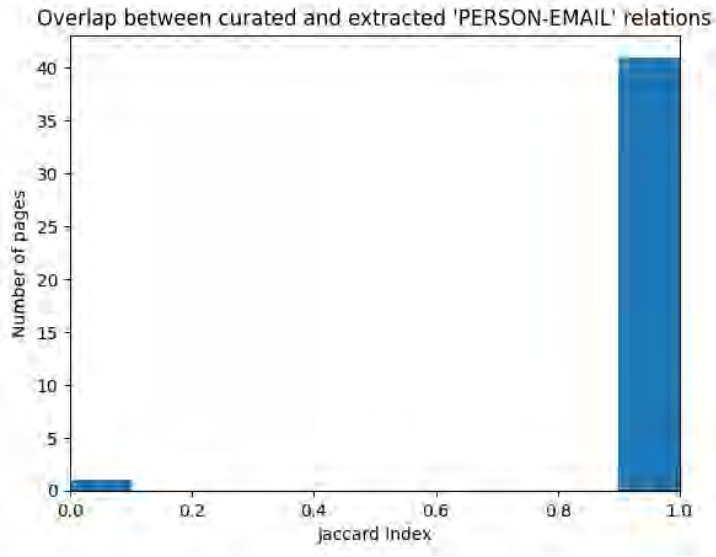


Extract – relation extraction performance

Public



Private

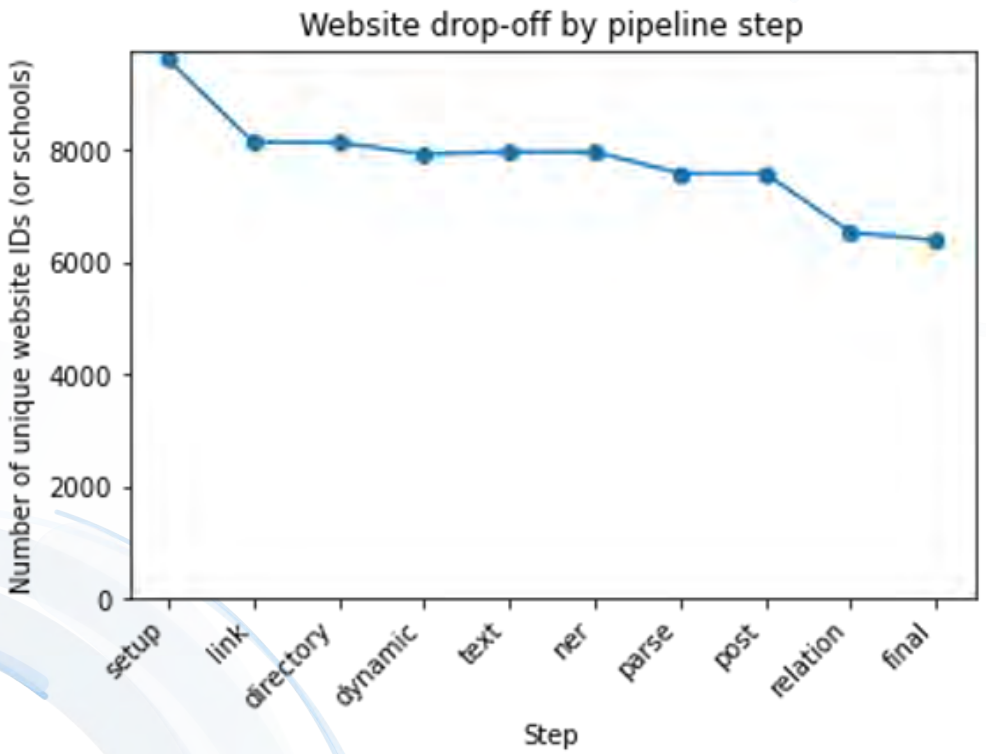


Total runtime for the full public school sample of 9,627 pages is ~4.6 days

Step label	Step description	Runtime (s)
etl	Extract transform load for google results	0
setup	Database setup and initial Google results load	2
link	Link harvesting of subpage links from in-sample website URLs	~6103
directory	Directory page detection with gazetteer	17342
dynamic	**Dynamic scraping, pagination detection, and page download**	**230675**
text	Convert HTML download files to text strings	337
ner	**Conduct named entity recognition (NER) on text strings**	**73744**
parse	Parse names, titles, and emails from the HTML downloads	4340
post	Post-process the parsed results to clean and filter	18241
interim	Construct an interim payload of school names to teacher names	254
rel	**Perform relation extraction to related names, titles, and emails**	**~50000**
final	Construct the final payload relating school names to related names, titles, and emails	38

Schools in sample drop off when websites are not found or no information is extracted

Public



Private