# Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing

Kylie L. Anglin

Journal of Research on Educational Effectiveness

Many education policy decisions are made at the local level. School districts make policies regarding hiring, resource allocation, and day-to-day operations. However, collecting data on local policy decisions has traditionally been expensive and time-consuming, sometimes leading researchers to leave important research questions unanswered.

This paper presents a framework for efficiently identifying and processing local policy documents posted online – documents like staff manuals, union contracts, and school improvement plans – using web-scraping and natural language processing.

**Step 1: Gather every *potentially* relevant document from local websites.**

There is no need to search every website for the policy document that interests you. Say we are interested in using teacher manuals to determine whether school districts have a salary schedule for their teachers. Searching every district website for their staff manual would take a long time. Instead, we will use a web-crawler to collect every potentially relevant web-page and document, and, in the next step, we'll throw out the irrelevant documents. A *web-crawler* is a software application that systematically browses the internet. In the context of school policy data, we feed the web-crawler a list of district websites. The crawler then searches each website on the list and extracts plain text from every web-page and embedded document that it can find (HTML, PDFs, Word Docs, etc.). After our web-crawler is done, we have text from lots of documents.

**Step 2: Narrow our collection of documents to those describing local policies of interest.**

Now, we narrow our collection of mostly irrelevant documents using a text classifier. A *text classifier* automatically assigns a document to a category based on its content. To use a text classifier, we need to label a subset of our documents as either relevant or irrelevant. (In our example, we label every document as either a staff manual or not a staff manual). Then, an algorithm learns the text features that best predict a document's relevance and we can apply that algorithm to the rest of our documents. Many text classification methods require their users to provide the classifier with a set of features (like word frequencies) describing the text, but state-of-art classifiers do not require this step. This means that some of the best classifiers are actually the easiest for researchers to use. After we apply our classifier to the full set of documents, we are left with only the documents that are most likely to contain relevant policy information.

**Step 3: Extract policy data.**

Now that we have the right documents, we need to extract the information we care about. In the simplest approach, we search the document for key words (like *salary schedule*). In the more nuanced approach, we train additional classifiers to further narrow the text to the portions we care about (for example, the portions of the teacher manual that discusses salaries) and to categorize the document according to the research question. (*Does the text indicate that teacher salaries follow a schedule?)* When the process is complete, we have analyzable policy data which can be incorporated into evaluations or be used to describe the policy landscape.

In an era where rich information is readily available on the Internet, the Gather-Narrow-Extract framework provides researchers with the capacity to collect important policy information efficiently and at scale.

*Full Article Citation*:
Anglin, Kylie L. (2019). Gather-Narrow-Extract: A Framework for Studying Local Policy Variation Using Web-Scraping and Natural Language Processing. *Journal of Research on Educational Effectiveness* 12(4): 685–706. https://doi.org/10.1080/19345747.2019.1654576.