How it Works

The Humans in the Loop project is looking at ways we can use machine learning methods to generate and enrich collection information and crowdsourcing methods to help train machine learning methods and verify the accuracy of their output.

TRAINING

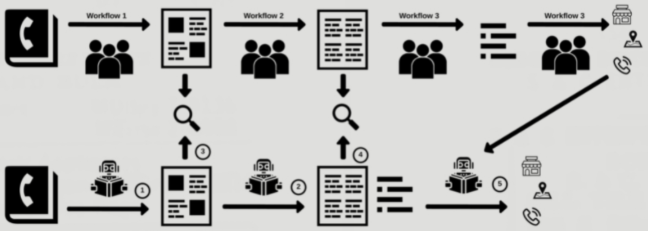
Machine learning methods look for patterns in data (text, images, sound, numbers—any digital information) based on patterns (or "models") that humans provide them as guidance. Humans "train" these models with the kinds of output they would like to see so the machine learning algorithm can produce similar output. (See the "About" page for more about machine learning.).

VERIFICATION

In order to improve, machine learning outputs need to be verified regularly against correct, human-generated outputs, or "ground truth." Humans can then use quantitative accuracy measurements or qualitative evaluation methods to determine how to make the machine learning process better. This might involve changing parameters in the algorithm or re-training the model with more data or different data.

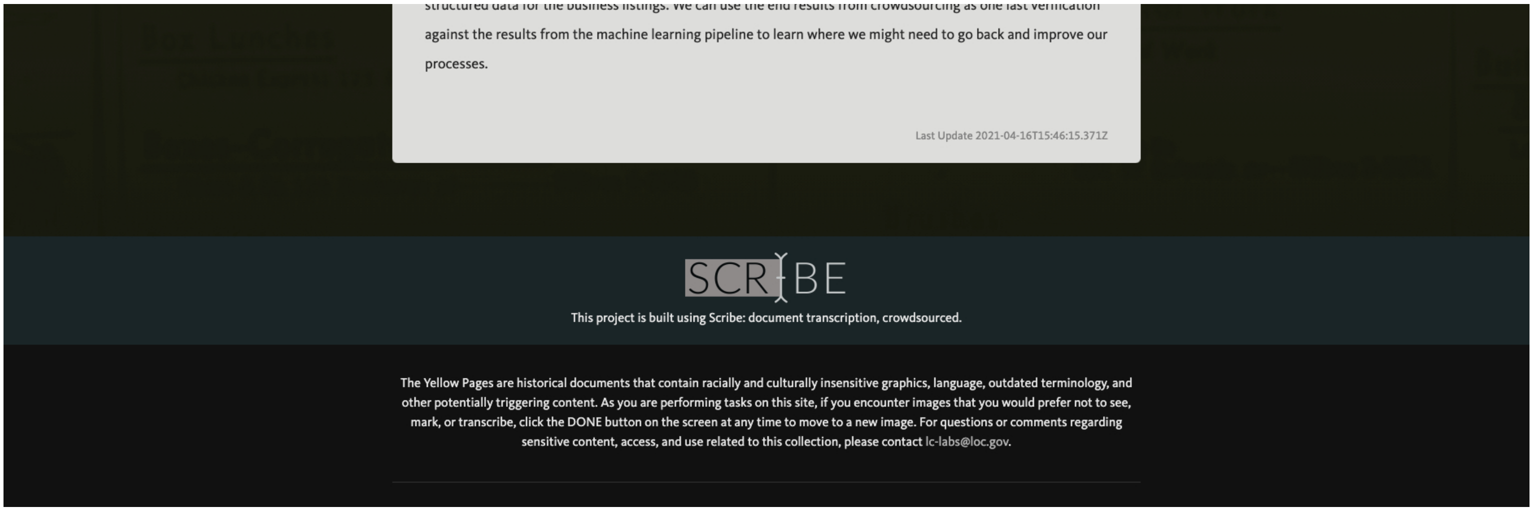
MACHINE LEARNING PIPELINE

For the Yellow Pages project, we are starting with digitized microfilmed page images and trying to extract structured data about the individual businesses--their names, addresses, phone numbers, business types, and other information--through a series of machine learning methods. The diagram below shows how your crowdsourcing efforts (upper pipeline) contribute to helping our machines learn:



1. An object detection algorithm looks at the "two-up" images of digitized microfilmed pages from the Telephone Directory collection and splits them into individual pages. Another object detection algorithm looks at each individual page for anything that looks like an advertisement. Using the page coordinates for advertisements found by the object detection algorithm, an OCR (optical character recognition) algorithm processes everything on the page that is not an advertisement into text and corresponding coordinates. Another algorithm parses the text and its page coordinates into business groupings and telephone tips. Now we know the coordinates of the advertisements, and the coordinates and text of the business groupings and telephone tips.
2. The same algorithm parses the text and its page coordinates into individual business listings and business grouping headings, or types (ex. "Decorating--Interior"). Now we know the coordinates and text of the business listings and business grouping types.
3. Page coordinates of advertisements, business groupings, and telephone tips marked by users in Workflow 1 are used as ground truth to verify the accuracy of the machine learning processes using object detection and OCR to find these same page segments.
4. Page coordinates and text of business types and listings marked by users in Workflow 2 are used as ground truth to verify the accuracy of the OCR algorithm trying to find these same page segments. Transcriptions generated from Workflow 2 are used as ground truth to help verify the accuracy of the OCR in recognizing the text itself.
5. A natural language processing technique called conditional random fields (CRF) reads the text of each business listing and identifies the business name, addresses, phone numbers, and any other information. Some of the business listing entities marked and transcribed in Workflow 3 are used to train a model for the CRF and others are used to verify the accuracy of the CRF's results.

At the end of this process we can output results from either pipeline—crowdsourcing or machine learning—as structured data for the business listings. We can use the end results from crowdsourcing as one last verification



structured data for the business listings. We can use the end results from crowdsourcing as one last verification against the results from the machine learning pipeline to learn where we might need to go back and improve our processes.

Last Update 2021-04-16T15:46:15.371Z

This project is built using Scribe: document transcription, crowdsourced.

The Yellow Pages are historical documents that contain racially and culturally insensitive graphics, language, outdated terminology, and other potentially triggering content. As you are performing tasks on this site, if you encounter images that you would prefer not to see, mark, or transcribe, click the DONE button on the screen at any time to move to a new image. For questions or comments regarding sensitive content, access, and use related to this collection, please contact lc-labs@loc.gov.