Data Science Project: Parsing Science Policy Issues Data

last edited: June 10th, 2021

Opportunity

Sci4NY wants to visualize the top science-related policy issues by each NYC community district in order to create dialogue that science knowledge can help create better policy about important community/borough issues. One of their biggest bottlenecks is extracting data from pdf reports/websites. Having to enter data manually would be very time-consuming.

Is it possible to generate an systematic method of parsing NYC community district reoports pdf documents or district profile websites so that it could ease the data collection process?

Current State

New York City Department of Health and Mental Hygiene (DOHMH) have published NYC Community Health Profiles pdf documents for 59 community district, each captures the health of 59 community districts across the city. They contain over 50 measures of neighborhood health. These reports highlight the disparities among neighborhoods and can be used by policymakers, community groups, health professionals, researchers and residents to encourage community engagement and action.

New York City Department of City Planning published Community District Profiles for each of 59 community districts. The profiles addresses the most pressing concerns, i.e. the top three pressing issues perceived and identified by that community district.

Google Cloud now provides Optical Character Recognition (OCR) technology that could help detect and extract text from images.

People, roles, and contacts

- Project Owner: Nancy Holt
- Data Scientists: Students
- Stakeholders: Sci4NY(Science For New York)

Requirements and metrics

- Deliverable
  - likely a deliverable is a manually labeled test set (of pdfs) so in later improvements we know if something breaks. You may want to create an algorithm to detect issues, e.g. compare the time series of these metrics and if one deviates too far then flag them.
  - A systematic algorithm that automatically extract the relevant data we need from Community Health Profiles pdf documents & Community District Profiles, and merge 59 community districts data.
  - A final dataset with each row represents one districts' information, and each column represents a relevant feature.

- Metrics
  - Websites:
    - Minimum Parsing Coverage: 90%

    - Minimum Parsing Correctness: 90%
  - Health Profiles:
    - Minimum Parsing Coverage: all the tabular formatted data
    - Ideal Parsing Coverage: 100%
    - Minimum Parsing Correctness: 90% for all tabular formatted data
  - Time your algorithm and make sure the running time is within a reasonable time
- Stretch Goal
  - An interactive Mapping to present multiple information.

Timeline and milestones

[Tentative]

Date	Description	People
Week 1	Kick-off & Data Delivery	Students and Mentor
Week 2	Successfully parsed a single website's top 3 issues; Automatically retrieve one data source from all PDFs	Students and Mentor
Week 3	[First round of results sharing] Successfully device an algorithm for parsing 59 websites' top 3 issues	Students, Mentor and Project Owner
Week 4	TBD	Students and Mentor
Week 5	TBD	Students and Mentor
Week 6	[Second round of results sharing] Anticipated deliverable to be shared with project owner	Students, Mentor and Project Owner
Week 7	TBD	Students and Mentor
Week 8	[Final round of results sharing] Final report and/or a final slide due	Students, Mentor and Project Owner
TBD	Presentation	Students

Data Specification

Description	Specification and attributes	Purpose	Source
NYC community district profiles data (59 districts)	- Top three pressing issues	Prioritize science policy issues	NYC government website
NYC health profiles (59 districts)	- Education Data - Social & Economic Condition Data - Housing Data - Maternal and Child Health Data - Healthy Living Data - Health Care Data - Health Outcomes Data	Rank the quantitative data to identify the top districts for each category and top issues for each district	NYC government website
Environment & Health Data Portal	- Indicators on environment & health topics	Supplemental data for mapping	NYC government website

- Links:
  - NYC Community District Profiles Data
  - NYC Health Profiles
  - Environment & Health Data Portal

Data Validation

- Please verify if the profile is reported uniformly across all the districts
- How are the missing values distributed over the fields?

[The profiles is formatted uniformly across all the districts.]

[Only Children avoidable hospitalization data is missing in certain districts.]

Hand-off and maintenance

The project is complete when

- The following code is documented
  - Web scrapping algorithm for extracting top three pressing issues in NYC district profiles websites.
  - PDF parsing algorithm for extracting main seven category data from Health Profile pdf documents.
  - Advise a way to validate the quality of the data extraction(human in the loop process might be necessary).
- A data frame with each row represents one districts' data and each column represents a feature is compiled from 59 districts' Data.