All Our Yesterdays: A toolkit to explore web archives in Colab

Tim Ribaric

Sam Langdon

Motivation?

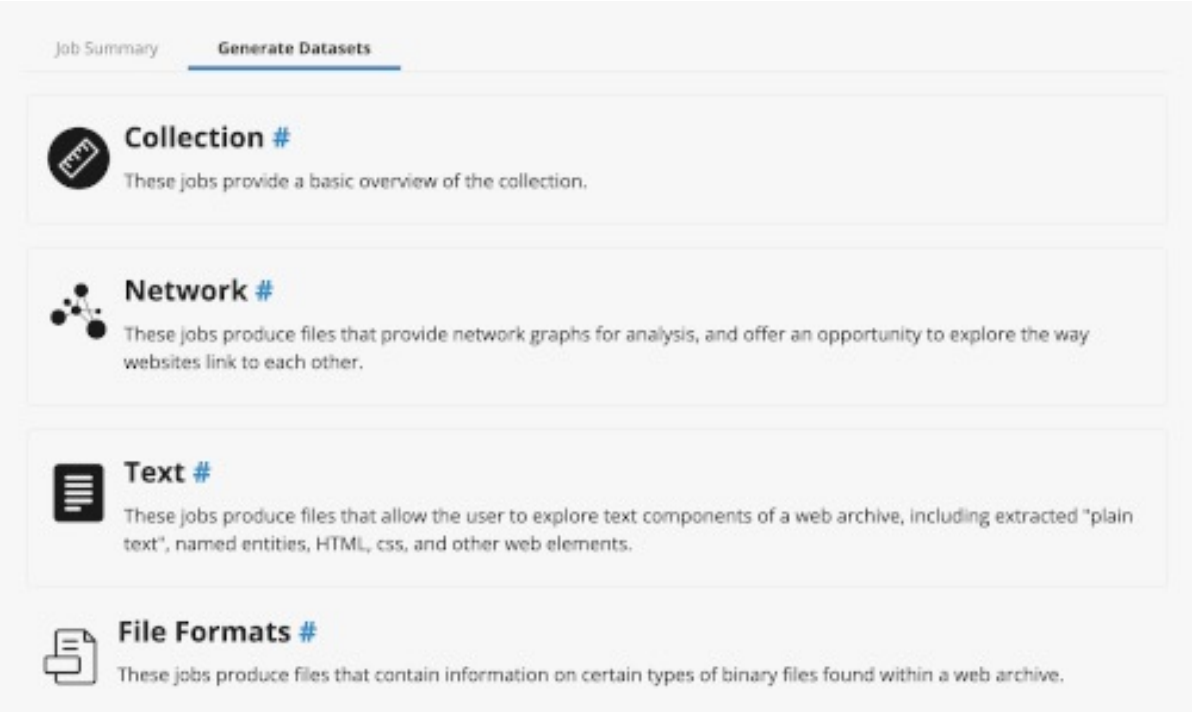
- WARCs are difficult to work with directly (especially if they are gigabytes in size)
- No standard desktop tool is suitable to explore them
- Often you don’t want to look at the whole archive anyway

Motivation?

- Looking to make a toolkit that scaffolds up an environment that will create the derivatives and plot out basic analysis steps
- Build on the new additions to the ARCH tool that is being developed as part of ARCHIVE-IT

Motivation?

- One of the things Archives Unleashed has been working on
- An integrated derivative generator
- Used these derivatives to perform some in-depth analysis



Motivation

- What do you do with your derivative? Particularly if you don't have much background / experience working with CSV Files
- Google Colab is a great place to start

ARCH_Data_Explore

Notebooks and datasets for the Archives Unleashed Cohort Grant for the **Covid-19 In Niagara** proje

More details on project site: https://brockdsl.github.io/archives_unleashed/

Notebook listing

ARCH Data Exploration	Notebook	Open in Colab
COMM 4P35 Tutorial	Notebook	Open in Colab
Hackfest notebook	Notebook	Open in Colab
Muni Data Export	Notebook	Open in Colab
Prep Domain Data	Notebook	Open in Colab
Twitter Data Export	Notebook	Open in Colab
Municipal Data Similarity	Notebook	Open in Colab
Another example of Municipal Data Similarity using SpaCy	Notebook	Open in Colab
Municipal Data Similarity using TF-IDF	Notebook	Open in Colab
Content size of pages over time	Notebook	Open in Colab
Frequency of page updates over time	Notebook	Open in Colab
Crawl frequency visualized	Notebook	Open in Colab

Demo

Derivative Generation

Demo

Analysis of a derivative

Roadmap

- Archives Unleashed Grant for ‘Niagara COVID Archive’
- Prototype of notebooks created for analysis
- Match of Mind Grant to develop AOYTK software

…

- Further Match of Mind Grant to do User Testing on user notebooks

More info

- Project is available on GitHub:
  - https://brockdsl.github.io/AOYTK/

- Looking for collaborators to test? dsl@brocku.ca