Web Archives Analysis at Scale with the Archives Unleashed Cloud

Nick Ruest (York University)
Ian Milligan (University of Waterloo)
Plan for The Talk

- Introduction
- The Problem
- Our Interdisciplinary Team
- Analysis at Scale with Archives Unleashed Tools
  - Toolkit
  - Cloud
  - Notebooks
- Caveats
- Conclusions
The Problem
Web Archives are Important

- How we preserve and disseminate cultural information has dramatically changed;
- Since 1996, and the advent of web archiving at the Internet Archive and national libraries, how we remember has dramatically altered:
  o In scope
  o In speed
  o In scale
… in other words ...

- More data than ever before is being preserved;
- And it’s being saved and delivered to us in very different ways…
All historians who want to study periods after the 1990s will have to use web archives.
You can’t study the 1990s without web archives.

And historians aren’t ready...
Access at scale has lagged.
Option One: The Wayback Machine

The Wayback Machine is a service provided by the Internet Archive. It allows users to search for and view historical versions of web pages. The Wayback Machine is useful for preserving information that might be removed from the internet or for verifying information that is now hard to find. It can be used for research, historical analysis, or simply to see how a website has changed over time.
Option Two: Working with the Underlying Data

WebARChive (WARC) File
Option Two: Working with the Underlying Data
Option Two: Working with the Underlying Data

- Potential
  - Text analysis at scale;
    - Finding particular mentions of keywords, people, organizations, concepts, etc. over time
    - Finding patterns over time (i.e. culturomics or other forms of cultural analytics)
    - Other text mining applications
  - Network analysis at scale;
    - Leveraging hyperlinks to see how people link to each other differently over time;
    - Finding pages of interest through historical applications of PageRank and other network concepts;
  - Moving between “distant” and “close” scales
Option Two: Working with the Underlying Data

- **Downsides**
  - Difficulty of tools to work with WARC files (humanists might be used to working with text at scale... they’re not used to WARC files);
  - Size of datasets (small web archives are in the tens of GBs; medium ones are in the 100GB-1TB range; large ones can easily begin to exceed 10TB);
  - Lack of a research community.
In other words, researchers need to explore web archives beyond the Wayback Machine... but the tools and infrastructure aren’t there.
Enter the Archives Unleashed Project

Archives Unleashed
Our Team

Ian Milligan
Historian, University of Waterloo

Nick Ruest
Librarian/Archivist, York University

Jimmy Lin
Computer Scientist, University of Waterloo
Our Mission Statement

Archives Unleashed aims to make petabytes of historical internet content accessible to scholars and others interested in researching the recent past.
So what do we do?
Archives Unleashed Projects

- Archives Unleashed Toolkit
- Archives Unleashed Cloud
- Archives Unleashed Datathons
Archives Unleashed Toolkit

- An open-source platform for analyzing web archives with Apache Spark;
- Scalable
  - Can work on a powerful cluster
  - Can work on a single-node server
  - Can work on a laptop (on MacOS, Linux, or on Windows with a Linux VM)
  - Can work on a Raspberry Pi for all your personal web archiving analysis needs
Using the Toolkit is based on the Filter-Analyze-Aggregate-Visualize (FAAV) Cycle
Filter

- Filter down content
  - Focus on a particular range of crawl dates;
  - Focus on a particular domain;
  - Content-based filter (“global warming”) or those who link to a given site
- Can be nested - i.e. pages from 2012 from liberal.ca that link to conservative.ca and contain the phrase “Keystone XL”
Analyze

- After filtering, want to perform analysis – extracting information of interest.
- Such as:
  - Links and associated anchor text?
  - Tagging or extracting named entities?
  - Sentiment analysis.
  - Topic modeling.
Aggregate

- Summarize the output of the analysis from the previous step.
  - Counting
    - How many times is Jack Layton or Barack Obama mentioned?
    - How many links are there from one domain to another?
  - Finding maximum (page with most incoming links?)
  - Average (average sentiment about “Barack Obama” or “Donald Trump”)

---
Visualize

- Output data as a visualization
  - Tables of results
  - External applications (i.e. GEXF files for Gephi)
Great!
So why doesn’t everybody use the Toolkit?!?!
Our Cutting Edge Interface

```scala
scala> :paste
// Entering paste mode (ctrl-D to finish)
import io.archivesunleashed._
import io.archivesunleashed.matchbox._
RecordLoader.loadArchives("example.arc.gz", sc)
 .keepValidPages()
 .keepDomains(Set("www.archive.org"))
 .map(r => (r.getCrawlDate, r.getDomain, r.getUrl, RemoveHTML(r.getContentString())))
 .saveAsTextFile("plain-text-domain")
```
In other words...

We have a wonderful platform that takes WARC files and converts them into formats that are familiar to digital humanists, computational social scientists, systems librarians, digital archivists, and beyond.

.. but you basically need to be a developer to run the simplest of commands (despite ample documentation and outreach... the command line interface is a bridge too far).
Enter the Archives
Unleashed Cloud
Archives Unleashed Cloud

- A web-based front end for working with the Archives Unleashed Toolkit;
- Runs on our central servers or you can run one yourself;
- Uses WASAPI – Web Archives Systems API – to transfer data
  - Currently Archive-It supported;
  - We are exploring integration with WebRecorder.io and other WASAPI endpoints
- Generates a basic set of research derivatives for scholars to work with
Archives Unleashed Cloud

- Download options for each collection
  - Full text of a web archive;
  - Full text of the top-ten most popular domains in a web archive;
  - Network diagram with characteristics pre-computed (Gephi);
  - Raw network diagram (origin/destination/weight);
  - Domain frequency statistics
How it works

- Sign-up
- Archive-It Credentials
- Populate User Dashboard (Filter)
- Analyze Collection
- Download WARCs

- Jupyter Notebook
- Download (Filter)
- Visualize
- Generate Derivatives (Aggregate)
- Analyze WARCs

- Analyze
- Aggregate
- Visualize
But where does our platform end... And the researcher begin?
Archives Unleashed Cloud Notebooks

- Jupyter Notebooks
- One for each derivative
- A “mad-libs” approach - fill in the blanks with the variables (domains, dates, collections, etc.) that you are interested in, and it does basic computations for you
- Still under development
- Bundled with data – download, run, explore data in your browser
Archives Unleashed Cloud Notebooks

By giving researchers these notebooks, with data, can we begin to jumpstart the process of research question creation and imagining what they can do with the data.
Welcome
Welcome to the Archives Unleashed Cloud Jupyter Notebook. This demonstration takes the main derivatives from the Cloud and uses Python to analyze and produce information about your collection.

Please feel free to create an Issue to let us know about any bugs you encountered or improvements you would like to see.

If you have Python experience, please feel free to change the provided code to suit your own needs.

We recommend that you use File > Make a Copy first before changing the code in the repository. That way, you can always return to the basic visualizations we have offered here. Of course, you can also just re-download the Jupyter Notebook file from your Archives Unleashed Cloud account.

How Jupyter Notebooks Work
If you have no previous experience with Jupyter Notebooks, the most important thing to understand is that that `\texttt{\textbackslash{shift}} + \texttt{\textbackslash{enter/return}}` will run the Python code inside a cell and output it to below the cell.

The cells that cover the required inputs, marked “Setup,” need to be run before the rest of the notebook will work. These cells will import all the libraries and set basic variables (e.g., where your derivative files are located) for the notebook. After that, everything else should be able to run on its own.

If you just want to see all results for your collection, use `\texttt{Cell > Run All}`.

Setup

The Collection ID
This variable is the most important, and the only variable you need to change to see a complete set of visualizations for your Archives Unleashed Cloud derivatives.

```
In [1]: COLLECTION_ID = "ARCHIVES UNLEASHED"
```

Using the collection id, the rest of the variables will be set up here. Also, some libraries like `pandas`, `numpy`, `networkx`, and `sklearn` will be imported so you can do the analysis.

```
In [2]: # Required packages.
from collections import Counter
import numpy
import matplotlib.pyplot as plt
from nltk.tokenize import word_tokenize, sent_tokenize
from nltk.corpus import stopwords
from nltk.sentiment import SentimentIntensityAnalyzer
from nltk.stem import PorterStemmer

# Setup Archives Unleashed Cloud data.
```
Archives Unleashed Usage Statistics

- Users: 144
- Collections: 792
- Files: 1,077,392
- Jobs completed: 4431

(as of this morning!)

- Job time: 10631h, 29m, 42s (1.25yrs!)
- Longest job: 590h, 31m, 49s
- Largest collection: 17.6T (compressed)
- Data analyzed: 159T (compressed)
Finally, we aim to build community around web archives.
Archives Unleashed Datathons

Helping to lower barriers;

Bringing people interested in web archiving (both collection + analysis) together;

Establishing a community through online communication and in-person work and social events;

Establishing a true community of practice around web archiving practice.
Archives Unleashed Datathons

- To date we’ve run (in this sequence) a series of datathons in **Toronto, Vancouver, and Washington DC**
  - a previous iteration had four events as well
- Gaining more experience with working with cultural heritage at scale
What’s Next?

- **Sustainability** has been baked into our grant from the very start (thanks Mellon!).
  - Costs USD$7/TB to process using the Archives Unleashed Toolkit.
What’s Next?

- Actual processing costs are relatively affordable – approx. US$7/TB to process WARCs and generate derivatives.
  - Large collection like University of Toronto’s “Canadian Political Parties and Interest Groups” would cost under US$30 to process and generate all of our derivative types seen in the Cloud.
- But of course, computing costs aren’t the crux...
What’s Next?

- Supported by Andrew W. Mellon Foundation; Compute Canada; Start Smart Labs; and some institutional support from Waterloo and York.
- Limitations (beyond computing costs):
  - Developer Time
  - Community Involvement
  - Sustainable Infrastructure
What’s Next?

- We know how much it costs;
- We’ve forged good partnerships with institutions, including the Internet Archive, datathon hosts (Simon Fraser, Toronto, George Washington), International Internet Preservation Consortium, and others;
- Held consultations with research libraries + consortias; and
- Are exploring tangible partnerships to bring web archive analysis to a broader audience.
We look forward to your questions and thoughts.
Thanks to our supporters!
Links

- archivesunleashed.org
- cloud.archivesunleashed.org
- github.com/archivesunleashed
- slack.archivesunleashed.org
- news.archivesunleashed.org
- twitter.com/unleasharchives