The Archives Unleashed Notebook: Madlibs for Jumpstarting Scholarly Exploration of Web Archives

Ryan Deschamps,[1] Nick Ruest,[2] Jimmy Lin,[3] Samantha Fritz,[1] and Ian Milligan[1]
[1] Department of History, University of Waterloo
[2] York University Libraries
[3] David R. Cheriton School of Computer Science, University of Waterloo

ABSTRACT

This paper introduces the Archives Unleashed Notebook, which is designed to work with derivative datasets from the Archives Unleashed Cloud, a platform for analyzing web archives. These datasets contain common starting points for scholarly inquiry, including full text content and the domain-level webgraph. Our notebooks interactively walk a scholar through the process of interrogating a collection using a fill-in-the-blanks 'madlibs' approach to promote engagement. Scholars start with a notebook populated with common analyses, in which they can make minor changes to variables to alter the subject of study in systematic ways.

1 INTRODUCTION

Over the past several years, our research team has been tackling the numerous hurdles associated with providing scholarly access to web archives—in particular, helping humanities scholars and social scientists cope with the daunting challenges of analysis at scale. This paper represents our latest effort, which involves building notebooks that interactively guide scholars through sample analyses and inviting further engagement using a fill-in-the-blanks 'madlibs' approach.[1]

In recent years, "notebooks" (Jupyter Notebooks being the most popular) have emerged as a popular tool for data science. Notebooks are typically provided on a web-based platform where code fragments and execution results (for example, graphs and figures) are placed side by side to support rapid interactions. Authors can intersperse descriptive text in a notebook, creating a coherent narrative around a particular analysis. Notebooks can be saved, shared, and reexecuted easily, supporting collaboration and reuse [4].

The Archives Unleashed Notebook brings notebook capabilities to web archives. Specifically, notebooks help scholars answer the perennial question "how do I start?" and provide natural jumping off points for subsequent analyses. This paper shares our experiences bridging the gap between vast web archive collections (hundreds of gigabytes or even terabytes) and intuitive, easily-accessible analytics tools to support scholarly inquiry.

2 TOOLKIT TO CLOUD TO NOTEBOOK

The first step to providing scholarly access to web archives is to tackle the challenge of scale; we have been working on this problem for several years now. The Archives Unleashed Toolkit (AUT),[2] which grew out of our earlier Warcbase project [2], engaged computer scientists and historians to co-design an analytics framework usable by humanities scholars and social scientists with no formal computer science training. AUT offers a Scala domain-specific language on top of the Apache Spark open-source data analysis platform, where scholars manipulate large web archives by defining data-parallel transformations over collections of records.

Over the past few years, our toolkit has been deployed in a number of "datathons" that have brought together librarians, scholars, computer scientists, and other stakeholders [3]. Through these sessions we have gained valuable feedback on what scholars really want, and these lessons have informed the technical direction of AUT. Based on these experiences, we've discovered that scholars are often unsure where to even begin in interrogating a web archive. To provide guidance, in previous work we proposed a model for scholarly interactions that starts with a question and proceeds iteratively through four main steps: filter, analyze, aggregate, and visualize [2]. Common analytics tasks, ranging from probing crawl statistics to visualizing webgraphs to analyzing frequent mentions of named entities all fit nicely into this model. Furthermore, we have discovered that scholars are frequently interested in the same types of derivatives as starting points to their analyses [1]. These include domain crawl distributions, full text and associated metadata, and the domain-to-domain network graph.

The Archives Unleashed Cloud was the next step in our efforts: instead of requiring individual institutions or scholars to procure computing resources and then to download, install, and configure AUT, we provide a cloud platform that, in essence, represents the canonical deployment of AUT. Of course, since AUT is open source, anyone can run their own copy—but it is fairly obvious why scholars would opt to use our platform. In the current deployment, we have collaborated with dozens of content partners who subscribe to Internet Archive's Archive-It service for content harvesting. The Archives Unleashed Cloud is able to automatically ingest these collections and create the derivative datasets described above. These derivatives are typically orders of magnitude smaller than the raw web archives [1], and we envision that scholars would download and further manipulate the data with tools they are already comfortable with, such as Python, R, or even Microsoft Excel.

The Archives Unleashed Notebook is our answer to the question: Can we make it even easier for scholars to analyze web archives?

3 WALKTHROUGH

Although notebooks are available in a variety of programming languages, we decided to use Python as it appears to be familiar to the most participants from our datathons. Furthermore, established libraries such as `networkx`, `nltk`, and `igraph` allow users to produce fairly sophisticated analyses with only modest effort.

[1] As noted in Wikipedia, "Mad Libs" is a phrasal template word game where one player prompts others for a list of words to substitute for blanks in a story, before reading the—often comical or nonsensical—story aloud.
[2] https://github.com/archivesunleashed/aut

978-1-7281-1547-4/19/$31.00 ©2019 IEEE
DOI 10.1109/JCDL.2019.00059

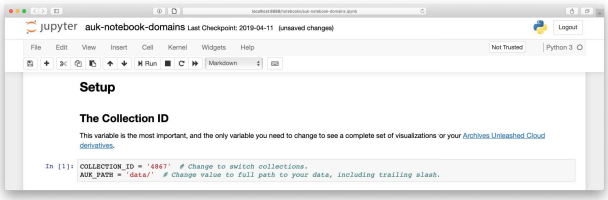

Figure 1: An example of a 'madlibs' input: the user provides the id of the collection to be analyzed.

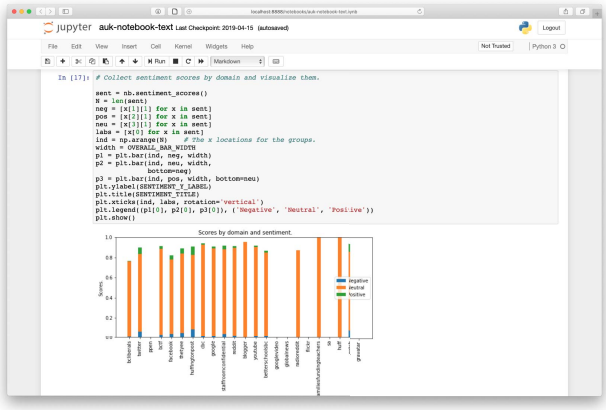

Figure 2: Sentiment analysis scores by domain. The blue represents "negative", whereas the green represents "positive".

Our current implementation is available on GitHub.[3] The notebook guides scholars through sample analyses, each corresponding to a potential research question, using a fill-in-the-blanks 'madlibs' approach. For example, to specify the collection to analyze, the scholar only needs to enter the collection id (see Figure 1), and upon reexecution all analyses and visualizations will update appropriately. Each analysis supports one or more parameterizations (e.g., time period of study) that scholars can adjust, supporting customization in specific ways. Once again, the scholar only needs to fill in the blank, and all analytical results will be updated.

We stress that our primary contribution is the concept of a 'madlibs'-style walkthrough, and not our current set of analyses per se (although we have found them to be useful). The following describes a few primarily for illustrative purposes:

Domain frequency provides a basic overview of the contents of a collection. We support filtering out particular domains that a scholar might suspect to be uninteresting, e.g., "facebook.com" URLs reflecting embedded widgets rather than real content.

Text analysis is provided using the popular Natural Language Toolkit (`nltk`). We provide examples that use text visualization to explore "what's going on" in a text. These include plotting the most frequently-used terms in the collection and showing where those terms appear in the archive using dispersion plots. The dispersion

[3] https://github.com/archivesunleashed/auk-notebooks

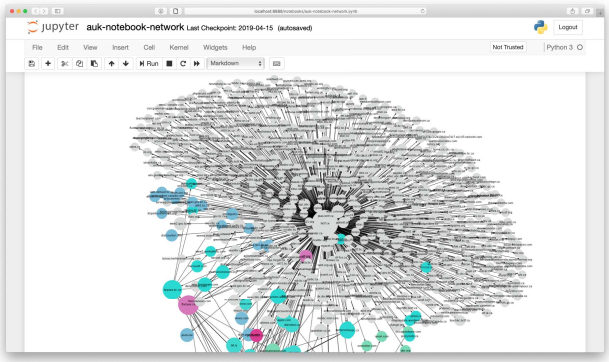

Figure 3: The ego network of a large node.

plots can provide comparisons using keywords of interest to a scholar, specified 'madlibs'-style. Another capability is sentiment analysis, which determines positive, neutral, or negative sentiments expressed in text—by way of evaluating each sentence and grouping the results by domain. This example is shown in Figure 2.

Network analysis is also provided in the current notebook. Here, we are able to take advantage of network visualization information already precomputed by the Archives Unleashed Cloud, which includes node sizing (based on degree), layout, and cluster-based coloring. This allows us to present network visualizations directly in the notebook without running computationally-intensive algorithms. Further manipulation with the `networkx` library can yield alternate views such as an ego network—a network visualization focused on a particular domain of interest—as seen in Figure 3.

4 CONCLUSIONS

Our goal is simple: to provide guidance on how a web archive collection can support scholarly inquiry. Ultimately, scholars will need to break free of the constraints imposed by our notebooks (and indeed any "canned" analysis) in order to engage with web archives in truly creative ways. However, our experience has been that the "gentler" we make this transition—from guided exploration to creative expression—the more scholars will be confident conducting independent inquiries. Such engagement is the key to building a robust community of scholars and content curators, which will help ensure that our digital world can be studied well into the future.

Acknowledgments. This work was primarily supported by the Andrew W. Mellon Foundation. The Social Sciences and Humanities Research Council of Canada provided additional support.

REFERENCES

[1] R. Deschamps, S. Fritz, J. Lin, I. Milligan, and N. Ruest. 2019. The Cost of a WARC: Analyzing Web Archives in the Cloud. In *JCDL*.
[2] J. Lin, I. Milligan, J. Wiebe, and A. Zhou. 2017. Warcbase: Scalable Analytics Infrastructure for Exploring Web Archives. *J. Comput. Cult. Herit.* 10, 4, Article 22 (July 2017), 30 pages.
[3] I. Milligan, N. Casemajor, S. Fritz, J. Lin, N. Ruest, M. Weber, and N. Worby. 2019. Building Community and Tools for Analyzing Web Archives through Datathons. In *JCDL*.
[4] B. Randles, I. Pasquetto, M. Golshan, and C. Borgman. 2017. Using the Jupyter Notebook As a Tool for Open Science: An Empirical Study. In *JCDL*.