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

Ryan Deschamps¹, Nick Ruest², Jimmy Lin¹, Samantha Fritz¹, and Ian Milligan¹

¹ University of Waterloo, ² York University

Motivations

- Web archives are a critical resource for those exploring the recent past.
- Humanists and social scientists face several challenges with web archives: scholarly access and scale.
- We engaged computer scientists and historians to co-design an analytics framework that would be usable by humanities scholars and social scientists with no formal computer science training. This is the FAAV cycle [1].
  - Filtering (i.e. by date, domain, subject)
  - Analyzing (finding pages, images, documents, etc. that fit certain criteria)
  - Aggregating (counting or performing statistical operations), and
  - Visualizing (from tables to graphs or Word Clouds)
- This project involves building notebooks that interactively guide scholars through sample analyses and inviting further engagement using a fill-in-the-blanks ‘madlibs’ approach.

Platform Development: Toolkit > Cloud > Notebook

The Archives Unleashed team has been tackling the challenge of scale by tools to bridge the gap between vast web archive collections (hundreds of gigabytes or even terabytes) and intuitive, easily-accessible analytics tools.

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. Community feedback and consultation with librarians, scholars, computer scientists, and other stakeholders [2] has:

- Provided environmental scan of needs
- Informed technical direction
- Identified that scholars are frequently interested in the same types of derivatives as starting points to their analyses: domain crawl distributions, full text and associated metadata, and the domain-to-domain network graph
- Demonstrated that scholars are often unsure where to even begin….

References