Workshop - HathiTrust Research Center Algorithms

This page is a companion to the Workshop for select graduate students on the HathiTrust Research Center (HTRC) Portal 4/8/2014.

Preparation

- **Please do bring a laptop!** If you have one, bring your own. If you do not have one, please feel free to check one out at the Olin circulation desk. Having a laptop will allow you to participate in the exercises and get the most out of class.
- **No special software will be needed.** All exercises will be done through a Web browser, without any special plugins. However, **Firefox** is the browser that will work best for rendering results.
- **You will need to obtain credentials to log in.** Please follow the steps on the HTRC wiki (see "How do I obtain an account?") for getting started with the **production portal**.
- **Think about** subject matter around which you might want to build a collection for analysis. The production portal is limited to the 3M+ items that have been digitized by Google and are in the public domain. These tend to be works that have been published before 1923, or are government documents, or items that are exceptions to various copyright restrictions.

Agenda

The class is a guided exploration in the HTRC portal. We will explore the algorithms and use them to discover their capabilities, limitations, and various strategies for addressing those challenges. This allows us to explore the HTRC portal in specific, and grapple firsthand with basic issues encountered in computational analysis of text.

Support tips for exercises

- **Book mark** this page for handy access. We will be referring to it at points in the workshop.
- **Log on** to the HTRC Production Portal with your personal credentials.
    - Once signed in, click on "Algorithms" in the black navigational bar at the top of the page. Once you have done so, we will be ready to begin.
- **Worksets** can be managed through the Production Portal's Blacklight instance. You can also access this by using the "Create Workset" link. Regardless of the avenue of entrance, **you will have to log in a second time with the same credentials** you are using in the Production Portal.
    - Worksets can be created through Uploading Worksets of a CSV file containing Volume ID.
    - Worksets with classification MUST be created through upload.
- I've made three collections for your use in this class. (You are not obligated to use these; you may want to use collections of your own in the exercises.)
    - ShakespeareComedies@MPaolillo - 58 dramas authored by William Shakespeare with MARC 655 field denoting "Tragedies."
    - ShakespeareTragedies@MPaolillo - 54 dramas authored by William Shakespeare with MARC 655 field denoting "Comedies."
    - ShakespearePlays@MPaolillo - A larger CSV formatted collection consisting of the contents of both of the other two collections.
- Use this page when referencing custom **stop word lists**. I can post lists relevant to your collections if you have them.
- There may be a **bug that prevents display** of the results of the algorithm "Meandre_OpenNLP_Date_Entities_To_Simile". You can display locally by following this **fix**:
    - Download the date_entity_simile.html to your machine and save.
    - Open with a text editor and change this line

      ```
      <script src="https://htrc2.pti.indiana.edu/HTRC-UI-Portal2/js/timeline-api.js" type="text
      /javascript"></script>
      ```

      to this line

      ```
      <script src="http://api.simile-widgets.org/timeline/2.3.1/timeline-api.js" type="text/javascript"><
      /script>
      ```

    - Save the html file and open with a web browser and the data should display. (Firefox works best with Simile.)

Resources

- HathiTrust is:
    - an international partnership of over 100 institutions.
    - a digital library containing over 13 million books, 38% of which are in the public domain. All items are fully indexed, allowing for full text search within all volumes. You can login with your Cornell NetID to
        - Create Collections (public or private)
        - Download PDF's of any item available in full text
    - a trustworthy preservation repository providing long-term stewardship, redundant robust backup, continuous monitoring, persistent identifiers for all content

    - where Cornell University Library deposits books it digitizes at scale.
- HathiTrust Research Center (HTRC) - a collaborative research center (jointly managed by Indiana University and the University of Illinois) dedicated to developing cutting-edge software tools and cyberinfrastructure that enable advanced computational access to large amounts of digital text.
    - HTRC Production Portal - a web-based user experience of the HTRC. The production portal makes available all the full-text indexes of the Google-digitized deposits to HathiTrust that are in the public domain.
    - HTRC User Community Wiki - home of the user support documentation, meeting notes, elist addresses and sign-up information, and FAQs.
- A summary of the algorithms is attached below.

File	Modified
Microsoft Word Document HTRCAlgorithmDescriptions.docx	Apr 03, 2014 by Michelle A. Paolillo