Providing Pin-point Page-level Precision to 1 Trillion Tokens of Text for Workset Creation

David Bainbridge
University of Waikato
Hamilton, New Zealand
davidb@waikato.ac.nz

J. Stephen Downie
University of Illinois
Urbana-Champaign, USA
jdownie@illinois.edu

Boris Capitanu
University of Illinois
Urbana-Champaign, USA
capitanu@illinois.edu

ABSTRACT

We report on the work undertaken developing a web environment that allows users to search over 1 trillion tokens of text—down to the page-level—of the HathiTrust Part-of-Speech Extracted Features Dataset to help produce worksets for scholarly analysis. We present an extended example of the web environment in use, along with details about its implementation.

KEYWORDS

Very Large Digital Libraries, Extract Feature Text Analysis, Workset Creation

ACM Reference Format:
David Bainbridge, J. Stephen Downie, and Boris Capitanu. 2018. Providing Pin-point Page-level Precision to 1 Trillion Tokens of Text for Workset Creation. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3197026.3203873

1 INTRODUCTION

In previous work the HathiTrust Research Centre (HTRC) has made publicly available a Part-of-Speech Unigram Extracted Feature Dataset of the full HathiTrust Digital Library (HTDL), containing over 5.7 billion pages of scanned and OCR'd content [1]. In this article we detail the work done in developing a searchable version of this resource that helps support workset creation. The implementation centres around a digital library architecture designed for large scale content with fine-grained access. In addition to metadata level field search, users can specify queries down to the page-level for words in 54 identified languages, and in the case of 6 languages, specify queries by part-of-speech. In total this constitutes over a trillion tokens of indexed text.

To illustrate the capabilities of the developed resource we start by giving an extended example of it in use: locating the pages that make reference to violins in the *Sherlock Holmes* book series. We then go on to provide details about its implementation.

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA

ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203873

2 WORKED EXAMPLE

Figure 1 shows a snapshot of the web page that a user encounters when they first visit the developed website.[1] Notice the tabs along the top of the page. The first tab is for searching at the page level, and the second for volume-level metadata searching. In the snapshot we are looking at the third tab, through which the user can specify a combined page and volume level query. The top search box is for the volume-level metadata—*Sherlock* has been entered in this case—and the search box below that for page-level text searching—where *violin* has been entered. The final tab is for more advanced use where users can enter their query directly using a fielded Boolean query syntax. Information icons are strategically positioned throughout the search tabs to provide contextualised user help.

Below the two search boxes is an area for controlling the languages and parts-of-speech that are queried, based on the mappings from the universal POS tagset developed by Petrov et al. [2]. This part of the interface makes extensive use of check-boxes, and defaults to searching by all parts-of-speech in the English language because, at 50%, this language represents the largest proportion of the DL's content. Clicking on *Show other languages ...* expands this area of the interface to reveal the other languages available.

For the query shown in Figure 1, the checkboxes are set to perform a title-level search for *Sherlock* with pages grouped by volume, where the pages themselves include the term *violin*. Matching is case-insensitive but—given the complexities of querying across multiple languages—not stemmed. Figure 2 shows the result of performing this query. The central feature produced is the paginated result set. Above this are collated details about the search, such as how many matches were found—1,413 pages in 251 volumes—and links to share the query through email and social media platforms. Through the facets on the left, the user can further refine the query, such as restricting items to be only in the public domain.

Different to many DLs, the produced result set is not the end-result of the search interface. Beyond sharing the URL of the query, the website supports the formation of worksets. The most direct way to do this is to export the query through the actions available in the upper-left corner of the shown snapshot. From here a list of volume identifiers can be downloaded, or alternatively a more fine-grained list exported that gives volume with HTDL sequence number (effectively its page number). For authenticated users, there is also the option to publish the list of IDs resulting from the query in the HTRC registry directly.

A shopping cart feature is provided to support the formation of more intricate worksets built out of multiple queries over a period of time. The icon for this appears in the upper portion of the result set area, on the right and is used to add items from the result set to

[1] https://solr1.ischool.illinois.edu/solr-ef/index.html

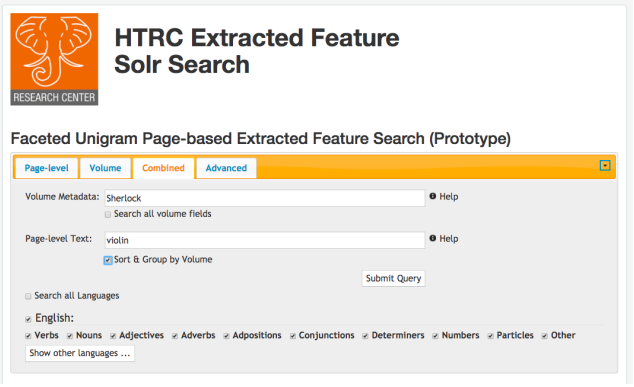

Figure 1: The Extracted Feature Solr search Page.

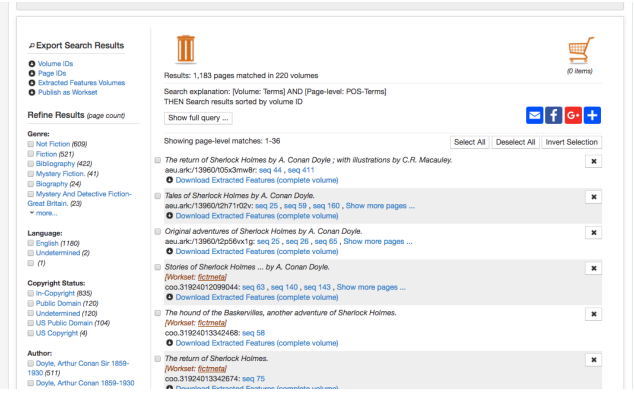

Figure 2: The Extracted Feature Solr search result set.

the cart. Complementing this, a trashcan icon (upper-left), allows the user to remove items to make it clearer to see what is left.

Three further features worthy of mention that occur within one matching item of the result set are: the *Unigram POS Page-viewer* which allows the user to view content—even in-copyright content—as a bag of words, sorted in various ways, such as alphabetically grouped by part-of-speech; the *Extracted Feature JSON Download* feature which provides a convenient, simple way to selectively download an item of interest from the the Extracted Feature dataset; and *Worksets Inclusion* which means the user can see which existing worksets an item is already in. The workset name is hyperlinked, so a user can explore this relationship further by accessing the named workset through the HTRC registry.

3 IMPLEMENTATION

Turning the 15+ million JSON Extracted Features files into a malleable, searchable resource was broken down into two key steps: transforming the data into Hadoop Sequence files, and developing a Spark framework program to ingest the documents into a Solr index. A 48 core machine was used for converting the JSON files mounted on a network disk into the Hadoop Sequence format. The process involved decompressing each JSON files (BZip2 format) on-the-fly, and then recompressing the content (again BZip2) within the formation of the Hadoop Sequence files. This process took 2 weeks to run.

The resulting files were then copied onto the Hadoop Disk File System (HDFS) operating across a 10 node, 120 core cluster machine set up to operate the Spark framework. This took slightly over a day to run, with the replication factor set to 3.

The Spark framework was where the ingest of the JSON files took place. This involved reading in the HTRC Extracted Feature JSON format and transforming it into the JSON format used by Solr for indexing. The location of the Solr server was itself separate from the machine running the Spark framework, this time a cluster computer with 2 nodes and 24 cores. The transformed JSON data was streamed to the Solr server through a dedicated switch. The Solr server was configured as a 20 core (aka shard) collection, with its replication value set to 1. The properties of the schema fields used in Solr were set to "do not store" for full-page text fields to conserve space. The ingest process took two weeks to run, and resulted in index files that cumulatively consume 7 TB.

The user interface shown in the previous section is a single-page web application. It uses AJAX calls to collect information from various servers in the HathiTrust orbit. The key call is to the Solr server. As the volume-level metadata is stored in the Solr index, this is also where it retrieves the bibliographic fields needed to present the search results. Other functionality provided by the search interface is provided through a dedicated Servlet-based server. The API to this server supports AJAX calls to access the original Extracted Feature JSON files, either individually or in Zipped format. There is also a URL shortening service provided (so long queries can be emailed out, for example) and where server-side storage of a user's session is kept, such as the items in their shopping cart.

4 CONCLUSION

The interface described here has already been used to satisfy a variety of use cases for scholars engaged with HTRC: from ones that can be satisfied with a single query term—such as on how many pages does the word *Canada* appear? (75,528,670)—through to more sophisticated ones—such as the worked example given above—and beyond.

For the most part, we have found that these use cases need to be formulated through the Advanced Query tab to provide the best response. While the worked example was constructed to give a flavour of what is possible with the interface, for instance, it is not the case that all the book titles in the *Sherlock Holmes* series include the name of the main character in it. A more accurate query is to search by author metadata and sift through the result set. This is, however, complicated by the fact that variations in how Arthur Conan Doyle's name is recorded means that an ORing of the variants (there are three) is needed, and from a "cold start" using the Advanced Query tab, what these three variations are, is not immediately apparent. In future work we seek to better support, through interactive features, the ways in which these more complex queries are constructed.

REFERENCES

[1] P. Organisciak, B. Capitanu, T. Underwood, and J. S. Downie. 2017. Access to Billions of Pages for Large-Scale Text Analysis. In *iConference Proceedings*, Vol. 2. 66–76. https://doi.org/10.9776/17014

[2] S. Petrov, D. Das, and R. McDonald. 2011. A Universal Part-of-Speech Tagset. *ArXiv e-prints* (April 2011). arXiv:cs.CL/1104.2086