国際シンポジウム「大学における研究データ管理の意義と支援人材育成」

Darch, Peter
イリノイ大学アーバナ・シャンペーン校情報学部 : 助教

Downie, J. Stephen
イリノイ大学アーバナ・シャンペーン校情報学部 : 教授

冨浦, 洋一
九州大学データ駆動イノベーション推進本部研究データ管理支援部門 : 部門長

石田, 栄美
九州大学データ駆動イノベーション推進本部研究データ管理支援部門 : 教授

https://doi.org/10.15017/6790816

出版情報：2023-06-19. University of Illinois Urbana Champaign
バージョン：
権利関係：

HATHITRUST research center

Big Data Use in the Humanities and Social Sciences:

The Case of HathiTrust Research Center

Presented by
Professor J. Stephen Downie
University of Illinois Urbana-Champaign
Email: *jdownie@Illinois.edu*
Twitter: *@profdownie*

School of Information Sciences
The iSchool at Illinois

Takes many hands

This slide deck is a communal effort of HTRC community members, and includes contributions from:

- Ryan Dubnicek
- Glen Layne-Worthey
- Niko Parulian
- Ming Jiang
- Ted Underwood
- Dan Evans
- Boris Capitanu
- David Bainbridge
- Peter Organisciak
- And, many, many other collaborators and colleagues

Challenges & opportunities

How to *manage* the scale and scope of the HathiTrust Digital Library?

How to *leverage* its scale and scope?

How to *facilitate research* given significant *copyright constraints*?

How better to enable work with *non-English* (and non-Latin, and R-to-L, etc.) languages?

How to *curate* a collection like this, to make it more *diverse, inclusive, & reflective* of the world of knowledge and books?

How best to *manage* the data that is generated by scholarly exploration

HathiTrust is…

HATHI TRUST

A partnership of member libraries
(Keio University the only one in Japan)

A collective digital library (the largest ever assembled)
(UIUC's books are among them:
about a million of them)

The fruits of large-scale digitization initiatives

A trusted digital repository

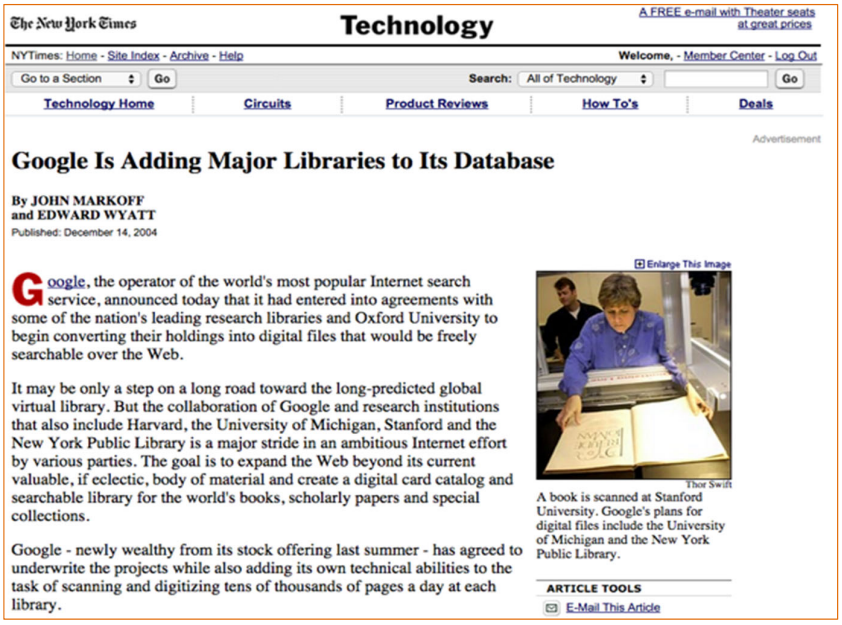
A provider of special services
(e.g., the "Emergency Temporary Access Service")

Host of the HathiTrust Research Center

Currently Digitized:

- 17,645,865 total volumes
- 8,484,623 book titles
- 469,920 serial titles
- 6,176,052,750 pages
- 791 terabytes
- 209 miles
- 14,337 tons
- 7,048,962 volumes(~40% of total) in the public domain

HathiTrust Origin stories



The New York Times — Technology

A FREE e-mail with Theater seats at great prices

NYTimes: Home - Site Index - Archive - Help — Welcome, - Member Center - Log Out

Go to a Section — Go — Search: All of Technology — Go

Technology Home | Circuits | Product Reviews | How To's | Deals

Advertisement

Google Is Adding Major Libraries to Its Database

By JOHN MARKOFF and EDWARD WYATT

Published: December 14, 2004

Google, the operator of the world's most popular Internet search service, announced today that it had entered into agreements with some of the nation's leading research libraries and Oxford University to begin converting their holdings into digital files that would be freely searchable over the Web.

It may be only a step on a long road toward the long-predicted global virtual library. But the collaboration of Google and research institutions that also include Harvard, the University of Michigan, Stanford and the New York Public Library is a major stride in an ambitious Internet effort by various parties. The goal is to expand the Web beyond its current valuable, if eclectic, body of material and create a digital card catalog and searchable library for the world's books, scholarly papers and special collections.

Google - newly wealthy from its stock offering last summer - has agreed to underwrite the projects while also adding its own technical abilities to the task of scanning and digitizing tens of thousands of pages a day at each library.

Enlarge This Image

Thor Swift

A book is scanned at Stanford University. Google's plans for digital files include the University of Michigan and the New York Public Library.

ARTICLE TOOLS

E-Mail This Article

December 2004:
Google & five research libraries announce massive book scanning project.

September 2005:
Authors Guild files a lawsuit against Google and the libraries for "massive copyright infringement."

The HathiTrust Digital Library

Around 50% English (but including over 450 languages)

From the 15th to 21st century (but with a strong plurality from the 20th century)

Around 61% in copyright or status unknown

Range of genres: fiction, history, science, government documents, and more

Items all contributed by HathiTrust member libraries

But it's not *perfect*…

- Mass digitization
  - (3+ billion page turns by thousands of scanning staff
- Minimal curation
- Uncorrected OCR

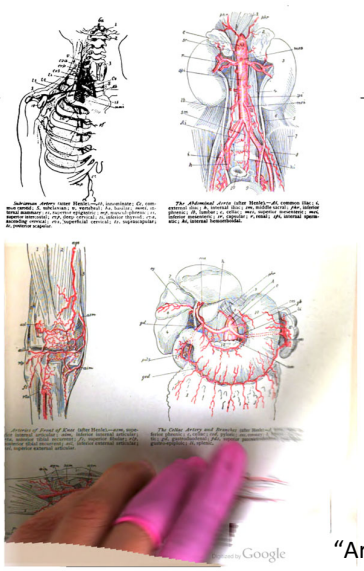
"You haven't hurt yourself, I expect," he said dryly, "so there's no harm done. I'll call that fellow with the lantern to give us a light." He did, and the vague shadow

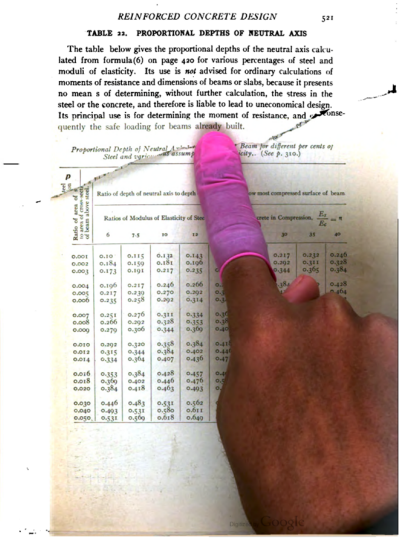
GT: I 'll call that fellow with the lantern to give us a light . "

OCR: I 'll call that fellow with ' Ehe 1i ht ' and the Vague shadow preced

CER = 0.484 WER= 0.667

Ming Jiang, et al. [Untitled manuscript], 2022





REINFORCED CONCRETE DESIGN 521

TABLE 22. PROPORTIONAL DEPTHS OF NEUTRAL AXIS

The table below gives the proportional depths of the neutral axis calculated from formula(6) on page 420 for various percentages of steel and moduli of elasticity. Its use is *not* advised for ordinary calculations of moments of resistance and dimensions of beams or slabs, because it presents no mean s of determining, without further calculation, the stress in the steel or the concrete, and therefore is liable to lead to uneconomical design. Its principal use is for determining the moment of resistance, and consequently the safe loading for beams already built.

"Art of Google Books"
http://theartofgooglebooks.tumblr.com/
Accessed October 25, 2018

…and it's not *comprehensive* or appropriately *diverse*

The HathiTrust Digital Library is *wholly dependent* on what member institutions provide, so…

…academic library acquisition patterns determine its content, e.g.,

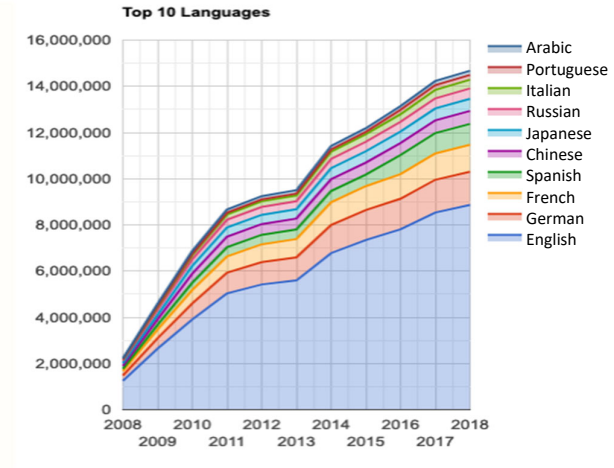
A dearth of romance novels
See Katherine Bode, "Why you can't model away bias," *MLQ*, 2019

Substantial gaps in Black speculative fiction
See Indiana University, "Corpus Completion Survey," https://sites.google.com/iu.edu/coverage/

…echoing the strengths & weaknesses of academic libraries writ very large: lots of copies of lots of editions of lots of titles by Jane Austen
… but not nearly enough Octavia Butler

HathiTrust collections over time

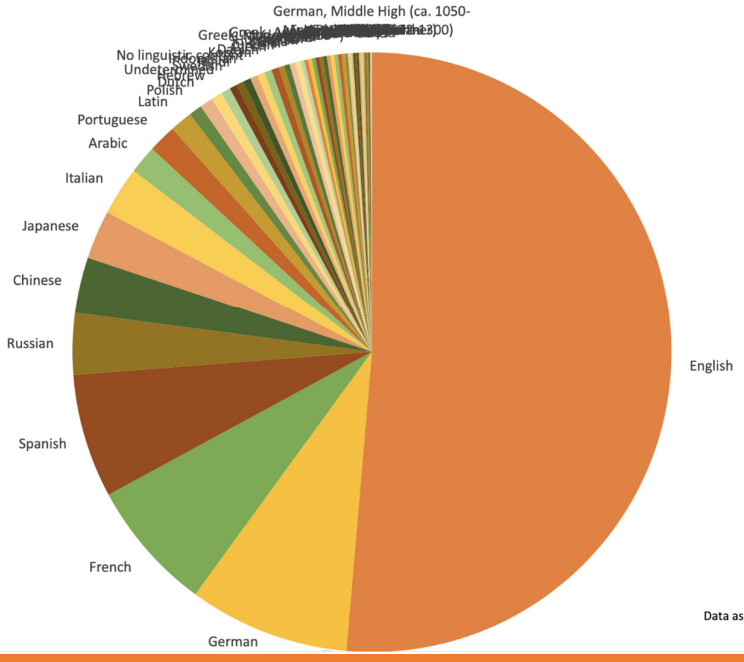
Publication dates of items in HathiTrust



Languages in HathiTrust
(by year of ingest, as of 2018)

HathiTrust languages (as of 2022)

Top 83 HathiTrust languages (>1,000 volumes each)



HathiTrust Research Center

Co-hosted by Indiana University and the University of Illinois

Computational analysis of the HathiTrust Digital Library collections

Development and promotion of ***non-consumptive research***

- Collection-building
- Data crunching
- Creation of new text-mining interfaces and tools
- Deeper understanding and enhancement of our collections

https://analytics.hathitrust.org/

Non-consumptive* research

From the rejected "Authors Guild v. Google Books" Settlement:

"'**Non-Consumptive Research**' means research in which **computational analysis** is performed on one or more Books, **but not research in which a researcher reads or displays substantial portions** of a Book to understand the intellectual content presented within the Book."

What and why?

- Complies with copyright law
- Foundation of HTRC work
- Related term: ****non-expressive use***

How?

- Partial Access
- Transformative Access
- Capsule Access

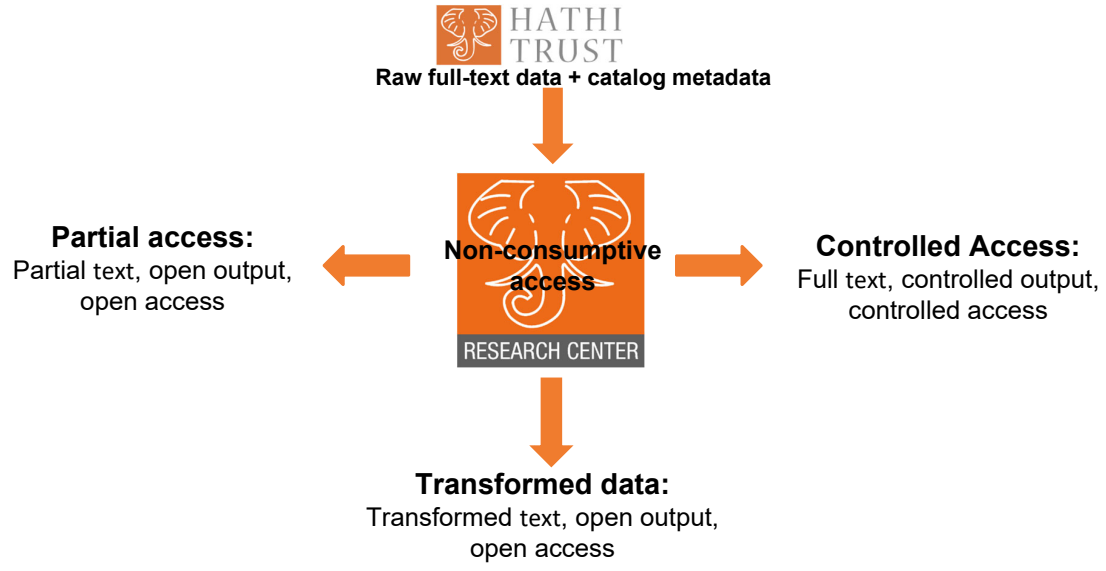
Non-consumptive research examples

- Machine learning and AI
- Text extraction
- Textual analysis and information extraction
- Linguistic analysis
- Image analysis
- File manipulation
- OCR correction
- Indexing and search

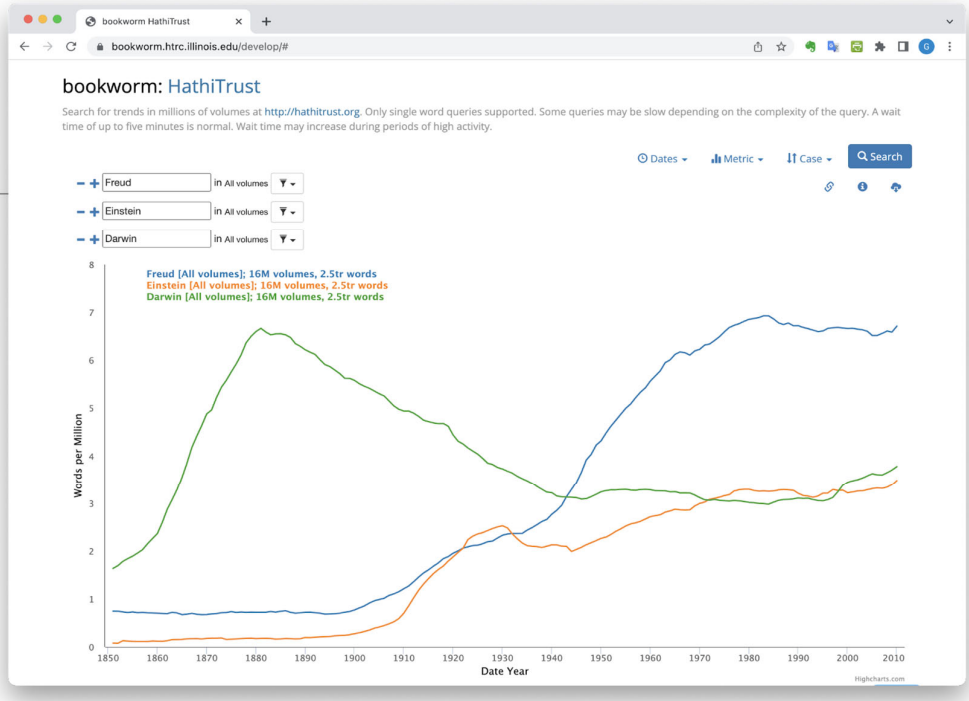
More here: https://www.hathitrust.org/htrc_ncup

HTRC Analytics

https://analytics.hathitrust.org



Bookworm



https://bookworm.htrc.illinois.edu

"Gendered Characterizations" visualization

Word usage in **English-language fiction** over time…

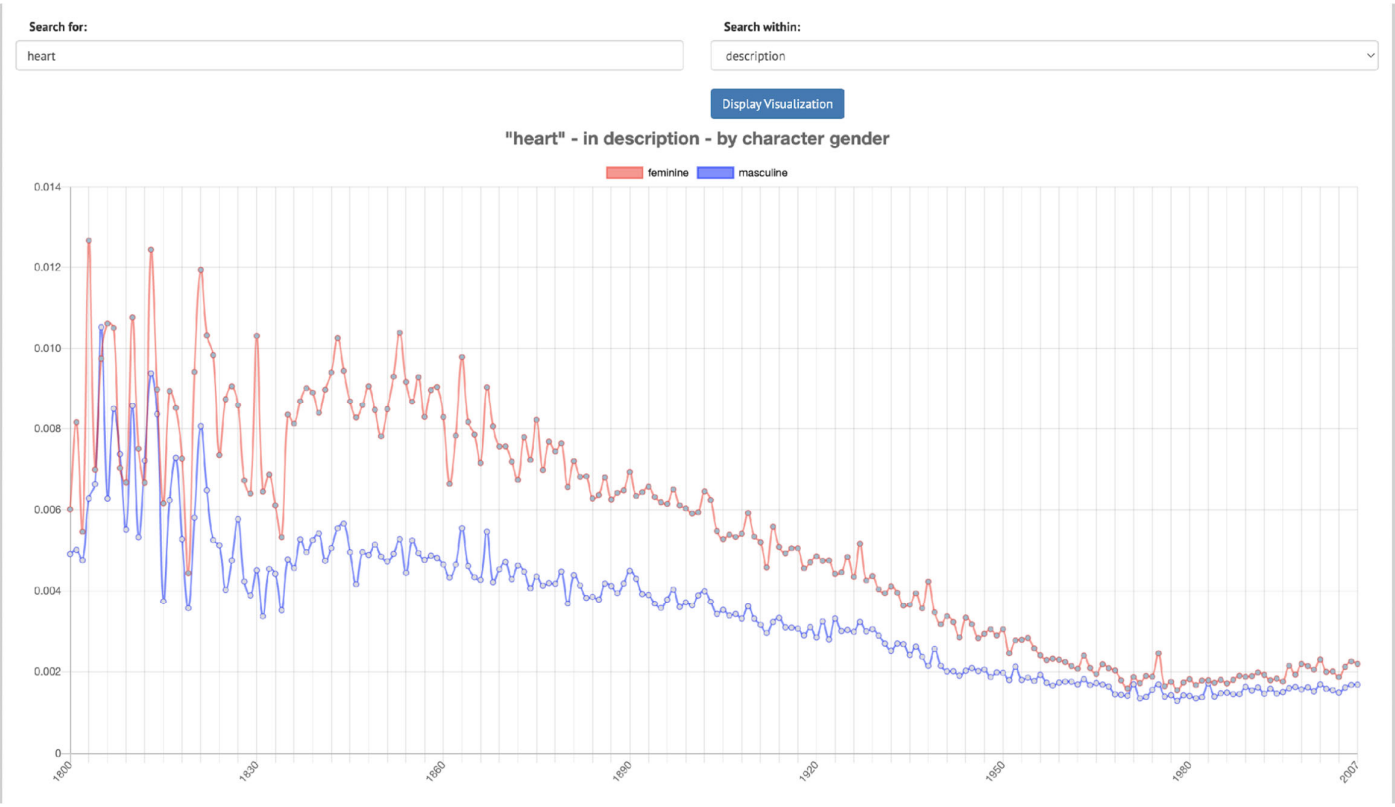
…in **descriptions** or in **dialogue**

…comparing the **genders** both of a word's associated fictional **characters**

…and of its **authors**.

Example: "heart"

https://tools.htrc.illinois.edu/genderviz/



HTRC Extracted Features

- A dataset derived from the entire HathiTrust corpus
- Volume-level, page-level, word-level data
- JSON format: structured data
- Copyright-free (downloadable in part or as a whole by anybody anywhere!)

https://analytics.hathitrust.org/datasets

HATHI TRUST Research Center

HTRC Extracted Features Dataset

Page-level features from 17.1 million volumes [v.2.0]

Attribution

Jacob Jett, Boris Capitanu, Deren Kudeki, Timothy Cole, Yuerong Hu, Peter Organisciak, Ted Underwood, Eleanor Dickson Koehl, Ryan Dubnicek, J. Stephen Downie (2020). *The HathiTrust Research Center Extracted Features Dataset (2.0)*.HathiTrust Research Center.https://doi.org/10.13012/R2TE-C227

17,123,746
10,550,952
6,572,794

6,221,631,336

2,906,819,723,689

Extracted Features dataset


```
{
  "id": "aeu.ark:/13960/t5x649n2b",
  "metadata": {
    "schemaVersion": "1.2",
    "dateCreated": "2015-02-12T20:51",
    "title": "Sermons delivered on various occasions by Matthew Richey.",
    "pubDate": "1840",
    "language": "eng",
    "htBibUrl": "http://catalog.hathitrust.org/api/volumes/full/htid/aeu.ark:/13960/t5x649n2b.json",
    "handleUrl": "http://hdl.handle.net/2027/aeu.ark:/13960/t5x649n2b",
    "oclc": "720289813",
    "imprint": "J. Ryerson, 1840."
  },
  "features": {
    "schemaVersion": "2.0",
    "dateCreated": "2015-02-19T17:14",
    "pageCount": 304,
    "pages": [
      {
        "seq": "00000001",
        "tokenCount": 60,
        "lineCount": 56,
        "emptyLineCount": 29,
        "sentenceCount": 6,
        "languages": [
          {
            "en": "1.00"
          }
        ],
        "header": {
          "tokenCount": 0,
          "lineCount": 0,
          "emptyLineCount": 0,
          "sentenceCount": 0,
          "tokenPosCount": {}
        },
        "body": {
          "tokenCount": 60,
          "lineCount": 56,
          "emptyLineCount": 29,
          "sentenceCount": 6,
          "tokenPosCount": {
            "6": {
```

```
      "body": {
        "tokenCount": 433,
        "lineCount": 89,
        "emptyLineCount": 22,
        "sentenceCount": 13,
        "tokenPosCount": {
          "1": {
            "CD": 2
          },
          "2": {
            "CD": 2
          },
          "3": {
            "CD": 2
          },
          "4": {
            "CD": 1
          },
          "5": {
            "CD": 1
          },
          "6": {
            "CD": 1
          },
          "est": {
            "NN": 3
          },
          ">»": {
            "NN": 1
          },
          "entirely": {
            "RB": 1
          },
          "quality": {
            "NN": 1
          },
          "cliché": {
            "FW": 1
          },
```

Other Derived Datasets

Word Frequencies in English-Language Literature, 1700-1922
(Ted Underwood)

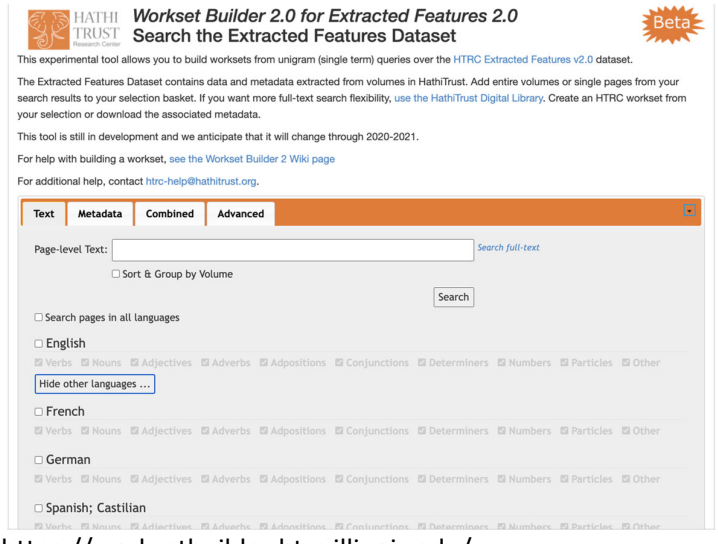
- Contains word frequencies for all English-language volumes of fiction, drama, and poetry
- Contains other volume metadata

Geographic Locations in English-Language Literature, 1701-2011
(Matthew Wilkens and Guangchen Ruan)

- Contains volume metadata as well as geographical locations
- Based on similar set of volumes in Ted Underwood's derived dataset

https://analytics.hathitrust.org/datasets

Workset Builder: fine-grained search



https://worksetbuilder.htrc.illinois.edu/

AN OPEN DATA APPROACH TO REVEALING INDIGENOUS TEXTS IN LARGE-SCALE DIGITAL REPOSITORIES: A CASE-STUDY OF LOCATING PAGES OF MĀORI TEXT IN THE HATHITRUST

XML

1. ABSTRACT

In this case study we report on our experiences in locating pages of M?ori text in the HathiTrust Digital Library (HTDL). Using traditional biographic metadata, *i.e.*, the language field, only 182 items were returned out of HTDL's 17.1 million volumes. Our Open Data approach is based on the freely available HathiTrust Extracted Features Dataset. We establish a collection of high frequency terms in Te Reo Māori, which we iteratively use as search terms to identify a group of candidate texts. We then apply NLP analysis to verify those texts that contain substantial amounts of the Māori language. Using this approach we were able to increase the number of volume returned to 598. This positive result suggests that scholars who want to analyse other low-resourced languages should be able to adopt our workflow to reveal otherwise hidden texts in their desired languages.

David Bainbridge (davidb@waikato.ac.nz), University of Waikato, New Zealand, J Stephen Downie (jdownie@illinois.edu), University of Illinois, USA and Hemi Whaanga , University of Waikato, New Zealand

Worksets

Workset = a **custom dataset** of HathiTrust volumes for analysis in HTRC

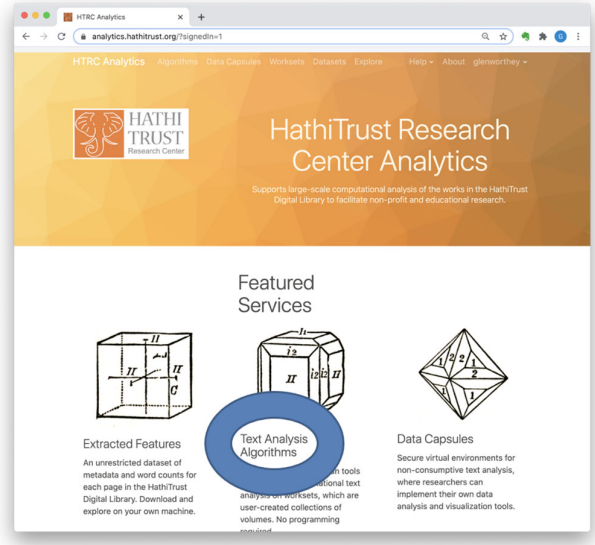
A simple list of volume IDs selected for analysis

- Transferable across HathiTrust systems
- Doesn't include the text

Citable, shareable, reproducible research

Analyze data from the HTDL without access to full text from HTDL

Researchers see only their workset and analyses

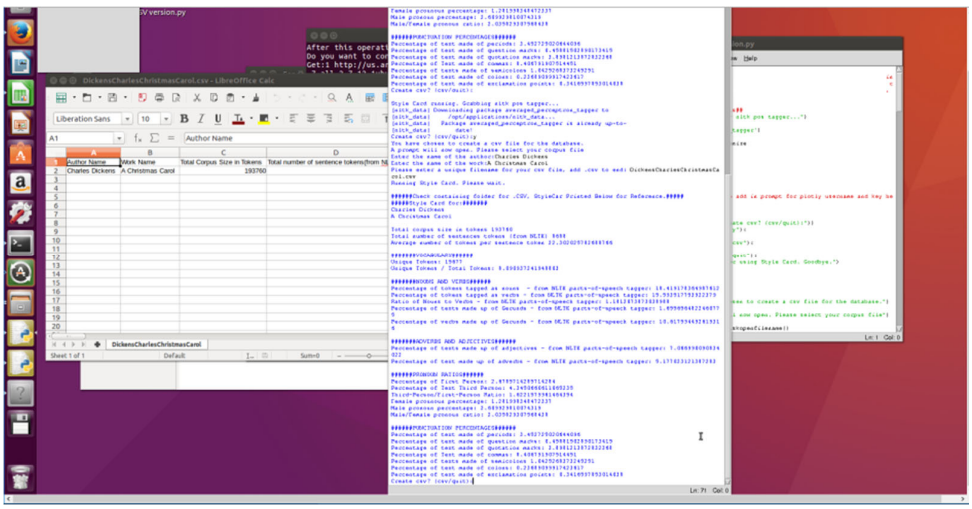


Data Capsules for controlled full-text access

Secure analysis environments

Linux virtual desktop

Protocols for data import & export



https://analytics.hathitrust.org/staticcapsules

Working on challenges

HathiTrust languages, a deeper dive: Japanese

- Deliberately targeting difficult cases: the **page-level (NLP) metadata** identified the text as **Japanese**
  But the the **volume-level (human catalog) metadata** was **anything BUT** Japanese

From manual classification of 400 randomly sampled pages:

- 6% were front cover images
- 6% were handwriting
- 19% were blank containing (some dirt marks)

Additionally:

- 19% of pages were horizontal in orientation, 59% vertical, remainder could not be determined
- 46% of the sampled pages contained Kanji script
- However only 1 of them was found to be in the Japanese language

David Bainbridge, Genna Hilbing, Ming Jiang, Yuerong Hu, Glen Layne-Worthey, J. Stephen Downie, *A Study on the Accuracy of OCR- and NLP-based Detection of Japanese Text in the HathiTrust Extracted Features v2.0 Dataset*, DH2022 (Tokyo, Japan)

Identifying front-matter pages algorithmically

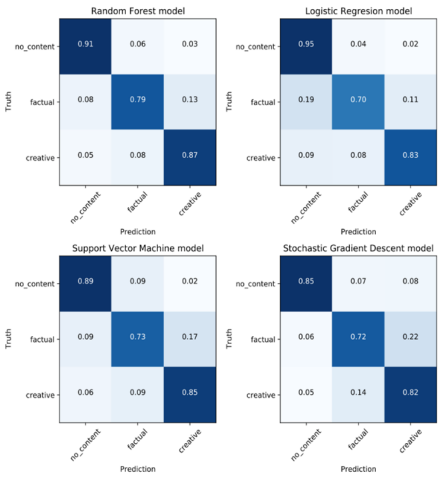
Identifying Creative Content at the Page Level in the HathiTrust Digital Library Using Machine Learning Methods on Text and Image Features

Nikolaus Parulian & Glen Worthey
HathiTrust Research Center
School of Information Sciences, University of Illinois Urbana-Champaign

iConference 2021

HATHI TRUST Research Center

School of Information Sciences
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN



Benchmarking NLP performance on uncorrected OCR

A Prototype Gutenberg-HathiTrust Sentence-level Parallel Corpus for OCR Error Analysis: Pilot Investigations

ABSTRACT

This exploratory study proposes a prototype sentence-level parallel corpus to support for the study of optical character recognition (OCR) quality in curated digitized library collections. Existing data resources, such as IC
ally aligned content
document-based or l
lenge of studying OC
tic features like senter
book-aligned corpus

1 INTRODUCTION AND BACKGROUND

The massive digitization of physical prints through machine scanning and optical character recognition (OCR) is of crucial importance to cultural heritage, knowledge preservation and general

GT: I stepped forward and looked through a chink between the logs .
OCR: I stepped forward an 1 r'ooked throuj^di a chink l)ctween tlie logs ,

```
I- stepped forward and l----ooked through---- a chink be---tween th--e logs .-
-1stepped forward an-1 r'ooked throu--j^di a chink--l)ctween t-lie logs-,
```

unaligned character positions aligned space adjacent to any unaligned character set

Figure 2: Visualization of GT-OCR character sequence alignment for token-based error detection. Symbol "-" denotes a gap character in the returned alignment. "|" denotes the position of aligned characters.

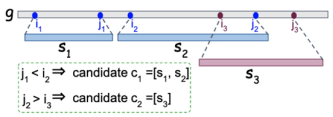


Figure 1: Aligned OCR candidate preparation, where g is a GT sentence, $\{s_1, s_2, s_3\}$ are OCR snippets, i, j denote the starting and ending position of any $g - s$ common string, respectively.

Algorithm 1: Sentence Alignment

Input: G: GT sentence list, H: OCR snippet list
Output: P=[(g, h)]: Sentence pair list

for *each sentence g in G* **do**
  Retrieve a sub-list of OCR snippets H'
  for *each snippet s in* H' **do**
    Compute similarity score $Sim_g = \frac{|g \cap s|}{|g|}$
    Compute similarity score $Sim_s = \frac{|g \cap s|}{|s|}$
    Find the longest common string $g[i:j]$
  end
  Get candidate pair list $C = [(c, \sum Sim_g, \{Sim_s\})]$ (see Figure 1)
  Find the optimal candidate with $Max(Sim_g)$ from C
  if $Max(Sim_g) > threshold\ A$ ***and*** $Avg[Sim_s] > threshold\ B$ **then**
    Add this candidate into P
  end
end

Ming Jiang, et al., JCDL 2022

What about under-represented literatures?

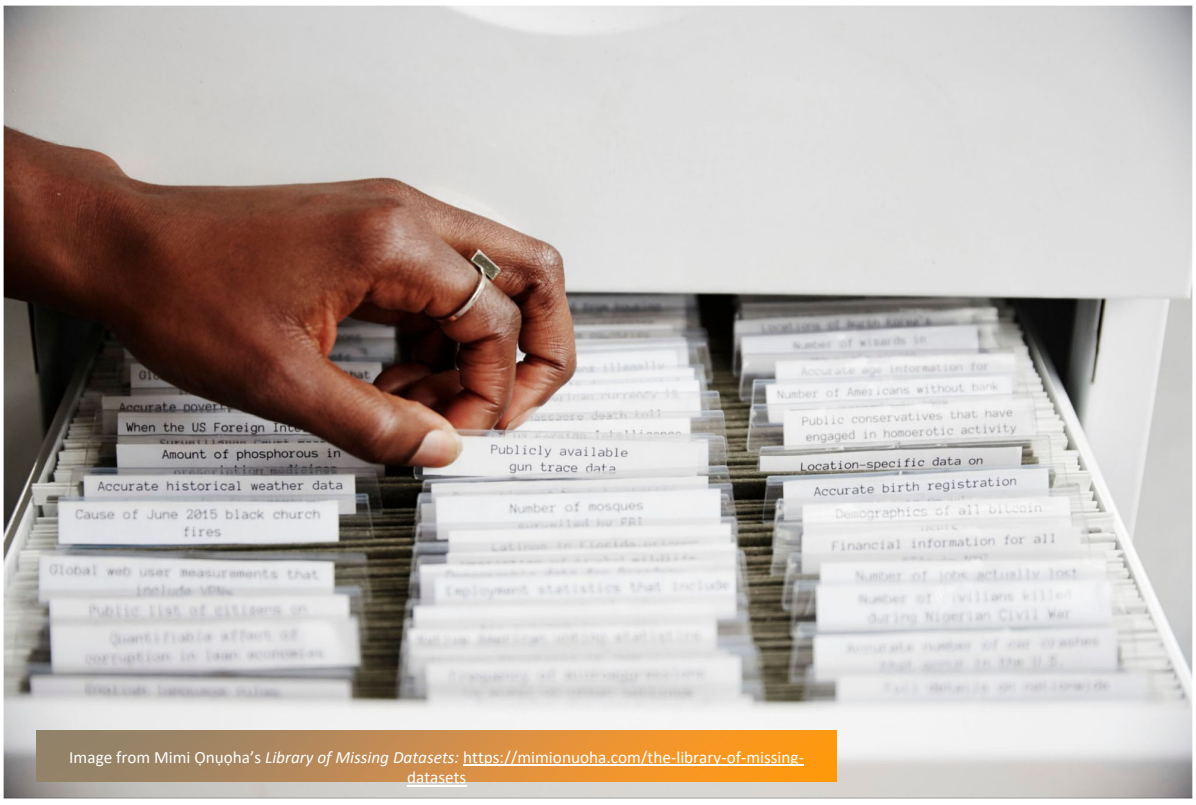


Image from Mimi Ọnụọha's *Library of Missing Datasets*: https://mimionuoha.com/the-library-of-missing-datasets

SCWAReD

SCWAReD

Scholar-Curated Worksets for Analysis, Reuse and Dissemination (SCWAReD)

A 3-year Mellon Foundation-funded project to collaborate with scholars to assemble and document worksets of traditionally under-resourced and/or marginalized textual communities.

Flagship partners **The Project on the History of Black Writing** at the **University of Kansas**, led by co-PI Dr. Maryemma Graham, will workset of all African-American fiction in HTDL, based on manually-verified list of Black authors and works at Project HBW.

Other partner projects:

- **Mining the Native American Authored Works in HathiTrust for Insights** Dr. Kun Lu, Dr. Raina Heaton (University of Oklahoma), and Dr. Raymond Orr (Dartmouth)
- **The Black Fantastic: Curated Vocabularies, Artifact Analysis and Identification** Dr. Clarissa West-White (Bethune-Cookman University) and Dr. Seretha Williams (Augusta University)
- **Creating Period-Specific Worksets for Latin American Fiction** Dr. José Eduardo González (University of Nebraska, Lincoln)
- **The National Negro Health Digital Project: Recovering and Restoring a Black Public Health Corpus** Dr. Kim Gallon (Brown University)

HISTORY OF BLACK WRITING
THE UNIVERSITY OF KANSAS

Mellon Foundation

SCWAReD

General SCWAReD workflow:

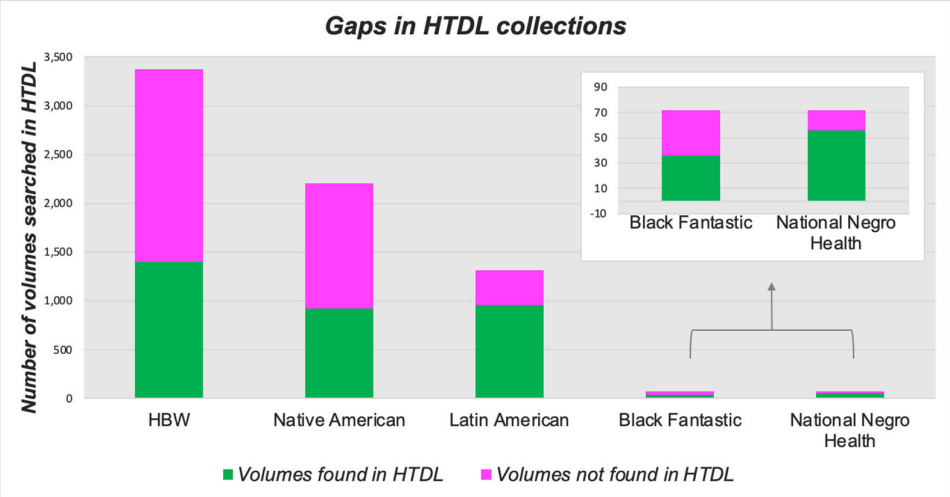
- Work with expert scholar collaborators to identify or generate a list of volumes relevant to their domain
- Search HTDL for these volumes using computational methods:
  - Author-title search of metadata records using regular expressions
  - Keyword analyses
  - Training machine learning classifier (more later!)
- Create workset of resultant volumes, including:
  - Workset inclusion rationale
  - Documentation of search methods, found and missing volumes
- Conduct exploratory data analysis of initial worksets:
  - Keyword + context analysis
  - Sentiment analysis
  - Entity extraction
  - Topic modeling
- Share results

We Are Here

SCWAReD

SCWAReD worksets

Searches have yielded the following found volumes from author-title lists that were searched:

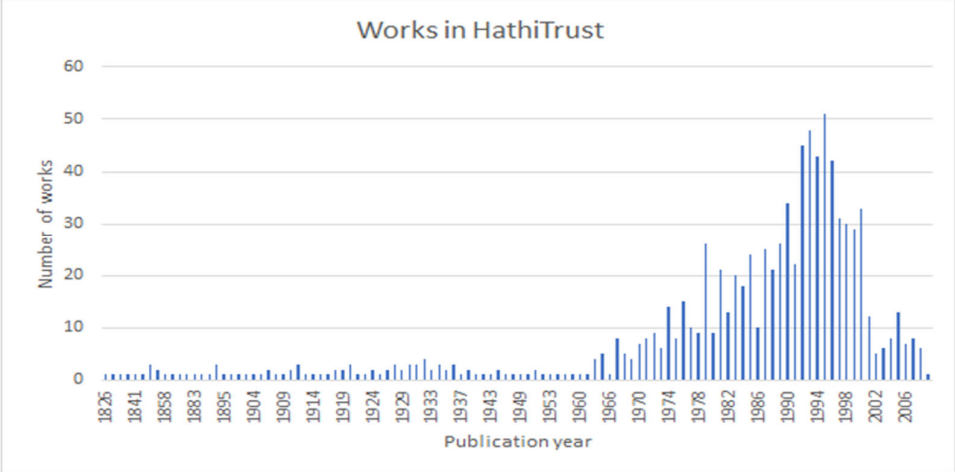


Work is ongoing to pilot how to fill gaps where possible and to do final verification on found volume lists. An announcement will come with final numbers and metadata!

SCWAReD

Native Authored Works - Early Exploratory Analysis

- The vast majority of works about Native peoples are not written by Native peoples (compare to 35,445 HT records for “American Indian”)
- Native authorship was very limited prior to 1960
- Most Native authors in North America represented in HTDL are writing in English
- Found authors from ~207 federally recognized tribes represented in the database (compare with ~1,208 recognized Native Nations and First Nations in the US and Canada)

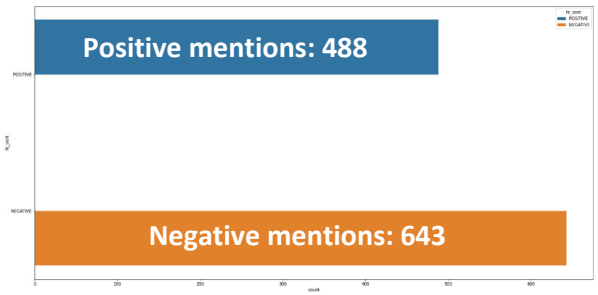


SCWAReD

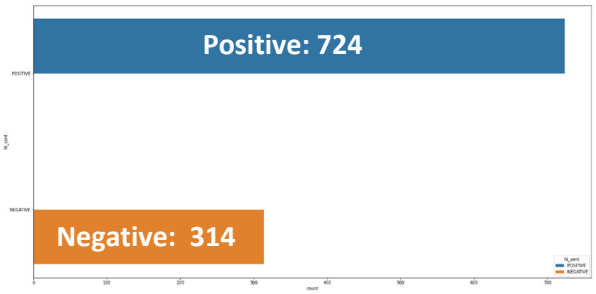
Native Authored Works - Early Exploratory Analysis

Sentiment discrepancy between "settler" and "pioneer" when extracted within their contexts (sentence level).

Still needs to be explored in-depth by domain experts!



"Settler" used more often in **negative** context



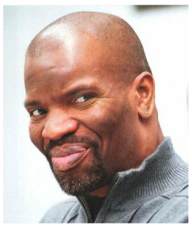
"Pioneer" used more often in **positive** context

SCWAReD

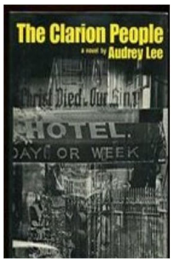
Black Fantastic - Early Exploratory Analysis

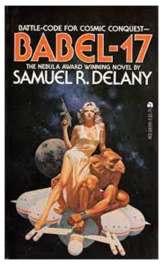
Black Fantastic literature:

- Is written by artists of the African Diaspora
- Can be defined using **Richard Iton's** definition: fiction engaged with the intersections of race and technology, as well as transcultural iterations of world-building [3]
- Is sometimes referred to as "Afrofuturism"



Richard Iton





3. Iton, Richard. In Search of the Black Fantastic. Oxford University Press, 2008. https://doi.org/10.1093/acprof:oso/9780195178463.001.0001.

SCWAReD

Black Fantastic - Early Exploratory Analysis

"Time" and time words are key terms in BF texts:

title	keyword	word_vol_rank	vol_words	tf-idf normalized
Night studies : a novel /	time	1	6426	0.04255064906
Over Edom, I lost my shoe /	time	1	3466	0.01673340132
I want a black doll.	time	2	3117	0.01154263193
Kindred /	time	3	2128	0.01891898843
The survivors	time	3	2311	0.01379651864
The salt eaters /	time	3	3137	0.01202072911
The landlord.	time	3	3002	0.01017663998
Tragic magic : a novel /	time	3	1239	0.009220445623
The militants /	time	3	1799	0.008947247234
Sweet whispers, Brother Rush /	time	4	1526	0.0122939275
The spook who sat by the door : a novel /	time	4	2073	0.009220445623
Light ahead for the Negro /	time	4	1040	0.004439473819

SCWAReD

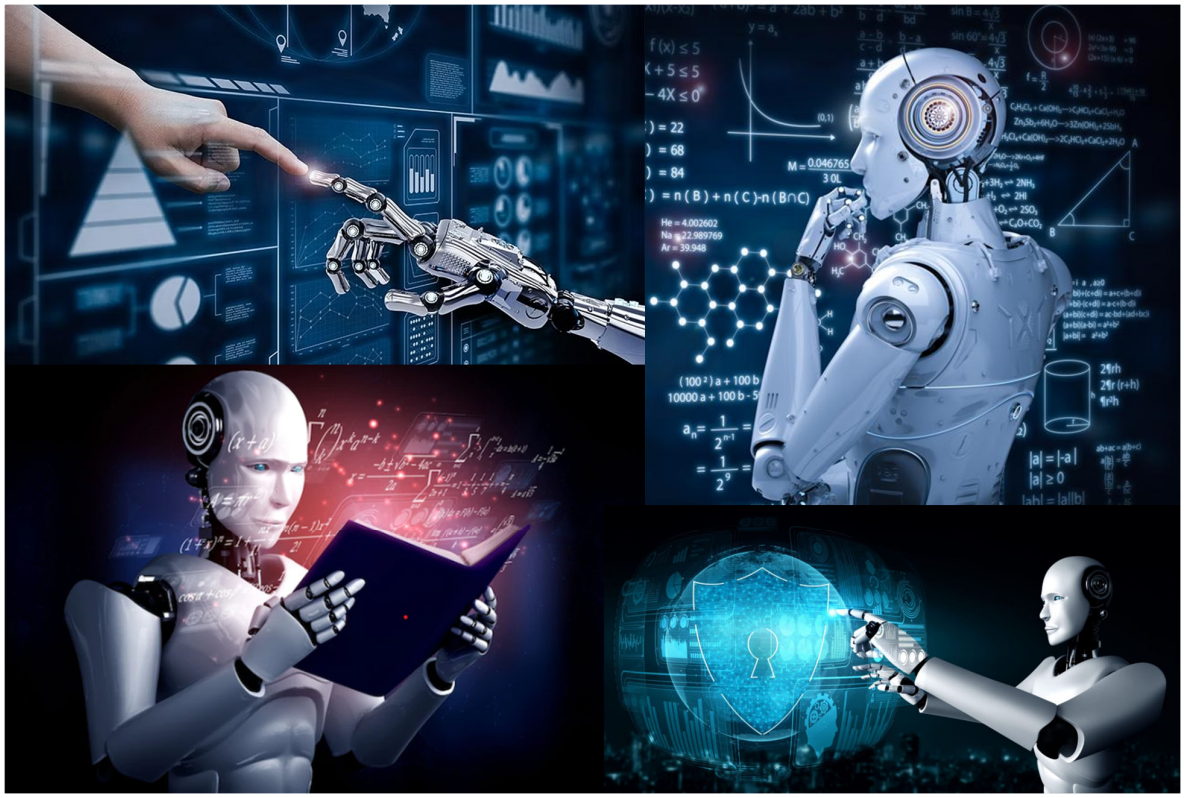
Black Fantastic - Early Exploratory Analysis

Black Fantastic as a genre is complex and encompasses many different styles and themes.

A.K.A. a number of our exploratory ideas didn't yield significant results

- Topic modeling was inconclusive
- Term analysis only surfaced strong links to time terms
- Black Fantastic genre was complex enough to fool a classifier (more on this later!)

But there remains a lot of data to be analyzed and questions to be asked!



LEVERAGING MACHINE LEARNING

ML & HTRC

HTDL's massive amount of varied data presents possibilities for testing machine learning (ML) approaches for practical library tasks

Two projects at HTRC leveraging machine learning:

- **Uncovering the Black Fantastic**
  - Led by Nikolaus Nova Parulian, and in conjunction with Dr. Clarissa West-White and Dr. Seretha Williams as part of SCWAReD project
  - Can ML identify genre fiction (using Black Fantastic as a case study) within larger sets of fiction?

- **Extending NovelTM Datasets for English-Language Fiction**
  - Working with Dr. Ted Underwood, co-author, with Patrick Kimutis and Jessica Witte, of the initial NovelTM dataset
  - Seeks to improve upon initial classification process and extend the dataset to include items added since initial dataset was created

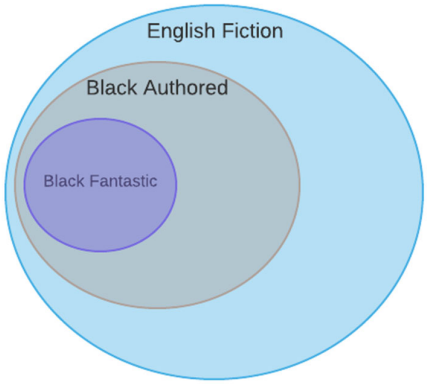
UNCOVERING BLACK FANTASTIC

The challenges of genre fiction:

- Variation between genres could/will be minute
- Genre is not something about which even humans agree
  - Complicates creating/assembling training data

Classification goal:

- Test results of training a classifier to differentiate between English-language fiction, Black-authored fiction and Black Fantastic fiction
- Benchmark performance of 4 predictive models:
  - Support Vector Machine (SVM)
  - Random Forest (RF)
  - Logistic Regression (LR)
  - Stochastic Gradient Descent (SGD)



UNCOVERING BLACK FANTASTIC

Uncovering Black Fantastic - The Data

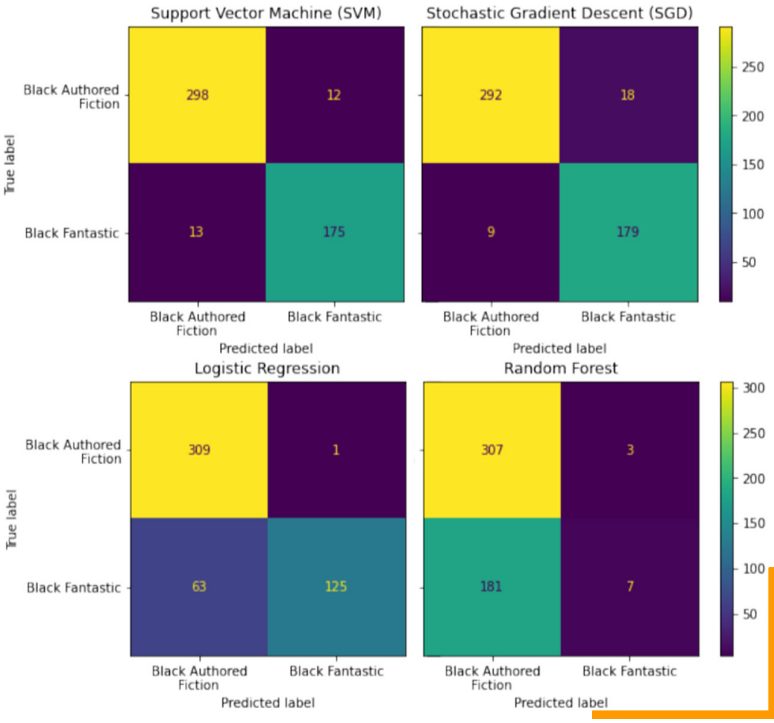
We train the classifier using HTRC EF data for each volume that is then vectorized (converted to numerical representations of relationships between words at the volume level) using TF-IDF

Workset Type	# Volumes	# Tokens	# Unique tokens	Average. volume occurrence count for token
Black Authored (BA) Fiction	1,556	325,214	213,547	12.21
General English-Language (EL) Fiction	1,524	255,053	141,443	11.14
Black Fantastic (BF) Fiction	37	94,034	28,204	4.33

UNCOVERING BLACK FANTASTIC

Uncovering Black Fantastic - The Results

- SVM performed the best
- However, overall there is not enough data for a task like this to be useful from a practical standpoint of finding unknown Black Fantastic volumes in the HTDL
- More work to be done with more training data



UNCOVERING BLACK FANTASTIC

Uncovering Black Fantastic - Word Feature Topic Clustering

EXTENDING NOVELTM DATASETS

Extending NovelTM Datasets for English-language fiction

The initial NovelTM project used ML to find English-fiction in the HTDL

Why?

- The HTDL is massive, and metadata is missing, inconsistent, and incomplete.
- Even if metadata is complete:
  - standards have changed and evolved over time, leading to inconsistent metadata
  - we did not start explictly cataloging fiction as a genre until the 1990s* and still today we often do not do so with sufficient granularity (e.g. to find “sci-fi” or “romance” easily)
- These problems are significantly exacerbated by scale, especially that of the massive HTDL

* Miller, David P. “Out from Under: Form/Genre Access in LCSH.” Cataloging & Classification Quarterly 29, no. 1–2 (June 1, 2000): 169–88. https://doi.org/10.1300/J104v29n01_12.

EXTENDING NOVELTM DATASETS

The satanic verses

LDR	00630nam a22002171a 4500
001	001073126
003	MiAaHDL
005	20210817000000.0
006	m d
007	cr bn ---auaua
008	890227s1989 nyu 00011 eng d
010	␣␣ ‡a 88040266
020	␣␣ ‡a 0670825379
035	␣␣ ‡a (MiU)990010731260106381
035	␣␣ ‡a sdr-miu.990010731260106381
035	␣␣ ‡a (OCoLC)18558869
035	␣␣ ‡a (RLIN)MIUG89-B7404
035	␣␣ ‡z (MiU)Aleph001073126
050	0 ␣ ‡a PR9499.3.R8 ‡b S28 1989
082	0 ␣ ‡a 823 ‡2 19
100	1 ␣ ‡a Rushdie, Salman.
245	1 4 ‡a The satanic verses / ‡c Salman Rushdie.
260	␣␣ ‡a New York, NY : ‡b Viking, ‡c 1989.
300	␣␣ ‡a [12], 3-546, [3] p. ; ‡c 24 cm.
538	␣␣ ‡a Mode of access: Internet.
590	␣␣ ‡a labd: In original dust jacket.
CID	␣␣ ‡a 001073126
DAT	0 ␣ ‡a 19890323000000.0 ‡b 20210817000000.0
DAT	1 ␣ ‡a 202109190607l2.0 ‡b 2021-09-19T18:28:26Z
DAT	2 ␣ ‡a 2020-07-20T17:30:02Z
CAT	␣␣ ‡a SDR-MIU ‡d ALMA ‡l prepare.pl-004-008
FMT	␣␣ ‡a BK
HOL	␣␣ ‡0 sdr-miu.990010731260106381 ‡a MiU ‡b SDR ‡c MIU ‡p mdp.39(
974	␣␣ ‡b MIU ‡c MIU ‡d 20210919 ‡s google ‡u mdp.39015046867852 ‡y
974	␣␣ ‡b MIU ‡c MIU ‡d 20210928 ‡s google ‡u mdp.39015038895879 ‡y
974	␣␣ ‡b MIU ‡c MIU ‡d 20211029 ‡s google ‡u mdp.39015038928274 ‡y
974	␣␣ ‡b MIU ‡c MIU ‡d 20210916 ‡s google ‡u mdp.39015054423457 ‡y

El Aleph

LDR	00531nam a22001931 4500
001	001522594
003	MiAaHDL
005	20210817000000.0
006	m d
007	cr bn ---auaua
008	890828s1952 ag llllllllllllllspald
035	␣␣ ‡a (MiU)990015225940106381
035	␣␣ ‡a sdr-miu.990015225940106381
035	␣␣ ‡a sdr-txu-1.b24665514
035	␣␣ ‡a (OCoLC)4654409
035	␣␣ ‡a (SAZTEC)179807260
035	␣␣ ‡a (CaOTULAS)176514151
035	␣␣ ‡z (MiU)Aleph001522594
040	␣␣ ‡c SAZTEC
100	1 ␣ ‡a Borges, Jorge Luis, ‡d 1899-1986.
245	1 3 ‡a El Aleph.
250	␣␣ ‡a 2. ed.
260	␣␣ ‡a Buenos Aires, ‡b Editorial Losada ‡c [1952]
300	␣␣ ‡a 157 p. ‡c 21 cm.
538	␣␣ ‡a Mode of access: Internet.
CID	␣␣ ‡a 001522594
DAT	0 ␣ ‡a 19890828000000.0 ‡b 20210817000000.0
DAT	1 ␣ ‡a 202109170607l2.0 ‡b 2021-09-17T19:43:30Z
DAT	2 ␣ ‡a 2020-04-27T17:30:02Z
CAT	␣␣ ‡a SDR-MIU ‡d ALMA ‡l prepare.pl-004-008
FMT	␣␣ ‡a BK
HOL	␣␣ ‡0 sdr-miu.990015225940106381 ‡a MiU ‡b SDR ‡c MIU ‡p mdp.390150164441563 ‡s MIU ‡1 990015225940106381
974	␣␣ ‡b MIU ‡c MIU ‡d 20210917 ‡s google ‡u mdp.39015016441563 ‡y 1952 ‡r ic ‡q bib ‡t non-US bib date1 >= 1926
974	␣␣ ‡b TXU ‡c TXU ‡d 20190701 ‡s google ‡u txu.059173024388081 ‡y 1952 ‡r ic ‡q bib ‡t non-US bib date1 >= 1926
974	␣␣ ‡b TXU ‡c TXU ‡d 20190701 ‡s google ‡u txu.059173024387941 ‡y 1952 ‡r ic ‡q bib ‡t non-US bib date1 >= 1926

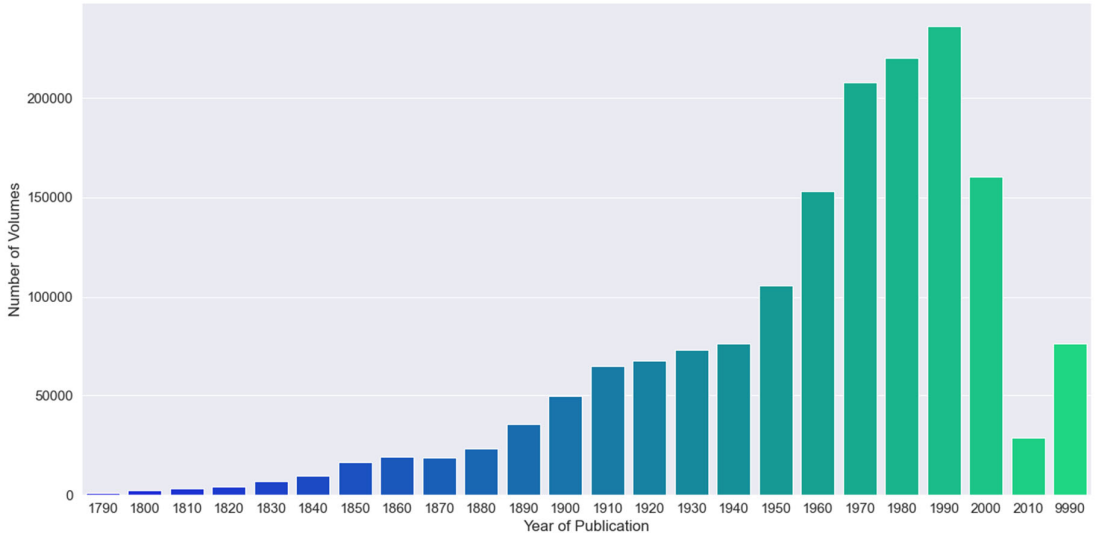
Madame Bovary; mœurs de province

LDR	00751nam a22002291 4500
001	001015020
003	MiAaHDL
005	20211114000000.0
006	m d
007	cr bn ---auaua
008	890208s1857 fr 00000 fre d
010	␣␣ ‡a 12027102
035	␣␣ ‡a (MiU)990010150200106381
035	␣␣ ‡a sdr-miu.990010150200106381
035	␣␣ ‡a (OCoLC)2782742
035	␣␣ ‡a (SAZTEC)095307540
035	␣␣ ‡a (CaOTULAS)176049956
035	␣␣ ‡z (MiU)Aleph001015020
040	␣␣ ‡c CarP
050	0 ␣ ‡a PQ2246 ‡b .M2 1857 (Rare Bk Coll)
100	1 ␣ ‡a Flaubert, Gustave, ‡d 1821-1880.
245	1 0 ‡a Madame Bovary; mœurs de province.
260	␣␣ ‡a Paris, ‡b M. Lévy frères, ‡c 1857.
300	␣␣ ‡a 2 v. (490 p.) ‡c 18 cm.
538	␣␣ ‡a Mode of access: Internet.
899	␣␣ ‡7 239950112300063 81 ‡a 39015024525076 ‡b Buhr A 798453
CID	␣␣ ‡a 001015020
DAT	0 ␣ ‡a 19890208000000.0 ‡b 20211114000000.0
DAT	1 ␣ ‡a 202111140605l9.0 ‡b 2021-11-16T18:31:24Z
CAT	␣␣ ‡a SDR-MIU ‡d ALMA ‡l prepare.pl-004-008
FMT	␣␣ ‡a BK
HOL	␣␣ ‡0 sdr-miu.990010150200106381 ‡a MiU ‡b SDR ‡c MIU ‡p mdp.39015024525076 ‡s MIU ‡1 990010150200106381
974	␣␣ ‡b MIU ‡c MIU ‡d 20211116 ‡s google ‡u mdp.39015024525076 ‡y 1857 ‡r pd ‡q bib ‡t non-US bib date1 < 1896

EXTENDING NOVELTM DATASETS

Extending NovelTM Datasets - But Why Now?

About 1.6 million new English-language items that might be fiction have been added to HTDL since the first NovelTM dataset was generated



EXTENDING NOVELTM DATASETS

Extending NovelTM Datasets - The Data

We use a sample of volumes to train and test the classifier, and we benchmarked 3 different samples:

- **Sample 1:** 10,108 random volumes, matching distribution of items added to HTDL since 2016, by decade, yielding 1,605 fiction, 8,503 non-fiction.
- **Sample 2:** 9,969 random volumes, with the same selection logic as sample 1, but incorporating as many manually-verified fiction vols from NovelTM dataset as possible, yielding 1,580 fiction, 8,389 non-fiction volumes
- **Sample 3:** 9,061 volumes, including 53 F and 211 NF volumes for every decade represented in items added to HTDL since 2016, creating a train/test with equal volumes for each decade, yielding 1,328 fiction, 7,733 non-fiction volumes

We also benchmarked 3 different models:

- Logistic regression
- Support vector machine
- Random forest (120 trees)

Each model was implemented using scikit-learn in Python

Extending NovelTM Datasets - The Results

Each model and sample did well, but the best overall (F1) was LR on the corrected Sample 3.

	Logistic Regression			Support Vector Machine			Random Forest		
	P	R	F1	P	R	F1	P	R	F1
Sample 1	0.7838	**0.9755**	0.8692	0.8384	0.9205	0.8776	0.8665	0.8930	0.8795
Sample 2	0.8589	0.9470	0.9008	0.8885	**0.9238**	0.9058	0.8824	**0.8940**	0.8882
Sample 3	0.8804	0.9199	0.8997	0.9286	0.8750	0.9010	**0.9697**	0.8889	**0.9275**
Sample 3 - Corrected	**0.9249**	0.9506	**0.9376**	**0.9702**	0.9043	**0.9361**	0.9689	0.8642	0.9135
Mean values	0.8620	**0.9483**	0.9018	0.9064	0.9059	**0.9051**	**0.9219**	0.8850	0.9022

EXTENDING NOVELTM DATASETS

Extending NovelTM Datasets - The Results

While reviewing errors, 4 main types were identified:

- **Incorrect ground truth:** volumes incorrectly tagged as fiction or not fiction in their metadata. Examples include Stephen Crane's *The Red Badge of Courage* or *Wuthering Heights* by Emily Bronte incorrectly being marked as not fiction.
- **Volumes that blur the lines of fiction, such as memoir, biography, or travel narrative:** volumes that look a lot like fiction or not fiction, but are the inverse. Examples include Daniel Defoe's *Robinson Crusoe* or John Hanning Speke's *Journal of the Discovery of the Source of the Nile*—the former being a "fake" travel narrative and the latter purporting to be authentic.
- **Non-prose fiction:** volumes that looked like fiction, but are a form separate from prose, and thus not correct for our dataset. Examples include books of poetry and dramas.
- **True errors:** the least frequent errors—volumes the model just got plain wrong. Examples include annotated volumes, containing fore- and/or afterwords that can influence the model, like*The Works of Dr. Jonathan Swift from 1751.* Other examples are Ward Greene's collection of prominent historical news stories, *Star Reporters and 34 of Their Stories*, or a bound anthology of *Frank Leslie's Lady's Magazine*.

COMING SOON

TORCHLITE

Tools for Open Research and Computation with HathiTrust: Leveraging Intelligent Text Extraction (TORCHLITE)

- A 2-year NEH-funded project to develop API infrastructure for access HTRC's Extracted Features (EF) Dataset and a web-based dashboard of visualization widgets
- Work is ongoing on dashboard and widget design, API development
- Will include incrementally updated EF data, easier mode of access to HTRC EF data, and the ability to better create custom visualizations of HathiTrust and HTRC data both inside the browser and in custom code.

BOOKNLP FICTION DATASET

BookNLP dataset for English-language fiction
Detailed derived data and metadata for each of ~213,000 English-language fiction volumes (from NovelTM set), in a fully open and public domain dataset.

Data included for each volume:

- Tokens and metadata about tokens *
- Entities–people, places, organizations–in the text
- Quotations, speakers and metadata
- Supersense tags - advanced semantic and entity tags based on WordNet
- .book file with the following info about every character mentioned 2+ times:
  - proper/common/pronominal references
  - referential gender
  - actions for the which they are the agent and patient
  - objects they possess
  - modifiers

David Bamman

BookNLP is a DH tool for book-length documents by David Bamman, Ted Underwood and Noah Smith. More info about BookNLP: https://github.com/booknlp/booknlp

* With some documented modification of standard output files to adhere to non-consumptive use policy

Big Release Announcement

HTRC Docs Spaces Search Log in Sign up

Pages / ... / HTRC Derived Datasets

HTRC BookNLP Dataset for English-Language Fiction

Created by Ryan Dubnicek, last modified on Jun 12, 2023

Work with rich, unrestricted entity, word, and character data extracted from over 200,000 volumes of English-language fiction in the HTDL

The **HTRC BookNLP Dataset for English-Language Fiction (ELF)** derived dataset was created using the BookNLP pipeline, extracting data from the NovelTM English-language fiction set, a supervised machine learning-derived set of around 213,000 volumes in the HathiTrust Digital Library.

BookNLP is a text analysis pipeline tailored for common natural language processing (NLP) tasks to empower work in computational linguistics, cultural analytics, NLP, machine learning, and other fields. This dataset has the potential to power exciting new computational research of English-language literature, along with more methods-focused work in the areas mentioned above, with minimal infrastructure support from HTRC. As with all derived datasets, a key goal of the project is also to lower the barrier for working with HathiTrust Digital Library (HTDL) and HTRC data, and allow users to leverage computational resources they may have personally or through their institutional affiliation. Specificities of the data, a discussion of its non-consumptiveness, and other pros and cons of release follow in the next sections.

This dataset is modified from the standard BookNLP pipeline to output only files that meet HTRC's non-consumptive use policy that requires minimal data that cannot be easily reconstructed into the raw volume to be released. Please see the Data section below for specifics on what files are included and their description.

Short URL for this page: https://wiki.htrc.illinois.edu/x/BgCUCQ

Dataset Stats

# of volumes represented	201,527
# of files	604,561
# files derived from in-copyright volumes	90,857
Size of full dataset (gigabytes)	451.2 GB

Jump to Section

- About the Data
  - Entities
  - Supersenses
  - Character data (.book files)
- Getting the Data
  - Download the data via rsync
  - Accessing the Dataset from an HTRC Data Capsule
  - External Resources

Attribution

Ryan Dubnicek, Boris Capitanu, Glen Layne-Worthey, Jennifer Christie, John A. Walsh, J. Stephen Downie (2023). *The HathiTrust Research Center BookNLP Dataset for English-Language Fiction.* HathiTrust Research Center. https://doi.org/10.13012/d4gy-4g41

This derived dataset is released under a Creative Commons Attribution 4.0 International License.

Short URL for this page: https://wiki.htrc.illinois.edu/x/BgCUCQ

Summary and conclusions

- Working at Hathitrust scale is a real big data challenge
- Combination of human expertise ***and*** AI/ML required to
  - Discover gaps
  - Build worksets
  - Correct metadata
  - Extract features
  - Overcoming OCR errors
  - Identify copyright issues
  - Analyze data
  - Curate and disseminate data and results

ありがとう

Thank you