Toward a Big Data Analysis System for Historical Newspaper Collections Research

Sandeep Puthanveetil Satheesan, Bhavya, Adam Davies, Alan B. Craig, Yu Zhang, and ChengXiang Zhai

Platform for Advanced Scientific Computing (PASC) Conference, Basel, Switzerland - June 27, 2022





Extreme Science and Engineering Discovery Environment



Introduction

- Large digitized historical newspaper collections are now available
 - Chronicling America
 - ProQuest Historical Newspapers[™]
 - Europeana Newspapers
 - The British Newspaper Archive
 - ο.
- We now can ask complex and broad **questions of historical relevance** as depicted in the **newspapers**
- Previously, one had to visit a library or an archive and perform **extensive** and **tedious manual research** to access this type of information
- Gap in availability of **historical newspaper collections** vs. **big data systems** for processing and analyzing such collections, and extracting relevant information
- We propose a scalable and customizable **big data analysis system** using **HPC** and **AI analysis tools** for images and text



Motivating Question

- How has the concept of *juvenile delinquency* been **socially constructed** over time in news media?
 - \circ $\,$ Our system has been developed in collaboration with Dr. Yu Zhang's research on this topic
 - Recorded rates of juvenile delinquency in the US and UK increased substantially in the early/mid 19th century. Was this due to a(n)...
 - objective increase in criminal juvenile behavior or resulting harm?
 - subjective change in social construction of juvenile criminality?
 - Millions of historical news articles could provide novel insights
 - Only feasible to analyze this volume of data with HPC and AI analysis tools for images and text





Challenges

- Designing efficient, intelligent algorithms that can quickly process **millions of newspaper records** (hundreds of terabytes of data)
- Segmenting historical newspaper images that comes with variations and imperfections, differentiating different content types (e.g. headlines, articles, ads), and extracting text through OCR/re-OCR
- Retrieving **relevant information** from heterogeneous newspaper collections
- Customizing **image** and **text analysis** tools to support researchers in **interactively analyzing** and **visualizing** information in real-time



Our Contributions

- Investigate design principles for developing a big data analysis system for newspaper collections research
- **Qualitatively evaluate** current AI techniques related to processing newspaper images and texts
- Use flexible text analysis operators to support diverse analysis workflows, enabling social scientists to investigate a wide variety of novel research questions
- Facilitate human-in-the-loop collaboration between social scientists and AI systems



Related Work

- The Valley of the Shadow (<u>https://valley.lib.virginia.edu/</u>)
 - Digital archive containing thousands of original letters and diaries, newspapers and speeches, census and church records from two U.S. counties in the Civil War era
 - Outcome of tireless manual effort of collecting, curating, and transcribing documents
- Newspaper Navigator (<u>https://news-navigator.labs.loc.gov/</u>)
 - Public dataset containing extracted headlines and visual content processed on 16.3 million historical newspaper pages on the LOC-CA
 - Multiple data products, a scalable pipeline, and a fine-tuned deep learning model to recognize content like photographs, cartoons, headlines, advertisement, etc.
 - Our system has many similar components, but also includes a variety of interactive text analysis operators and visualizations (e.g. temporal semantic analysis, human-in-the-loop text classification, topic modeling, etc.)
- ShiCo (Kenter et al., 2015; Martinez-Ortiz et al., 2016)
 - Uses historical word embeddings to find and visualize changing vocabulary related to a concept over time
 - We extend this tool and integrate it with other text analysis operators



System Design & High-level Architecture

- Design Principles
 - **Reuse** existing software
 - Modularize system
 - **Move compute to data** via application of Docker containers
 - Integrate innovative and existing text analysis operators that users can flexibly combine
 - Support **human-in-the-loop** Al/human collaboration



Data Collection

- Collect data from relevant sources and download it into suitable data and compute resources
- Library of Congress Chronicling America (LOC-CA)
 - Historical American newspapers in the public domain from 1777 to 1963
 more than 19.6 million pages of digitized U.S. newspapers
 - Downloaded 18.4 million text files and metadata files and 500,000 scanned newspaper images LOC-CA Bulk Data API
- ProQuest Historical Newspapers[™]
 - Digital archive offering full-text and full-image articles for newspapers dating back to the 18th Century
 - Digitized newspaper data that are not in the public domain due to restrictions on the data, we had to take additional steps to ensure compliance
 - Using about **58 million documents** from this collection



Image Analysis

- Methods used to perform OCR or newspaper segmentation or any other operation performed on the image to retrieve some useful information from it
- Analyzed OCR metadata to understand the data and to recreate the reading order and segment articles from text files (<u>Satheesan et al.</u>, <u>2019</u>). Later focussed on newspaper segmentation of the LOC-CA collection and re-OCR
- Newspaper region segmentation Mask-RCNN model
 - ResNet-101 backbone
 - Pre-trained on MS COCO dataset
 - **Training set: 60 newspaper images**; 1728 manually annotated regions, **validation set: 25 newspaper images**; 801 manually annotated regions (latest count)
 - Classes: Advertisement, article, banner, masthead, photo, and run head
 - With data augmentation horizontal flipping, rotation, added brightness and blurring pixels





Image Analysis: Newspaper Segmentation

Output of our current best performing Mask-RCNN model on some example newspaper images showing different segmented regions



	The hazel hoe is considered by for- ist-fire fighters as the best single ool for digging fire line or trench. It is a strong broad-bladed hoe suitable or cutting weeds and small brust- is well as digging in the soil.	extraastion of child life, which i immo-power capital. "England and Prance began as th United States has been tempted to begin—by letting the child rabo ico heavily for the way in child labo nerves, weakened bodies and prema- tured decite, but England and Franc- tured to life war burdens from th children by giving them a chance to play. There is urgent need to give our boys and girls an America equare deci-their safety valve o play.	INFORM THE ONE OF CONSERVATION	Too Much. "It's hard to believe that she's as in- telligent a woman as they say she is. She's going to marry a Jap, you know." "Oh, well, love is blind, they say." "Yes, but there's no excuse for its being color-blind." — The Catholic Standard and Times.	Getting Him in Line. Dr. Younger—You're a little bit un der weight. You don't eat enough. Mr Mussel—Nonsense! If 1 were to ea more I'd have dyspepsia. Dr. Younger —I know that, but that's my specialty you know.	<text><text><text></text></text></text>
--	---	---	--	---	--	---

Doing Good

Few medicines have met with more favor or accomplished more good than Chamberlain's Colle and Diarrhoea Remedy. John F. Jantzen, Delmeny, oj Sack., says of it, "I have used Chamberlain's Colle and Diarrhoea Remedy myself and in my family, and can recommend it is being an exceptionally fine preparation."

ALMOST ENTIRE WORLD

NOW ENGAGED IN WAF

Of the 1,600,000,000 people who

opulate the earth only 120,000,000

r less than one-tenth, are at peace

The Tentonic allies have 160 000 000

eople and the entente nations 1,

60,000,000. If the war should con

inue two or three years longer it i

juite possible that only a few mil

ions of all the people in the work

ould be on a peace footing, and

hey will be suffering all the rigors o

war except those of actual combat.

American soldiers engaged in th French forests write, "there hav been no fires in these forests fo years." NOTICE TO WOOD CONTRACT Bids will be received by the t ty Court of Columbia County, on an until August 71/918 at 2 of p. m., at which time bids will opened for 100 cords of four abab or cordwood, delivered courthouse in St. Helen, Org The right is reserved to rejee and all bids.

The Limit 5f Speed. The fool and his money are soo arted, but for quick operation Wa treet holds the record.—Detroit Fro Press.

FOREST NOTE:

R. CON

S FIRE! FIRE! FIRE! FIRE! See E. E. Quick, St. Helens, and insure your property in the Oregon Fire Relief Association of McMinn ville, Oregon. Don't put it off. Mch6tf W. I. WARREN, Agent

Bastile Day, July 14, the French Republic national holiday, was observed and celebrated this year on land and see by American land and navy forces under special orders islike manner as is observed the American Fourth of July. NOTICE TO CREATIONS THE COLUTY COULT OF THE COLUTY OF COLUTY OF THE COLUTY OF COLUMELA. In the Column of the Column The underlaged laying lease a pmined by the Conserv Using of the sector of the Basis of William Cotent of the Basis of William Coled as resh, notice in interfu gives the column of the Column of the column of the Column of the sector of the Basis of William Cotent of the Basis of William Cotent of the Basis of William Coled as resh, notice in interfu gives the column of the Column of the them with the sport solution meets and first published James to James of the Column of the Column James of the Column of the Sector Restored the Sector of the Sector Restored the Sector of the Column of the Column of the Sector of the Sector Restored the Sector of the Sector Restored the Sector of the Sector of the sector of the Sector of the Sector Restored the Sector of the Sector of the Sector Restored the Sector of the Sector of the Sector Restored the Sector of the Sector of the Sector Restored the Sector of the Sector of the Sector Restored the Sector of the Sector of the Sector Restored the Sector of the Sector of the Sector of the Sector Restored the Sector of the Sector

<text><text><text><text><text><text><text>

Legal Information

It Is Astonishing tively Few T Certain great food s	How Compara-
tively Few T Certain great food s	her Are.
Certain great food s	H12.4 (A.F.C.)
ed thomeolyne within i	Analas have seen
The second	dapies have prov-
en incluserres within i	the age long expe-
rience of humanity to	possess a larger
amount of nutritive v	alue, digestibility
and other good qualit	ies and a smaller
proportion of undesi	rable properties
than any others. The	se, through an ex-
ceedingly slow and e	radual process of
the survival of the f	ttost have come
to form the stanles of	food in come
to round the staples of	rood in common
use by the human r	ace all over the
world. It is really	astonishing how
comparatively few th	tere are of them
when we come to com	sider them broad-
ly-the flesh and the	milk of three or
four domesticated ani	mais, the flesh of
three or four and the	eggs of one spe-
cles of domesticated	birds, three great
grains, wheat, rice an	d maize, and half
a dozen smaller and n	uch less frequent
ones, a hundred or so	enorias of fishes
and shellfish two sne	are a dozon or so
starah containing poot	and tabana anda
two of which the not	ato and the man
two of which, the pot	ato and the man-
loc, are of real into	ernational impor-
tance, twenty or thirt	y fruits, forty or
nity vegetables-make	e up two-thirds of
the food supply of the	he inhabitants of
the world.	
Instead of wonderi	ng at the variety
and profuseness of the	human food sup-
ply the biologist is 1	ather inclined to
ejaculate with the Lo	ndon footman im-
mortalized by John I	Leech, who, when
told by the cook the	t there would be
mutton chops for ding	tor and roast hoot
for support orelained	1. "Nothink been
tor supper, exclaimed	r. Nothink but
beet, mutton and po	rk-pork, mutton
and over: Hin my	opinion, hit's 'igh

Example newspaper article segments cropped using Mask-RCNN model output.

TURESINUIES

Most forest fires are due to ignorance, carelessness or maliciousness. Lookout men, whose sole business, is to detect and report forest fires, live on many of the high peaks of Oregon and Washington, during the dry season and have within their range of vision most of the timberland of these two states.

An Aid to Ambition.

"I shall devote a great deal of time to study for my new career," said the statesman who had been promoted in congressional prominence,

"Do you think that study necessarily prepares a man for the future?" "No. But it helps him to forget the past."—Wasaington Star. The TLEURANCE TABLE The transmission of the product of the the transmission of the theory of the t

Text Analysis Architecture

- Primary Modules:
 - Text Preprocessing
 - Text Analysis
 - Queries and Visualizations
- Submodules implemented as Python Flask microservices
- Human-in-the-loop (HITL) approach to provide user feedback



Document-level Text Analysis

- Identify articles about a given concept
 - Unsupervised approach: automatically **discover** related terms and topics
 - Query expansion with Word2vec embeddings
 - **Topic modeling** with LDA
 - Online/supervised approach: **human-in-the-loop** document classification

Word	fulltext: "delinquency" OR fulltext: "drug addiction" OR fulltext: "drug abuse" OR fulltext: "teen pregnancy" OR fulltext: "alcoholism" OR fulltext: "substance abuse" OR fulltext: "child
juvenile_delinquency	abuse" OR fulltext: "crime prevention" OR fulltext: "prevention" fulltext: "hivaids" OR fulltext: "elder abuse" OR fulltext: "venere
Models	disease" OR fulltext: "hiv aids" OR fulltext: "juvenile crime" OR fulltext: "public awareness" OR fulltext: "homelessness" OR fulltext: "awareness" OR fulltext: "homelessness" OR
1990.w2v	"among negroes" OR fulltext: "occupational hazards" OR fulltext "arony negroes" OR fulltext: "occupational hazards" OR fulltext "oro vontion" OD fulltext: "domestic violence" OD fulltext: "anvi
Number of words	0
10	
GET THE SIMILAR WORDS	
Model: 1990.w2v	
Word: juvenile_delinquency	
delinquency 0.7948	
drug addiction 0.79210	
drug abuse 0.7775	
teen pregnancy 0.7653	_
GENERATE THE QUERY	

LOCAL & AND PROVINCIAL INTELLIGENCE

LOCAL AND PROMINCIAL Det Fair Our annual fair which beam the above captivating this and which by the type has not been as here from det for number of year as it has been on this accession was held on Tuesday last in Chapol ethed Safford OF pigs there was great abundance and they were atmost Rendy given away very free purchasers being turned hack. If they any thing in the shape of money free purchasers being turned hack of both cattle and former but prices were emailed by the purchasers being turned hack of a they are the solution of money free as a first numerous show of facth cattle and formers but prices were emarked by two Of homes the quality was the word we over recollect to have asser to are seen any numerous and imposing depley of aboves for the juvenile class of fair relaterts comprising deaths are garits of every height and emorged after most excellents comprising deaths and the series percey for there was not single drice leager that charged mone for right of her an althout the suppore to accommodate them to the degreewait of the times were atthe these or four initials to faced lables. The series not all how were of one brinks as one of them initial to faced lables. They were not all how proves of one brinks as one of them initial to faced lables. They were not all how prices were of one brinks as one of them initial to faced lables. They were not all how proves of one brinks as one of them initial to faced lables.

Word-level Text Analysis

- Characterize historical semantic dynamics
 - **Term-frequency analysis**: find proportion of documents that use given terms each year
 - Word embedding (Word2vec) analyses:
 - Shico: automatically find and visualize related terms over time
 - Cross-Context Lexical Analysis (CCLA): quantify semantic consistency over time to detect periods of stability or change





Preliminary Experiments

- Recall **Motivating Question**: How has the concept of **juvenile delinquency** been **socially constructed** over time in news media?
 - Start by...
 - **retrieving** news articles describing *juvenile delinquency*
 - detecting when social construction of *juvenile delinquency* changes
- Experiments:
 - **Human-in-the-loop (HITL) Text Classification**: manually annotate news articles that describe *juvenile delinquency*, then train machine learning classifiers to find more
 - Cross-Context Lexical Analysis (CCLA): quantify periods of semantic change in *juvenile delinquency*-related terms, and look for periods of rapid change

HITL Text Classification

	Accuracy	Precision	Recall	F1
Random	0.466±0.05	0.136±0.02	0.38 ±0.12	0.199 ± 0.04
Stratified Random	0.688±0.05	0.175±0.1	0.184 ± 0.1	0.176±0.09
Logistic Regression	0.842±0.04	0.631 ± 0.34	0.276±0.19	0.354 ± 0.23
XGBoost	0.868±0.02	0.754±0.14	0.442±0.12	0.541±0.1

Human-in-the-loop text classification helps identify relevant articles

- Unsupervised search method identified 237/92k articles (1790-1830)
- 19/237 manually annotated as "relevant"
- Trained machine learning models on the annotated dataset
- Optimize Effort vs. Accuracy tradeoff

CCLA (Cross-Context Lexical Analysis)

- How have terms denoting *juvenile delinquency* changed over time?
 - **Aggregate** terms that describe concept of *juvenile delinquency*
 - **Train** Word2vec models (for each decade)
 - **Compute** (normalized) CCLA score for each successive pair of decades
 - Visualize CCLA time series
 - Peak/plateau: (relative) semantic stability
 - Dip: (relative) semantic change



Computational Performance and Scalability (1/2)

- **Downloading Datasets (LOC-CA)**: developed specialized program to download data batches in parallel using the LOC-CA Bulk Data API
 - Used 1 compute node from Pittsburgh Supercomputing Center Bridges-2 (AMD EPYC[™]7742 128 cores and 256GB RAM per node)
 - 128 parallel download tasks with 0.01 seconds sleep time after each HTTP request
 - Full download took ~15.3K core-hours
- Image Analysis: fine-tuned Mask R-CNN for newspaper article segmentation using LOC-CA
 - Used 1 compute node from NCSA Hardware Accelerated Learning (HAL) Cluster (NVIDIA® Tesla® V100 1 GPU, 16 CPU cores per node, 1.2GB RAM per CPU core)
 - Avg. image resolution is 5595 x 7653 px; rescaled to 800 x 1024 px during training
 - Training time \approx 8-10 hours

Computational Performance and Scalability (2/2)

- Text Analysis: preprocessed and analyzed text of ProQuest Historical Newspapers™
 - Used *in-house* Intel Xeon Silver 4210 CPU (*due to licensing restrictions*)
 - Offline operations:
 - Indexed 175 GB data took ~1wk on 2 cores
 - Trained LDA topic model in ~2d (on 8 cores), word embedding models in ~2d (on 4 cores), and text classifiers in ≤30min (on 1 core)
 - Online operations:
 - Text classifiers and CCLA operate almost instantaneously (<10ms on 1 core)
 - Kibana or ShiCo can take up to 4 minutes for some complex queries (e.g. containing many terms or search operators)

Conclusion & Future Work

- Presented the **design**, **implementation**, and **preliminary evaluation** of a big data analysis system for historical newspaper collections
- Applied AI analysis tools to historical newspapers to enable **investigation of novel social science research questions**
- Remaining Challenges:
 - **Support many users**: improve efficiency and parallelization of online analysis operators to support queries, visualizations, and feedback from many users simultaneously
 - **Integrate system**: unify all modules into single system, use results of downstream analyses to improve upstream preprocessing/analysis modules, etc.
 - **Improve robustness:** adapt term-level text analysis operators (e.g. ShiCo, CCLA) to be less sensitive to imperfect inputs (e.g. errors in OCR and word segmentation)
 - **Study social construction**: apply our system to study the historical social construction of *juvenile delinquency*

Acknowledgements

- Extreme Science and Engineering Discovery Environment (XSEDE) and Bridges System at Pittsburgh Supercomputing Center - Primary data storage and compute resources
- **XSEDE Extended Collaborative Support Service (ECSS)** Research software development and consulting expertise
- Illinois Campus Cluster Additional data storage and compute resources
- Hardware Accelerated Cluster (HAL) at NCSA Deep Learning model training and evaluation
- Cline Center for Advanced Social Research and University of Illinois at Urbana-Champaign Access to ProQuest Historical Newspapers™
- **Benjamin Meier** Correspondence regarding newspaper article segmentation

Questions?

Feel free to contact {sandeeps, bhavya2, adavies4}@illinois.edu



