A Historical Big Data Analysis To Study The Social Construction Of Juvenile Delinquency - Latest Progress

Presenters: Yu Zhang, Sandeep Puthanveetil Satheesan, Bhavya, and Adam Davies

XSEDE ECSS Symposium - May 17, 2022
Project Team

Yu Zhang, Ph.D., Assistant Professor, Dept. of Criminology, California State University, Fresno,

ChengXiang ("Cheng") Zhai, Ph.D., Donald Biggar Willett Professor in Engineering, Dept. of Computer Science, University of Illinois

Alan B. Craig, Ph.D., XSEDE

Sandeep Puthanveetil Satheesan, M.S., Senior Research Programmer, NCSA, University of Illinois

Bhavya, Ph.D. Candidate, Dept. of Computer Science, University of Illinois

Adam Davies, Ph.D. Candidate, Dept. of Computer Science, University of Illinois

Matthew Jin, Undergraduate Student, Dept. of Computer Science, University of Illinois
Introduction

1. **Why Social constructionism?** It contrasts the *external world* with the human created or constructed world.

2. **Human constructed world is constituted by “mental facts”** (Searle & Willis, 1995)
   a. “purely mental facts” (e.g., intention, feeling, desire, happiness, joy, etc.);
   b. “social facts” (e.g., friendship, violence, custom, etc.)
   c. “institutional facts” (e.g., money, language, religion, empire, democracy, leadership, etc.).

3. **Crime/Juvenile delinquency are “social facts”**.

   Studying crime from the perspective of social constructionism provides a historical and contextual view.

1. **Why “Juvenile Delinquency”?**

   *Knowing these historical and contextual factors, we may have a comprehensive understanding of this criminal behavior, and helps us to predict the future.*

Challenge and Project Vision

● Challenge
  ○ Humanly impossible to go through millions of historical newspaper files to find answers to broad social science research questions

● Vision
  ○ Use historical newspaper collections and image + text analysis methods to address novel social science research questions
  ○ Design and develop a generic big data analysis system for social science research with a focus on juvenile delinquency
Progress

- Work started on this project at the end of 2017/early 2018

- Major project phases/modules
  - Data Collection & Storage
  - Image Preprocessing & Analysis
  - Text Preprocessing & Analysis
  - Query and Visualization

- Major milestones during the first two years:
  - Started gathering and analyzing data from the Library of Congress Chronicling America (LOC-CA) newspaper collection
  - Identified some limitations of available OCR text files - no article segmentation, OCR text quality was not great
  - Worked on improving OCR quality, identifying article boundaries as well as identified software for getting better text
  - Gave our first ECSS Symposium talk, May 2018
  - Got access to new collection - ProQuest Historical Newspapers™
  - Published a poster paper at the IEEE eScience conference, October 2019
Data Collection

- Library of Congress Chronicling America (LOC-CA)
  - Historical American newspapers in the public domain from 1777 to 1963 - contains more than 19.6 million pages of digitized U.S. newspapers
  - Downloaded 18.4 million text files and 500,000 scanned newspaper images using LOC-CA bulk API
  - 1 compute node from PSC Bridges-2 - Used about 15.3K SUs

- ProQuest Historical Newspapers™
  - Digital archive offering full-text and full-image articles for newspapers dating back to the 18th Century
  - Contains 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

- Focused on newspaper segmentation of the LOC-CA collection and re-OCR
- Transfer learning using Mask-RCNN model pre-trained on MS COCO
- Training Set
  - 60 scanned newspaper images
  - 1728 manually annotated regions
- Validation Set
  - 25 images
  - 801 manually annotated regions
- Used 1 compute node from NCSA HAL Cluster (NVIDIA® Tesla® V100 - 1 GPU, 16 CPU cores per node, and 1.2GB RAM per CPU core)
- Training time ≈ 8-10 hours
Image Analysis: Newspaper Segmentation

Output of Mask-RCNN model on some example newspaper images showing different segmented regions.
The hazel hoe is considered by forest-fighters as the best single tool for digging fire line or trench. It is a strong broad-bladed hoe suitable for cutting weeds and small brush as well as digging in the soil.

American soldiers engaged in the French forests write, “there have been no fires in these forests for years.”

FIRE! FIRE! FIRE! FIRE! See E. E. Quick, St. Helens, and insure your property in the Oregon Fire Relief Association of McMinnville, Oregon. Don’t put it off.

Almost entire world now engaged in war. Of the 1,000,000,000 people who subscribe the war, 15,000,000 are in the United States, a fact which is significant.

Aid to Ambition.

“Terror of the forest is to detect and report forest fires, live on many of the highest peaks of Oregon and Washington, during the dry season and have within their range of vision most of the timberland of these two states.

Too Much.

“It’s hard to believe that she’s as intelligent 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 it being color-blind.” — The Catholic Standard and Times.

The Limit of Speed.

The fool and his money are so curtained, but for quick operation W. I. Warren holds the record.—Detroit Free Press.

NOTICE TO WOOD CONTRACT Bids will be received by the Board of Commissioners of Columbia County, on or before August 20, at 2 p.m., at which time they will be opened. For 200 cords of firewood,乱 or cordwood, delivered at the courthouse in St. Helens, Oregon. The right is reserved to reject and all bids.

NOTICE TO CHIPPERS CORDWOOD CUTTER OF THE COUNTY OF COLUMBIA: The Board of Commissioners of Columbia County, Oregon, will receive bids on the 20th day of August, 1920, at 2 o'clock p.m., for 200 cords of cordwood to be delivered at the courthouse in St. Helens, Oregon. The right is reserved to reject and all bids.
Image Analysis: Current + Next Tasks

- Currently applying Tesseract on segmented newspaper images to evaluate OCR output quality
- Currently working on getting the newspaper segmentation program run on Bridges 2
- Complete newspaper segmentation of the LOC-CA collection
- Perform re-OCR of segmented newspaper images using Tesseract or Google Cloud Vision API
Text Analysis Goals

- Semantic search of historical articles about “juvenile delinquency”
  - *Human-in-the-loop* retrieval
  - Topic Modeling

- Analyze changes in meaning and usage of “juvenile delinquency” terms over time
  - Shico, CCLA
Kibana Search

Proportion of articles

Year

“juvenile”

“juvenile delinquency”
Human-in-the-loop Retrieval

Query expansion + Kibana Search

Text Classification

Human Annotation
Topic Modeling

- Leveraged a probabilistic model: LDA

Sample “crime”-related topics

Sentence 1: ... The chairman charged the grand jury as follows, ... to the grand jury that the number of prisoners...

Topic Coverage: 98%

Shico

- Visualizes changes in neighbors of query terms (i.e., “juvenile”, “delinquency”)
- Based on learned Word2vec representations: words that share common contexts have similar representations
  - Can also mine related terms for downstream tasks

CCLA: Approach

- Goal: quantify, visualize change in historical meaning and use of “juvenile delinquency” terms (in news media)
- Setup
  - Split ProQuest by time period (data-sparse 1790-1830, then each subsequent decade)
  - Word2vec embeddings (inc. phrases)
- Cross-Context Lexical Analysis (CCLA)
  - Nearest neighbors → “know a word by the company it keeps”
  - How similar are the nearest neighbors between decades?

CCLA: Results

<table>
<thead>
<tr>
<th></th>
<th>1840s</th>
<th>1890s</th>
<th>1910s</th>
</tr>
</thead>
<tbody>
<tr>
<td>criminals</td>
<td></td>
<td>reformatory</td>
<td>defectives</td>
</tr>
<tr>
<td>vagrancy</td>
<td></td>
<td>hardened</td>
<td>criminals</td>
</tr>
<tr>
<td>prisons</td>
<td></td>
<td>criminals</td>
<td>truancy</td>
</tr>
<tr>
<td>gaols</td>
<td></td>
<td>reformatories</td>
<td>reformatories</td>
</tr>
<tr>
<td>separate confinement</td>
<td></td>
<td>incorrigibles</td>
<td>offenders</td>
</tr>
</tbody>
</table>
Remaining Work

1. Retrieve large collection $D$ of documents covering juvenile behavior (including but not limited to delinquency) across all time periods

2. Run topic models on $D$ to discover $k$ topics (“behaviors”) across time periods
   a. Identify topics corresponding to group-level social construction (e.g. advantaged, contenders, dependents, deviants)

3. Analyze relationship between aggregated “concept” embedding and each social construction topic

Takeaway: flexible, general-purpose “big data” image and text analysis toolkit for social science research, with a specific focus on social construction and juvenile delinquency
Conclusion

- Highly interdisciplinary project that combines large collections of image and text data to study the social construction of juvenile delinquency
- Data Collection: Obtained data from two historical newspaper collections and performed different data processing techniques
- Image Analysis: Developed methods for generating segmented newspaper data and working towards better OCR generated text
- Text Analysis: Developed methods for performing semantic search of historical articles about juvenile delinquency, explored and developed tools to analyze changes in the meaning of juvenile delinquency-related terms
- Query and Visualization: Developed query capabilities and generated basic visualizations based on the Text Analysis results and working towards answering specific social science questions
Additional References

- https://github.com/sandeep-ps/Mask_RCNN/tree/master/samples/newspaper
- Stock images (Microsoft Office) and images from Google Slides Search
Acknowledgement

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC) through allocation TG-SES170016.

We would like to thank the XSEDE Extended Collaborative Support Service (ECSS) program.

This work also made use of the Illinois Campus Cluster, a computing resource that is operated by the Illinois Campus Cluster Program (ICCP) in conjunction with the National Center for Supercomputing Applications (NCSA) and which is supported by funds from the University of Illinois at Urbana-Champaign.

This work also utilized resources supported by the National Science Foundation’s Major Research Instrumentation Program, grant number 1725729, as well as the University of Illinois at Urbana-Champaign.

We would like to thank the Cline Center for Advanced Social Research and the University of Illinois at Urbana-Champaign for providing access to ProQuest Historical Newspapers™.
Thank you!

Questions?

yuzh@csufresno.edu, {sandeeps, bhavya2, adavies4}@illinois.edu

Upcoming Publication: