From Chronicling America to Newspaper Navigator: Improving Access to Historic Newspaper Photos at the Library of Congress through Machine Learning

Benjamin Charles Germain Lee & Nathan Yarasavage

chroniclingamerica.loc.gov
news-navigator.labs.loc.gov
National Digital Newspaper Program (NDNP)

https://www.loc.gov/ndnp

- United States newspaper digitization pre-1963
- Collaboration since 2005
  - 50 state partners
  - Library of Congress (LC)
  - National Endowment for the Humanities (NEH)
- Shared technical specification
- Open access
Chronicling America

https://chroniclingamerica.loc.gov/

- 17+ million pages
- 3,000+ newspaper titles
- Content coverage:
  - 48 states, 2 territories, and D.C.
  - 23 languages
  - 1777-1963
Chronicling America Data Visualizations

https://www.loc.gov/ndnp/data-visualizations/
Chronicling America Data Access

https://chroniclingamerica.loc.gov/

- Open access / public domain
- API & bulk data access
- Metadata & search results from site can also be returned in JSON format

Advancing Research and Scholarship

- Projects using Chronicling America data: https://www.loc.gov/ndnp/extras/
Choose the best transcription or enter your own!

Original prompt: Enter picture's caption, illustrator or photographer and category.

<table>
<thead>
<tr>
<th>Caption</th>
<th>Doings of the Duffs—A Flover Will Pass Anything on the Road.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artist</td>
<td>W. R. Allman</td>
</tr>
<tr>
<td>Category</td>
<td>Comics/Cartoon</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Caption</th>
<th>Doings of the Duffs—A Flover Will Pass Anything on the Road.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artist</td>
<td>F. W. Allman</td>
</tr>
<tr>
<td>Category</td>
<td>Comics/Cartoon</td>
</tr>
</tbody>
</table>
Can we “throw open the treasure chest” by training a machine learning algorithm to process all of Chronicling America?
• facilitate research in **computer science**, **library science**, and the **digital humanities**
- facilitate research in **computer science**, **library science**, and the **digital humanities**
- Engage the public and enable a wide range of **creative computing** and **public humanities** projects
NAZI COUNTERATTACKS REPULSED NEAR CAEN

Allied Infantry Pushes 10 Miles North of Rome. Air and Sea Aid Advance. German Scout Plane Smashed in D-Day Ton.

Allies Clear Enemy From Landing Areas, Battle Raging Inland. Beachheads Extended Over 50 Miles; Both Sides Confuse in One's Vision.

1,000,000 People in Captivity. German Troops Massed. US Troops on the Move. American Flag Nears North Coast.

Highways Open. Sailors Aren't Quiescent. Air Raids on London. USAAF Bombs German Factories.

New National Cemetery in D.C. for 50,000 is Sought by Array.

Viewing Room Open. US Troops on the Move. American Flag Nears North Coast.

Highways Open. Sailors Aren't Quiescent. Air Raids on London. USAAF Bombs German Factories.
7 Different categories:

- Photographs
- Illustrations
- Comics
- Editorial Cartoons
- Maps
- Headlines
- Advertisements
Plus Captions!
Faster-RCNN
Running the Pipeline
Running the Pipeline

- 16,368,041 Chronicling America pages processed in total
Running the Pipeline

- 16,368,041 *Chronicling America* pages processed in total
  - ~100 TB of image & XML data
Running the Pipeline

- 16,368,041 *Chronicling America* pages processed in total
  - ~100 TB of image & XML data
- ~19 days on 2 g4dn.12xlarge Amazon AWS EC2 instances
Running the Pipeline

- 16,368,041 *Chronicling America* pages processed in total
  - ~100 TB of image & XML data
- ~19 days on 2 g4dn.12xlarge Amazon AWS EC2 instances
  - 96 vCPUs & 8 NVIDIA T4 GPUs
Running the Pipeline

- 16,368,041 *Chronicling America* pages processed in total
  - ~100 TB of image & XML data
- ~19 days on 2 g4dn.12xlarge Amazon AWS EC2 instances
  - 96 vCPUs & 8 NVIDIA T4 GPUs
- All pipeline code in Python, PyTorch
Running the Pipeline

- 16,368,041 *Chronicling America* pages processed in total
  - ~100 TB of image & XML data
- ~19 days on 2 g4dn.12xlarge Amazon AWS EC2 instances
  - 96 vCPUs & 8 NVIDIA T4 GPUs
- All pipeline code in Python, PyTorch
- notebooks/process_chronam_pages.ipynb in Newspaper Navigator Repo
Newspaper Navigator

Beyond Words

Common Objects in Context

QR code to Github repo:

news-navigator.labs.loc.gov

#NewspaperNavigator
- HTTPS, S3 requests
- Pre-packaged datasets
QR code to Github repo: news-navigator.labs.loc.gov

#NewspaperNavigator
All in the public domain!

[QR code to Github repo: news-navigator.labs.loc.gov #NewspaperNavigator]
Data Archaeology

Iowa state bystander.
(Des Moines, Iowa)
14 Oct. 1910

Franklin's paper the statesman.
(Denver, Colorado)
15 Oct. 1910

The broad ax.
(Salt Lake City, Utah)
15 Oct. 1910

The broad ax.
(Salt Lake City, Utah)
26 Nov. 1910
Data Archaeology

Iowa state bystander. (Des Moines, Iowa) 14 Oct. 1910
Franklin's paper the statesman. (Denver, Colorado) 15 Oct. 1910
The broad ax. (Salt Lake City, Utah) 15 Oct. 1910
The broad ax. (Salt Lake City, Utah) 26 Nov. 1910

http://dx.doi.org/10.17613/k9gt-6685
Newspaper Navigator Search Application
Newspaper Navigator Search Application

- A search user interface for 1.5 million photos from the Newspaper Navigator dataset
Newspaper Navigator Search Application

- A search user interface for 1.5 million photos from the Newspaper Navigator dataset
- Empowers users to train AI navigators to retrieve visually similar content
Newspaper Navigator Search Application

- A search user interface for 1.5 million photos from the Newspaper Navigator dataset
- Empowers users to train AI navigators to retrieve visually similar content
- Exposes the machine learning training to the user
Newspaper Navigator Search Application

- A search user interface for 1.5 million photos from the Newspaper Navigator dataset
- Empowers users to train AI navigators to retrieve visually similar content
- Exposes the machine learning training to the user
- Training and predicting on all 1.5 million photos takes just a couple of seconds
Search with Newspaper Navigator!

Search for: dog

displaying 1 - 100 records in total 5199

Update Reset Filters

Sorting by: date, oldest to newest

Your AI Navigators
None defined yet
Select an AI navigator:

(+) The AI navigator will try to find photos like the ones selected below!

sailboats

+ New AI Navigator

Train Your AI Navigator!

(-) The AI navigator will try to avoid photos like the ones selected below!

sailboats

Below, your AI navigator has found visually similar photos based on the selections on the left. You can re-train your AI navigator by adjusting your selections on the left and making new selections below.
Newspaper Navigator + the Digital Humanities
Q1: How can Newspaper Navigator help us “see” editors?

In collaboration with: Jim Casey, Molly Hardy, Joshua Ortiz Baco, Sarah Salter
A Heat Map of Advertisements on the Front Pages of *The Opelousas Courier, 1870-1909*
Q2: How can Newspaper Navigator help us study Ladino newspapers?

In collaboration with: Devin Naar
כמנים ומכירים אלינו אנונימים
לאקוניות אינפרנים
אני לא אנונימי או מ的带领, שא
przedsiębiorst'וים מאה קאזה, פדריה
נוד שלארנסקי ומרית און טופר, מת
מאבק בעור ואון טופר, ד"ר"א פאולא
(לאקוניות, תומס) ד"ר אראס ביד
בירבח, ק"הינידה" ממק"ש אצ"פ
די פאפאיל על גנרל ד"ר פ"אפא
ר פאפול ואון טופר, פ"א פאפול
ענדה בירבח, ענדה בירבח, ענדה
איבושר ממעד
ל Vara, 1930
<table>
<thead>
<tr>
<th>Newspaper Title</th>
<th>Number of Images</th>
<th>Number of Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>El Instruktör revista siyentifika i literaria</td>
<td>331</td>
<td>331</td>
</tr>
<tr>
<td>El jugeton, Jurnal umoristiko</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>El Kirbatch Americano (1915 - 1917)</td>
<td>206</td>
<td>206</td>
</tr>
<tr>
<td>El Luzero Sefaradi (October, 1926)</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>El Luzero Sefaradi (May, 1927)</td>
<td>28</td>
<td>28</td>
</tr>
<tr>
<td>El Luzero Sefaradi (Total)</td>
<td>56</td>
<td>56</td>
</tr>
<tr>
<td>El Progresso / Yosef Daat</td>
<td>332</td>
<td>332</td>
</tr>
<tr>
<td>El Tiempo (1890-1891)</td>
<td>926</td>
<td>926</td>
</tr>
<tr>
<td>El Tiempo (1896-1897)</td>
<td>1,138</td>
<td>1,138</td>
</tr>
<tr>
<td>El Tiempo (1900)</td>
<td>584</td>
<td>584</td>
</tr>
<tr>
<td>El Tiempo (1900-1901)</td>
<td>981</td>
<td>981</td>
</tr>
<tr>
<td>El Tiempo (Total)</td>
<td>3,629</td>
<td>3,629</td>
</tr>
<tr>
<td>La boz de Oriente (April, 1931 - April, 1932)</td>
<td>860</td>
<td>860</td>
</tr>
<tr>
<td>La Vara (1922 - 1941)</td>
<td>5,071</td>
<td>10,142</td>
</tr>
<tr>
<td>Total</td>
<td>10,489</td>
<td>15,560</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Visual Content Type</th>
<th># &gt; 90%</th>
<th># &gt; 70%</th>
<th># &gt; 50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photographs</td>
<td>348</td>
<td>770</td>
<td>1060</td>
</tr>
<tr>
<td>Illustrations</td>
<td>52</td>
<td>378</td>
<td>960</td>
</tr>
<tr>
<td>Maps</td>
<td>27</td>
<td>182</td>
<td>300</td>
</tr>
<tr>
<td>Comics</td>
<td>10</td>
<td>39</td>
<td>184</td>
</tr>
<tr>
<td>Editorial Cartoons</td>
<td>8</td>
<td>31</td>
<td>111</td>
</tr>
<tr>
<td>Advertisements</td>
<td>18,381</td>
<td>31,523</td>
<td>42,505</td>
</tr>
</tbody>
</table>
T-SNE embedding visualization of visual content with confidence scores > 90% (2,518 images in total)
Acknowledgments

BCGL would like to thank:

- Jaime Mears, Eileen Jakeway, Meghan Ferriter, Laurie Allen, Abbey Potter, Leah Weinryb Grohsgal, Kate Zwaard, Deb Thomas, Nathan Yarasavage, Robin Butterhof, Chris Adams, Tong Wang, John Foley, Megan Halsband, Brian Foo, Mark Sweeney, the entire NDNP staff at the Library of Congress, and everyone who has made *Chronicling America* + *Beyond Words* possible

- Dan Weld at the University of Washington

- Gabe Pizzorno (Harvard), Michael Haley Goldman (USHMM), Robert Ehrenreich (USHMM), Jim Casey (Penn State), Molly Hardy (NEH), Joshua Ortiz Baco (UT Austin), Sarah Salter (TAMUCC), Devin Naar (University of Washington)
Acknowledgments

This research is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-1762114, the Library of Congress Innovator in Residence position, and the Richard Willner Memorial Fellowship as part of the University of Washington’s Jewish Studies Graduate Research Fellowship program.
Feel free to reach out!
bcgl@cs.washington.edu
Questions?