Newspaper Navigator: Hosting the Dataset and Deploying the Search Application

Chris Adams & Benjamin Charles Germain Lee

news-navigator.labs.loc.gov
Can we “throw open the treasure chest” by training a machine learning algorithm to process all of Chronicling America?
I. The Newspaper Navigator Dataset

news-navigator.labs.loc.gov
Running the Pipeline

- 16,368,041 *Chronicling America* pages processed (99.998%)
  - ~100 TB of image & XML data
- ~19 days on 2 g4dn.12xlarge Amazon AWS EC2 instances
  - 96 vCPUs & 8 NVIDIA T4 GPUs

x 16.3 million pages
1) How do we publicly host the dataset and make it accessible to a wide audience?

2) How do we do so without incurring unpredictable costs?
2 primary forms of access:

- For computation: S3 requests under the “requester pays” model
- For public use: prepared datasets (no coding or S3 necessary)
II. The Newspaper Navigator Search Application

news-navigator.labs.loc.gov/search
Newspaper Navigator Search Application

- A search user interface for 1.5 million photos from the Newspaper Navigator dataset
- Empowers users to train open facets ("AI navigators") to retrieve visually similar content
- Exposes the machine learning training to the user
- Training & predicting on all 1.5 million photos is nearly instantaneous
Select an AI navigator:

baseball players

Save

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

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

baseball players

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.

https://news-navigator.labs.loc.gov/search
Challenges Surrounding the Search Application’s Deployment

- How do we create a sustainable solution for deploying the search application with low maintenance requirements?
- How do we handle issues surrounding load at launch without overpaying the rest of the time?
- How do we ensure that the interactive machine learning component is fast and is not a challenge for upkeep?
Solution: Containerization & AWS Fargate!

- Containers are designed to be ephemeral, with no state to preserve except where it’s explicitly configured.
- Automated container build process (see GitHub repo), making it easy to ship security updates and for new people to use the project.
- We deployed the container on AWS ECS Fargate: no host servers or configuration to manage, auto-scaling policy adds and removes instances based on load or failures, health checks avoid accidentally deploying a broken build.
18 Months since Deployment...

- Running continuously for a year and a half with virtually no issues!
- Supported tens of thousands of user sessions to date
- Engagement from scholars, teachers, students, genealogists, and members of the American public
- More usage statistics are forthcoming in a *Signal* blog post with LC Labs!
Newspaper Navigator Resources

- Dataset: https://news-navigator.labs.loc.gov/
- Search Application: https://news-navigator.labs.loc.gov/search
- GitHub repository: https://github.com/LibraryOfCongress/newspaper-navigator
We 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, 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
Acknowledgments

- This research is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant DGE-1762114 and the Library of Congress Innovator in Residence position.
Questions?