Text analysis, data requests, and the academic library

IASSIST/CARTO 2024

Who are we?

David Beales
rdb104@case.edu

Digital Scholarship Partner, Digital Scholarship
Case Western Reserve University

Jen Ferguson
j.ferguson@northeastern.edu

Head, Research Data Services
Northeastern University

Amy Kirchhoff
amy.kirchhoff@ithaka.org

Senior Manager
Constellate

David Lowe
davelowe@tamu.edu

Assistant Professor, Global Languages and Cultures
Texas A&M University

Todd Suomela
todd.suomela@bucknell.edu

Associate Director, Digital Pedagogy and Scholarship
Bucknell University

Agenda for our hour

What is text analysis?	5 minutes
Case Western Reserve University experiences	10 minutes
Northeastern University experiences	10 minutes
Texas A&M University experiences	10 minutes
Bucknell University experiences	10 minutes
Questions	15 minutes

What is text analysis?

Amy Kirchhoff

Senior Manager

Constellate

Amy.Kirchhoff@ithaka.org

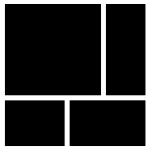
Amy Kirchhoff
Senior Manager
Constellate
Amy.Kirchhoff@ithaka.org

ITHAKA

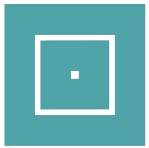
We are a not-for-profit with a mission to improve access to knowledge and education for people around the world. We believe education is key to the well-being of individuals and society, and we work to make it more effective and affordable.

JSTOR

CONSTELLATE



ARTSTOR



PORTICO

ITHAKA S+R

CONSTELLATE

Constellate is the only text and data analysis platform that integrates access to scholarly content and open educational resources into a cloud-based application and lab to help faculty, librarians, and other instructors easily teach text and data analysis.

With Constellate, learners across all disciplines can apply text analysis methods to content, hone their skills with our on-demand tutorials, attend live classes taught by the experts at Constellate, and engage with our inspiring user community.

Text analysis is the practice of extracting information from collections of text to discover new ideas or answer research questions (these collections can be small or large).

In text analysis, natural language processing and machine learning algorithms are used to classify, sort, and compile data to identify patterns, relationships, sentiments, and create new knowledge.

Text and data analysis is a core competency within data literacy as a whole and it is connected to digital humanities, data mining, data analytics, and big data.

youtubekids.com

Apps Bookmarks TDM I P Art Me

baby

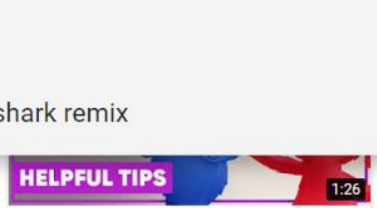
baby shark

baby songs

baby

baby shark remix

Lucas the Spider - It's Hot



What To Do When You're Afraid | The Monster at the End of This...

[All Episodes] Baby Shark's Big Show! | +Compilation |...

Ball Pit Party | Kids Song for Learning Colors - Giant Ball Pit...

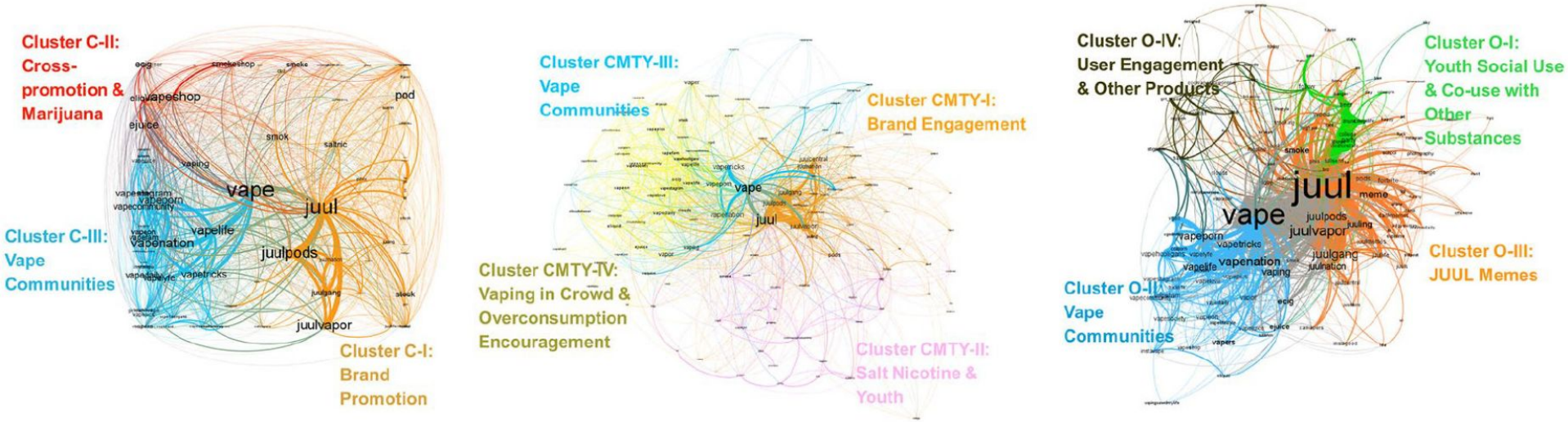
The Itsy Bitsy Spider

Sesame Street Monster Meditation #2: Goodnight Body...

Baby Shark | Baby Shark

> A total of 701,891 tweets have been retrieved and included in the daily sentiment analysis. The sentiment regarding Pfizer and Moderna vaccines appeared positive and stable throughout the four months, with no significant differences in sentiment between the months. In contrast, the sentiment regarding the AstraZeneca/Oxford vaccine seems to be decreasing over time, with a significant decrease when comparing December with March.

Robert Marcec, Robert Likic, Using Twitter for sentiment analysis towards AstraZeneca/Oxford, Pfizer/BioNTech and Moderna COVID-19 vaccines, *Postgraduate Medical Journal*, Volume 98, Issue 1161, July 2022, Pages 544–550, https://doi.org/10.1136/postgradmedj-2021-140685



Kostygina, Ganna, Miao Feng, Lauren Czaplicki, Hy Tran, Shreya Tulsiani, Siobhan N. Perks, Sherry Emery, and Barbara Schillo. “Exploring the Discursive Function of Hashtags: A Semantic Network Analysis of JUUL-Related Instagram Messages.” Social Media + Society, (October 2021). https://doi.org/10.1177/20563051211055442.

Decision Support Systems
Volume 152, January 2022, 113651

A text-mining based cyber-risk assessment and mitigation framework for critical analysis of online hacker forums

Baidyanath Biswas [a], Arunabha Mukhopadhyay [b], Sudip Bhattacharjee [c], Ajay Kumar [d], Dursun Delen [e, f]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.dss.2021.113651

Highlights

- Novel text-mining based cyber-risk assessment and mitigation framework.
- Identify hacker expertise using explicit and implicit features on online forums.
- Expert hackers demonstrate leadership in online forums.
- Compute financial impact for every {hacker expertise, attack-type} combination.
- Prioritize hacker mitigation strategies.

Springer Link

Research Article | Published: 13 January 2022

Text visualization for geological hazard documents via text mining and natural language processing

Ying Ma, Zhong Xie, Gang Li, Kai Ma, Zhen Huang, Qinjun Qiu ✉ & Hui Liu

Earth Science Informatics (2022) | Cite this article

Abstract

An increasing number of geological hazard documents about the mechanism and occurrence process of geological disasters contain unstructured geoscientific data that are not fully utilized. Text mining and visualization techniques offer opportunities to leverage this wealth of data and extract valuable information from dense, abstract geological disaster reports to quickly focus on the core information in geological reports and improve the efficiency of report usage. In this research, a flow framework for the automatic extraction of key information and its transformation to a simple and intuitive form for managers/researchers to quickly navigate, understand and make more informed decisions based on the key information are described. To automatically extract key information from text, an optimized term frequency-

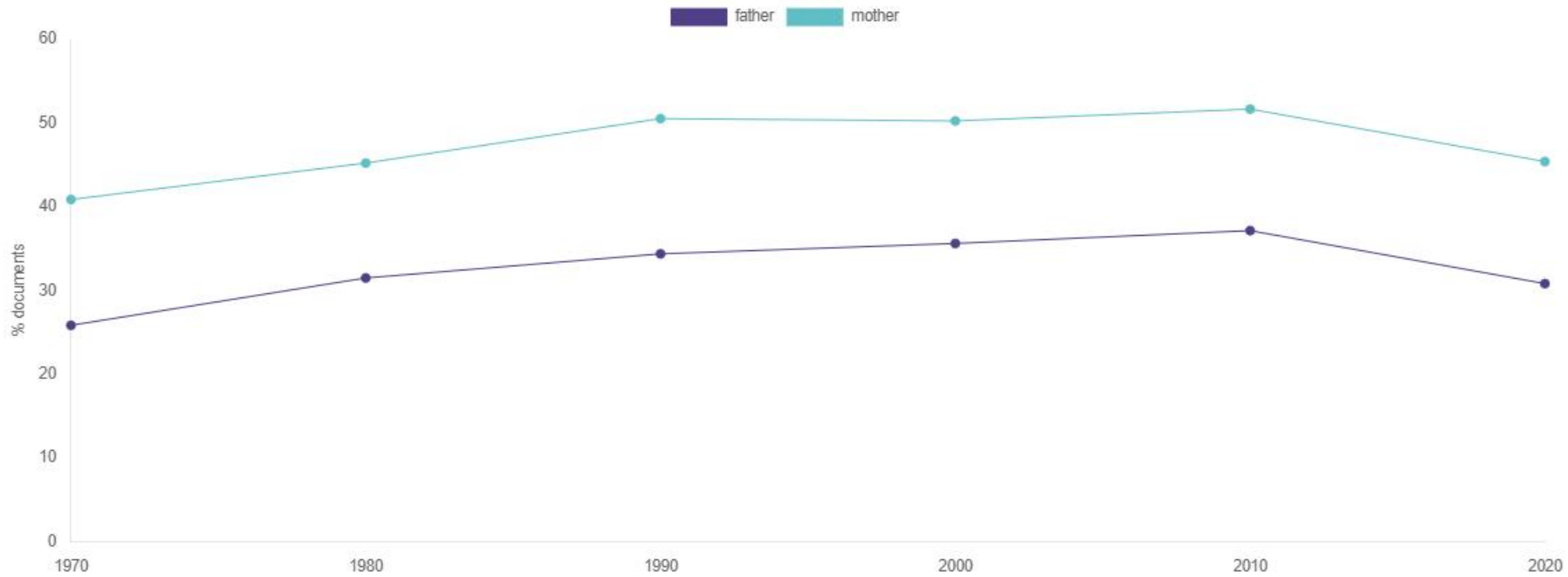
Integration of Automatic Text Mining and Genomic and Proteomic Analysis to Unravel Prostate Cancer Biomarkers

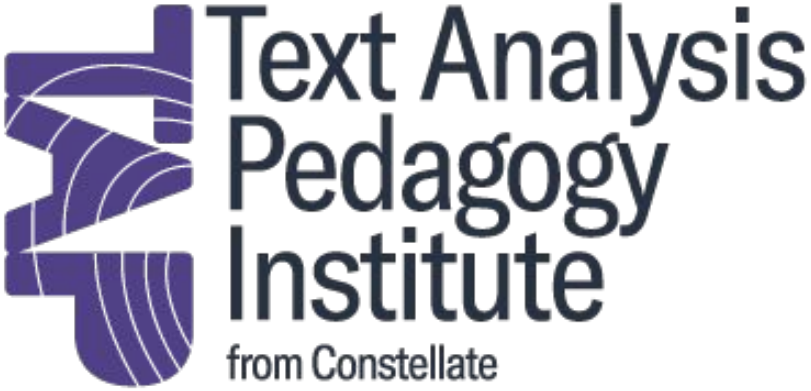
Tânia Lima, Rita Ferreira, Marina Freitas, Rui Henrique, Rui Vitorino, and Margarida Fardilha*

Cite this: *J. Proteome Res.* 2022, 21, 2, 447−458
Publication Date: January 4, 2022
https://doi.org/10.1021/acs.jproteome.1c00763

CONSTELLATE

"foster care" limited to document type(s) article, chapter, book, research_report from 1970 - 2024 - (45,773 total documents)





https://constellate.org/tap-institute

Case Western Reserve University

David Beales

rdb104@case.edu

Digital Scholarship Partner
Kelvin Smith Library
Case Western Reserve University

The Freedman Center for Digital Scholarship

KELVIN SMITH LIBRARY
CASE WESTERN RESERVE UNIVERSITY

My Path to Digital Scholarship

University Photographer

+

Digital Asset Management

=

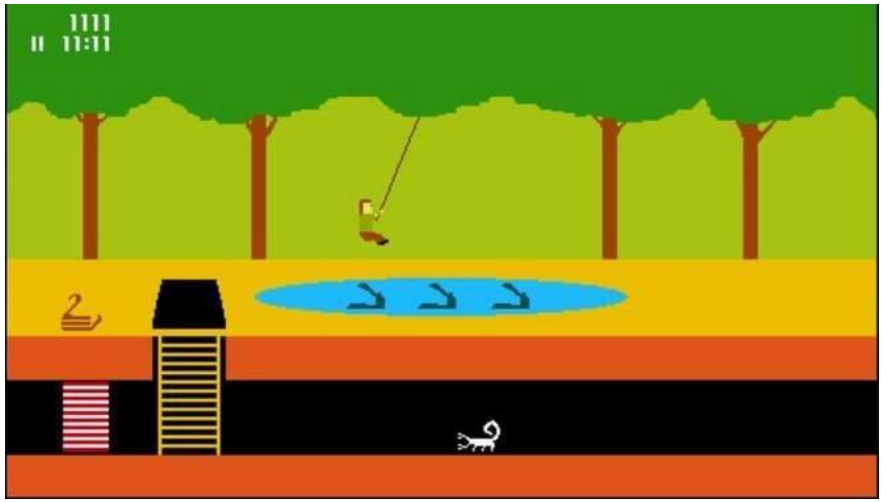
Libraries!

KELVIN SMITH LIBRARY
CASE WESTERN RESERVE UNIVERSITY

Digital Scholarship at KSL

Challenges!

Boons!



KELVIN SMITH LIBRARY
CASE WESTERN RESERVE UNIVERSITY

Outreach and Exploration

Workshops ->

Project Support ->

More Individual/Group Instruction ->

More projects!

KELVIN SMITH LIBRARY
CASE WESTERN RESERVE UNIVERSITY

Where am I now?

Known as the Text Analysis person?

LLM Hype is providing opportunities.

New team lead means new outreach opportunities

KELVIN SMITH LIBRARY

CASE WESTERN RESERVE UNIVERSITY

So...

Separate your text analysis support from subject liaison work.

Build an interdisciplinary learning community to support diverse need.

KELVIN SMITH LIBRARY

CASE WESTERN RESERVE UNIVERSITY

Northeastern University

Northeastern University
Library

Research Data Services

We Help People Use Data!

We can help you analyze, map, visualize, and manage data. We can guide you to create compelling projects for your courses and research. We work with Northeastern faculty, staff, and students at all levels of experience and in any location.

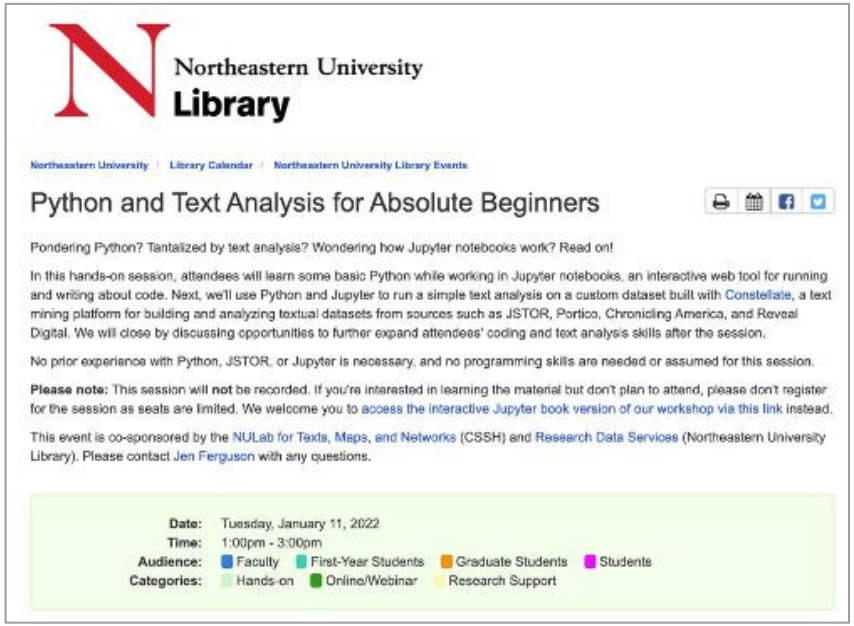
We provide:

- Consultations with specialists online or in person
- Visits to classes, student or research groups, guest lectures, and workshops
- Online tutorials and resources

Contact: library-rds@northeastern.edu.

RDS supports data analysis, data management, data visualization, GIS, & text analysis.

1. Interest in text analysis is robust.



Northeastern University Library

Northeastern University / Library Calendar / Northeastern University Library Events

Python and Text Analysis for Absolute Beginners

Pondering Python? Tantalized by text analysis? Wondering how Jupyter notebooks work? Read on!

In this hands-on session, attendees will learn some basic Python while working in Jupyter notebooks, an interactive web tool for running and writing about code. Next, we'll use Python and Jupyter to run a simple text analysis on a custom dataset built with Constellate, a text mining platform for building and analyzing textual datasets from sources such as JSTOR, Portico, Chronicling America, and Reveal Digital. We will close by discussing opportunities to further expand attendees' coding and text analysis skills after the session.

No prior experience with Python, JSTOR, or Jupyter is necessary, and no programming skills are needed or assumed for this session.

Please note: This session will **not** be recorded. If you're interested in learning the material but don't plan to attend, please don't register for the session as seats are limited. We welcome you to access the interactive Jupyter book version of our workshop via this link instead.

This event is co-sponsored by the NULab for Texts, Maps, and Networks (CSSH) and Research Data Services (Northeastern University Library). Please contact Jen Ferguson with any questions.

Date: Tuesday, January 11, 2022
Time: 1:00pm - 3:00pm
Audience: Faculty, First-Year Students, Graduate Students, Students
Categories: Hands-on, Online/Webinar, Research Support

To date:

300 attended
500 turned away

2. Interest in text analysis spans disciplines and roles.

Are published recipes becoming more diverse over time?

Art & Design

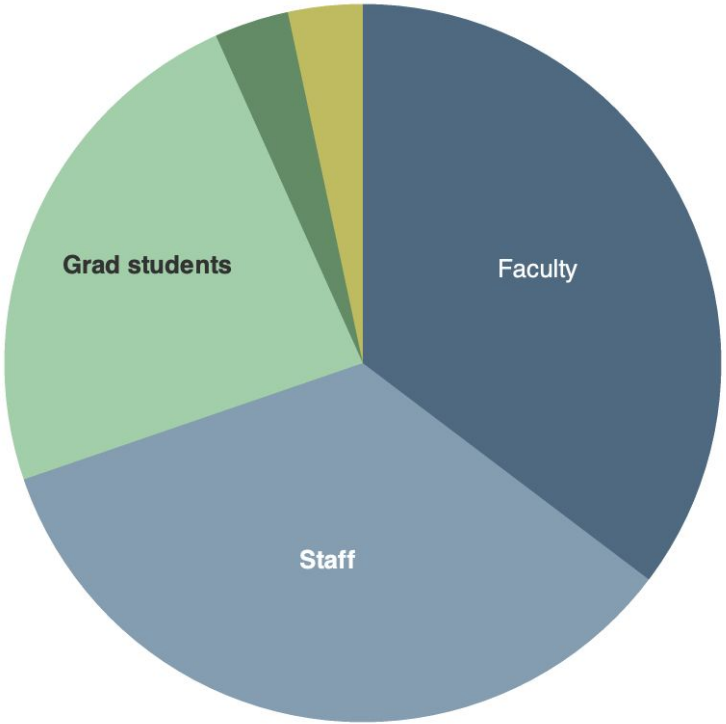
How does language used in apologies impact public opinion?

Political Science

Can news articles inform us about construction project risks?

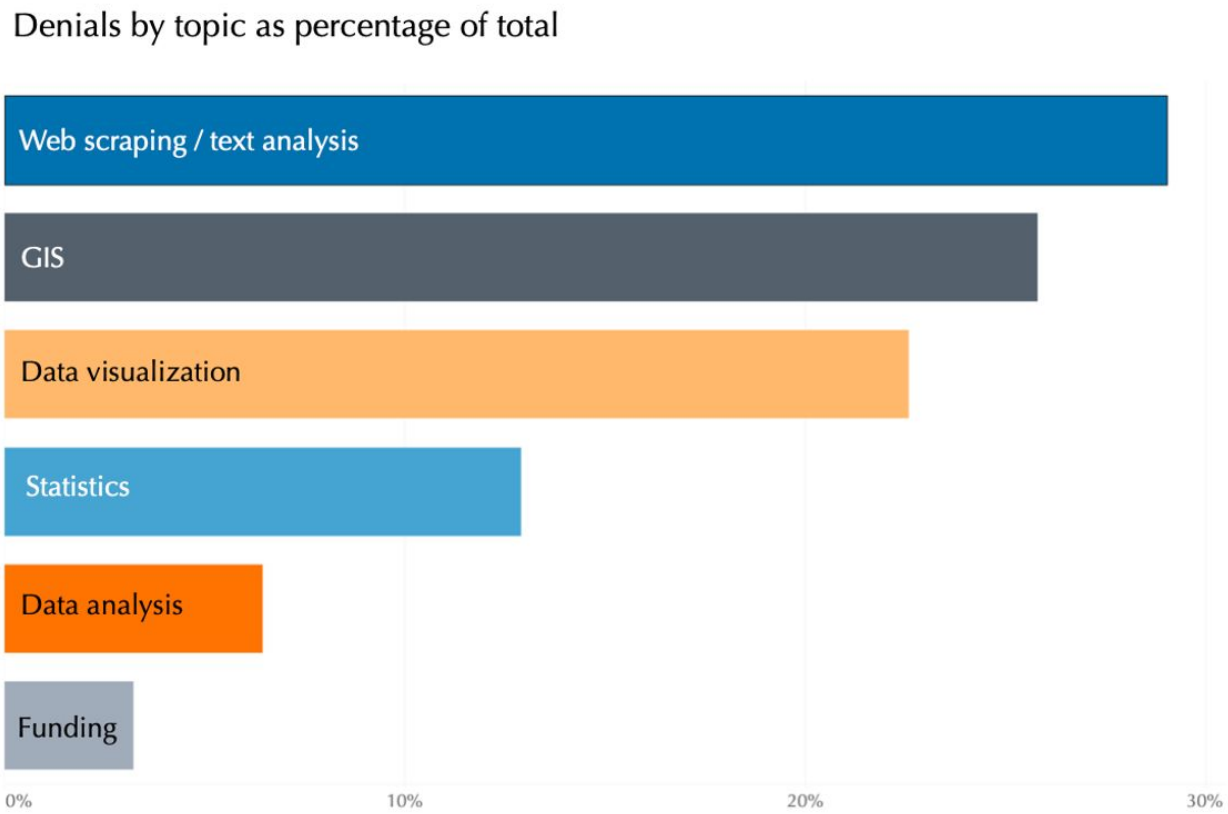
Civil Engineering

Who is coming to us for text analysis support?

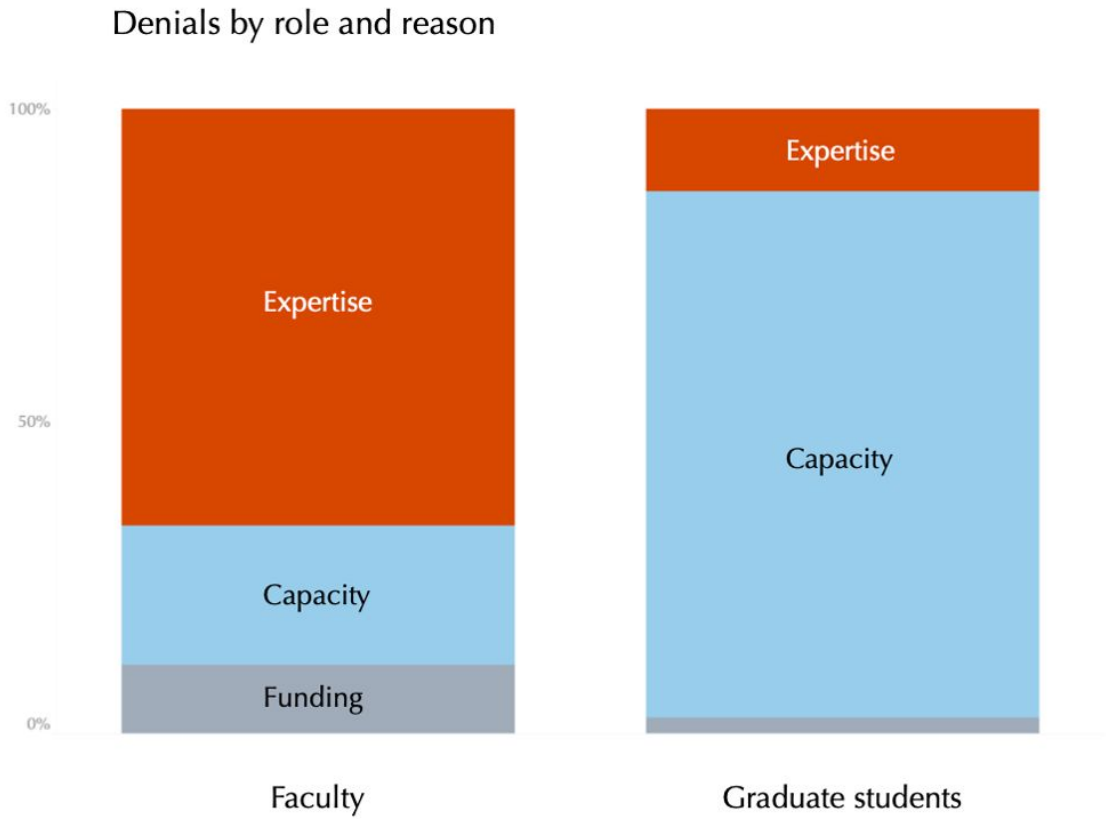


3. What does our turnaway data show?

What services are we unable to provide?



Why are we unable to provide these services?



1. Interest in text analysis is robust.

2. Interest in text analysis spans disciplines and roles.

3. What does our turnaway data show?

= A text analysis staff line moved to the top of our wishlist.

Texas A&M University

Leveraging Constellate in a Semester-long Text Mining Course at TAMU

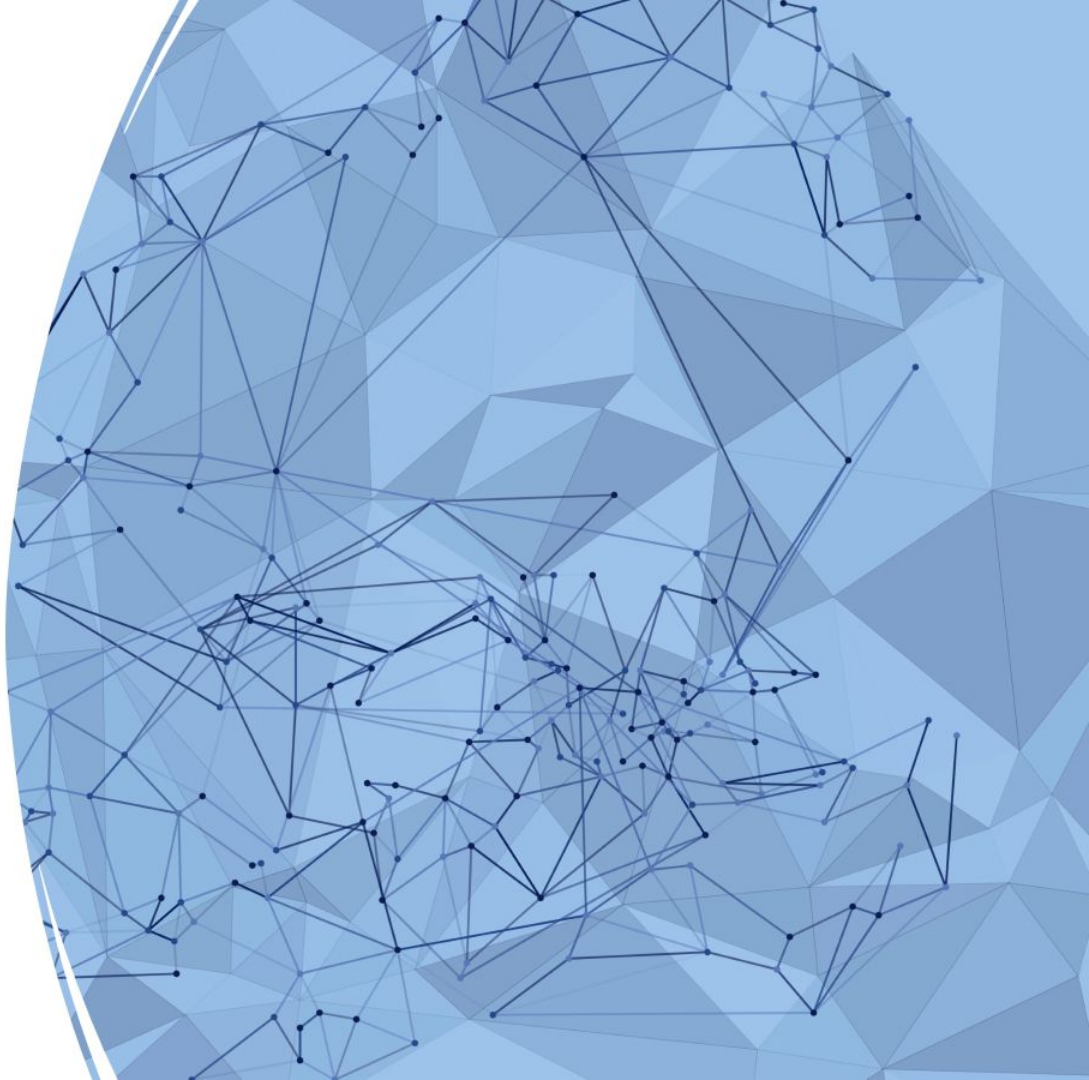
David B. Lowe

Assistant Professor

Dept. of Global Languages & Cultures

Texas A&M University

IASSIST and CARTO Conference 2024, Halifax, Nova Scotia



The Instructor

- Librarian in sheep's clothing: one of 23 TAMU librarians who became departmental faculty in recent reorganization
- Academic background in Russian
- Library professional background in digital projects and scholarly communication
- AY 2023/2024 Teaching load of 2 courses/semester:
  - Russian 101
  - Modern Languages 489 (Special Topics): "Interrogating Cultural Texts as Data"

The Course

- Syllabus: based on Melanie Walsh's (UW) "Intro to Cultural Analytics"
  - Replaced her github assignments with Constellate tutorials
  - Topical flexibility based on student project needs toward semester end
- Had strong support from Digital Humanities-inclined English & History Departments who advocated cross-listing
- For students to register, listed as a Special Topics course (*i.e.*, offered 3x max)
- Recently submitted for curricular approval for Fall 2025 Catalog
- Departmental listing was MODL = Modern Languages; will soon become Global Studies with new major
- Largely drew upperclassmen looking for an upper-level English course

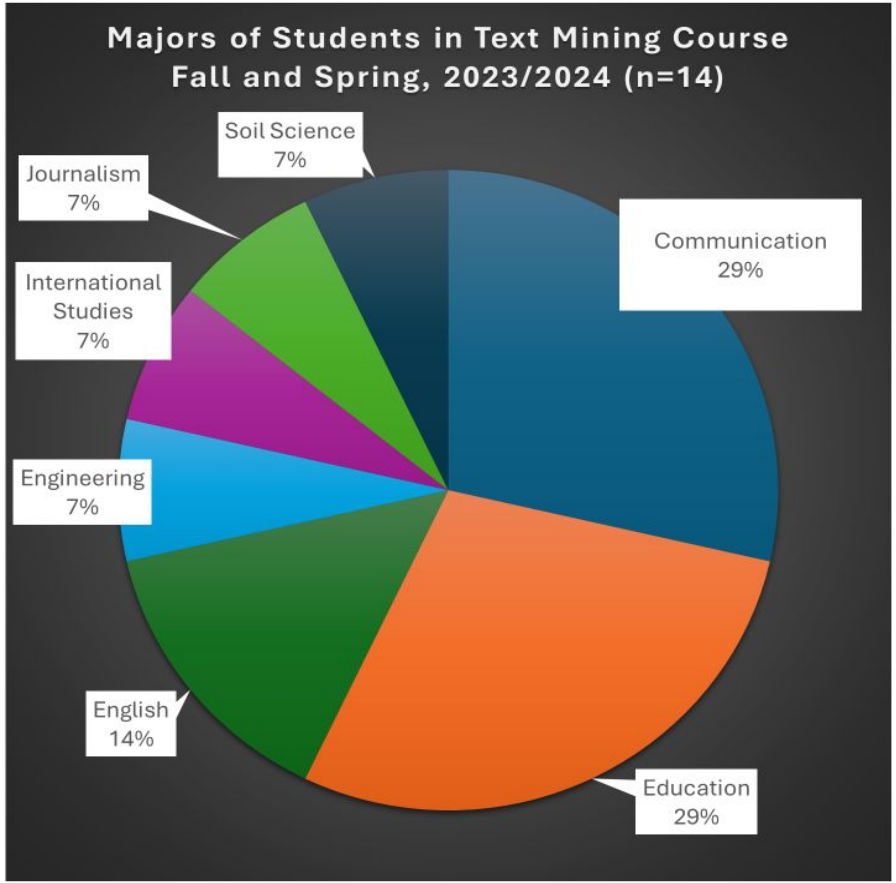
The Course Deliverables

Grading Policy

Quiz 1	20 points	Week 2
Quiz 2	20 points	Week 4
Quiz 3	20 points	Week 5
Quiz 4	20 points	Week 6
Lead discussion	20 points	TBA; signups
Dataset Biography	100 points	Week 8
Midterm Project	100 points	Week 10
Final project	170 points	by end of finals
& presentation	30 points	TBA
Total:	500 points	

The Students

- Expected broader range of majors
- Planned to do group projects that spread programming skills across all the projects
- In actuality, mostly majors in humanities
- Had mix of group and individual course projects



The Students' Projects: Topics (Data Source)

- Groups:
  - Banned Books (HathiTrust)
  - Texas Arrest Records (State data)
- Individuals:
  - Singapore demographics (Country data)
  - Texas State Legislature Committee Hearings and NGOs (State data)
  - Biblical Wisdom Literature and Gender Roles (biblegateway.com scrape)
  - Fantasy Novel Features vs. Popularity (GoodReads data)
  - "Main character" in Game of Thrones (Rotten Tomatoes data)
  - Education vs. Earnings (US Census data)
  - Looted Art and 20th Century Events (JSTOR/Constellate)

Take-aways

- Differences between ***library workshop*** and ***departmental for-credit course***, not least of which is marketable deliverable
- Compare: author's task for short story vs. novel
- Tricky to find a way for academic libraries to be entrusted with full-throated for-credit courses, but would be a positive thing for libraries to be wired into this aspect of teaching
- Experience with MODL 489 at TAMU proves that it is possible for those with librarian backgrounds to participate in this critical, for-credit coursework aspect of the mission of the university

Bucknell University

Bucknell University Facts

- 3500 undergraduate students
- Across three colleges: Arts and Sciences, Management, Engineering
- 450 faculty
- Library: 30 staff in reference services, archives, acquisitions, and digital pedagogy and digital scholarship

A data services puzzle

Audience

- Faculty, students, public

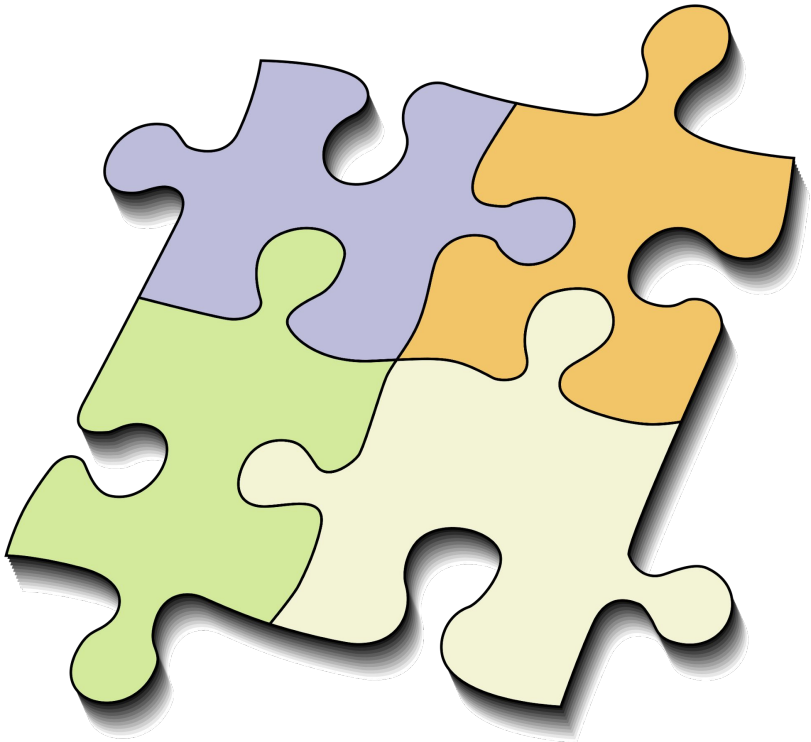
Technology

- Constellate, LWIC, Voyant, LEAF

Staff

- Text analysis, visualization

Data



Digitizing Suzette

Audience

- French faculty and students

Technology

- LEAF, Transkribus

Staff

- Digital humanities specialist, librarians

Data

- Digitized 19c French textbook

Heresies

Audience

- Art history and women's studies faculty and students

Technology

- LEAF writer, Transkribus

Staff

- Digital humanities specialist, librarians

Data

- Digitized feminist zine from 1970s and 1980s

Latinx Theater and DH Instruction

Audience
- English faculty and students

Technology
- Google sheets, GIS

Staff
- GIS specialist, librarians

Data
- Social media data and mapping information

Audience
- Classroom introduction to text analysis

Technology
- Constellate, Python

Staff
- Digital scholarship instructor

Data
- Constellate data examples

Your questions for us

David Beales
rdb104@case.edu

Digital Scholarship Partner, Digital Scholarship
Case Western Reserve University

Jen Ferguson
j.ferguson@northeastern.edu

Head, Research Data Services
Northeastern University

Amy Kirchhoff
amy.kirchhoff@ithaka.org

Senior Manager
Constellate

David Lowe
davelowe@tamu.edu

Assistant Professor, Global Languages and Cultures
Texas A&M University

Todd Suomela
todd.suomela@bucknell.edu

Associate Director, Digital Pedagogy and Scholarship
Bucknell University