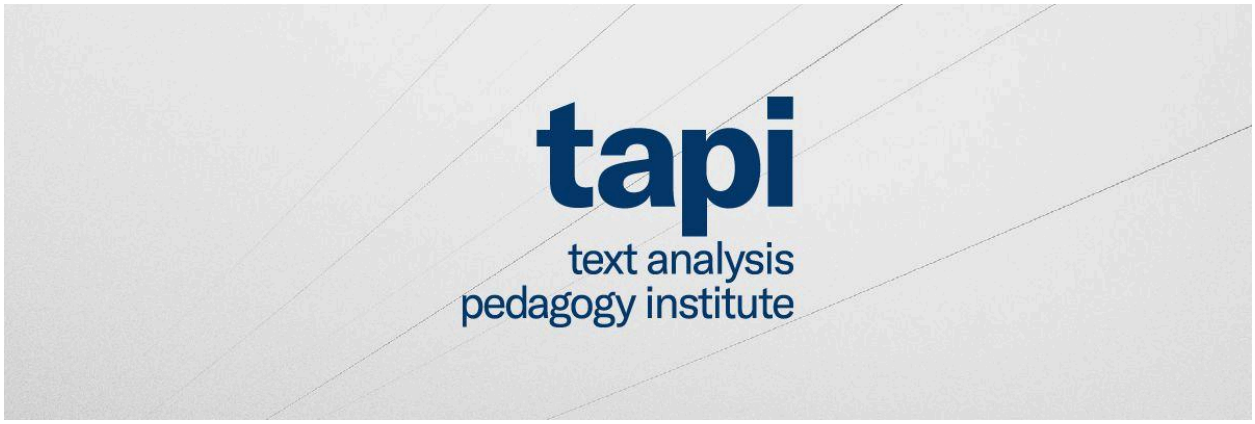

White Paper
Institutes for Advanced Topics in the Digital Humanities
Grant HT-272566-20
Project Director: Nathan Kelber
Institution: Ithaka Harbors
Reporting Period: 09/01/2020-11/30/2022

Project Summary

The Text Analysis Pedagogy (TAP) Institute, offered in 2021 and 2022, sought to address the humanities needs for greater community support, technical infrastructure, and open educational resources. In 2021, we partnered with The University of Virginia Library and The University of Virginia School of Data Science. In 2022, we partnered with The University of Arizona Library. By offering 20 courses by 11 expert instructors, the TAP Institute trained over 400 higher education faculty, librarians, research staff, and graduate students in humanities-focused text analysis. Many participants attended multiple courses, leading to over 1000 course seats for a combined offering of 90 hours of instruction.

The TAP Institute was a meeting place for the humanities text analysis community to gather and share resources. The TAP Institute's expert instructors generated 57 open educational resources, including 51 Jupyter Notebooks and R Markdown documents. The community created a record of practitioners and existing resources that are shared publicly on the website.

Project Origin and Goals

Today, the humanities have a valuable research role to play in the era of data science, big data, and machine learning, especially confronting social issues like algorithmic oppression, data privacy, and social media manipulation. The academic, social, and commercial need for humanities data literacy is greater than ever, leading to a growing number of institutes, workshops, books, and online courses.[1] These resources have sprung up to address a knowledge gap between the learning needs of the next generation of humanists and the technical skills of current humanities faculty. In practice, this means emerging humanities text analysts rarely learn the necessary skills through traditional coursework.[2] ITHAKA has spoken with more than 200 digital humanities faculty, librarians, and researchers across the United States—from Research 1 to Research 3, Ivy league, small liberal arts colleges, and community colleges—and a handful of digital humanities educators in England, Canada, and Australia. These conversations have confirmed research that argues the greatest challenges for supporting text analysis are not simply access to educational resources, but additionally access to data, community support, and technical infrastructure.[3]

The TAP Institute was inspired by existing models for digital humanities education such The Digital Humanities Summer Institute (DHSI), The Humanities Intensive Learning and Teaching (HILT), and the Digital Humanities Research Institute (DHRI). While these institutes often introduce text analysis courses, the TAP Institute focused squarely on humanities text analysis and included intensive programming training. This focus responds to a growing desire for teaching with executable code environments, seen in the curriculum of The Programming Historian, The Carpentries, and individual instructors (such as Melanie Walsh's course "Introduction to Cultural Analytics & Python").

Our primary goal was to help faculty, librarians, and research staff offer humanities text analysis at their home institutions. This train-the-trainers model relied on the following objectives:

- Offer teacher workshops in advanced text analysis techniques
- Create new open educational notebooks with executable code
- Reach a wide, diverse audience
- Network the community of professional and aspiring teachers

Project Activities, Team, and Participants

Project Activities

Work Schedule

Each TAP Institute followed a similar work schedule:

- Planning (October-January)
    - Plan theme
    - Recruit teachers
    - Course concepts
- Recruit Participants (February-April)
    - Issue Call For Participants (CFP)
    - Publicize and share CFP
    - Review Applications
    - Send out acceptances
- Course planning (May-June)
    - Instructors plan courses
    - Instructors test lessons in Constellate
- Institute (June-July)
    - Create and send pre-institute survey
    - Send invitations for Zoom and Slack
    - Organize and support classes
    - Create and send post-institute survey

Participant Recruitment

The TAP Institutes were advertised through new and existing channels including:

- Large community listserv lists
- Social media (Twitter, Facebook, etc.)
- Direct emails to institutions for reaching local listserv lists
- JSTOR Labs website

Institute Scheduling

The 2021 and 2022 TAP Institute both offered 10 courses. Most courses were three 90-minute sessions spread across Monday/Wednesday/Friday. We offered the courses during the summer when we figured the greatest number of academic staff would have time for professional development. The courses and their descriptions are below:

Python Basics 1-5
Nathan Kelber
This course is appropriate for complete beginners who have never programmed or done text analysis before.
If you've never programmed before, this course is a great introduction. Taught from a humanist perspective, this course will help you start writing your first code and unlock the potential of text analysis.

Introduction to R Programming
Jacalyn Huband
This course is appropriate for complete beginners who have never programmed or done text analysis before.
This course is a gentle introduction to R programming. With an emphasis on text analysis, this course will help you begin your adventures in programming.

A Gentle Introduction to Optical Character Recognition with PyTesseract
Hannah Jacobs
Python Basics required.
This course will introduce the concept of “Optical Character Recognition” (OCR), various tools available for performing OCR, and important considerations for successfully OCRing digitized text. Using Tesseract in Python, we’ll walk through the entire process using a variety of examples to show the range of challenges scholars can face when performing OCR. By the end of the course, participants should be able to use the course’s Jupyter Notebooks to perform OCR on their own; they should be able to identify possible technical challenges presented by specific texts and propose potential solutions; and they should be able to assess the degree of accuracy they have achieved in performing OCR.

A Practical Guide to Text Data Curation
Xanda Schofield
Python Basics required.
No matter how exciting your research question is or how fancy your models are, all text analysis projects depend on having text data that is tidy enough to analyze. This course surveys some practices of text data curation to filter out irrelevant text, refine a corpus vocabulary, and identify text artifacts in real world text collections. We will explore how to approach these tasks using Python libraries such as NLTK and spaCy, as well as explore how some text models, like LDA topic models, can actually serve as a tool for diagnosing recurring corpus issues.

Web Scraping and Text Analysis in Bilingual Social Media

Rubria Rocha De Luna
Requires Facebook account. No prior programming experience required.
This course is designed for attendees to learn how to web scrape social media posts, as well as download the information in csv format, clean it, and do basic analysis such as word frequency. To achieve this, we will rely on exercises with posts in Spanish, English or Spanglish, taken from Facebook pages belonging to organizations of migrants returned to Mexico. We will use some tools like Facepager, Notepad, Word, and RStudio.

Python Intermediate 1-4
Nathan Kelber and Zhuo Chen
Python Basics required.
An introduction to intermediate Python skills including comprehensions, working with .txt, .csv., and .json files, navigating filepaths with pathLib, and object-oriented programming (OOP).

Data Analysis with Pandas
Melanie Walsh
Python Basics required.
This workshop will introduce students to a popular Python package known as Pandas, a tool for data analysis and manipulation that is widely-used among data scientists. Participants will learn how to work with CSV files and JSON files, how to filter and aggregate data, how to make bar charts and time series plots, how to merge datasets with common values, and more. All case studies and examples will feature data relevant to the humanities, such as (potentially) library circulation data, screenplay data, and social media data.

Visualizing Humanities Data
Zoe LeBlanc
Python Basics required.
This course will introduce participants to some of the foundations and horizons of visualizing humanities data. To help us generate datasets we will lightly explore some text analysis methods, and then focus on some of the possibilities and pitfalls of visualizing data derived from these methods. In particular, this course will introduce participants to the principles of the grammar of graphics and exploratory data analysis through using the Python library Altair and Jupyter Notebooks. The goal of this course is to help participants learn how to incorporate visualizing humanities data into their research workflows, for both sharing aggregated information and making arguments.

Text Analysis in Ancient/Medieval Languages
William Mattingly
Python Basics required.
This workshop will introduce students to natural language processing (NLP) and text analysis in ancient and medieval languages. We will use Latin as a case study. Day 1 will focus on the basics of NLP and spaCy, one of the leading NLP libraries for Python. Day 2 will address the textual problems of working with ancient/medieval languages, including how to handle highly-inflected languages; lemmatization without a lemmatizer; and accounting for textual,

geographical, and temporal variances of the language. Day 3 will address a single text analysis problem: named entity recognition (NER) in Latin. On this final day, we will develop a workflow for solving this problem. Students will leave this workshop with a strong understanding of NLP and NER. They will also have an understanding of how to solve text analysis problems in highly-inflected or dead languages. Students will be provided with the resources for further learning. Finally, students will leave the workshop with a working NER model that they can use and improve in the future.

Working with Twitter Data
Melanie Walsh
Python Basics and command line experience recommended.
This course will prepare students to collect, analyze, and visualize Twitter data. Students will learn how to work with the Twitter API and with the Python library twarc, one of the most popular tools for Twitter data. We will also introduce basic text analysis methods that are appropriate for short documents like tweets. Participants who are eligible for the Academic Research Track of the Twitter API will have the opportunity to work with the entire historical archive of tweets (2006-2022).

Introduction to Natural Language Processing with spaCy
William Mattingly
Python Basics required.
This course will introduce the key concepts of natural language processing (NLP) and an NLP Python library, spaCy. SpaCy allows users to cultivate robust pipelines for text analysis. In Day 1 we will learn about NLP concepts and how to install and use the spaCy library generally. On Day 2, we will learn how to use spaCy to identify linguistic features within a document. On Day 3, we will learn about how to apply those features to solve real-world problems for information extraction.

Multilingual Newspaper Data and Visualizations
Sylvia Fernández Quintanilla
No prior programming experience required.
This course is designed for attendees to learn close reading text analysis with bilingual (Spanish and English) newspapers hosted in various digital repositories; create bilingual datasets and clean the data; select images from the newspapers and edit them; adapt these datasets for visualizations (mapping, timelines and networking) approaching it through time, space, cultural and historical contexts. We will use tools like Excel, Open Refine, Carto, Timeline JS, and GraphCommons.

Introduction to Pandas (William Mattingly)
William Mattingly
Python Basics required.
This course introduces students to working with tabular data in Python through the Pandas library. On Day 1, you will learn how to install and import Pandas; you will also learn about some of its basic features, such as the DataFrame. Day 2 will focus on finding, organizing, and sorting

data. Day 3 will focus on advanced searching methods, such as filtering, querying, grouping, and GroupBy. A few additional lessons will be provided on plotting data in Pandas.

Intro to Machine Learning
William Mattingly
Python Basics required. Introduction to NLP with spaCy is recommended.
This workshop will introduce students to machine learning (ML), from its early beginnings to its modern applications; students will also be introduced to a branch of ML known as deep learning. We will specifically address how ML can be used to solve text-based problems. Day 1 will focus on the basics of ML, the key concepts and terms that practitioners must know. Day 2 will be dedicated to a common ML problem: text classification. Day 3 will focus on an adjacent problem: topic modeling. On both days, students will be provided a worfklow for solving these problems. Students will leave this workshop with a firm understanding of ML conceptually and a basic understanding of how to engage in ML via Python. Finally, students will be provided with the resources for further learning.

Intro to Machine Learning
Grant Glass
Python Basics required. Knowledge of Pandas recommended.
This course will introduce you to many techniques available to analyze textual data with different Machine Learning techinques in Python. You will be introduced to the theory and method of Machine Learning and given some practical skills on how to write and execute machine learning code in Python. Some basic experience with Python will be required for participation in the class coding projects, but feel free to join us if you want to have a better understanding of what Machine Learning techniques can do for humanists. Generally speaking, this class will help you think about humanities problems through the lens of Machine Learning.

Named Entity Recognition
Zoe LeBlanc
Python Basics required.
This course will introduce participants to one of the core areas of natural language processing - named entity recognition. While annotating datasets with set standards is one of the oldest areas of DH research (particularly with the Text Encoding Initiative), this course will focus on some of the newer approaches for identifying and annotating objects of interest in any given text. The course will focus on using the Python library Spacy with both it's built-in functionality, and also learning how to expand upon it for more specific uses. While this course is taught in English, participants are encouraged to bring sources in multiple languages. Ultimately, participants will learn both how to leverage NER in their research and how to tailor NER to their specific textual sources.

Machine Learning for Humanists
Grant Glass
Python Basics required. Knowledge of Pandas recommended.

This course will introduce students to the variety of machine learning (ML) algorithms available for textual analysis. Throughout the three days of the course, we will address how ML can be used to solve text-based problems. Day 1 will focus on the basics of ML and students will use supervised learning to work through a research question. Day 2 will be dedicated to a common ML technique: Topic Modeling. Day 3 will focus on more advanced techniques such as using language models to classify text. Everyday students will be provided a workflow for using these techniques on their own research questions.

Introduction to Multilingual Named Entity Recognition
Python Basics required. Introduction to NLP with spaCy is recommended.
This course will introduce students to named entity recognition with emphasis placed on multilingual documents. In Day 1, we will address some of the common issues one faces in handling multilingual documents, such as inconsistent text encoding and text standardization, and some of the current state-of-the-art transformer-based language models. We will also meet some of the key features of spaCy's NER pipelines. On Day 2, we will jump into rules-based NER with spaCy. On Day 3, we will explore machine learning (ML) based NER in spaCy. Here, we will learn the essentials of creating good datasets for training NER models.

How to do Things with Topic Models
Rafael Alvarado
Python Basics and Python Intermediate recommended.
This workshop will introduce students to the concept of topic models and how they have been used to advance humanistic research. Topics to be covered include topic models as a general task in text analytics, creating topic models from scratch using Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF), visualizing their results, evaluating their performance, and interpreting their results. In addition, students will be exposed to examples of how topic models have been used in humanistic and social science research. Work will be conducted using Python 3 and Jupyter Notebooks.

Team

The TAP Institutes relied on a wide variety of individuals for its support and success. The TAP Institute team falls into three groups:

- Institutional Partners and Organizers
- Support Staff
- Instructors

Institutional Partners and Organizers

Our institutional partners at The University of Virginia and The University of Arizona were instrumental for helping craft the vision behind the initial grant proposal. Then the Covid-19 pandemic struck, and we were forced to pivot from intimate, in-person institutes for about 20 people to large, virtual institutes. Our partnerships shifted from planning planes, hotels, and food to helping define how each virtual institute could meet their values, goals, and identities. That

led us to re-envision each institute's formats, goals, instructors, and course materials. Our team for this work at the University of Virginia included:

- **Dr. John Unsworth**, University Librarian and Dean of Libraries at the University of Virginia. Dr. Unsworth is a resident of Virginia and an American citizen. His support was mainly through the connection of interested parties and planning.
- **Dr. Phil Bourne**, the founding Dean of the School of Data Science at the University of Virginia. Dr. Bourne is a resident of Virginia and an American citizen. His support was mainly through the connection of interested parties and planning.
- **Dr. Rafael Alvaredo**, the Program Director of the M.S. in Data Science and an Associate Professor at the School of Data Science at the University of Virginia. His organizational support was helpful in pivoting from the in-person to the virtual institute. He also helped review applications.
- **Dr. Jacalyn Huband**, Computational Research Consultant in Research Computing at the University of Virginia. Her organizational support was helpful in pivoting from the in-person to the virtual institute.

Our team at the University of Arizona included:

- **Sarah Shreeves**, Vice Dean of University Libraries. Her support was helpful in the planning stages of the institute, including setting the theme and goals of the event.
- **Megan Senseney**, Head of Research Engagement at University of Arizona Libraries. Her support was essential for setting the theme of the institute, including the courses and instructors.

Support Staff

Ithaka Harbors helped to support the institutes through organization and technical expertise:

- **Alex Humphreys**, who is the Associate Vice President of JSTOR and the Director of JSTOR Labs, helped supervise the project's director, Nathan Kelber.
- **Ted Lawless** (who has since moved to a new position), was a Software Developer and ran technical support for the conference teaching platform, Constellate.
- **Amy Kirchhoff**, who is the Archive Service Product Manager for JSTOR and Portico, helped organize the conference. She helped organize the Slack workspace, the Zoom channel invitations and recordings, and served as an application reviewer.
- **Matt Lincoln**, who is a Senior Software Engineer, helped with technical support with the Constellate platform.
- **Ian Desjardins**, who is a Senior Software Engineer, helped with technical support with the Constellate platform.

Instructors

The 2021-2022 TAP Institutes relied on an instructional staff of 11 expert instructors:

- Dr. Rafael Alvaredo, University of Virginia
- Dr. Sylvia Fernández Quintanilla, University of Texas at San Antonio
- Dr. Grant Glass, University of North Carolina at Chapel Hill
- Hannah Jacobs, Duke University

- Dr. Jacalyn Huband, University of Virginia
- Dr. Nathan Kelber, JSTOR Labs
- Dr. Zoe LeBlanc, University of Illinois
- Dr. William Mattingly, Smithsonian Data Science Lab
- Dr. Rubria Rocha De Luna, Texas A&M University
- Dr. Melanie Walsh, University of Washington
- Dr. Xanda Schofield, Harvey Mudd College

Participants

The participants for the TAP Institutes were primarily academic faculty, librarians, research staff, and graduate students from the USA. We received applications from every region of the country, with the most from the Mid-Atlantic, the South, the Southwest, and the Midwest. We also received a handful of applications each year from every continent except Antarctica. Our international applications came from Australia, Austria, Canada, China, England, Estonia, France, Germany, Greece, India, Israel, Italy, Japan, Luxembourg, Malawi, Netherlands, Nigeria, Peru, Singapore, Switzerland, South Africa, and Spain. In total, we received 484 applications.

Project Outcomes

The TAP Institute trained over 400 higher education faculty, librarians, research staff, and graduate students in humanities-focused text analysis. Many participants attended multiple courses, leading to over 1000 course seats for a combined offering of 90 hours of instruction. The number of course seats doubled in size year-over-year.

2021 TAP Institute (337 total seats)
- Python Basics (79)
- Introduction to R Programming (20)
- A Gentle Introduction to OCR (21)
- How to do Things with Topic Models (21)
- Data Analysis with Pandas (27)
- Machine Learning A, instructor Mattingly (30)
- Machine Learning B, instructor Glass (34)
- Visualizing Humanities Data (40)
- Text Analysis in Ancient/Medieval Languages (30)
- Named Entity Recognition (35)

2022 TAP Institute (672 total seats)
- Python Basics (A and B Sections) (89)
- Working with Twitter Data (101)
- A Practical Guide to Text Data Curation (99)
- Introduction to NLP with spaCy (92)
- Intro to Multilingual Named Entity Recognition (37)

- Web Scraping and Text Analysis in Bilingual Social Media (55)
- Machine Learning for Humanists (76)
- Introduction to Pandas (56)
- Multilingual Newspaper Data and Visualizations (67)

The TAP Institute's expert instructors generated 57 open educational resources, including 51 Jupyter Notebooks and R Markdown documents. The community created a record of practitioners and existing resources that are shared publicly on the project website. Additionally, the community assembled a public list of participants and existing humanities text analysis resources (courses, notebooks, research projects, books, and videos).

Most courses were taught on Zoom using Jupyter Notebooks with the Constellate platform. This allowed participants to write and execute code in real-time. All the lessons with code are available through the project website and hosted on GitHub.

Project Evaluation and Impact

The primary impact for the humanities was the training of over 400 potential text analysis teachers and the development of new, open educational materials to support teaching. Many of the learning materials created for the 2022 TAP Institute are unique within the humanities, focusing on specialized knowledge, such as:

- Multilingual Named Entity Recognition
- Web Scraping in Bilingual Social Media
- Machine Learning for Humanists
- Multilingual Newspaper Data and Visualization
- Practical Text Data Curation
- Twitter Data Analysis

While text analysis techniques are found within digital humanities research, there are few resources for learning them, particularly in computational environments which combine process with executable code. Our Institute helped to fill a gap by providing learning materials that are responsive and applicable to the needs of humanists.

In addition to to the Humanities and Information Studies, we received applications from scholars in a dizzying variety of disciplines including: Computer Science, Linguistics, Educational Technology, Data Science, Health, Medicine, Social Sciences, Law, Journalism, Government Information, Information Technology, Human Computer Interaction, Artificial Intelligence, Machine Learning, Psychology, Archives, Communication, Publishing, Digital Collections, Anthropology, Social Psychology, Engineering, Social Work, and Accounting.

Participation from a wide range of scholars and fields ultimately benefits humanities scholars by overcoming disciplinary silos. It also allows for opportunities to enrich a wide range of fields with humanities-based insights into text analysis.

The overall impact of the institute teaching and lessons needs should be understood in the context of the deep academic need for this training. From 2021 to 2022, the reach of the TAP Institute expanded from 213 attendees to 250. We collected data using Google Forms before and after each institute to get a sense of The TAP Institute's impact.

2021 Insights

Before the 2021 institute, most participants said they had little familiarity with programming:

- 64% had never programmed with Python
- 64% had never used a Jupyter Notebook
- 69% were not comfortable explaining basic Python concepts

Before the institute, participants (n=107) were asked whether they felt ready to teach text analysis, they reported:

Yes (14%)
No (65%)
Maybe (21%)

After the institute, participants (n=43) were again asked if they felt ready to teach text analysis. They responded:

Yes (47%)
No (28%)
Maybe (25%)

Even with this tremendous growth, participants were eager for more courses. After the institute, 68% of participants indicated that "growing their skills" was the most significant challenge for starting teaching. Participants indicated that their "highest priority for improving or starting teaching text analysis" was attending "more TAP Institutes" (46%). All respondents indicated interest in attending TAP Institute 2022 with either "Yes" or "Maybe".

In 2022, 50% of returning participants indicated they had taught text analysis since the last institute.

2022 Insights
In 2022, when asked if participants were ready to teach text analysis, they reported:

<u>Pre-Institute (n=67)</u>
Yes (9%)
No (79%)
Maybe (12%)

Post-Institute (n=44)
Yes (46%)
No (27%)
Maybe (27%)

Participants were asked to rate their confidence in their coding skills on a scale of 1-10.

Score	Pre-Institute	Post-Institute
1	7.5%	0
2	13.4%	2.3%
3	14.9%	9.1%
4	20.9%	11.4%
5	16.4%	9.1%
6	9%	15.9%
7	6%	27.3%
8	7.5%	11.4%
9	1.5%	4.5%
10	3%	9.1%

When asked, “Would you attend TAP Institute in 2023?,” participants responded:

Post-Institute
95.5% Yes
2.3% No
2.3% Maybe

In addition to generally positive feedback:

> *Thank you for this awesome program and making your content open access!*
> *The institute is a leading academic institution.*
> *The Institute is new but already well organized and full of activities*
> *Well organized*
> *Gap-filling work.*
> *Continued great work - TAP represents one of my favorite professional development activities.*
> *It was excellent!*

Please do continue online offerings.
You are doing amazing work. Thank you so much for helping us out :)
This was a great institute and I really appreciated the virtual offering.
I learned so much. I would like to retake the courses so that I can solidify what I learned.
Fills a huge need for many experienced professionals
Thank you for offering these courses!
Very good (and generous) initiative.
Very inclusive, helpful series of workshops.
the lessons being available online/github is helpful
There's a lot of potential
Fantastic work you all are doing!!!

We also received feedback on desired courses for future institutes:

Intermediate Python
Twitter Analysis, Using Document the Now for ethical web scraping, bilingual social media analysis
Advanced level of python basics, machine learning analyses with various types of data.
Text curation
content similar to 2022
maybe train the trainer sessions? teach us methods for teaching it to others and designing courses or workshops for it
NER, SpaCy, topic modeling/MALLET
Machine learning and Introduction to Pandas
sentiment analysis
Introduction to stylometry and authorship attribution (e.g. advanced use of the stylo programme)
Maybe something involving transformers and creating curated text data sets
More moderate level courses
More machine learning, more intro sessions (especially for NLP) that really break down best practices instead of trying to run through a project, database design
Even more exploration into text analysis with multilingual content! Delving into training your own models would be cool too.
Similar to what was offered this time, since we could only take 3.
A slightly gentler introduction to NLP, and a more advanced Python for humanists
Something on sentiment analysis and emotions analysis please!
I think a similar slate makes sense.
I wish there was a geospatial tool for Python class.
Python Basics and maybe an intermediate Python course
Regular Expressions; by then, maybe transformer models.
multilingual text analysis, web scraping
Advanced Python, Theory + coding for important NLP methods
Next step in Python and/or pandas. Go beyond intro to some more complicated uses (objects, __main__, groupby, merge)
More Pandas courses.

Multi-lingual text modelling
Something about setting up an environment / notebook to facilitate a session, and pedagogical concerns
Web scraping
Pandas, Spacy, R
Tableau, data visualization & plotting tools
More general workshop beyond the introduction to Python but before spaCy or machine learning.
Machine Learning, web scrapping, text mining
NER
making pipelines, using copyrighted data, combining text and network analysis
More on social media analysis, web scraping, visualizations
Twitter Data, Corpus Linguistics, R workshops
Perhaps a project-based course, with early instructional information that requires advanced work and a discussion-based classroom environment to discuss problems/lessons learned
It would be great to see Deep Learning and Large Language Models courses.
network analysis with either R or Python

Our overall takeaway from the TAP Institutes feedback is that we have only begun to meet the significant need for the humanities community, and there is a great desire for more institutes and learning opportunities.

Project Continuation and Long-Term Impact

Though we do not yet have data to measure long-term, downstream impacts, we hope the TAP Institute breaks down disciplinary barriers between the humanities and sciences, encouraging deeper conversations on the social implications of algorithmic decision-making. We hope the communities, conversations, and skills we are helping to build can help shape a better future, not just for commercial success but for social progress and equality. The TAP Institute can help by improving data literacy education, enabling citizens to call out and dismantle data practices that reinforce social inequality. The concepts of data literacy and data justice are increasingly essential, particularly as universities across the world are developing digital humanities and data science degrees, departments, and even schools (such as at The University of Virginia). We believe it is essential that humanists have the skills to join conversations on data, and that humanities training is essential for engaging with a society increasingly run by “intelligent” systems.

At a scale of over 400 participants over two years, we believe the TAP Institute’s train-the-trainers model has the potential to make a significant contribution to humanities knowledge. Combined with our open educational and computational resources, potential humanists looking to learn and teach text analysis have significantly more resources at their disposal than existed two years ago.

Although our funding from the National Endowment for the Humanities has come to a close, we plan to continue the TAP Institutes in 2023. Depending on our ability to find new partners and/or funders, our model for the institute may need to be adjusted. The Slack workspace for the institutes remains open, but unfortunately Slack has changed their policy for retaining conversations beyond 90 days. We may explore a new conversation space for courses next year. We have already created a mailing list (https://ithaka.groups.io/g/tdm-jstor-portico) for those interested in continuing to participate going forward. We look forward to sharing our white paper and open educational resources with the wider humanities community, holding open the door for learners across the nation.

Endnotes

1. Institutes include the Digital Humanities Summer Institute, the Digital Humanities Research Institute, and Humanities Intensive Learning and Teaching. Workshops include those offered by THATCamps, The Carpentries, and those funded by this very grant. Outside of the popular journal *The Programming Historian*, relevant books and guides include *Python Programming for the Humanities* by Folgert Karsdorp and Maarten van Gompel; *Computational Historical Thinking with Applications in R* by Lincoln A. Mullen; *The Rubyist Historian* by Jason Heppler; *Exploring Big Historical Data: The Historian's Macroscope* by Shawn Graham, Ian Milligan, and Scott Weingart; *Text Analysis with R for Students of Literature* by Matthew L. Jockers; *The Art of Literary Text Analysis* by Stéfan Sinclair, Geoffrey Rockwell, and Melissa Mony; *Humanities Data in R* by Lauren Tilton and Taylor Arnold; *Digital Research Methods with Mathematica* by William J. Turkel; *Exploratory Programming for the Arts and Humanities* by Nick Montfort; *An Introduction to Text Analysis: A Coursebook* by Brandon Walsh and Sarah Horowitz; *Hacking the Humanities Video Tutorials* by Paul Vierthaler; *Six Septembers: Mathematics for the Humanist* by Patrick Juola and Stephen Ramsey; *Text Mining with R: A Tidy Approach* by Julia Silge and David Robinson; and *Statistics for the Humanities* by John Canning.
2. The issue is more complicated than a need for faculty training, but largely stems from the slow— and slowing of—hiring in humanities for tenure track positions. Fewer secure faculty positions means institutions struggle to replace tenure lines and most humanities departments lack someone on staff qualified to teach text analysis. At the same time, graduating students with expertise in text analysis are either scooped up by (or forced into) industry. (Freshly-minted PhDs have also moved into libraries, where digital humanities and innovation centers often represent a more fertile space for digital research.) As budgets contract, there is less room for additions to the general curriculum, particularly for courses based on methods instead of period or content.
3. See Miriam Posner's "Here and There: Creating Dh Community" (2016); Vivian Lewis, Lisa Spiro, Xuemao Wang, and Jon E. Cawthorne's "Building Expertise to Support Digital Scholarship: A Global Perspective" (2015); and Neil Fraistat's "Data First: Remodeling the Digital Humanities Center" (2019).