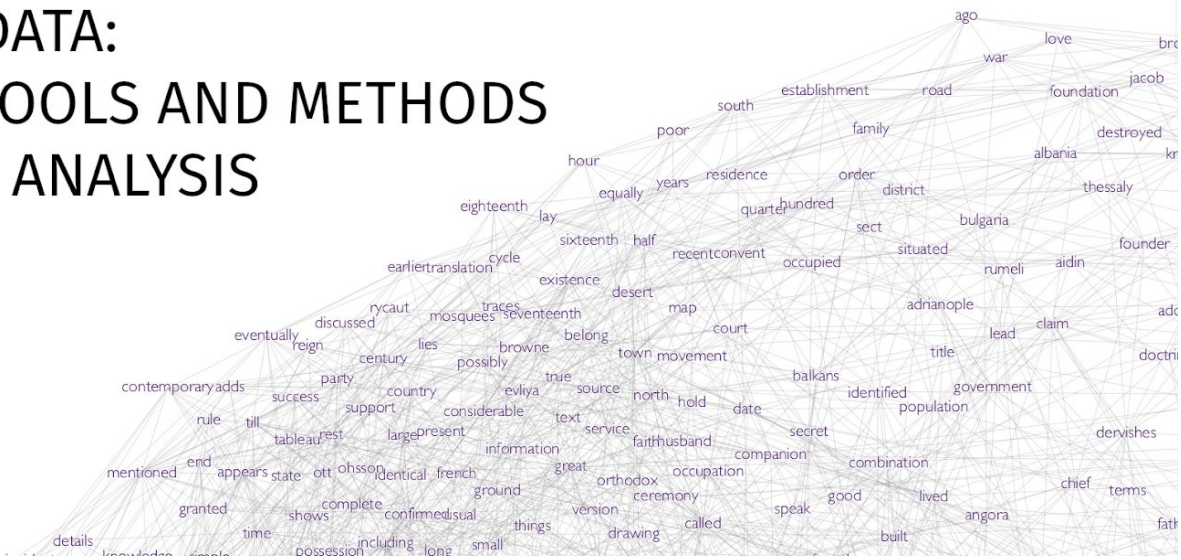
NATIONAL ENDOWMENT FOR THE HUMANITIES

The City College of New York | Division of Humanities and the Arts

TEXT AS DATA: DIGITAL TOOLS AND METHODS FOR TEXT ANALYSIS



NEH CARES DIGITAL HUMANITIES WORKSHOPS

Instructor: Dimitris C. Papadopoulos
NEH CARES Instructional Technologist, Division of Humanities and the Arts, CCNY
Email: dpapadopoulos@ccny.cuny.edu
https://dhccny.commons.gc.cuny.edu

Zoom meeting link (for all sessions):
https://ccny.zoom.us/j/97878629707

WEDNESDAY, OCTOBER 28
10:00 am - 12:00 pm:
Foundations: creating and compiling text-based data

1:00 pm - 3:00 pm:
Patterns and insights: basic tools for quantitative text analysis

FRIDAY, OCTOBER 30
10:00 am - 12:00 pm:
Visualizing texts: maps, graphs, networks

1:00 pm - 3:00 pm:
Sustainable texts: diversity, collaboration, and sustainability in text analysis and publishing projects

This workshop has been made possible in part by a major grant from the National Endowment for the Humanities: NEH CARES
Any views, findings, conclusions, or recommendations expressed in this workshop
do not necessarily represent those of the National Endowment for the Humanities.

OUTLINE

Digital humanists have been using and developing, for decades, a wide range of digital methods and software tools for text analysis that, in recent years, tend to be more user friendly and less demanding in terms of cost and computing resources. At the same time, mass digitization projects and new web-based publishing platforms and environments have opened up new possibilities for the analysis and exploration of text as data in different historical, cultural and linguistic contexts. These trends seem to remove many of the technological entry barriers for scholars interested in the analysis in literary and historical texts and text archives and resources but at the same they pose new problems and challenges in terms of the biases and assumptions, ethical issues, and socio-political implications of compiling, quantifying, analyzing visualizing texts. This series of workshops will introduce participants to some basic computational tools and methods for text analysis and exploration such as Voyant Tools, AntConc, Jstor's Text Analyzer, Recogito, the Natural Language Toolkit, and will cover different stages in the lifecycle of text analysis projects, from compiling corpora and cleaning data sources, statistical analysis, visualization and publishing, with sustainability in mind. In order to test digital tools and methods, we will be using examples of different texts (by genre or subject) and datasets including some multilingual resources. No coding skills or previous experience with computational text analysis is required.

WEDNESDAY, OCTOBER 28

Foundations: creating and compiling text data

10:00 am - 12:00 pm

In this session participants will be introduced to methods of creating, merging and compiling text-based datasets. We will discuss open source datasets, text corpora tools, text mining and data cleaning methods. We will also address the often problematic nature of corpora and textual resources and discuss issues of algorithmic bias, representation, exclusion and fragmentation, in creating and assembling text-based datasets for analysis. At the end of the session participants will be familiar with key methods and tools involved in the process of converting "raw" literary, historical and other types of text into machine readable and analysis-ready, text-based datasets.

Patterns and insights: basic tools for quantitative text analysis

1:00 pm - 3: 00 pm

This workshop has been made possible in part by a major grant from the National Endowment for the Humanities: NEH CARES
Any views, findings, conclusions, or recommendations expressed in this workshop
do not necessarily represent those of the National Endowment for the Humanities.

In this session we will explore some key concepts, methods, and tools for basic, quantitative text analysis that can help us reveal certain patterns in our text-based datasets. We will discuss terms such as "word frequencies," "clusters" or "collocates," and we will try software tools such as Voyant Tools, Antconc, Jstor's Text Analyzer, Ngram Viewer, or Textal that can help us get some initial insights into our texts.

FRIDAY, OCTOBER 30

Visualizing texts: maps, graphs, networks

10:00 am - 12:00 pm

In this session we will discuss different approaches to the concept of "distant reading," examine some key techniques such as topic modeling and explore software tools for analyzing and visualizing texts, from command line tools and programming languages such as R and Python to graphical user interface environments and simple tools for the statistical analysis and visual exploration of texts. We will also discuss mixed methodological approaches in exploring entities and relations in texts through certain projects and examples.

Sustainable texts: diversity, collaboration, and sustainability in text analysis and publishing projects

1:00 pm - 3: 00 pm

Although a plethora of software tools have simplified the process of quantitative text analysis, scholars and students often have to deal with issues of reliability, continuity and support for less represented fields, communities, and languages. In this session, we will discuss problems and challenges of collaborating on text analysis and publishing projects with minimal resources, in cross-disciplinary teams, or with multilingual sources, in what often tends to be an Anglocentric digital humanities environment.

This workshop has been made possible in part by a major grant from the National Endowment for the Humanities: NEH CARES
Any views, findings, conclusions, or recommendations expressed in this workshop
do not necessarily represent those of the National Endowment for the Humanities.