

IST800 – Text as Data for Dissertation and Thesis Projects (3 credits)
(Brief Version of Syllabus Draft Only; Subject to Change)

Spring 2023, Mondays, 9:30 AM – 12:15 PM
Hinds Hall 120, Enrollment Capacity: 12
Instructor: Jeffrey Stanton
Contact: jmstanto@syr.edu
Office hours by request: Hinds Hall 323E

Catalog description: Selected areas within the information field, emphasizing related disciplines and their relationships to the diagnosis of information needs and the collection, storage, management, regulation and dissemination of information. Repeatable.

Brief Overview for Spring 2023: Although classic quantitative data collection methods such as surveys and experiments still play a vital role in sociotechnical research, natural language text also serves as an increasingly important data source. Part of this trend is a consequence of social media: Various online platforms have provided both the technology and the social community in which people can create and share massive amounts of text. Researchers have created mechanisms to harvest text from these platforms and subject the text to systematic analysis.

Natural language text is complex and replete with ambiguities that make it much more challenging to analyze than quantitative data. As a result of widespread interest in analyzing text, however, there is an ever-expanding set of tools and techniques useful for integrating text analysis into research projects. Many of these techniques are quite straightforward to use and interpret, even without a background in math or programming. For students planning a dissertation or thesis project, this course will provide the background, philosophy, and skills needed to integrate a systematic, computer-aided text analysis into a research project. Students who successfully complete the course can expect the following outcomes:

- Essential knowledge of linguistic considerations when using text as data
- Essential knowledge of the data structures associated with textual data along with the diagnostic and visualization tools needed to set up and conduct an analysis
- Increased awareness of research literature from various fields of study that incorporates text analysis into efforts to address research questions
- Practice with conducting analyses of and writing about analytical results for text
- Improved familiarity with Python, Jupyter Notebooks, and the ecosystem of add-on Python libraries, leading to the capability of independently undertaking systematic analysis of natural language text on future projects

Prerequisites: There are no course prerequisites for this course and I will strive to make the course accessible for doctoral students who don't consider themselves to have strengths in math or coding. This is a research methods course rather than a course to learn the applied programming skills needed for professional work in Natural Language Processing (you should feel free to sign up for IST664 if this is what you are seeking). You will use some of your high school math (e.g., algebra). You will certainly find the Python easier if you have previously learned a programming language, but I will provide tutorial resources suitable for all prior levels of coding experience (including resources for those who have no previous background in coding).

Class Structure: We will use 11 weeks of the academic semester to complete an exploration of a variety of text analysis techniques and the research context that surrounds them. The other three weeks will be used for student presentations (all proposals in Week 9, final papers in Weeks 13 and 14). Keeping in mind that many of the individual topics we examine could consume a whole semester of study, we will work on developing a sufficient understanding of the fundamentals to make continued, independent study more efficient. Here's a rough layout of the semester:

- Before the class starts: Materials to support a quick introduction to the Jupyter Notebook environment offered by Google Colab; introduction or refresher on essential elements of the Python language; additional resources for continuous learning of Python
- First two weeks: Orientation to text as data for social science and humanities research; experimenting with Python, Jupyter Notebooks, and Python libraries; where to find existing datasets and how to harvest new data
- Two weeks each: 1) data representations and analytical techniques; 2) part of speech tagging and use of pretrained models; 3) integrating text data into predictive analyses
- One week for brief term-project proposals
- One week each: 1) simple neural models; 2) recurrent neural models; 3) attention-based neural models; development of the analysis and write-up of your project occurs in the background
- Last two weeks: Student project presentations (note that this assumes a class size of 10-12; with fewer students the presentations will all occur in Week 14)

Textbook: I will draw readings from a few different textbooks that are available as eBooks from Syracuse University Libraries. You are welcome to purchase paper copies of any of the books you like, but that is not required for the course.

Code: I will provide extensive code examples using Python and Jupyter notebooks in Google's Colab environment. Colab and Python provide the ideal environment both for learning about coding and for manipulating text. Some parts of the notebook code will be modified from examples provided in the readings. You can modify and extend these code examples to serve as the basis for the text analysis in your own project.

Evaluation: Each class will include some discussion of the research context, some "tech clinics" where we work with code and some test data, lots of Q&A, an opportunity to "check-in" on the progress of your term project, and a few other kinds of activities. At each class session we will discuss the next step in progress on your project: this will be the basis of a homework assignment for the following week. Homework for each week is geared to helping you make step-by-step progress on your project.

Class participation provides one basis for evaluation of your performance in the course, so your absence from more than one class session across the semester will be problematic. Aside from class participation, the primary basis for evaluation will be your term-project and the resulting write up of your text analysis. The goal is for you to complete a unit of work that is of appropriate scope and quality such that you can efficiently refine it into a conference paper. In parallel, the work you do for this class could easily provide a preliminary study or "Study 1" for your dissertation/thesis project. The final presentation of your project write-up in Week 13 or 14 of the course will serve as the capstone of the semester and as a way of providing your fellow students with additional tutorial information about the techniques you used. Each presentation will include an overview walkthrough of the Python code that helped you conduct your analysis.

Unlike masters or undergraduate level classes, in doctoral classes grades are not a critical component of the class dynamic. You are in this class to learn, to make the most progress you can, and to feel as competent, skilled, and prepared as possible at the end in service of completing your dissertation and/or other research projects more efficiently and effectively. Your participation in every class session and your development of a credible research project that analyzes text are essential and required.

Week-by-Week Course Guide:

<i>Week</i>	<i>Assignments and Activities</i>
0	Before the course begins: Guidance on how to get started with Colab (Jupyter Notebooks) and Python
1	Overview of Text as Data in the context of social sciences and humanities; Practice using Python and Jupyter Notebooks – Data structures, procedures, libraries; Web-scraping and other harvesting techniques; Project preparation 1: research interests and possible data sources;
2	Example research literature from using text as data; Handling text in Python, using NLTK and spaCy Project preparation 2: narrow down research interests, expand possible data sources;
3	Choosing the unit of analysis for using text in research projects; Creation of term-document matrices with Gensim; Project preparation 3: Preliminary description of data sources;
4	Extraction of text from sources; Research ethics considerations; Analysis using sklearn to process DTMs; Project preparation 4: One page mini-proposal
5	Part-of-speech basics; Using spaCy for POS and dependency tagging; Project preparation 5: Revised mini-proposal with data specification;
6	Origins of Word2vec and GloVe; Using spaCy to extract vector representations; Project preparation 6: Expansion of mini-proposal part 1 (expanded lit review)
7	Introduction/refresher on predictive models; Clustering, classifiers, and sentiment; Project preparation 7: Expansion of mini-proposal part 2 (preliminary method)
8	The curse of dimensionality and data reduction; Using penalized predictive models; Project preparation 8: Development of slide deck for proposal presentation;
9	Proposal presentations and feedback
10	Origins of deep learning Introduction to Keras; one hot encoding; CNN models Project work 1: Data setup and cleaning;
11	Grammatical dependencies; Using LSTM models; Project work 2: Data analysis
12	Attention and its applications; Using BERT models; Project work 3: Write up
13	Completed project presentations; optional module on using chatbots for data collection and studying user interaction with chatbots
14	Completed project presentations

