Data Analysis for Linguistic, Cultural, and Historical Research

Syllabus

DIGS 20004 / 30004

Instructor: Jeffrey Tharsen
tharsen@uchicago.edu

MWF 10:30-11:20

Office Hours: Fridays noon-2pm, or by appt.
Regenstein Library 216

Social Sciences Research Building 401

Office Phone: (773) 834-5534

Course Description

Data analysis is a rapidly expanding field with broad applicability throughout the hard sciences, social sciences, and the humanities. The ability to use and analyze data effectively in linguistic, cultural, and historical research provides a highly desirable and broadly applicable skill set: in academia, in government, and in the private sector.

This course is intended as a theoretical and methodological introduction to a number of the most widely used and effective techniques, strategies and toolkits for textual, qualitative and quantitative data analysis, with a primary focus on those most suited to the analysis of complex data sources.

We will also consider how the creation of large digital corpora and the ubiquity of large-scale digital data sources has now changed how scholars engage with the archive, and what opportunities new types of repositories might offer for new computational approaches to the study of literature and history.

In addition to evaluating new digital methodologies in the light of traditional approaches to data analysis, students will gain extensive experience in using R to conduct their analyses, and by the end of the course, will have developed their own individual projects, thereby gaining a practical understanding of data analysis workflows and tools and methods for evaluating the results achieved through various data analytical and visualization strategies.

Throughout this course, the sources, methodologies and tools we will focus on will be in part decided by student interests and goals, so as we progress, please take note of and send to me any specific types of toolkits or approaches you find might be useful or relevant for your work and analyses. Suggestions or ideas you have on approaches to data analysis and other topics we will address in the course are welcome at any time.

Course Goals

Students who complete this course will gain a foundational understanding in computational data analysis for linguistic, cultural and historical research. Students will gain practical experience in the variety of data analytical toolkits available in the R platform and refine their abilities to use, evaluate and develop new digital resources for analysis.

No prior knowledge of digital technologies or computer programming is required for this course but all students should plan to develop final projects or papers featuring original work related to one or more of the data analysis strategies we will employ.

Required Texts and Readings

(to be distributed in PDF format via Canvas)

Taylor Arnold and Lauren Tilton, *Humanities Data in R* (Springer, 2015)

Robert Kabacoff, *R in Action* (Manning, 2015)

Brett Lantz, Machine Learning with R (Packt, 2015)

All required readings for the course will be provided via the online Canvas platform at canvas.uchicago.edu . Any students without access to Canvas must inform the instructor so we can set up alternate methods for you to access the readings.

Further Reading and Digital Resources

Dan Toomey, *R for Data Science* (Packt, 2014)

Course Plan and Policies

Monday and Wednesday sessions will mainly focus on reviewing the content of the assigned readings and include lectures on and discussions of specific topics. Friday sessions will be dedicated to open discussion and R programming strategies, allowing for free-flowing, detailed and individualized discussions directly relevant to the week's assignments and topics.

Assignments

Weekly assignments will primarily be comprised of programming exercises in R. The code and output is to be submitted to the instructor for evaluation by email or unless otherwise directed.

A formal Final Project and Final Exam will be required of all students (see below).

The Final Exam will be comprised of multiple-choice questions and written responses and will be given at the time and date designated by the University's exam schedule. If for any reason you will not be able to take an exam as scheduled, you must gain prior approval from the instructor for alternate means to take the exam.

Final Project / Final Paper :

An initial proposal for the dataset(s) to be used in the final project will be due at the end of the second week, to be finalized by the end of Week 5.

Topic(s) for the final project/paper (or project white papers) are to be developed in consultation with the instructor and are to be **submitted in writing** (minimum of one paragraph in length) **by the end of Week 7. All projects/topics must have received written preapproval** (email is fine) **and will be set by the end of Week 8.**

Final projects should center on the analysis of a specific data source and include at least some of the methods we will cover and use in the course. Final projects that employ new and/or unique datasets and reach innovative conclusions will receive the highest scores.

A full written explanation of the scope and utility of the project, at least 3 pages in length (Times 12pt, double-spaced), will be required by the due date of the final project. All project coding and use of data sources will be closely reviewed, and the potential impact of the project will play a major role in its assessment. No group projects will be allowed.

All students will be given space and service units for analyses on Midway, the university's high-performance computing (HPC) cluster, depending on the needs and dependencies of each individual project, developed and maintained in consultation with the instructor. Students will be responsible for all administration and content management associated with their projects.

Students may choose to do a Final Paper instead of a Final Project. The paper must be between 10 and 15 pages in length (Times 12pt, double-spaced), and should provide detailed evaluations of and research into at least one digital resource, methodology and/or toolkit directly related to those covered in the course readings and discussions, and must include discussion of at least one programming toolkit and/or algorithm. Proper spelling, grammar and construction of your paper (thesis, argumentation, transitions, conclusions) will be strongly considered in its evaluation.

All Final Project Reports and Final Papers are due by midnight on the Friday of Exams Week. Penalties for late projects/papers will be assessed at a rate of one letter grade per day. If you will need an extension and/or to take a course grade of Incomplete, you must have received

approval for this in writing (email is fine) from the instructor by midnight on Friday of Exams Week.

Attendance

The success of our course discussions depends upon your active participation, so your contributions are important to me. Please note that your attendance isn't enough to make this course successful; I expect that you will also participate regularly in class by sharing your own observations and ideas, comments and critiques.

Absences may be excused on account of documented illness, religious observances, participation in university-sponsored athletic events, and serious emergencies. Please let me know in advance if you will be missing class for any reason. You can miss up to 3 classes without penalty. After that, your final grade will be lowered one-third of a grade for each additional absence (A- becomes B+; B becomes B-, etc.).

Grading / Evaluation

Attendance and participation:	20%
Short projects and exercises (Handouts):	20%
Final exam:	20%
Final project or paper:	40%

Special Needs

Students with any form of special needs, physical, learning or otherwise, are welcome in my courses. It is University policy to provide, on a flexible and individualized basis, reasonable accommodations to students who have disabilities that may affect their ability to participate in course activities or to meet course requirements (see http://disabilities.uchicago.edu/). All students with disabilities should contact me to discuss their individual needs for accommodations.

Course Schedule

All readings are to be completed prior to the course time under which they are listed.

Students should also feel free to bring up any of the works in the "Further Reading" list during in-class discussions or in the Short Project reports.

"Arnold", "Kabacoff" and "Lantz" refer to the course texts listed above; all other readings will be made available in PDF format in the Canvas course system. All "Handouts" will be posted to the Canvas system and also distributed in class.

Week.Class	Themes / Topics	Readings & Assignments
1.1	Introduction & Syllabus	
1.2	Data Analysis and Digital Storytelling	*Kabacoff : Preface & About Kabacoff pp.3-19*
1.3	Workflows, RStudio, Introduction to HPC	*Arnold pp. 3-24* ***Handout 1***
2.1	Data Types : Sources and Formats "What's in a Name?"	*Kabacoff pp. 20-44*
2.2	Text Mining, Text Processing and NLP Basics	*Arnold pp. 131-175*
2.3	Digital Textual Analysis	***Handout 2*** ***Dataset initial proposal due***
3.1	Data Visualization I : Graphs, Charts and Statistics	*Kabacoff pp. 46-70, 117-164*
3.2	Data Visualization II : Advanced Graphics	*Arnold pp. 63-79* *Kabacoff pp. 255-278*
3.3	Harnessing and Creating Data Visualizations	***Handout 3***

4.1	Univariate, Bivariate, Multivariate Data Analysis	*Arnold pp. 25-60*
4.2	Regressions I	*Kabacoff pp. 167-211*
4.3	Analytical Approaches to Complex Data Systems	***Handout 4***
5.1	Advanced Variance Analyses	*Kabacoff pp. 212-238*
5.2	Regressions II	*Kabacoff pp. 301-318*
5.3	*No class*	***Handout 5*** ***Final dataset proposal due***
6.1	Principal Component Analysis and Time Series	*Kabacoff pp. 319-367*
6.2	Networks and Network Theory	*Arnold pp. 81-93*
6.3	Practical Regressions & PCA	***Handout 6***
7.1	Social Network Visualizations	*TBD*
7.2	What's in a Network? (Network Optimization)	*TBD*
7.3	Advanced Network Visualizations	***Handout 7*** ***Final Project/Paper Topic Due***

8.1	Geospatial Data	*Arnold pp. 95-110*
8.2	Geospatial Data Analysis	*TBD*
8.3	*No Class – Thanksgiving Break*	***Handout 8*** ***Final Project / Paper Topic finalized***
9.1	Machine Learning Basics, Clustering and Classification	*Lantz pp. 1-26* *Kabacoff pp. 369-413*
9.2	Advanced Machine Learning and Neural Networks	*Lantz (selections)*
9.3	Beyond "Big Data"	
10.1	Student Presentations	
10.2	Student Presentations	
	Final Exam	
	Final Papers / Projects Due	