1 Humanities Text Mining/Meaning

Instructor	Evan Donahue
Email	evan.donahue@duke.edu
Day/Time	Tues/Thurs 3:30-4:45 EST
Location	Zoom
Office Hours	TBA
Syllabus Version	March 9, 2021

1.1 Course Overview

The increasing centrality of digital media in modern society has led to an abundance of digital text, including websites, blogs, scientific literature, news, fan fiction, emails, UFO reports, social media posts, product reviews, forums, government reports, and even algorithmically generated text. The computational study of such textual datasets promises to offer an unprecedented window into questions of language, culture, science, politics, and contemporary society. However, realizing those promises will require contributions from the statistical, computational, and social sciences as well as from the humanities. This course offers a hands-on introduction to text mining, the art and science of analyzing textual data using computation. Students will gain the ability to work with large textual datasets using a variety of core text mining techniques including clustering, topic modeling, parsing, stylometry, and sentiment analysis. At the same time, they will also be challenged to think deeply about how to interpret the results of these computational techniques. They will learn to apply a critical lens to the methods of text mining, combining statistical, algorithmic, and critical knowledge to understand what such techniques reveal, what they obscure, and perhaps to imagine new modes of computational reading.

There are no prerequisites for this course. It is open to all majors and all necessary skills can be learned within the course. Students will learn to conduct extended text mining research projects at a level commensurate with their background and expertise. Students will also gain experience communicating their findings to the rest of the class across disciplinary lines and receiving feedback and suggestions from a range of different perspectives.

This class fulfills the requirements for the R, STS, ALP, and SS curriculum codes.

1.2 Course Objectives

By the end of this course, you will be able to:

- *Recall* the major text mining methods covered in this course.
- *Explain* how those methods work algorithmically.
- *Interpret* the results of text mining analyses.
- *Propose* new text mining research questions that can be answered by the methods discussed in the course.
- *Produce* a complete end-to-end text mining project of a novel corpus using one or more of those methods.

1.3 Structure of the Class

This course is organized around a semester-long final project. Each week will begin with students informally presenting on their latest progress and receiving questions, ideas, and suggestions from the rest of the class. There will then be a lecture on any new topics followed by a presentation and discussion of various examples of recent or classic text mining work that exemplify the new topic to generate ideas for how the topic may be applied to student projects. The discussion will then be followed by an applied lab where students learn to apply any new techniques either to their own projects or to sample data, as appropriate. Finally, each week will end with a discussion about how these new techniques might be applied to each project in order to generate ideas for discussion during the subsequent week.

The final project will broken up into a series of milestone deliverables. Rubrics detailing these deliverables will be made available as they become relevant. Because this course is open to all approaches to text mining, including scientific, literary, and cultural analysis as well as digital art, rubrics may need to be adjusted on a case by case basis, and approved with the consent of the instructor.

Collaboration is allowed and encouraged. Students may work on the same dataset and share code and other resources freely. However, each student must bring their own set of questions or objectives to the work and produce their own write-ups for the milestone deliverables. If two or more students interested in precisely the same questions wish to form a group produce a joint write-up, this is also allowed, and the details can be worked out with the instructor on a case-by-case basis.

1.4 Readings

All readings will be made available on the course Sakai in time for their assigned dates as listed on the schedule. There are no additional books to purchase.

1.5 Forums

In addition to hosting the readings, the Sakai site also hosts two forums we will use throughout the course.

1.5.1 Technical Troubleshooting Forum

In light of the fact that we are all using the same tools, and that those tools are computational in nature, it is very likely that 1) those tools will break, and 2) they will often break in similar ways for multiple people in the course. In order to best facilitate the dissemination of technical fixes and knowledge, feel free to ask and answer questions here. I will monitor this forum as well and hopefully common questions can be answered once and easily accessed by everyone.

1.5.2 Anonymous Feedback Forum

Particularly because this class is new, it may not (yet) be a well-oiled machine where everything works perfectly. If you have ideas for improving the class or want to flag issues you think might be candidates for reconsideration, please feel free to drop your thoughts here. The forum is both anonymous and moderated, so only you and the instructor will see your posts, and your identity will remain anonymous. I have found that anonymous feedback is easier to solicit than de-anonymized feedback, and I would much rather find out sooner rather than later if there is something I can improve. You can't hurt my feelings. I promise.

1.6 Disability Statement

Students with disabilities who believe that they may need accommodations in the class are encouraged to contact the Student Disabilities Access Office at 919.668.1267 or disabilities@aas.duke.edu as soon as possible to better ensure that such accommodations are implemented in a timely fashion.

1.7 Academic Integrity

Duke University is a community dedicated to scholarship, leadership, and service and to the principles of honesty, fairness, respect, and accountability. Citizens of this community commit to reflect upon and uphold these principles in all academic and non-academic endeavors, and to protect and promote a culture of integrity. To uphold the Duke Community Standard:

- I will not lie, cheat, or steal in my academic endeavors;
- I will conduct myself honorably in all my endeavors; and
- I will act if the Standard is compromised.

1.8 Preferred Contact

Please do not hesitate to contact the instructor via email (evan.donahue@duke.edu) with any questions or concerns. Expect a response within two business days of email delivery. If you do not hear a response in that time, please feel free to reach out again—it may have been eaten by the internet.

1.9 Assignments, Attendance, & Evaluation

Your grade in this class will be a function of three things: attendance, weekly presentations, and the semester project.

1.9.1 Attendance & Participation (20%)

20% of your final grade will be based on your number of unexcused absences. Medical, athletic, or other official excused absences will not count as unexcused absences. Other reasonable absences (such as for job interviews) may be excused at the instructor's discretion, and will not count as unexcused absences **as long as** arrangements are made with the instructor **prior** to the absence.

Number of Unexcused Absences	Participation Grade
0	A+
1-2	A
3-4	B
5-6	C
7-8	D
9+	F

1.9.2 Weekly Blog & Presentations (20%)

20% of your final grade will be based on weekly status updates posted on the class blog. Your update should answer three questions: what work did you do on your project this week and what if anything have you found? What are your plans for your immediate next steps? And what aspects of your project would you most like class feedback on? These blog entries need not be long and will be graded only for completion, not for content. Their primary purpose is to help you organize your thoughts before you informally present your update to the class and get feedback, suggestions, or generate other discussion. They will also help me review everyone's progress and offer help when necessary. Note that it is to be expected that sometimes technical issues will get in the way of substantive conceptual work. If you encounter technical problems you cannot solve, the course forum for technical trouble shooting should be the first place you turn, and failing that the instructor is available by email and can schedule a Zoom office hour as needed. If you still cannot resolve the issue by the beginning of the next week, come prepared to describe it in detail so the class can help you troubleshoot or devise alternative approaches to your questions if one technical route is at an impasse.

Blog entries should be posted using the Sakai Blog tool and labeled with the appropriate week (eg Week 1). Blog entries are due at 11:59 PM every Monday.

Number of Missed Blog Posts	Reading Response Grade
0	A+
1-2	A
3-4	B
5-6	C
7-8	D
9+	F

1.9.3 Final Project (60%)

60% of your final grade in this course will be based on the completion of a semester long text mining project, which will be divided into 5 "milestones" that will be graded separately and used to compute your final score for the overall project. The project must be some type of novel computational analysis, interpretation, or production using an unstructured, real world corpus, but within that framework, the specifics of the type of corpus and type of analysis can vary according to the nature of the material and the student's prior technical expertise.

Like attendance and the reading responses, these milestones will be graded by contract. This means that no individual component, including the final draft, will receive an evaluative score. Instead, if you hand in a draft that meets the requirements for that milestone (which will be outlined at appropriate times throughout the semester), you will receive full credit for it, as well as detailed feedback that pertains to building on that milestone in preparation for the next one. Assignments that are late will receive no credit. One caveat is that in order to complete a milestone, the instructor must *accept it* as meeting the requirements of that particular milestone. If a draft does not meet its requirements, you may receive a "revise and resubmit" along with comments about how to fix the write-up. As long as the initial draft was handed in on time, you may revise as many times as is necessary to reach an "accept," at which point you will receive full credit for the milestone. You should therefore have no fear of trying something creative or experimental.

Number of Milestones Completed	Final Project Grade
5	A+
4	A
3	B
2	D
0-1	F

2 Schedule

2.1 Unit 1 - Language as Data

This unit introduces the field of text mining. In this unit, students will see a range of examples text mining work from various fields to illustrate what text mining can be used for and to serve as inspiration for their own projects. We will cover the tools needed to obtain datasets and prepare them for exploratory analysis in the next unit. By the end of the unit, students will have obtained a dataset and prepared it for analysis.

2.1.1 Week 1 (Jan 21): Introduction

2.1.2 Week 2 (Jan 26 & 28): Data Collection

- **Reading:** Janssens, *Data Science at the Command Line*. Ch 2.
- **Milestone 1 (Feb 3):** Project Proposal

2.1.3 Week 3 (Feb 2 & 4): Data Cleaning

- **Reading:** Rawson & Muñoz, "Against Cleaning"
- **Milestone 2** (Feb 8): Data Preparation

2.2 Unit 2 - Exploratory Data Analysis

This unit introduces the basic tools of exploratory data analysis for language data. Students will learn to work with data, produce informative statistics and visualizations, and begin to ask and answer questions using text mining. By the end of the unit, students will have command of a range of tools for working with their datasets and be prepared to use them to facilitate and supplement the more advanced forms of analysis in the next unit.

2.2.1 Week 4 (Feb 9 & 11): Data Exploration & Visualization

- **Reading:** Tukey, *Exploratory Data Analysis*. Ch 1.

2.2.2 Week 5 (Feb 16 & 18): Natural Language Processing

- **Reading:** Tukey, *Exploratory Data Analysis*. Ch 2.

2.2.3 Week 6 (Feb 23 & 25): Sentiment Analysis & Classification

- **Reading:** Jurafsky et al., "Linguistic Markers of Status in Food Culture: Bourdieu's Distinction in a Menu Corpus"

2.3 Unit 3 - Advanced Text Mining Techniques

This unit introduces more advanced statistical techniques that summarize, cluster, and classify texts. Students will learn the theoretical underpinnings of these techniques and gain an intuition for applying them and interpreting their results using the more basic forms of exploratory data analysis from the previous unit. By the end of the unit, students will have subjected their datasets to various forms of analysis as appropriate for their data and will begin to build an argument based on their findings in preparation for the final write up.

2.3.1 Week 7 (Mar 2 & 4): Information Retrieval & Summarization

- **Reading:** Roland, "Topic Modeling: What Humanists Actually Do With It"

2.3.2 Week 8 (Mar 11): Classification & Domain Adaptation

- **Reading:** Binder, "Alien Reading: Text Mining, Language Standardization, and the Humanities"

2.3.3 Week 9 (Mar 16 & 18): Topic Modeling

- **Reading:** Hoover, "Word Frequency, Statistical Stylistics and Authorship Attribution"

2.3.4 Week 10 (Mar 23 & 25): Many-to-Many Network Analysis

- **Reading:** Schöch, "Principal Component Analysis for Literary Genre Stylistics"

2.4 Unit 4 - Special Topics

This unit will cover special topics associated with text mining. The precise topics covered will vary in order to showcase other possibilities than those that have been yet explored in student projects. Although topics covered here may apply to student projects, they are primarily meant to broaden students' sense of where and how text mining might be applied. For the end of the unit, students should focus on pursuing the questions that have emerged from their analyses in the previous unit in preparation for the final write up.

2.4.1 Week 11 (Mar 30 & Apr 1): Data Visualization

- **Reading:** Bamman, "Validity"
- **Milestone 3 (Apr 5):** Preliminary Analysis

2.4.2 Week 12 (Apr 6 & Apr 8): Association Rule Mining

- **Reading:** Swanson, "Undiscovered Public Knowledge"

2.4.3 Week 13 (Apr 13 & 15): Social Network Analysis

- **Reading:** Dekker et al., "Evaluating named entity recognition tools for extracting social networks from novels"

2.4.4 Week 14 (Apr 20 & 22): Final Presentations

- **Reading:** TBA
- **Milestone 4 (Apr 22):** Draft Project Write-up
- **Milestone 5 (Apr 29):** Final Project Write-up Due