Introduction to Computational Text Analysis

Jae Yeon Kim

Updated: 2021-04-29

Overview

This advanced course introduces students to computational text analysis. Text is everywhere, from open-ended questions in surveys to newspaper articles to social media posts. Recent developments in natural language processing and machine learning allow us to collect and analyze text data faster, easier, and at scale. This course will teach you about these tools and techniques and their promises and limitations in social science applications. In addition, we will explore how we can be more creative about taking the text-as-data approach by treating the computational text analysis as one component of a larger research project.

The course is divided into four sections:

- We will learn the promises and limitations of the text-as-data approach.
- We will learn how to collect a large volume of text data using web scraping and application programming interfaces.
- We will learn how to analyze text data using natural language processing and machine learning.
- We will learn how to combine the text-as-data approach with other means of collecting and analyzing qualitative and quantitative data.

Objectives

By the end of the semester, students will create a data science project using computational text analysis. They will learn what aspects of research questions they can or can't answer using computational text analysis.

- Students will *describe* what the text-as-data approach is and its limitations and promises.
- Students will *understand* the essential data science workflow applied to computational text analysis.
- Students will *practice* collecting text data using web scraping and an application programming interface.
- Students will *practice* analyzing text data using natural language processing and machine learning.
- Students will *create* a data science project using computational text analysis.

Logistics

Contributors

- Instructor: Jae Yeon Kim

Time and location

Office hours

Office hours will be held ...

Slack & GitHub

- Slack for communication (announcements and questions). It would be best if you asked questions about class material and assignments through the Slack channels so that everyone can benefit from the discussion. I encourage you to respond to each other's questions as well.
- GitHub for course materials. All course materials will be posted on GitHub at ..., including lecture notes, code demonstrations, sample data, and assignments.

Accessibility

This class is committed to creating a safe and inclusive environment in which everyone can participate. If you have a particular concern (e.g., disability), please come to me as soon as possible (ideally within the first two weeks) so that I can make special arrangements.

It's Okay Not to Know

Asking questions is your privilege in an academic environment (we are all here to learn). There is no such thing as a stupid question.

Audting

No auditing is permitted.

Late policy

No late assignments are accepted.

Course requirements and grades

Students will form a team of 3-5 people. Note that all of the following activities are **individual-based.** You are encouraged to work in team, but you should submit your assignments and final project independently.

I will assign you to teams based on the responses to the survey, which I will circulate at the beginning of the semester. The assignment will be based on an algorithm that optimizes diversity among team members.

Students must complete two take-home assignments, one focusing on digital data collection (due week 4) and the other focusing on computational text analysis (due week 8). You are also required to submit a one-page research proposal by week 5. Once approved by the instructor, you will present the key aspects of your final project in the form of a lightning talk by weeks 8-10. After incorporating the feedback received from the instructor and other students, you will hand in the final project by week 12.

This is a graded class based on the following:

- Participation (10%)
- First assignment: data collection (15%)
- Second assignment: data analysis (15%)
- Final project proposal (10%)
- Pitch (10%)
- Final project (50%)

Class participation The class participation portion of the grade can be satisfied in one or more of the following ways:

- attending the lectures
- asking and answering questions in class

- contributing to class discussion through the Slack workspace

Assignments

1. Digital data collection assignment (due **week 4**)
2. Computational text analysis assignment (due **week 8**)

Research proposals and final projects

1. Short research proposal (due **week 5**): You will submit a one-page description of your final project. For more information, see this guideline. I will provide a template for the final project soon.
2. Lightning talk (elevator pitch) (due **weeks 8-10**): You will present your project in a maximum of 5-minute talk, with 5 minutes for class Q&A. When you prepare slides, think about spending one minute for each slide. For more information, see this guideline.
3. Final project (due **week 12**): You will submit the final project applying the framework, tools, and techniques we learned in class to your problem of interest.

Computer requirements This course is hands-on (meaning you need to type and run things using a programming language). We will learn not only the ideas of computational text analysis but also how to do it step-by-step. For this reason, you are required to install the following programs:

- Access to the UNIX command line (e.g., a Mac laptop, a Bash wrapper on Windows)
- Git
- R and RStudio (latest versions)
- Anaconda and Python 3 (latest versions)
- Pandoc and LaTeX

All of the required software is **free** and should be installed in the first week of the class.

See install.md for more information.

Course schedule

- Week 1: Text-as-Data: Promises and Limitations
- Week 2: Parsing Semi-Structured Data at Scale
- Week 3: Web Scraping
- Week 4: Application Programming Interface
- Week 5: Couting Texts at Scale
- Week 6: Modeling Texts Using Algorithms
- Week 7: When We Know Labels: Text Classification
- Week 8: When We Don't Know Labels: Topic Modeling
- Week 9: Text Networks and Word Embeddings
- Week 10: How to Combine Text and Other Data Analysis
- Week 11: Reading Week

- Week 12: Final Project Due

Questions, comments, or suggestions

Please create issues if you have questions, comments, or suggestions.

Acknolwedgement

The course structure is influenced by Chris Bail's Text as Data course. I have also cited all the other references in the course materials whenever I am aware of related books, articles, slides, blog posts, or YouTube video clips.