Text-as-Data

DACSS 697D
University of Massachusetts Amherst
2022

Contact Information

Dr. Eunkyung Song
Data Analytics and Computational Social Science
University of Massachusetts Amherst
E-mail: eunkyungsong@umass.edu

Course Description

With the recent explosion in availability of digitized text, social scientists increasingly are turning to computational tools for the analysis of text as data. In this three credit course, students will first learn how to convert text to formats suitable for analysis. From there, the course will introduce and proceed through tutorials on a variety of natural language processing approaches to the treatment of text-as-data. This will include relatively simple dictionary approaches for measurement, supervised learning approaches for document classification, vector representations, contextualized embeddings, and more.

Course Objectives

- Equip students to become knowledgeable consumers of text data research, capable of critically analyzing research that employs text data and text-as-data techniques.

- Provide students with the tools to design and complete basic and advanced text-as-data research, from converting text to formats appropriate for analysis to estimating text-as-data models.

- Develop students ability to work individually and collaboratively on subjects with important real-world relevance.

- Enable students to communicate — both orally and in written format — clearly and appropriately the results or shortcomings of text-as-data research.

Text

There is no required text for this course. All required readings are posted to the course website. This syllabus outlines general areas of study throughout the semester, as well as listing specific reading assignments on a daily basis. It is vital that you keep current with the readings, as they will provide the basis for in-class lectures and discussions.

Grading

Grades are calculated as follows:

Modules & Quizzes (30%)

Blog Posts (30%)

Research Project (30%)

Participation (10%)

Modules & Quizzes Students are required to complete a series of 10 short interactive modules on line through Google Colab. Each will familiarize students with a different text-as-data approach, and how to implement said approach in R. After completing the module, students will complete a short quiz. Quizzes are primarily graded on completeness rather than having correct responses to each question; therefore, it is imperative that you complete each quiz.

Blog Posts Students are required to complete blog posts every other week during the course of the semester. The blog posts should detail your progress working with text-as-data with the corpus that you are using for your research project. As such, they should serve as ongoing documentation of (a) your growing expertise with text-as-data, and (b) your growing progress on the final project. Your first blog post should detail your general interests and the research question(s) you plan to explore. After that, blog posts should simply reflect the material for that week. Therefore, you might detail in one blog post your experience with getting the corpus for your final project into R and formatted to your liking. The blog posts could include plots and screen caps of successful work, but might also include details on unforeseen challenges, terrifying error messages, or just what you think is an unbelievably ugly plot you mistakenly created. The goal is to document and reflect on your progress.

Research Project The class is tailored around aiding you in producing a research paper appropri ate for submission and presentation at an academic conference. Students are expected to complete an original research project that features both a strong theoretical grounding, research design, and original analysis that features a text-as-data component.

Participation As a group, we will meet for lecture every Tuesday. During this time, we will cover the central concepts for that week and discuss examples of research engaging in that type of text as-data analysis. Students must contact the professor *before* the class session they will miss in order to ensure the absence is excused. In class, students are expected to participate regularly, and participation should reflect careful consideration of the readings and topic.

Final letter grades are assigned using the University's Plus-Minus Grading Scale according to following rubric:

A (94-100%)

A- (90-93%)

B+ (86-89%)

B (81-85%)

B- (77-80%)

C+ (74-76%)

C (70-73%)

F (Below 70%)

Software

Students in this class will use R and RStudio. The software is free and available online; the course website includes a guide for installing both on your machine. The course assumes no familiarity with the R programming language, though that is helpful. We will also discuss and utilize Python but at a much smaller scale.

Academic Honesty

Since the integrity of the academic enterprise of any institution of higher education requires honesty in scholarship and research, academic honesty is required of all students at the University of Massachusetts Amherst.

Academic dishonesty is prohibited in all programs of the University. Academic dishonesty includes but is not limited to: cheating, fabrication, plagiarism, and facilitating dishonesty. Appropriate sanctions may be imposed on any student who has committed an act of academic dishonesty. Instructors should take reasonable steps to address academic misconduct. Any person who has reason to believe that a student has committed academic dishonesty should bring such information to the attention of the appropriate course instructor as soon as possible. Instances of academic dishonesty not related to a specific course should be brought to the attention of the appropriate department Head or Chair. The procedures outlined below are intended to provide an efficient and orderly process by which action may be taken if it appears that academic dishonesty has occurred and by which students may appeal such actions.

Since students are expected to be familiar with this policy and the commonly accepted standards of academic integrity, ignorance of such standards is not normally sufficient evidence of lack of intent.

For more information about what constitutes academic dishonesty, please see the Dean of Students' website:
http://umass.edu/dean_students/codeofconduct/acadhonesty/

Statement on Disabilities

The University of Massachusetts Amherst is committed to making reasonable, effective and appropriate accommodations to meet the needs of students with disabilities and help create a barrier free campus.

If you are in need of accommodation for a documented disability, register with Disability Services to have an accommodation letter sent to your faculty. It is your responsibility to initiate these ser vices and to communicate with faculty ahead of time to manage accommodations in a timely man ner. For more information, consult the Disability Services website at http://www.umass.edu/disability/.

Course Schedule

Daily reading assignments are listed in parentheses. Note that reading assignments are listed according to the day on which the subject matter will be discussed; they should therefore be read prior to that date.

Week	Topic	Assignments
1	Introduction	Tutorial 1 & Quiz 1
2	Using Text as Data	Tutorial 2 & Quiz 2 Blog Post 1
3	Acquiring Texts: Scraping & APIs	Tutorial 3
4	Natural Language Processing	Tutorial 4 & Quiz 3 Blog post 2
5	Preprocessing	Tutorial 5 & Quiz 4
6	Representing Texts 1: Bag(s) of Words	Tutorial 6 & Quiz 5 Blog post 3
7	Representing Texts 2: Word Embeddings	Tutorial 7 & Quiz 6
8	Dictionary Approach and Sentiment Analysis	Tutorial 8 & Quiz 7 Blog post 4
9	Supervised Learning (1)	Tutorial 9 & Quiz 8
10	Supervised Learning (2)	Blog post 5
11	Unsupervised Learning (1)	Tutorial 10 & Quiz 9
12	Unsupervised Learning (2)	Blog post 6

13	Causal Inference	
14	Practicum and Presentation	