Syllabus: Textual data analysis with R

Linguistics 460 (Moreton)*

2024 August 19 (M)

Time:	MW 3:35–4:50
Places:	Global Education 1005 or: Zoom: Meeting ID: 963 8032 6331, passcode: 543792[a] *This is an in-person class. Zoom links are provided in case an emergency forces us to switch to remote instruction.*
Instructor	Elliott Moreton, `moreton@unc.edu`
Office hours:	(Tentatively:) T 2–3, F 2:30–3:30, and by appointment
Office:	Smith 101 or: Zoom: Meeting ID 973 4347 1746, passcode 531096[b]
Textbooks:	Navarro, *Learning statistics with R: a tutorial for psychology students and other beginners*[c] Sanchez, *Handling and processing strings in R*[d] Silge and Robinson, *Text mining with R: a tidy approach*[e] Wickham, Çetinkaya-Rundel, and Grolemund, *R for data science*, 2nd ed.[f] *These books are available free on line.*

[a] `https://unc.zoom.us/j/96380326331?pwd=MWl6MHo4S2RVT214UlIrOEZFUFVLQT09`
[b] `https://unc.zoom.us/j/97343471746?pwd=cGRMUElraDAyRVY4UmcxT1RFcmxiZz09`
[c] `https://open.umn.edu/opentextbooks/textbooks/559`
[d] `https://gotellilab.github.io/Bio381/Scripts/Feb07/HandlingAndProcessingStringsInR.pdf`
[e] `https://www.tidytextmining.com/`
[f] `https://r4ds.hadley.nz/`

This is an in-person class. Zoom links are provided in case an emergency forces us to switch to remote instruction.

1 Target audience, course goals, and learning objectives

LING 460 is intended to introduce methods for working with textual data (corpora, databases, etc.) to an audience of linguists and others interested in using linguistic data to test hypotheses. The main topics include

- The R programming language[1]
- Statistical hypothesis testing
- Processing data in text form

These include both conceptual knowledge and practical skills. By the end of the course, students will have learned:

[1] `https://cran.r-project.org/index.html`

- How to organize, structure, and clean textual data
- How to use data-visualization tools to find patterns in data
- How to use regular expressions to search for patterns in text data
- The logic behind some of the most commonly-used statistical tests, and how to use them
- How to do some basic text-mining tasks (e.g., sentiment analysis, n-gram analysis, document classification)
- How to use classification models to make predictions based on patterns in your data

The course will culminate in a student-designed data project.

2 Prerequisites and requirements

There are no prerequisites. This class is designed for people with little or no experience in programming, statistics, or linguistics. **Those with extensive background in statistics or computer science are encouraged to take a more advanced course** so that (a) they can make the most efficient use of their limited time in college, and (b) actual beginners who are on the waiting list for this class can enroll in it.

LING 460 fulfills the following requirements:

- Research and Discovery requirement in IDEAS In Action
- Data Science minor core requirement (similar to STOR 120)
- Cognitive Science minor requirement
- Elective credit for Linguistics major and minor

3 Approximate semester schedule

Unit 1 (Weeks 1–5): R basics. Descriptive statistics.	R concepts: Data types, data frames, functions, base R graphics. Statistical concepts: probability distributions, sampling theory, confidence intervals.
Unit 2 (Weeks 6–8): Statistical hypothesis testing.	R concepts: branches, loops, iteration. Statistical concepts: Central Limit Theorem, logic of frequentist hypothesis testing (null-hypothesis significance testing), z-, t-, and χ^2-tests.
EXAM 1 (midterm)	
Unit 3 (Weeks 9–11): Working with text data.	R concepts: tidy format, document-term matrices. Text-mining concepts: term frequency and tf-idf, Zipf's Law, lexical diversity, sentiment analysis, topic modelling (time permitting)
Unit 4 (Weeks 12–14): Classification models (dictionary-based, naive Bayes, LASSO)	Work on final projects
Project presentations (week before Finals Week)	
EXAM 2 (final)	

4 Normal weekly schedule

Here's what will happen in a normal week:

	Monday	Wednesday
⋮	⋮	⋮
Week n	Lab n due Reading n assigned	Start Lab $n+1$ Quiz n due
Week $n+1$	Lab $n+1$ due Reading $n+1$ assigned	Start Lab $n+2$ Quiz $n+1$ due
⋮	⋮	⋮

In some weeks, there won't be a lab, or there won't be a quiz, so the serial numbers on the labs and quizzes will eventually get out of alignment with the serial number of the week.

Some weeks are abnormal, in that they only have a Monday or only have a Wednesday.

5 Class in the time of coronavirus

As I write this, the University's plan is to remain open for a full 15-week semester, and to hold classes in person. If the virus situation deteriorates, we may go to all-Zoom classes, at the Zoom link on p. 1 of this syllabus. Otherwise, we will meet in person and the class will not be streamed or recorded.

If you are too ill (with Covid or otherwise) to attend class safely, please contact me in advance to arrange to attend by Zoom. Others will not be admitted into the Zoom room.

6 Where to find course components

The main tools we will be using to communicate in this course are the following:

1. The class log[2], on the World Wide Web, updated after each class. Here you will find
   (a) A brief outline of what was covered each day
   (b) A list of any assignments made that day
2. The Canvas site[3]. Everyone who is enrolled in the class should already have access to it. Our class's ID, if you need it, is 65589. The main things we will need there are
   (a) Course materials like slides, handouts, and readings (under `Modules`)
   (b) The place to pick up lab assignments (under `Assignments`).
   (c) The place for on-line quizzes (under `Quizzes`)
   (d) A discussion forum for asynchronous collaboration (under `Discussions`)
   (e) The gradebook (under `Grading`)

[2] `http://users.castle.unc.edu/~moreton/Lin460/460log.html`
[3] `http://canvas.unc.edu`

3. The Zoom meeting links (see p. 1 of the syllabus). If Zoom is not already installed on your computer, please go to `zoom.unc.edu`[4] to get it. *This is an in-person class. The Zoom links are provided in case an emergency forces us to meet on-line.*

7 Course components and grading

Numeric grades will be converted to UNC's letter-grade system by mapping the numeric range from 60 to 100 onto the 10 passing letter grades from D to A, with four numeric points per step (except that A has 5 points, 96 to 100).

Reading quizzes (10%) These are meant to track progress in doing and understanding the readings and attending class. They will be done through the `Quizzes` section of the Canvas site, i.e., on-line, with two attempts and unlimited time. Normally, these will be due on Wednesday before class.

Lab homeworks (20%) There will be about 8 of them, spread over the first 2/3 of the semester. They consist of small programming (coding) projects. They will normally be assigned on Wednesday, when we will start working on them *in class*, and due Monday before class.

Each lab homework will be graded pass/fail. "Pass" means "made a serious attempt to answer each question correctly, even if the answer turned out to be wrong". *Simply putting superficial and obviously wrong answers will not count.*

You may drop or miss *one* of the lab homeworks.

Attendance and participation (10%) You are expected to come to class in person, having done the reading and thought about it until either (a) it makes sense, or (b) you can express precisely what about it doesn't make sense; either way, you'll have something to talk about in class.

If I start getting the impression that people aren't doing the readings, I'm going to institute pop quizzes. These are annoying because they waste class time, but coming to class without having done the reading wastes even more class time.

Missing classes will make it hard to keep up. It will also lower your participation grade. If you miss a class, it is your responsibility to get missed materials from me or other students. Always check the website if you have been absent.

Exams (30%) There will be two, a midterm and a final, weighted equally and both cumulative from the beginning of the course. They will be done during the scheduled class or final-exam time. They will be open-book and open-computer, and will require using R.

Final project (30%) In the last five weeks of the class, students, working in groups of three, will choose a research question, then design, execute, and analyze an experiment to answer it, and finally present the question and the results to the class. This will take place in several steps, and I'll be giving details as each one comes up. All members of the group must contribute to coding and analysis. The grade will have a group component and an individual component.

[4] `http://zoom.unc.edu`

8 Policies

Attendance. If you must miss class because of a medical or family emergency, you should let me know *beforehand* by emailing or buttonholing me in person. If you miss a class, it is your responsibility to get missed materials from me or other students. Always check the class log and the Canvas website if you have been absent.

Late assignments. Homework solutions will normally be discussed in class the day the assignment is due. Therefore, as a general rule, NO LATE ASSIGNMENTS WILL BE ACCEPTED FOR CREDIT. To accommodate for emergencies, one lab-homework grade will be dropped (as noted above in the section about lab homeworks). For anything more than that, please see the University Approved Absences Office[5] and let me know by email as soon as you are able to arrange alternatives.

Collaboration and citation. It is a really good idea to discuss assignments with other people in the class and solve the problems together. However, each person (or each partnership) should write up their solution alone. If you work with other people, or look up information in sources that aren't officially part of this course, you are required to acknowledge them in the writeup. There is no shame in collaborating with other people, or in digging out information independently, but you need to give credit where it is due.

Generative artificial intelligence. "People" in the preceding paragraph means "humans". Use of generative AI for this class is governed by the guidelines developed by UNC-CH's Generative AI Committee, which can be found at this link[6] starting at the heading "Syllabus guidelines for generative AI" and continuing through the end of the page. The instructor (and the Honor Court, if it comes to that) will expect you to have read these guidelines. They apply to every aspect of the course, as long as they are not superseded by explicit written instructions from the instructor in an assignment.

Recording. Permission to make audio or video recordings of class will be given only in special circumstances (e.g., to students with hearing impairments). If you don't have written permission from me, you should not make audio or video recordings.

Dates are still tentative at this point. I'll give at least two weeks' notice of the midterm, and will hand out an exam syllabus (a study guide) one week before each exam.

The Carolina Honor Code[a] is in effect in this class, and I will treat violations seriously. You should review it. If you have questions about interpretation, you should bring them to me.

[a] `http://instrument.unc.edu`

9 General UNC-CH course policies and resources

Accessibility Resources UNC-Chapel Hill facilitates the implementation of reasonable accommodations for students with learning disabilities, physical disabilities, mental health struggles, chronic medical conditions, temporary disability, or pregnancy complications, all of which can impair student success. See the ARS website for contact and registration information: `https://ars.unc.edu/about-ars/contact-us`

Attendance Policy No right or privilege exists that permits a student to be absent from any class meetings, except for these University Approved Absences:

[5] `https://uaao.unc.edu/`

[6] `https://provost.unc.edu/student-generative-ai-usage-guidance/`

1. Authorized University activities
2. Disability/religious observance/pregnancy, as required by law and approved by Accessibility Resources and Service and/or the Equal Opportunity and Compliance Office(EOC)
3. Significant health condition and/or personal/family emergency as approved by the Dean of Students, Gender Violence Service Coordinators, and/or the Equal Opportunity and Compliance Office (EOC).

Absences for reasons other than those listed above must be approved in advance by the instructor. If a student misses class for any reason, the student is responsible for finding out what material has been missed and is encouraged to speak to the instructor.

University Testing Center The College of Arts and Sciences provides a secure, proctored environment in which exams can be taken. The center works withinstructors to proctor exams for their undergraduate students who are not registered with ARS and who do not need testing accommodations as provided by ARS. In other words, the Center provides a proctored testing environment for students who are unable totake an exam at the normally scheduled time (with pre-arrangement by your instructor). For more information, visit `http://testingcenter.web.unc.edu/`.

Counseling and Psychological Services CAPS is strongly committed to addressing the mental health needs of a diverse student body through timely access to consultation and connection to clinically appropriate services, whether for short or long-term needs. Go to their website: `https://caps.unc.edu/` or visit their facilities on the third floor of the Campus Health Services building for a walk-in evaluation to learn more.

Honor Code Statement Students are bound by the Honor Code in taking exams and in written work. The Honor Code of the University is in effect at all times, and the submission of work signifies understanding and acceptance of those requirements. Plagiarism will not be tolerated. Please consult with the instructor if you have any questions about the Honor Code.