Special Topics: Text Analysis

36-468/668 · Fall 2019 · TR · 10:30-11:50 · SH 125

Instructor Information:	David Brown dwb2@andrew.cmu.edu Baker Hall 245N
Office hours:	TR 1:00-2:30 I am happy to meet at other times—just email for an appointment
TA Information:	Nil-Jana Akpinar nakpinar@andrew.cmu.edu TBD
	Benjamin LeRoy bpleroy@andrew.cmu.edu TBD
Office hours:	Nil-Jana: TBD Benjamin: TBD

Overview:

The analysis of language is concerned with how variables relate to people (their gender, age, and location, for example), how variables relate to use (such as writing in different academic disciplines), and how variables change over time. While we are surrounded by data that might potentially shed light on many of these questions, working with real-world linguistic data can present some unique challenges in sampling, in the distribution of features, and in their high dimensionality. In this course, we work through some of these issues, paying particular attention to the aligning of the statistical questions we want to investigate with the choice of statistical models, as well as focusing on the interpretation of results. Analysis will be carried out in R and students will develop a suite of tools as they work through their course projects.

Course Goals:

In this class, you will learn

- A range of statistical methods related to linguistic analysis and how to effectively pair those methods with research questions (i.e. to know when/why a particular method will/will not work based on research goals and/or data type)
- How to processes text in R
- The basic design principles of quantitatively based linguistic research
- How to develop a project and report it clearly and rigorously for a variety of audiences

Basic Expectations:

About the project: This is a hands-on, project-based class. While there will be shorter homework assignments (particularly early in the term as we build our familiarity with methods, concepts, and tools), you should also be thinking about and planning your project very early on. Working with linguistic data has a bit of a different rhythm than what you might be used to. It takes time to gather data; it takes time to get to know it; and it takes time to figure out what story it tells.

About the grammar: This is not a linguistics class. That said, you will be encountering linguistic terminology and concepts during the course. The goal is for you to develop your statistical reasoning, problem solving, and facility in effectively applying models and techniques to your research questions. Doing so is going to require you to understand some of the fundamental properties of linguistic data. For example, it isn't likely that you'll need to know off-hand what an object predicate is. However, you might use a dependency parser to extract noun phrases from a corpus of texts, for example. In that case, you should have a good understanding of what noun phrases are and why they might be interesting.

About the code: All of the coding in the course will be carried out in R. Often in statistics, we are provided reasonably well-formed data and questions. Our task is to engage those questions, applying the appropriate models. R is ideally suited for this kind of thing. We may need to "push around" our data and transform our data structures, but R does that very efficiently. Language data are different. Instead of data tables, we are usually presented with some combination of metadata and text. This requires us to process the latter – to turn text into variable measures (usually word counts, but not always). In this class, we will learn some Natural Language Processing tools and techniques. These, too, will be carried out in R. However, if you have an interest or background in other programing languages (particularly Python and C++), you may want to write/make use of those languages' functionalities. If you choose to do so, you must integrate those functions into your final R code so that your code can be run and tested. (Packages like *reticulate* and *Rcpp* make such integration relatively easy.) At the end of the term, you should have suite of NLP tools, which you can use in the future.

Class Schedule:

Our daily class schedule will be maintained on Canvas, where you will also find class readings and materials for daily assignments. Check Canvas on a regular basis.

Attendance and Lateness:

Learning how to work with corpus data and corpus tools is most effectively done by just digging in. With that in mind, much of what we do in class will involve you actively working with R and data. You will, therefore, need to bring a laptop to class or, if that is not possible, share one with a colleague.

Course Materials:

Statistics in Corpus Linguistics is required for the course. Brezina's text will provide us some important background and will serve as a useful jumping off point. Because the book isn't necessarily targeted for statisticians, it will be supplemented by additional readings. Those readings will be posted as PDFs or links on the Canvas course web site. You should always bring copies (paper or electronic) of the readings to class.

Work Commitments

Project:

Much of the class, particularly after mid-term, will be devoted to helping you design and execute a research project of your choosing. For this assignment, you will analyze a corpus of texts (at least part of which you have collected yourself) using some combination of statistical methods. The specifics will depend on the research questions that you are interested in pursuing, thus being able to explain and justify your statistical and methodological decisions will be important. The final write-up will be submitted in two parts. The first will be a 6-page report including an executive summary. The second will be your R code in a markdown file. You may collaborate on the second (see below), but the report is to be submitted individually.

Coffee-break experiments:

These mini-projects are inspired by Kibbitzers, which originated with Tim Johns at the University of Birmingham. A Kibbitzer explores a specific linguistic question or problem that emerges from the data we work with in class, from questions that arise in discussion, or from your own readings or curiosity. Mike Scott has preserved the UB Kibbitzer here. This assignment is also inspired by Mark Liberman's Breakfast Experiments™, which he publishes on Language Log. In the spirit of these, yours will take the form of short examinations of a linguistic feature or structure that you think is interesting. I would encourage you to look at the links above, and I will put up examples for you on Canvas. The length of these will vary, should be in very condensed IMRD form (~ 500 words). These assignments are intended to give you some space to practice and try things out, which can sometimes mean interesting failure. A failed experiment can give us just as much information as a successful one. With that in mind, don't be afraid if something doesn't work in the way you predict. The most important piece of your write-up in that case would be your Discussion: What went wrong? And why?

All coffee-break experiments are to be submitted as R markdown files (.Rmd) and corresponding PDFs on Canvas.

Midterm Mini:

Around the mid-term, you will be provided some textual data and a question. You will have a week to conduct your analysis and draft a 4-page (maximum) report. You may consult any of the course materials or other reference materials, but not other people.

Collaboration:

Outside of the classroom setting, research of the kind we'll be pursuing in this class is often collaborative. I want to provide you the option of working through the collaborative process if you so wish. Thus, you have the option of teaming up with one (and only one) colleague on the final project. For that project, you may combine forces in the building of a corpus (or sub-corpus) and in developing the R code needed to carry out your data processing and statistical procedures. However, you must write individual reports. That means you need to come up with separate (though perhaps related) research questions.

Grading:

Your work in this class will be weighted as follows:

Coursework/Participation	20%
Coffee-Break Experiments	15%
Midterm Mini	20%
Final Project	45%

Other Policies

Late Policy:

Work submitted late is subject to penalty. If for any reason you cannot meet the deadline for an assignment, contact me in advance so we can figure out how to best and most fairly handle your situation.

Resources

Communication Tutoring at the GCC:

You can receive one-on-one tutoring for any academic communication project for any class at the Global Communication Center (GCC). The GCC is a free service located in Hunt library. You can schedule a tutoring appointment directly on the website, or you can drop by first floor of the Hunt library for a walk-in appointment. See http://www.cmu.edu/gcc for appointments and information.

Accommodations for Students with Disabilities:

If you have a disability and have an accommodations letter from the Disability Resources office, I encourage you to discuss your accommodations and needs with me as early in the semester as possible. I will work with you to ensure that accommodations are provided as appropriate. If you suspect that you may have a disability and would benefit from accommodations but are not yet registered with the Office of Disability Resources, I encourage you to contact them at access@andrew.cmu.edu.

Important On-Campus Resources:

All of us benefit from support during times of struggle. You are not alone. There are many helpful resources available on campus, and an important part of the college experience is learning how to ask for help. Asking for support sooner rather than later is often helpful.

If you or anyone you know experiences any academic stress, difficult life events, or feelings like anxiety or depression, we strongly encourage you to seek support. Counseling and Psychological Services (CaPS) is here to help: call 412-268-2922 and visit their website at http://www.cmu.edu/counseling/. Consider reaching out to a friend, faculty or family member you trust for help getting connected to the support that can help.

If you or someone you know is feeling suicidal or in danger of self-harm, call someone immediately, day or night:

CaPS: 412-268-2922
Re:solve Crisis Network: 888-796-8226

If you or someone you know is the victim of sexual misconduct or assault:

CMU Title IX Office: 412-268-7125, http://www.cmu.edu/title-ix/reporting/index.html
PAAR (Pennsylvania Action Against Rape): 1-888-772-7227,
http://www.pcar.org/help-pa/victim-support

If the situation is life threatening, call the police:

On campus: CMU Police: 412-268-2323
Off campus: 911

If you have questions about this or your coursework, please let me know.