1 The Basics

1.1 Overview

Never before in human history has so much information been so easy to access. The promise of this wealth of information is immense, but because of its pure volume it is difficult to summarize and interpret. However, a burgeoning array of algorithms and statistical methods are beginning to make analysis of this information possible. These new forms of data and new statistical techniques provide opportunities to observe behavior that was previously unobservable, to measure quantities of interest that were previously unmeasurable, and to test hypotheses that were previously impossible to test.

In this course we will introduce a social science logic for how text can be included in every stage of the research process. Our goal is to describe the prevalence of a social behavior or phenomenon and make inferences about its origins. We explain how the abundance of text and new statistical methods facilitate these inferences. The goal of inference in social science research is qualitatively different than the goals that have been often used to evaluate text analytic methods, which often focus on performing a specific task. The focus on inference will push us to reconsider when and how some methods are useful, suggest new ways to evaluate methods, and will present new open questions in the use of text as data.

This ten week course is organized around three large components of the research process: discovery, measurement, and testing. Discovery is the process of hypothesis generation and often where scholars begin a research project. We will discuss methods that suggest ways of organizing texts that are specific to the task of discovery and help the researcher through the process of understanding the contours of the data. Measurement is the process of capturing the degree

---

*The development of this course has been influenced by my frequent collaborators: Justin Grimmer, Brandon Stewart and Dustin Tingley. Last Edited: March 31, 2019*
or extent of some behavior. We will introduce methods specifically focused on measurement, but also explain how we modify methods used for discovery to methods that are more specific to the goal of measurement. Testing is the least established area, in which we use text for prediction and causal inference. Text provides an opportunity for granular causal inferences and opens a wide range of questions previously impossible. The pairing of causal inference and text has produced new methodological questions that we highlight and provide initial answers to.

The goal of the course is to provide students with an overview of the literature while developing an understanding of what is possible. While the time scale does not permit a deep mathematical understanding of every approach, students will learn tools for analyzing texts quantitatively and a framework of how these methods can be used for social science inference.

1.2 Prerequisites

The most important prerequisite is a willingness to work hard on possibly unfamiliar material. Students with a prior course which covers maximum likelihood estimation or Bayesian inference will be most comfortable with the statistical material. Students also should have some basic background in computer programming. Others should consult the instructor. Gary King’s *Unifying Political Methodology* and Larry Wasserman’s *All of Statistics-A Concise Course in Statistical Inference*, in addition to Murphy’s *Machine Learning* listed below, offer excellent primers on the MLE framework underlying some of these methods.

1.3 Auditing

We are happy to have auditors in the class as space permits but we ask that auditors complete a subset of the assignments which are detailed in the Assignments section below, specifically the reading and research memo reviews but not the flash talks or end of semester project. Please email me to be added to Piazza.

2 Materials

2.1 Computational Tools

The best way, and often the only way, to learn new statistical procedures is by doing. We will cover use of many of these computational tools with programming done in R.

2.2 Books

The reading will primarily be from relevant articles in the field. Some chapters from the book manuscript by Grimmer, Roberts and Stewart which will be made available on Perusall.

Suggested It is often helpful to see the same material in alternative ways. Thus here are some other texts you might consult.

Natural Language Processing

2.3 Articles

Readings will be posted on Perusall. Perusall is a new e-book platform with collaborative annotation that allows you to post and answer questions directly in the text itself. This gives us the opportunity to answer questions outside of class in the text itself. So asking good questions not only helps you, it helps your classmates. If you know the answer to a question that another student posted, please make a contribution to the class and try to answer it!

You can enroll using access code ROBERTS-PTQNK.

3 Assignments

There are five types of assignments in the course. They are listed below with how many times they will be done:

1. **Collaborative Annotation and Reading**: This will be done every week in the Perusall online system.

2. **Problem Sets**: (2×) Each section of the course will have an associated problem set.

3. **Research Memos**: (3×) For the initial question, discovery section, and measurement section, you will submit a short (max 800 words) summary of how you used the methodology described that week for your own research project.

4. **Flash Talk**: (1×) For the discovery and measurement memos, one-half of the students will present a flash talk in class on their research memo. Flash talks will be five minutes.

5. **Memo Reviews**: (2×) Each week that you do not do a flash talk you will respond/review a research memo of one of your classmates which you have been assigned.
6. **Poster Session:** (1x) The last day of class students will present a finished research project as a poster.

7. **Final Research Memo:** (1x) Students will write up a final project memo that accompanies their poster.

We describe each of these assignments in more detail below.

### 3.1 Collaborative Annotation and Reading

The majority of the readings will be from articles in the field. The interdisciplinary nature of these methods means that the articles will be drawn from a variety of different fields and the lack of a single unifying text book treatment means that these articles will often implicitly demand different backgrounds. We will use the Perusall system to help each other out and work together on these readings through collaborative annotations. This will not only allow classmates to help each other understand the material, it will also highlight to the instructor what material would best be covered in class time.

### 3.2 Problem Sets

Each section of the class (discovery, measurement, and testing) will be accompanied by a problem set. The problem set is designed so that you can work through the topics we have discussed in class. Problem set will include both conceptual and data analysis problems.

### 3.3 Final Project

The class is designed to work up to a final project. Ideally, this would be a project that will contribute to an ongoing project – whether it be an article, dissertation, or other publication. Students are encouraged to work with a co-author on the project. Final projects will be presented in a poster format on the last day of class. The poster should be submitted with a short memo that contains more detail about the project and a replication file for the project.

### 3.4 Research Memos

Each research memo will be one stepping stone toward a final project. In at most 800 words, your goal is to lay out apply the methods discussed in the current section’s readings to the dataset for your final project. The more figures you use the better! This is a great opportunity to get feedback on your project from both the teaching staff and your fellow students. During the course you will write three memos, one which you will present as a five minute flash talk in front of the class.

### 3.5 Memo Reviews

For the memos for which you do not give a flash talk, you will submit a review of a classmate’s memo. The goal of the review is to provide constructive feedback on the proposed research project. The comments could address core areas such as corpus selections, the applicability of the methods or the theoretical relevance. Good comments will also go through the colleagues codes and point our suggestions.
3.6 Grading

Final grades will be assessed based on the assignments described above according to the following breakdown:

1. Class Participation, Collaborative Annotation and Reading: 20%
2. Research Memos: 20%
3. Research Memo Reviews: 10%
4. Problem Sets: 25%
5. Final Project: 25%

3.7 Additional Help

3.7.1 Piazza

We encourage you to post questions about the readings to Piazza. You will not be required to post, but the system is designed to get you help quickly and efficiently from classmates, the preceptors, and the professor. Unless the question is of a personal nature or completely specific to you, you should not e-mail teaching staff; instead, you should post your questions on Piazza. The course staff will be monitoring the page, but we encourage you to help your classmates as well.

4 Course Outline

The course takes place over ten weeks. We may adjust the schedule due to comprehension, time, and interest. Please note also that readings are subject to change, in particular you should expect that we will add individual articles for discussion in class.

4.1 Introduction (April 1)

This first week will cover basic details of the course and motivate the use of text data in the social sciences. We will outline the organizing framework of the class based on the four steps for analyzing text: (1) Identification of Text/Population of Study, (2) Discovery, (3) Measurement, (4) Testing.

April 1: Introduction to Text as Data

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapter 2.


Optional Reading

This second week will cover strategies for selecting, acquiring, and representing texts. We will cover how to think about selection bias and generalizability in text datasets and how to turn language into numeric data.


- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. Text as Data, Chapter 3.


- Dodds, Peter and Christopher Danforth. 2009. “Measuring the Happiness of Large-Scale Written Expression: Songs, Blogs, and Presidents”. Journal of Happiness Studies 11, 4. 441-456

Optional Reading On Sampling and Representativeness


Optional Reading On Selecting Texts


Optional Reading On Preprocessing


Optional Reading On Word Embeddings


- Richard Socher’s lecture on word embeddings: [https://www.youtube.com/watch?v=T8tQZChniMk&index=2&list=PLo0lw6BstMGYXGeVpJy0yHUAdEUE7BsUp]


Optional Reading on Dictionary Methods


April 15: Research Memo 1 Due – Selecting Projects

4.3 Discovery: Uncovering what we want to study (April 15-22)

This section will discuss methods of discovery: or ‘How do we organize our texts and generate hypotheses for our work?’ We will discuss methods for identifying discriminating words and unsupervised clustering/embedding methods. Throughout we will emphasize that in discovery we are less concerned with assumptions of model holding and merely generating interesting hypotheses and questions. On April 29, we will have flash talks where students describe discovery in their own datasets.

April 15 Reading: Clustering and Topic ModelsI

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. Text as Data, Chapter 4, Sections 1-3.


- DiMaggio, Paul, Manish Nag, and David Blei. "Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of US government arts funding.” Poetics 41.6 (2013): 570-606.

Optional Reading on Clustering

- 14.3. Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer.
- Chp 9. Bishop, Christopher. 2006. *Pattern Recognition and Machine Learning* (Sections 2.1, 2.2 especially)
- Chuang, Jason, Christopher Manning and Jeff Heer. “Without the Clutter of Unimportant Words: Descriptive Keyphrases for Text Visualization ” (2012) *ACM Transactions on Computer-Human Interaction*

**Optional Reading on Topic Models**

- Hannah, Lauren A. and Hanna M. Wallach “Summarizing Topics: From Word Lists to Phrases” (2014) *NIPS Workshop on Modern Machine Learning and Natural Language Processing*

**April 22: Problem Set 1 Due**

**April 22 Reading: Topic Models II and Identifying Distinctive Words**

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. *Text as Data*, Chapter 4, Sections 3 and 5.

**Optional Reading on Topic Models II**

- Nelson, Laura K. “Political Logics as Cultural Memory: Cognitive Structures, Local Continuities, and Women’s Organizations in Chicago and New York City” Manuscript


Optional Reading on Distinctive Words


- Chuang, Jason, Christopher Manning and Jeff Heer. “Without the Clutter of Unimportant Words: Descriptive Keyphrases for Text Visualization ” (2012) ACM Transactions on Computer-Human Interaction


April 29: Research Memos Due – Discovery
April 29: Discovery Flash Talks

4.4 Measurement: Supervised Methods (April 29-November 14)

Shifting into measurement, this section will cover how to take an organizational structure for our data and measure quantities of interest such as individual document classifications or proportions over a corpus. We will cover dictionary methods (and their limits), hand coding procedures and methods to learn classifiers from hand coding. We will also discuss the importance and difficulties of validation. On November 14, we will have flash talks students describe measurement in their own datasets

April 29: Classifiers Part I

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. Text as Data, Chapter 5 Sections 1,2, and 4.

- 3.4. Hastie, Tibshirani, and Friedman. The Elements of Statistical Learning Springer

Optional Reading

- Montgomery, Jacob M. and David Carlson. “Human computation scaling for measuring meaningful latent traits in political texts.”

May 6: Research Memo Reviews Due – Discovery

May 6 Reading: Classifiers Part 2

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. Text as Data, Chapter 5 Sections 4 and 5.


Optional Reading


May 13: Problem Set 2 Due

May 13: Validation

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. Text as Data, Chapter 5 Sections 7 and 8.

- Section 2.2 and 5.1. James, Witten, Hastie and Tibshirani. An Introduction to Statistical Learning Springer.

4.5 Testing: Causal Inference and Prediction (May 20)

In our final section, we discuss testing our theories using the framework of causal inference and prediction. We will discuss how text work fits into the Rubin Causal Model and how to think about text as response, treatment and confounder. We will also discuss text as a predictor of non-text events.

May 20: Research Memos Due – Measurement

May 20: Measurement Flash Talks

May 20: Causal Inference and Text

- Grimmer, Justin, Margaret Roberts, and Brandon Stewart. Text as Data, Chapter 6.

Optional Reading

- Fong, Christian and Justin Grimmer (2016) “Discovery of Treatments from Text Corpora”


- Roberts ME, Stewart BM, Nielsen R. (2016) “Matching Methods for High-Dimensional Data with Applications to Text”

**May 27: Research Memo Reviews Due – Measurement**

**May 27: Memorial Day No Class**

**June 3: Poster Session**