Prerequisites

At the very least, students should have a first class in statistics and/or inference under their belt before taking this course. In particular, basic knowledge of calculus, probability, densities, distributions, statistical tests, hypothesis testing, the linear model, maximum likelihood and generalized linear models is assumed. The core language and software environment of this course is R. If you are not familiar with R, you will struggle with the assigned exercises. Please check with the instructor if you are unclear as to whether you are qualified for this course.

Overview

The availability of text data has exploded in recent times, and so has the demand for analysis of that data. This course introduces students to the quantitative analysis of text from a social science perspective, with a special focus on politics. The course is applied in nature, and while we will give some theoretical treatment of the topics at hand, the primary aim is to help students understand the types of questions we can ask with text, and how to go about answering them. With that in mind, we first explain how texts may be modeled as quantitative entities and discuss how they might be compared. We then move to both supervised and unsupervised techniques in some detail, before dealing with some ‘special topics’ that arise in particular lines of social science research. Ultimately,
the goal is to help student conduct their own text as data research projects and this class provides the foundations on which more focussed, technical research can be built.

While many of the techniques we discuss have their origins in computer science or statistics, this is not a CS class: we will spend relatively little time on traditional Natural Language Processing issues (such as machine translation, optical character recognition, parts of speech tagging etc). Other offerings in the university cover those matters more than adequately. Similarly, this class will not much deal with obtaining text data: again, there are excellent classes elsewhere dealing with e.g. web-scraping.

Structure

As of January 8, 2021, the plan is that this course will provide the following on a weekly basis:

- approximately 70 minutes of pre-recorded lecture material (from the professor)
- approximately 30 minutes of “in-person” flipped material (from the professor) in room 101 of 19W4th Street. This will be recorded and livecast.
- a remote 50 minute section/lab with the TA, which will be recorded.

Enrolled students must attend live or watch remotely (live or later) all content we provide. The information and skills that you need to complete your homework assignments and term projects will be provided by the Professor or the TA. Note that the pre-recorded lecture/flipped session is subject to demand: if turnout is low, we will simply revert to ‘live’ (but recorded) lecturing. During the first week of this class, everything will be remote (nothing in person).

Sections: your TA will hold section Thu 2.00 PM - 2.50 PM, remotely. The Zoom link is https://nyu.zoom.us/j/2990706756?pwd=OGdpUU5ack1kdEhZUjI3ektaZUx0QT09 The TA’s materials will be stored on the NYU Classes site.

Assessment

There are no written exams in the class, and your grade will be based on a combination of:

- **Homeworks (50%)**: There will be (at least) three homeworks, all of which will involve modeling and coding of text data, and some theoretical work. Intellectual honesty is important at NYU: you may confer with colleagues, but all work on the homework must be your own. If you copy work or allow another student to copy your work, the homework will be graded zero and your case will be passed to appropriate authorities in the university. Homeworks must come in electronically as a single RMarkdown PDF/HTML document of both answers and code.

- **Final Paper (50%)**: There will be a final written paper of not longer than 10 double spaced pages of text, which explores an original research project or idea. This may be substantive or technical in nature. You are encouraged to work in teams of up to two people on this paper. The deadline for the paper will be May 12, 2021 with no extensions or exceptions.
Software

We will be using R, a statistical package. You can download and install R for free, from here:

[https://cran.r-project.org/](https://cran.r-project.org/)

To write and edit R code, you can use any software with which you are familiar and/or enjoy using. We suggest R Studio, which is free:


Textbooks and Reading

There are no required textbooks for the course. We will draw from some of the following (and other places!), and will make efforts to provide the readings online where appropriate:


Because the class is focussed on answering substantive questions with the techniques on offer, many of the readings are applied in nature.

COURSE SCHEDULE

1  Feb 2: Introduction and Overview (REMOTE ONLY THIS WEEK)

This class is great: take it. Note that this week, we will be entirely remote: there will be no flipped session, and the professor will use his zoom link to lecture.

2  Feb 9: Representing Text

- vector space model of a document
- feature choices/representation
• pre-processing: stemming and stopping
• bag of words (and alternatives)
• sparseness

Reading
• MRS ch 6 “Scoring, term weighting and the vector space model”

3 Feb 16: Descriptive Inference I
• word distributions: Zipf’s Law/Heap’s Law
• co-occurrence, collocations and phrasemes
• key words in context
• dis(similarity) measures and testing for differences

Reading
• MRS, Ch 5

4 Feb 23: Descriptive Inference II
• lexical diversity
• sophistication/readability/complexity
• linguistic style and author attribution
• sampling distributions for estimates

Reading


5 Mar 2: Supervised Techniques I

• dictionary based approaches
• sentiment (and other) dictionaries, LIWC
• Goldman-Sachs case study
• event extraction
• lie detection

Reading


6 Mar 9: Supervised Techniques II

• classification of documents
• evaluation of techniques: precision, recall
• crowdsourcing
• Naive Bayes Classification, estimating proportions
• ideological scales with ‘wordscores’
Reading

- MRS. “Text classification and Naive Bayes”.
- Michael Laver, Kenneth Benoit, and John Garry. 2003. Extracting policy positions from political texts using words as data American Political Science Review 97(2)

7 Mar 16: Supervised Techniques IIIA

- basics/varieties of machine learning
- support vector machines

Reading


8 Mar 23: From Supervised to Unsupervised

Supervised Techniques IIIB

- k-NN models
- random forests/tree techniques
- ensembles

Reading

Unsupervised Techniques I

- fundamentals of unsupervised learning
- (principal) components and data reduction
- singular value decomposition

Reading


9 Mar 30: Unsupervised Techniques II

- clustering (documents)
- Latent Semantic Analysis/Indexing
- parametric scaling of political speech
- count models: ‘wordfish’
- basics of semi-supervised techniques

Reading


10 Apr 6: Unsupervised Techniques III

- plate notation/graphical model
- basics of Bayesian methods
- Latent Dirichlet Allocation and Topic Modeling
- Variational Inference
- model selection/choosing k
Reading

- DM Blei and MI Jordan, 2006. Variational inference for Dirichlet process mixtures, Bayesian Analysis, Volume 1, Number 1, 121–143.

Apr 13: Work on Final Project, no lecture (will set up consulting time)

11 Apr 20: Unsupervised Techniques IV

- Correlated Topic Model
- Dynamic Topic Model
- Structural Topic Model
- Embeddings: Word2Vec (theory/overview)

Reading


12 Apr 27: Special Topics I

- modeling debate and discourse
- networks of communication
- bursts and memes

Reading


13 May 4: Special Topics II

- word embeddings for applied social science research
- Word2Vec vs Glove
- Turing tests: intrinsic v extrinsic assessment
- embeddings for time series problems

Reading


14 Wed, May 12: Final Projects Due