Text as Data

August 15 - August 24, 2017

Tuesday, Wednesday, Thursday 9-12; Seigle 306

Instructor: Justin Grimmer
Email: jgrimmer@stanford.edu
Office: Seigle 283
Course website: https://github.com/justingrimmer/WUSTL

Course Description

Language is the medium for politics and political conflict. Candidates debate during elections. Representatives write laws. Nations negotiate peace treaties. Clerics issue Fatwas. Citizens express their opinions about politics on social media sites. These examples, and many others, suggest that to understand what politics is about, we need to know what political actors are saying and writing.

This course introduces techniques to collect, analyze, and utilize large collections of text for social science inferences. The ultimate goal of the course is to introduce students to modern quantitative text analysis techniques and provide the skills necessary to apply the methods in their own research. In achieving this ultimate goal, students will also learn about core concepts in machine learning and statistics, developing skills that are transferable to other types of data and inference problems. They will also have the opportunity to develop their programming abilities and develop an original research project or to participate in an ongoing research project.

Prerequisites

At a minimum, students should have completed coursework on univariate inference and linear regression. The ideal student will have also taken a course on model based inference. The course will develop student's programming skills. Prior experience with R, Python, or a related language is strongly recommended.

Teaching Assistant

Jonathan Homola, homola@wustl.edu. Seigle 277.

List of Topics

- Day 1: Introduction, Regular Expressions, and Preprocessing
    - * Homework: Parsing, preprocessing, analyzing, and visualizing a presidential debate
- Day 2: Discovery (Unsupervised Learning I)
    - Measuring (Dis)Similarity
    - Fully Automated Clustering Methods
    - Discovering Discriminating Words
        - * Homework: Comparing press releases of Richard Shelby and Jeff Sessions (unigrams, trigrams, word separating algorithms, document similarity)
- Day 3: Measurement Part 1 (Supervised Learning)
    - Dictionary Methods
    - Coding Rules and Manual Coding
    - Naive Bayes, LASSO
    - Evaluating Classifiers
        - * Homework: Analysis of New York Times news stories (clustering methods, dictionary classification, Naive Bayes, LASSO, Ridge, KRLS)
- Day 4: Measurement Part 2
    - Ridge, LASSO, Elastic Net
    - Loss Functions and Model Complexity
    - Cross-Validation and Parameter Selection
    - Ensembles, CART, Random Forest, Super Learning
        - * Homework: Predicting Portuguese students' drinking habits (OLS, LASSO, Ridge, Elastic Net, Cross-Validation, Super Learning, Ensembles)
- Day 5: Measurement Part 3 (Unsupervised Learning II)
    - Topic Models (Vanilla LDA, Expressed Agenda Model, Dynamic Topic Model, Structural Topic Model)
    - Principal Components, Multidimensional Scaling

  - Ideal Points via: IRT and Principal Components
    * Homework: NYT data (`stm` package, vanilla LDA), Machiavelli's Prince (PCA, multidimensional scaling)

- Day 6: Causal Inference
  - Marginal Effects of Latent Treatments (AMCE)
  - Discovering Treatments from Text Corpora (Text as Treatment)
  - Discovering Dependent Variables (Text as Outcome)
    * Homework: Causal inference applications