PPOL 628: Text as Data – Computational Linguistics for Social Scientists

Class 1: Course Overview and String Manipulation

Course Info

- Instructor: Matt Denny -- md1723@georgetown.edu

- TA: Ted Ellsworth -- tedellsw@terpmail.umd.edu

- Office Hours:
  - Matt – 5:45-6:15 Tuesdays before class, appointments preferred.
  - Ted -- remote/email, by appointment

- Website: github.com/matthewjdenny/PPOL_628_Text_As_Data

Course Overview

This course seeks to arm its participants with the **theoretical background, practical experience, and technical capacity** to pursue cutting edge social science research using text data.

This course is designed to cover the key technical aspects of conducting research with text data: from **data collection and preprocessing, through to description and inferential analysis**.

We will cover various techniques from computational linguistics such as parts-of-speech tagging and sentiment analysis, term-category associations, supervised learning with text, topic modelling, and word embeddings, to name a few.

Prerequisites

- **I expect that students can load in and manipulate data in R**. I also expect that students have at least some basic familiarity with concepts like conditional statements (if/then) and looping (for and while loops).
- I expect that students are familiar with basic concepts in statistics such as the normal, uniform, multinomial distribution, and what it means to sample from a distribution; linear and logistic regression and the interpretation of parameter estimates from these models; basic statistical/mathematical concepts such as mean, variance, expected value, logarithmic and exponential functions.

Evaluation

- **Participation (25 %) --** I expect students to do the readings each week, to pay attention in class, to participate in discussing the articles we read, and to try out any example code we go over for themselves.
- **Homework (25 %) --** Each week you will responsible for applying and/or extending the example code we went over in the previous week to your own data, and providing a short write-up of what you found.
- **Final Project (25 %) --** Building off of one of the analyses you perform as part of your homework for the class, you will be asked to write an 8-10 page report fleshing out this analysis to learn something substantively interesting from your data
- **Final Exam (25%) –** A one-hour exam where I will ask you to answer 2-3 broad understanding questions drawn from the topics we cover in class.

Academic Integrity, Conduct, etc.

- Do your own work. You are here to learn, don’t waste your time, or my time. (Also, I am an expert in text re-use methods).

- I will follow Georgetown’s policies: honorcouncil.georgetown.edu/system/policies/standards-of-conduct/

- Respect your colleagues, instructor, TA.

Schedule (Part 1)

- 01/14/20 – Introductions and logistics.
- 01/21/20 – An overview of the field and collecting your own data.
- 01/28/20 – Text preprocessing (data collection due).
- 02/04/20 – NO CLASS (Travel Conflict).
- 02/11/20 – Basic NLP: Parts of speech.
- 02/18/20 – NO CLASS (Holiday).
- 02/25/20 – Dictionary-based methods, basic sentiment analysis.
- 03/03/20 – Corpus Description, Word counts, TF-IDF.

Schedule (Part 2)

- 03/10/20 – NO CLASS (Holiday).
- 03/17/20 – Term-category associations.
- 03/24/20 – Text Reuse.
- 03/31/20 – Supervised Learning with Text.
- 04/07/20 – Introduction to Topic Models.
- 04/14/20 – Assessing Topic Models and the Structural Topic Model.
- 04/21/20 – Word Embeddings.
- 04/28/20 – Make up class/special topics.
- 05/02/20 – Final Exam (7:00-9:00 pm).
- 05/08/20 – Final Project Reports Due.

Things I want to know about you!

- Your name and pronouns.
- Where you want to be in 5 years (professionally).
- Something you hope to get out of this experience.
- Your favorite food (or one of your favorites).

Data Collection

- You will be asked to collect your own corpus (collection of documents) for this class. You may use a dataset you find online, one your adviser already has, one you have already collected, or one you collect for this class.
- You dataset must contain:
  - 100+ documents (come see me if you really want to use one with fewer).
  - Dataset should be 100+ pages (~30,000+ words) in total.
  - Your dataset should have one categorical variable and one continuous variable per document for at least 100 documents. You can hand-code these for a subset of documents.

R Exercise

- github.com/matthewjdenny/PPOL_628_Text_As_Data