This version: May 29, 2017

University of Tokyo
An Introduction to Text-as-Data

June 3–4, 2017

Arthur Spirling
Department of Politics
Center for Data Science
New York University
arthur.spirling@nyu.edu
https://www.nyu.edu/projects/spirling/

Overview and Structure

There is a github website for the course here: `https://github.com/ArthurSpirling/UTokyo-TextAsData`.

The availability of text data has exploded in recent times, and so has the demand for analysis of that data. This short course introduces students to the fundamentals of 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, this course explains how texts may be modeled as quantitative entities and discusses how they might be compared. We then move to supervised and unsupervised methods—including topic models. 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).

Generally speaking over the two days, the course will have a 'sandwich' structure, with practical software exercises either side of theory presentation.

Software

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

`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:

`https://www.rstudio.com/products/RStudio/`

We will be using an `R` package designed by a team lead by Prof Ken Benoit, specifically engineered for social scientists working with text. The package is called `quanteda`. You can install it from the command line via

```
install.packages("quanteda")
```

You will also need a package specifically for data intake, called `readtext`. You can install it from the command line via

```
install.packages("readtext")
```

Textbooks and Reading

There are no required textbooks for the course. Some material will draw on this work ('MRS'):

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schtze, Introduction to Information Retrieval, Cambridge University Press. 2008.

This book has an online edition here `http://nlp.stanford.edu/IR-book/`

There will be several articles which are available in the 'usual places' (i.e. on line repositories). Ask the instructor if you can't track down what you need via e.g. google.

COURSE SCHEDULE

1 June 3

AM: Introduction and Overview: Representing Text

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

Reading

- MRS ch 6 "Scoring, term weighting and the vector space model"
- Denny, Matthew and Arthur Spirling. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. Available here

PM: Descriptive inference I

- dis(similarity) measures and testing for differences
- key words in context
- lexical diversity
- sophistication/readability/complexity
- literary styles and author attribution
- sampling distributions for text
- bursts and "burstiness"

Reading

- MRS, ch 5
- Benoit, K., Laver, M. and Mikhaylov, S. 2009. Treating Words as Data with Error: Uncertainty in Text Statements of Policy Positions. American Journal of Political Science, 53: 495513.
- A Spirling. 2016. Democratization and Linguistic Complexity, Journal of Politics.
- F Mosteller and D Wallace. 1963. Inference in an Authorship Problem, Journal of the American Statistical Association, Volume 58, Issue 302, 275–309.
- R Peng and N Hengartner. 2002. Quantitative Analysis of Literary Styles, The American Statistician, Volume 56.
- J. Kleinberg. Bursty and Hierarchical Structure in Streams Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining, 2002.

2 June 4

AM: Basic Supervised Learning and Coding

- dictionary based approaches: sentiment and other concepts
- event studies
- supervised scaling: Naive Bayes and `WordScores`
- supervised models for proportions (`ReadMe`)

Reading

- Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (pages 1–27 only).
- Gary King and Will Lowe. 2003. An Automated Information Extraction Tool for International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design. International Organization, 57, pp 617–642.
- Michael Laver and John Garry. 2000. Estimating Policy Positions from Political Texts. American Journal of Political Science Vol. 44, No. 3, pp. 619-634
- Yla R. Tausczik and James W. Pennebaker. 2009. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. Journal of Language and Social Psychology. March 2010 vol. 29 no. 1 24-54.
- D Hopkins and G King. 2010. A Method of Automated Nonparametric Content Analysis for Social Science American Journal of Political Science, Vol. 54, No. 1, January 2010, 229–247.
- W Lowe. 2008. Understanding Wordscores, Political Analysis, 16 (4): 356-371.
- Benoit, Kenneth, Conway, Drew, Lauderdale, Benjamin E., Laver, Michael and Mikhaylov, Slava. 2015. Crowd-sourced text analysis: reproducible and agile production of political data. American Political Science Review.

PM: Unsupervised Learning

- clustering
- unsupervised scaling (`Wordfish`)
- topic models, LDA
- structural topic model, `stm`

Reading

- Justin Grimmer and Gary King. 2010. General purpose computer-assisted clustering and conceptualization. Proceedings of the National Academy of Sciences. Vol 108, No 7. 2643–2650.
- Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. A Scaling Model for Estimating Time-Series Party Positions from Texts. American Journal of Political Science 52(3): 705-722.
- DM Blei, AY Ng and MI Jordan, 2003. Latent Dirichlet Allocation, Journal of machine Learning research 3, 993-1022.
- Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H. and Radev, D. R. (2010), How to Analyze Political Attention with Minimal Assumptions and Costs. American Journal of Political Science, 54: 209–228.

- Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B. and Rand, D. G. (2014), Structural Topic Models for Open-Ended Survey Responses. American Journal of Political Science, 58: 10641082