Course Syllabus: Text Analysis in Social Science
Max Planck Summer School 2017

Instructor: Elliott Ash

1 Overview

Welcome to the summer course on text analysis! The course lectures will include presentation of methods and applications, with live demonstrations using Python. See Section 3 below for instructions on installing Python and necessary packages. The two "exercise" meetings will consist of more lecture slides, and tutorials on the assignment, as needed.

2 Readings

Some recommended readings are posted here. They provide useful background for the lectures and will help you get the most out of the course material. I will not ask any questions on the exam from the readings that are not otherwise included in the slides.

2.1 Methods (Python)

I have read most of these materials, and can answer questions about them.

- Featurization:
  - *Natural Language Processing in Python* (`nltk.org/book`).
- Topic models:
  - Shivam Bansal, "Beginner's guide to topic modeling in Python," `https://www.analyticsvidhya.com/blog/2016/08/beginners-guide-to-topic-modeling-in-py`
- Word embeddings:
  - Taylor Brown, "Introduction to Word Embedding Models with Word2Vec", `https://taylorwhitten.github.io/blog/word2vec`.
- Document similarity and clustering:
  - Brandon Rose, "Document clustering in python", `http://brandonrose.org/clustering`
  - Vlad Niculae and Matt Kusner, "Word Mover's Distance in Python", `http://vene.ro/blog/word-movers-distance-in-python.html`
- Supervised learning:

  - Kevin Markham, "A friendly introduction to linear regression using Python", https://github.com/justmarkham/DAT4/blob/master/notebooks/08_linear_regression.ipynb
  - Aarshay Jain, "A complete tutorial on ridge and lasso regression in Python," https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-ridge-lasso-re
  - J. Warmenhoven and R. Jordan Crouser, "PCR and PLS in Python", http://www.science.smith.edu/~jcrouser/SDS293/labs/lab11/Lab%2011%20-%20PCR%20and%20PLS%20Regression%20in%20Python.pdf
  - Yhat, "Random forests in Python," http://blog.yhat.com/posts/python-random-forest.html
  - OPIG, "Using random forests in Python with Scikit-learn", http://www.blopig.com/blog/2017/07/using-random-forests-in-python-with-scikit-learn/

- Visualization:
  - Sebastian Raschka, "Turn your Twitter Timeline into a Word Cloud", http://sebastianraschka.com/Articles/2014_twitter_wordcloud.html.
  - Dan Saber, "A dramatic tour through Python's data visualization landscape", https://dsaber.com/2016/10/02/a-dramatic-tour-through-pythons-data-visualiza

2.2 Methods (R)

I can't vouch for or answer questions about these materials.

- General:
  - *Text Mining with* R, http://tidytextmining.com/
- Featurization:
  - Ken Benoit, "Getting started with quanteda", https://cran.r-project.org/web/packages/quanteda/vignettes/quickstart.html
  - Matt Denny, "Getting started with phrasemachine", http://www.mjdenny.com/getting_started_with_phrasemachine.html
- Topic models:
  - *Text Mining with* R, chapter 6.
- Supervised learning:
  - Margaret Roberts, Brandon Stewart, and Dustin Tingley, "stm: R packge for Structual Topic Models", https://cran.r-project.org/web/packages/stm/vignettes/stmVignette.pdf

2.3 Statistics and Machine Learning

- Topic models:
  - Blei and Lafferty, "Topic Models", http://www.cs.columbia.edu/~blei/papers/BleiLafferty2009.pdf.
- Word embeddings:
  - Yoav Goldberg and Omer Levy, "Word2Vec explained: Deriving Miolov et al's Negative Sampleing Word Embedding Method", https://arxiv.org/pdf/1402.3722.pdf
  - Piero Molino, "Word embeddings: Past, present, and future", http://w4nderlu.st/teaching/word-embeddings
  - Matt Kusner, Yu Sun, Nicholas Kolkin, and Killian Weinberger, "From word embeddings to document distances", https://mkusner.github.io/publications/WMD.pdf.
- Supervised learning:
  - Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *Elements of Statistical Learning*, http://www.statedu.ntu.edu.tw/bigdata/The%20Elements%20of%20Statistical%20Learning.pdf.
- High-dimensional econometrics:
  - Alexandre Belloni, Daniel Chen, Victor Chernozhukov, and Christian Hansen, "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain", https://arxiv.org/abs/1010.4345.
  - Susan Athey and Guido Imbens, "Recursive partioning for heterogeneous causal effects", https://arxiv.org/abs/1504.01132.
  - Stefan Wager and Susan Athey, "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests", https://arxiv.org/abs/1510.04342
  - Jason Hartford, Greg Lewis, Kevin Leyton-Brown, Matt Taddy, "Counterfactual Prediction with Deep Instrumental Variables Networks", https://arxiv.org/abs/1612.09596.
  - Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, James Robins, "Double/Debiased Machine Learning for Treatment and Causal Parameters", https://arxiv.org/abs/1608.00060.

2.4 Social-Science Applications

- General:
  - Matt Gentzkow, Bryan Kelly, and Matt Taddy, "Text as Data," https://web.stanford.edu/~gentzkow/research/text-as-data.pdf.
  - Justin Grimmer and Brandon Stewart, "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts," https://web.stanford.edu/~jgrimmer/tad2.pdf.
  - Elliott Ash, "The political economy of tax laws in the U.S. states", http://elliottash.com/wp-content/uploads/2013/08/tax-laws-abstract.pdf
- Featurization:
  - Matthew Denny and Arthur Spirling, "Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It," https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2849145.
- Topic models:
  - Kevin Quinn, Burt Monroe, Michael Colaresi, Michael Crespin, and Dragomir Radev, "How to analyze political attention with minimal assumptions and costs," http://clair.si.umich.edu/~radev/papers/AJPS2010.pdf.
  - Stephen Hansen, Michael McMahon, and Andrea Prat, "Transparency and deliberation with the FOMC: A computational linguistics approach", https://www2.warwick.ac.uk/fac/soc/economics/staff/mfmcmahon/research/fomc_submission.pdf
- Word embeddings:
  - Maya Rudolph and David Blei, "Dynamic Bernoulli Embeddings for Language Evolution," https://arxiv.org/pdf/1703.08052.pdf
- Supervised learning:
  - Matt Gentzkow and Jesse Shapiro, "What drives media slant?"
  - Zubin Jelbeh, Brice Kogut, and Suresh Naidu, "Political language in economics", https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2535453.
  - Matt Gentzkow, Jesse Shapiro, and Matt Taddy, "Measuring polarization in High-Dimensional Data: Method and application to congressional speech", https://web.stanford.edu/~gentzkow/research/politext.pdf
- Causal inference methods using text:

- Margaret Roberts, Brandon Stewart, and Richard Nielsen, "Matching Methods for High-Dimensional Data with Applications to Text", http://www.margaretroberts.net/wp-content/uploads/2015/07/textmatching.pdf
- Christian Fong and Justin Grimmer, "Discovery of treatments from text corpora," https://stanford.edu/~jgrimmer/SE_Short.pdf.

3 Programming in Python

There is no way around it – text analysis research requires some use of programming languages.

The example code used in class will be in Python. You can use another programming language but I might not be able to help with it. If you are using Python, please install Anaconda (Python 3.6 version), and download newest versions of nltk, gensim, spacy, unidecode, and wordcloud packages.

If you are using R, please install quanteda, spacyr, and stm packages. There will be no R examples during lectures, but I have provided some related links in the readings section.

4 Course Assignment

You will get the most out of the course if you produce a new corpus and dataset for use in a preliminary empirical analysis. If feasible, identify a corpus of documents, with accompanying metadata, that can be used to answer an interesting research question. For example, one good option is to use chatroom text from a lab experiment.

For those who have not chosen a corpus, I will help you download a sample of cases from courtlistener.com. See the Python script 'download-cases', which uses the CourtListener API to download a corpus of case text snippets with metadata. You will see on line 241 where the search query "death penalty" is specified; feel free to download your own corpus with any word or phrase, e.g. "contract breach", "mental health", "deterrence", etc. You also need to change the working directory (look for the "os.chdir" line). This data is on the state-year level, so there are many metadata data sets one could merge with.

The completed course assignment can be submitted to fulfill the exam requirements.

5 Written Exam

The exam will consist of conceptual questions based on the slides. I will provide some practice questions on Friday. The exam can be taken instead of, or along with, the course assignment. If you do both, you will get two chances to pass.

6 Acknowledgements

Thanks to Chris Bail, Brandon Stewart, Michael McMahon, and Piero Molino for useful slide decks.