Please use on the on-line version of this syllabus: faculty.washington.edu/jwilker/559/559Syllabus2015.pdf
Updated 2/26

POLS 559: Text as Data

Winter 2015

John Wilkerson (jwilker@uw.edu)
Andreu Casa-Salleras (acasas2@uw.edu

Class meets 4:40-7:20 Thursdays, MGH 097
Office Hours: Wilkerson 9:30-11 W Casas-Salleras: M 2-4, W 4:30-5:30
Smith 221

What information can we get from text? How do we get it? What are the different options for analyzing words as data? How do we know whether a method is doing a good job of capturing what's in a document?

This class introduces computational approaches to collecting, organizing, and analyzing text as data. Computational methods can assist the process of forming impressions and formulating hypotheses when a lot of data is involved, and in applying classification schemes at scale. Like most methods, computational methods are never a substitute for careful theorizing, and using them will not rescue a poorly conceived project. For this reason, we begin by learning about the practice of content analysis more generally.

Python is the go to language for large scale text processing. R also includes a number of valuable packages, but has important limits where processing is concerned. We will learn the basics of the Python programming language and apply some important text analytic methods. We will be working with code, so you either need to have some minimal experience using a coding language (e.g. you have written scripts in STATA or R), or you need to recognize that you will probably have some catching up to do. If you are already a Python user but have not done much with text, we would be happy to have you participate.

Readings and Resources

General comments.

A central virtue of text methods is that we have access to new data (text!). So you should be striving for a non-obvious 'transformative' opportunity in your area of interest. I would much rather see ambitious, albeit half-completed research projects than projects that merely demonstrate your ability to apply what we covered in class.

Andreu Casa-Salleras (acasas2@uw.edu) is being paid for 10 hours per week of POLS 559 consulting. He is your resource but please also be considerate of his time. Most of the problems you will encounter are common problems that have been answered many times (in manuals or via google). So please try to make a general policy of starting the homework early and giving your brain a little time to work things out before you seek outside assistance. You'll gain more confidence and more independence. Thanks!

What particular part of your code is causing the problem?

- Test parts of your code in sequence.
- Test that your code works with a simple dataset
- Consult common error messages.
- The Python Tutorial offers useful examples and Stack Overflow has a lot of answers

Seminar Slides (posted weekly)

Books (additional articles and papers are listed on the weekly schedule)

- Saldana, *The Coding Manual for Qualitative Researchers*
  - *Covers potential objectives of many content analysis projects, how to develop a coding scheme, and how to assess validity and reliability. We will not be using the software Saldana references.*
- Bird, Klein, Loper, *Natural Language Processing with Python*
  - *There is a free downloadable version of this book, but it can also be purchased*
- Lutz, *Learning Python (a very readable, big, introduction to the Python programming language)*

On-line Resources

- The Python Tutorial http://docs.python.org/2/tutorial/ *(comprehensive resource for python commands and syntax)*
- Learn Python http://www.learnpython.org/ *(code testing)*
- Pythex https://pythex.org/ *(Testing regular expressions)*
- Python Tutor http://www.pythontutor.com/visualize.html *(see how the programming language actually works)*

Grading

- Participation (20%)
  - Come to class prepared and participate in class activities (including assisting classmates). Please refrain from spending class time answering emails etc.
- Homework (40%)
  - Coding homeworks shouldn't take hours. If you are stuck, you are welcome to seek advice (from classmates or Andreu).
- Research Project (40%) - *Must be pre-approved by the instructor.*
  A central purpose of this class is to open up a whole new world of data opportunities – text as data. Be ambitious! Your project will pose an interesting question, gather a substantial amount of text, convert it to data, and analyze it using the tools of this course (or beyond).

  Papers should be about 15 pages in length. Substance is more important than the length. Papers can be light in terms of theory and literature review, but should be heavy in terms of having an ambitious and well supported research design. Be sure to include your carefully annotated iPython Notebook as a separate attachment.

(see Glossary at bottom)

Week of	Readings to be completed before the class in which they are listed Reading Homework due before the class in which they are listed Coding Homework due after the listed class on the following Tuesday. If there is a problem with a link, homework, etc. please let me know asap!
Jan. 8	**Topics: Computational Social Science; Introduction to Python** **Readings** Lazer, David et al. 2009. "Computational Social Science" *Science*, pp.721-23 Grimmer, Justin and Brandon Stewart. 2013. "Text as Data: The Promises and Pitfalls of Automated Content Analysis." *Political Analysis*, 1-31. **Reading Homework 1**: Briefly describe the four principles articulated in Grimmer and Stewart **Coding Homework 1:** Install Anaconda and test by opening this Notebook in I-Python. (You just need to confirm that you can open it, but feel free to do more) **Coding Homework 5**: Manual Annotation Instructions Datafile (don't take more than a couple of hours)
Jan. 15	**Topics: Content Analysis; Python basics; Crowdsourcing** **1. Content Analysis Readings:** Saldana, Chapters 1-5 *This is a very useful book for thinking about content analysis. You should reference these readings in the tentative research proposal due next week.* *Baumgartner, Jones, McLeod, Lessons from the Trenches, The Political Methodologist (optional)* **Reading Homework 2**: Lessons of Saldana **2. Crowdsourcing Readings:** Kenneth Benoit, Drew Conway, Michael Laver, and Slava Mikhaylov "Crowd-sourced data coding for the social sciences: massive non-expert coding of political texts," New Directions in Text As Data Conference (October 2012). Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design Jeffrey Heer, Michael Bostock ACM Human Factors in Computing Systems (CHI), 203–212, 2010 "15 Tips to get the most out of Mechanical Turk" AJ Kohn **Coding Homework 2 :** Crowdsource dabble **3. Python Basics Readings:** Erlich, Aaron. "Python for R Users" "Python for R – Strings" "Python for R – Dicts and Tuples" Consider watching the Khan Academy *Introduction to Python* videos. They are very helpful. Consider creating an IPython notebook and trying the same things (no need to turn it in).

	Coding Homework 3: Python Basics; Importing Data into Python --- **Resources** *IP1: Python Basics* *(refers to I-Python Notebook 1)* *IP2: Importing and Preparing Text* *[FOMC2_datafile]*
Jan. 22	**Topics: Some Research Opportunities; Obtaining text** **1. Some research opportunities** Andreu Casas Salleras, Social media Emily Gade, The Internet Archive and .gov John Wilkerson, Government documents and AUTH AUTH White paper **2. Obtaining Text** **Coding Homework 4:** Obtaining and Preparing Text **---** **Resources** *IP2: Importing and Preparing Text* *IP3: Scraping URLs and APIs* *IP4: Scraping PDFs* *R: Capturing Twitter Feeds* *Hottest/Warmest Year* *(5mb zip file)*
Jan. 29	**Topics: Manual and Automated Classification; Reliability and Validity; Obtaining Text** **Reading Homework 3:** Ambitious research proposal, version #1 (Instructions are on Catalyst. No more than 1 page in length, the more focused the better!) 3. **Manual Annotation Readings:** Baumgartner, Frank, Bryan D. Jones, Michael C. MacLeod. 1998 "Lessons from the Trenches: Quality, Reliability, and Usability in a New Data Source.". *The Political Methodologist*. 8(2):1-11 Boystun, Amber, Justin Gross, Phillip Resnick, Noah Smith, "Identifying Media Frames and Frame Dynamics within and Across Policy Issues." **Coding Homework 5**: Manual Annotation Instructions Datafile (don't take more than a couple of hours!) **4. Supervised Machine Learning** NetFlix Prize

	Hillard et al. 2007. "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research" *Journal of Information Technology and Politics*. Vol. 4(4) 2007 (see also) Collingwood, Loren and John Wilkerson. *"Tradeoffs in Accuracy and Efficiency in Supervised Learning Methods."* Journal of Information Technology and Politics. 9:3 (2012) Hopkins, Daniel and Gary King/ 2010. "A Method of Automated Non-Parametric Content Analysis for Social Science" *American Journal of Political Science* 54(1): 229–247 **Coding Homework 6**: Successfully run example in RTextTools Getting Started document. Construct a Confusion Matrix for the results. **5. Reliability Reading:** Klaus Krippendorf. "Reliability in Content Analysis: Some Common Misperceptions and Recommendations." *Human Communication Research* 30(3) 411-433
Feb. 5	**Topic: Converting Text to Data** **Coding Homework 7:** Normalization; Counting words, NLTK, Term Document Matrices *[FOMC0508.csv]* --- **Resources** *IP5: Preprocessing and Summarizing* *IP6: NLTK Tokenizing* *IP7: NLTK Corpus and Exporting*

Feb. 12	**Topic: Unsupervised Machine Learning** **Readings**: Mining the Dispatch David Blei, Probabilistic Topic Models, *Communications of the ACM* 2012 (see also Blei and Lafferty's corpus browsers based on topic models) Trey Causey, "The Battle for Bystanders: Information, Meaning Contests, and Collective Action in the Egyptian Uprisings of 2011" (we'll be replicating this) (some additional prominent Political Science articles) Quinn, Kevin, Burt L. Monroe, Michael P. Colaresi, Michael Crespin and Dragomir Radev. 2010. **"**How to Analyze Political Attention with Minimal Assumptions and Costs" *American Journal of Political Science***,** Vol. 54, No. 1, January 2010 Grimmer, Justin, and Gary King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108(7): 2643–2650. Roberts, Molly, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Gadarian, Bethany Albertson, David G. Rand. 2014. "Structural Topic Models for Open-ended Survey Responses" *American Journal of Political Science*. Jason Chuang, John D. Wilkerson, Rebecca Weiss, Dustin Tingley, Brandon M. Stewart, Margaret E. Roberts, Forough Poursabzi-Sangdeh, Justin Grimmer, Leah Findlater, Jordan Boyd-Graber, and Jeffrey Heer. Computer-Assisted Content Analysis: Topic Models for Exploring Multiple Subjective Interpretations. *Conference on Neural Information Processing Systems (NIPS)*. *Workshop on Human-Propelled Machine Learning*. Montreal, Canada, 2014. **Assessing model fit** Chuang et al. "Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment" **Coding Homework 8**: Supervised and Unsupervised Machine Learning ---------- **Resources** *IP8: LDA Topic Modeling* May need to install the gensim package (the introduction is worth reading) *LDA_presentation.R* Download and unzip Causey Data to a directory on your machine. TERMITE (dynamic visualization of topic models by Jason Chuang)
Feb. 19	**Topic: Semantic Analysis** **Readings**: Wilkerson, John, David Smith and Nicholas Stramp "Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach." Shaffer, Robert "Power in Text: Grammar and Language in Comparative Delegation Dynamics" Working Paper: University of Texas Austin.

	Coding Homework 9: Information Extraction and Text Reuse ---------- **Resources** *IP9: NLTK Information Extraction* *IP10: Text Reuse* *(Smith Waterman local alignment algorithm)* WCopyFind NewsDiffs MemeTracker Wordnet
Feb. 26	**Topic: Sentiment Analysis and Scaling** **Reading Homework 4:** Ambitious research proposal, version #2 • **Crowdsourcing** • **Dictionary** • **Supervised** • **Semi-supervised** 1. **Sentiment Readings:** Pulse of the Nation Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan "Thumbs up: Sentiment Classification using Machine Learning Techniques." (2002) Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing. 10: 79-86. 2. **Scaling Readings** Laver, M. and Garry, J. (2000). "Estimating policy positions from political texts." *American Journal of Political Science*, 44(3):619–634. Lowe, Will 2008. Understanding Wordscores *Political Analysis* (2008) 16:356–371 Proksch and Slapin, Wordfish manual **Coding Homework 10:** Scaling Comparison ---------- **Resources** Wordscores (STATA) Wordfish 1.3 *[wordfishdataexample.csv]* The Movie Review corpus (Lee et al) is one of the most extensive resources: See also Jan Weibe for sentiment word lists but there are lots of them out there for different needs. General Inquirer is one of the earliest
Mar. 5	**Topic: Cloud Computing** **Readings:** http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/get-set-up-for-amazon-ec2.html http://docs.aws.amazon.com/AWSEC2/latest/UserGuide/EC2_GetStarted.html
Mar. 12	Project Presentations
March 19	Final Papers due

Glossary of Terms

Precision – proportion of predicted cases of X's that are true Xs. (errors are false positives)
Recall – proportion of true Xs that are predicted Xs (errors are false negatives)
F-Score – the harmonic mean of precision and recall
Validity – a measure is valid if on average it accurately captures the concept to be measured
Reliability –a measure is reliable to the extent that it produces the same result each time
Bias – Reliability is not validity; a measure can reliably invalid
Confusion matrix – crosstabulation of *actua*l versus *predicted* results. Used to examine prediction success (precision, recall) overall and within specific categories.
Annotate = Classify = Code = Label (verb) Annotation = Class = Code = Label (noun)
Token – any element of a document (e.g. a word; space; semicolon).
Tokenization (aka text segmentation) - the process of breaking up a stream of text characters into meaningful elements (e.g. presence of a space is used to designate ' the bird' as two tokens,' the' and 'bird'
Feature – a token that the researcher judges to be relevant to the text task.
Parsing – Generally, the process of systematically disassembling a text into meaningful components (such as sentences or words). In NLP, a formal methodology for labeling specific words in a sentence according to linguistic rules (see *Stanford Parser*).
Normalization – eliminating differences in punctuation, such as removing capitalization
Stemming - process for reducing words to their stem, base or root form (e.g. fishing–fish)
Stop Word – common words that are not considered to be valuable features of a text and are therefore excludes (e.g. the)
Concordance, Collocation or Cooccurence – incorporating the context in which a word is used into its meaning, for example by examining word sequences (n-grams) instead of just words in isolation.
Regular expression – a concise but often flexible pattern intended to recognize strings of text (such as any date or any url)
Disambiguation – process of linking references to a single entity or topic. For example, in blogs, references to President Obama might take different forms 'Barack,' ' Obama,' 'The One,' 'the President' etc. Alternately, reconciling different spellings of the same word.
Named entities - elements of text that are to be classified into predefined categories (e.g. person names, organizations, locations, subjects, percentages, etc).
Semantic – Broadly, text that has meaning (e.g. a word rather than a hyperlink). Usually used in reference to natural language processing approaches that are concerned with linguistic structure
Sentiment – refers to polarity in classification (e.g. from hate to love, liberal to conservative, etc). Not necessarily single-dimensional.
Algorithm – a mathematical set of instructions about how to convert a set of inputs to an output. In automated content analysis, researchers select from a wide variety of off the shelf algorithms suited to different tasks (one of the most popular is SVM).
Machine learning – Generally, the ability of a computer program to get better at a task with more information. The clearest examples are supervised machine learning algorithms. A set of hand labeled examples (e.g.) is used to train an algorithm (how does the text of the examples in one category differ from the text of the examples in other categories?). The algorithm then predicts the categories of unseen events based on their text.

Bag of Words (BoW) –text analysis approaches that consider words as features in isolation (as opposed to NLP or concordance approaches that value relationships among words)

Natural Language Processing (NLP) –text analysis approaches that value grammatical (linguistic) information such as word order or sentence structure (subject-verb-object).