Department of Sociology
THE UNIVERSITY OF CHICAGO

SOCIOLOGY 40133

Computational Content Analysis

Friday 1:00 – 3:50pm
Winter 2017-2018
Classroom: Harper Memorial 130
http://chalk.uchicago.edu/

James A. Evans
Office: McGiffert 210
Tel.: 834-3612; jevans@uchicago.edu
Office Hours: Thursday 12:30-2:30pm

TAs: Reid McIlroy; reidmcy@uchicago.edu
Office Hours: Monday 11:00-1:00pm
Linzhuo Li; linzhuoli@uchicago.edu
Office Hours: Wednesday 11:00-1:00pm
(also) Office: McGiffert 210

I. The Course

A vast expanse of information about what people do, know, think, and feel lies embedded in text, and more of the contemporary social world lives natively within electronic text than ever before. These textual traces range from collective activity on the web, social media, instant messaging and automatically transcribed YouTube videos to online transactions, medical records, digitized libraries, and government intelligence. This supply of text has elicited demand for natural language processing and machine learning tools to filter, search, and translate text into valuable data. The course will survey and practically apply many of the most exciting computational approaches to text analysis, highlighting both supervised methods that extend old theories to new data and unsupervised techniques that discover hidden regularities worth theorizing. These will be examined and evaluated on their own merits, and relative to the validity and reliability concerns of classical content analysis, interpretive concerns of qualitative content analysis, and interactional concerns of conversation analysis. We will also consider how these approaches can be adapted to content beyond text, including audio, images, and video. We will simultaneously review recent research that uses these tools to develop social insight by exploring (*a*) collective attention and reasoning through the content of communication; (*b*) social relationships through the process of communication; and (*c*) social states, roles, and moves through heterogeneous signals within communication.

The course is structured around gaining understanding and experimenting with text analytical tools, deploying those tools and interpreting their output in the context of individual research projects, and assessment of contemporary research within this domain. Class discussion and assignments will focus on how to use, interpret, and combine computational techniques in the context of compelling social science research investigations. Python is the programming language in which assignments are modeled. Some programming experience is assumed.

II. Readings and Computational Tools

Readings will be circulated in class and through the course's Canvas site.

Computational tools required to fulfill the exploration assignments are modeled in python, and will most easily be replicated and extended in python, although other languages may be used for final projects (e.g., R, Julia, Matlab, etc.) Students must perform weekly homework in the course within notebooks students can download online and run from their own computers and/or run from the Research Computing Center's machines.

III. Course Requirements

A. READING and DISCUSSION (20%)

Content analysis is a seminar. Students are expected to read and reflect on assigned readings before class. Participation in class discussion is expected of all class members. Each class, there will typically be (1) one conceptual or orienting reading; (2) three examplary content analyses, of which students will be asked to peruse one, and skim the others (read abstract, methods section, tables/figures, conclusion); and (3) one more technical reading that details issues regarding the topic under investigation. Students will post a question (posed as an "issue") on GitHub relating to each category of reading each week, and vote five other questions a "thumbs up" by 10 am the Friday morning of class. Specifically, this means that students will post a question about the orienting reading; one of the three exemplary readings; and the fundamentals reading(s). They will also "vote up" 5 other questions in each category (orienting, the particular exemplary reading they perused, and the fundamentals reading(s)).

Once per quarter, students will also sign-up to perform a 5-minute summary, critical analysis, and alternative or extension analysis possibility in association with each of the three "exemplary" content analyses, presented in Ignite Talk format—5 minutes, 20 PDF slides, auto-advancing every 15 seconds (see a description and example here). Once per quarter, students will also sign-up to perform a 5-minute discussion of preliminary (juicy) results from analyzing their corpus using the Jupyter notebook code and requested extensions (described below) with a focus on analysis outputs, visualizations and content-driven stories. This is also to be presented in Ignite Talk format, where slides should include *interesting* screenshots from your Jupyter notebook homework assignment. All presenting students will post 20-PDF slide decks prior to class. Only three reading presentations and three code presentations will be (randomly) selected for presentation in class. Those remaining will post a 5-minute video of their presentations on Canvas.

Students (and my own) presentations should respond to highly up-voted questions posed on GitHub, which is why they are to be posted online three hours before class. The reading and discussion grade will be a function of questions asked, up-votes received, and the quality of code and reading presentations.

B. WEEKLY EXERCISES (40%)

Tuesday prior to each class, we will publish a Jupyter interactive Python notebook containing code examples and requested extensions associated with the skills we are learning that week. Students are expected to have familiarized themselves with the code before class to facilitate exploration with their own corpus during class.

Students are expected to complete a weekly homework and short memo intended to invite experimentation with the techniques we consider. Homework will consist of performing all operations and requested extensions from the published notebook on a corpus or corpora associated with students' anticipated final projects (see below), followed by a short, structured memo that effectively (1) summarizes results from preliminary analysis using content associated with your anticipated final project topic (e.g., with text and graphics or succinct tables), (2) identifies and interprets textual examples that facilitate qualitative validation of the patterns summarized, (3) critically evaluates the method's drawbacks and scope conditions for its beneficial deployment. As noted earlier, three students will sign-up to prepare the substance of these assignments and lead a discussion about their results before the assignments are formally due.

Student notebooks must be pushed at midnight on Wednesday following the relevant class (although five or six students will have presented results from this assignment on the prior Friday). Students are encouraged to work on these assignments in groups (although you will turn them in individually) and to key assignments to the substance of the final project.

Grading will follow the simplified emoji scheme: =) (~A), 8| (~B+), -_- (~B-), 0 (~F). I will drop the two lowest memo grades and so students will be required to produce between 6 and 8 memos over the course of the quarter for full credit.

D. FINAL PROJECT (40%)

Students will perform a substantial content analysis for social insight based on approaches and tools developed/explored over the course of the quarter. These projects must incorporate techniques from at least four weeks over the course of the quarter, and must validate inferences with qualitative interpretation and assessments. These projects may be performed as a group (of no more than 3 students), but the work should be proportional to the effort demonstrated in the project.† The motivation, process and findings from this project will be presented in a 5-minute Ignite style talk on Thurday, March 15 from 4:30pm. Students will submit (1) final presentation slides by Thursday, March 15 @ 4:30pm; (2) annotated code, and a (3) detailed 20 page research appendix that contains methodological justification and description, descriptive data and analysis, interpretation of findings and conclusion by Friday, March 16 @ 5pm.

I encourage students to develop these projects based on your own project and corpus, if sufficiently mature. We will also work to prepare some interesting corpora that groups may use for class projects, including (1) more than a century of English-language sociology, (2) more than a century of English-language news (from the U.S., England and India), (3) a corpus of 17th Century dialog, (4) English Wikipedia, (5) Reddit threads, (6) Clinical cancer abstracts, (7) Physics abstracts.

†Group projects should demonstrate more overall quality—closer to publication or conference presentation—because more iterations have been performed with more or better collections of text, more cleaning/refining of the relevant corpus, more iterations on the analysis to optimize the parameters, more meaningful results, better visualizations to convey the stories those results hold, etc. I do *not* mean that group project should be longer or (needlessly) perform more techniques.

IV. Calendar of Reading Assignments.

Week 1.

Jan.5: *General Introduction; Measuring Meaning & (Computationally) Reading Text*

Evans, James and Pedro Aceves. 2016. "Machine Translation: Mining Text for Social Theory". *Annual Review of Sociology* 42:21-50. DOI: 10.1146/annurev-soc-081715-074206

Using Regular Expressions in Python
https://regexone.com/lesson/introduction_abcs
https://regexone.com/references/python

General Python Tutorial
https://docs.python.org/3/tutorial/

Goodfellow, Ian, Yoshua Bengio & Aaron Courville. 2016. *Deep Learning*. MIT Press: Chapter 2-3 "Linear Algebra" and "Probability and Information Theory": 29-76.

Python Tutorial/Refresher led by TA Reid McIlroy on Monday, January 8, 11-1pm (e.g., during office hours) in SSR 404.

Week 2.

Jan.12: *Accounting for Words & Phrases [Corpus Linguistics]*

Michel, Jean-Baptiste et al. 2010. "Quantitative Analysis of Culture Using Millions of Digitized Books." *Science* express, December 16.

E1. Danescu-Niculescu-Mizil, Cristian, Lillian Lee, B. Pang, and John Kleinberg. 2012. "Echoes of Power: Language Effects and Power Differences in Social Interaction." *Proc. 21st Int. Conf. World Wide Web*, pp. 699–708.

E2. Gentzkow, Matthew & Jesse M. Shapiro. 2007. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78(1): 35–71.

E3. Chen, Keith. 2013. "The Effect of Language on Economic Behavior: Evident from Savings Rates, Health Behaviors, and Retirement Assets." *American Economic Review* 103(2): 690-731. (But you must also skim the quasi-retraction in Roberts, Seán G., James Winters, and Keith Chen. 2015. "Future Tense and Economic Decisions: Controlling for Cultural Evolution." *PLOS ONE* 10(7): e0132145.)

Manning and Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press: Chapter 5 ("Collocations"): 151-189.

Week 3.
Jan.19: *Discovering higher-level Patterns [Clustering and Topic Modeling]*

Timmermans, Stefan and Iddo Tavory. 2012. "Theory Construction in Qualitative Research: From Grounded Theory to Abductive Analysis." *Sociological Theory* 30(3) 167–186.

E1. Grimmer, Justin and Gary King. 2011. "General purpose computer-assisted clustering and conceptualization." *PNAS* (Feb. 3).

E2. Grimmer, Justin. 2013. "Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57(3): 624-642.

E3. Nelson, Laura K. 2015. "Persistent Political Logics: Geographical Differences and Temporal Continuities within the Women's Moevements in Chicago and New York City." Working paper. (But you must also skim the framework she draws from her analysis in Nelson, Laura K. 2017. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods & Research DOI: 10.1177/0049124117729703:* 1-40.)

Manning, Christopher, Prabhakar Raghavan and Hinrich Schütze. 2008. "Flat Clustering" and "Hierarchical Clustering." Chapters 16 and 17 from *Introduction to Information Retrieval.*

Blei, David. 2012. "Probabilistic Topic Models". *Communications of the ACM* 55(4):77-84.

Week 4.
Jan.26: *Exploring Semantic Spaces [Word & Document Embeddings]*

Osgood, Charles E., George J. Suci, Percy Tannenbaum. 1957. *The Measurement of Meaning*. Chapter 1 ("The Logic of Semantic Differentiation"):1-30.

E1. Caliskan, Aylin, Joanna J. Bryson, Arvind Narayanan. 2017. "Semantics derived automatically from language corpora contain human-like biases." *Science* 356(6334):183-186.

E2. Hamilton, William H., Jure Leskovec, Dan Jurafsky. 2016. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change." *arXiv preprint arXiv:1605.09096*. (All code and data available here. But also see Nikhil Garg, Londa Schiebinger, Dan Jurafsky and James Zou's follow-on article, 2017. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *arXiv preprint arXiv:1711.08412*.)

E3. Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings." ArXiv.org: 1607.06520.

Jurafsky, Daniel and James H. Martin. 2015. Speech and Language Processing. Chapters 15-16 ("Vector Semantics", "Semantics with Dense Vectors")

Week 5.
Feb.2: *Sampling, Crowd-Sourcing and Reliability [Modeling Accuracy]*

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to its Methodology*. Thousand Oaks, CA: Sage: "Sampling" 111-124.

E1. Roscigno, Vincent J. and Randy Hodson. 2004. "The Organizational and Social Foundations of Worker Resistance." *American Sociological Review* 69(1): 14-39. Also read coding protocol, available on the chalk site.

E2. Dodds, Peter Sheriden et al. "Human language reveals a universal positivity bias." *Proceedings of the National Academy of Sciences* 1112(8):2389–2394, doi: 10.1073/pnas.1411678112

E3. Salesses, Philip, Katja Schechtner, and César Hidalgo. 2013. "The Collaborative Image of The City: Mapping the Inequality of Urban Perception." *PLoS ONE* 8(7):e68400. doi:10.1371/journal.pone.0068400

Dawid, A. P., and Skene, A. M. 1979. "Maximum Likelihood Estimation of Observer Error-rates using the EM Algorithm." *Applied Statistics* 28(1): 20-28.

****Pitch Texts/Context for Final Papers Due in this Week's Homework****

Week 6.
Feb.9: *Classifying Concepts, Accounts, and Documents [Machine Classification]*

Hopkins, Daniel J. and Gary King. 2010. A Method of Automated Nonparametric Content Analysis for Social Science. *American Journal of Political Science* 54(1): 229-247.

E1. Cheng, Justin, Michael Bernstein, Cristian Danescu-Niculescu-Mizil and Jure Leskovec. 2017. "Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions." *WWW* 2017: 1-14.

E2. So, Richard and Hoyt Long. 2015. "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning." *Critical Inquiry* 42(2): 235-267.

E3. Klingenstein, Sara, Tim Hitchcock, and Simon Dedeo. 2014. "The Civilizing Process in London's Old Bailey." *PNAS* 111(26):9419-9424.

Manning, Christopher, Prabhakar Raghavan and Hinrich Schütze. 2008. "Text Classification and Naïve Bayes", "Vector Space Classification," and "Support Vector Machines." Chapters 13-16 from *Introduction to Information Retrieval*: 234-320.

Witten, Ian H., Eibe Frank, Mark A. Hall, Christopher J. Pal. 2017. "Ensemble Learning" Chapter 12 from *Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition*: 351-371.

Week 7.
Feb.16: *Extracting & Modeling Claims, Arguments, Events [Computational Linguistics]*

Franzosi, Roberto. 1994. "From Words to Numbers: A Set Theory Framework for the Collection, Organization, and Analysis of Narrative Data." *Sociological Methodology* 24:105-136.

E1. Mohr, John W., Robin Wagner-Pacifici, Ronald L. Breiger, and Petko Bogdanov. 2013. "Graphing the grammar of motives in National Security Strategies: Cultural interpretation, automated text analysis and the drama of global politics" *Poetics* 41(6):670-700.

E2. Ignatow, Gabriel. 2004. "Speaking Together, Thinking Together? Exploring Metaphor and Cognition in a Shipyard Union Dispute." *Sociological Forum* 19(3): 405-433.

E3. Schein, Aaron, David Blei, John Paisley, Hannah Wallach. 2015. "Bayesian Poisson Tensor Factorization for Inferring Multilateral Relations from Sparse Dyadic Event Counts." *KDD '15* August 11–13, 2015, Sydney, NSW, Australia.

Manning and Schütze. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press: Chapter 3 ("Linguistic foundations"): 81-113.

Jurafsky, Daniel & James H. Martin. 2017 (3rd Edition). *Speech and Language Processing.* Singapore: Pearson Education, Inc.: Chapter 22 ("Information Extraction"): 739-778.

Week 8.

Feb.23: *Content Relations [Network Analysis]*

Carley, Kathleen. 1993. "Coding Choices for Textual Analysis: A Comparison of Content Analysis and Map Analysis." *Sociological Methodology* 23:75-126.

E1. Goldberg, Amir, Sameer B. Srivastiva, V. Govind Manian, William Monroe and Christopher Potts. 2016. "Fitting In or Standing Out? The Tradeoffs of Structural and Cultural Embeddedness". American Sociological Review 81(6): 1190-1222.

E2. Aral, Sinan and Marshall Van Alstyne. 2011. "The Diversity-Bandwidth Trade-off." *American Journal of Sociology* 117(1): 90-171.

E3. Feng, Shi, Jacob G. Foster, James A. Evans. 2015. "Weaving the fabric of science: Dynamic network models of science's unfolding structure." *Social Networks* 43 (October 2015):73–85.

Easley, David and Jon Kleinberg. 2010. Chapter 4.4 ("Tracking Link Formation in Online Data") and 13 ("The Structure of the Web") from *Networks, Crowds, and Markets: Reasoning about a Highly Connected World.*

Week 9.

Mar.2: *Beyond Text*

1) *Discourse, Conversation, and Music [Sound/Signal Processing]*
2) *Scenes, Visual events and Art [Machine Vision]*

Discourse & Scenes

Goffman, Erving. 1981. Conclusion (**only pps. 308-311**) from "Replies and Responses." *Language in Society* 5 (3): 257-313.

Collins, Randall. 2009. "The Micro-sociology of Violent Confrontations" and "Confrontational Tension and Incompetent Violence" (beginning of Chapter 2) from *Violence: A Microsociological Theory*: 37-43.

E1. McFarland, Daniel A., Dan Jurafsky, and Craig Rawlings. 2013. "Making the Connection: Social Bonding in Courtship Situations." *American Journal of Sociology*, 118(6):1596-1649.

E2. Pentland, Alex. 2012. "The New Science of Building Great Teams." *Harvard Business Review*, April 2012. http://hbr.org/2012/04/the-new-science-of-building-great-teams/ar/1

E3. Naik, Nikhil, Scott Duke Kominers, Ramesh Raskar, Edward L. Glaeser, César A. Hidalgo. 2017. "Computer vision uncovers predictors of physical urban change." PNAS 114(29):7571–7576.

Jurafsky, Daniel & James H. Martin. 2017 (3rd Edition). *Speech and Language Processing*. Singapore: Pearson Education, Inc.: Selections from Chapter 9, "Automated Speech Recognition", 291-312.

LeCun, Yann, Yoshua Bengio & Geoffrey Hinton. 2015. "Deep Learning." *Nature* 521: 436-444.

Week 10.
Mar.9: Reading Days – No Class

Week 11.
Mar.15: Thursday, 4:30pm, Content Fair and presentation of projects

****Turn in Final Project due Friday, March 16 at 5pm****