Soc 211 Computational Methods in Social Science
Spring 2014, Wednesdays, 3-5:30pm, SSB101
Instructors: Edward Hunter and Akos Rona-Tas
TA: K. Sree Harsha
edwardhunter@ucsd.edu
aronatas@ucsd.edu
kharsha@cs.ucsd.edu

Overview

The rapid proliferation of the internet, mobile devices, social media and related technologies in daily life has presented a profound opportunity for social scientists to pose and test theories of social phenomena in novel advantageous ways. Our lives are now conducted online in such scope and variety, from social interactions to political blogging, topical web searching and consumption of digitally archived and accessed media, that online digital artifacts are beginning to be probed to reveal patterns of social organization and behavior in situ and at scales impossible to achieve using traditional social scientific methodology. A new cross-disciplinary field has emerged, Computational Social Science (CSS), with the aim of exploiting this opportunity. This course provides an overview of the current state of CSS and a practical, hands- on introduction to some of the most common tools employed in CSS. It is intended that students will complete the course with sufficient knowledge and experience to use existing tools to conduct CSS research, and with sufficient preparation for further formal or self-directed study into methods development and evaluation.

The course consists of three components: (1.) In class presentations and discussions of the most current CSS literature covering methods, applications, validation, and philosophical issues; (2.) Lectures covering background material and the most common machine learning tools applicable to social science research; (3.) Homework laboratories exercising the methods using standard text corpora from the machine learning community.

Lecture and Laboratories Outline

Week	Lecture Subject	Lab Assigned
4/2	Overview and Intro to CSS	Lab 0: Verify Course Account
4/8	Intro to Python	Lab 1: Python Basics
4/16	Overview of Probability	None
4/23	Naive Bayes Models	Lab 2: Naive Bayes Classification
4/30	Nearest Neighbor Models	Lab 3: kNN Classification
5/7	Support Vector Machines	Lab 4: SVM Classification
5/14	k-Means, Spherical k-Means	Lab 5: Spherical k-Means Clustering
5/21	Hierarchical and/or Spectral Clustering	Lab 6: Clustering 2
5/28	Non-Negative Matrix Factorization	Lab 7. NMF Topic Identification and Clustering
6/4	TBD	None

Possible Items for TBD Lecture(s)

- Student Interest.
- Topic Models, Latent Dirichlet Allocation (no verified Python impl. yet).
- Ideological Scaling.
- EM Algorithm in General and on the Unit Hypersphere.
- Twitter API Data Harvesting.
- Analysis of Student Datasets.
- Databases Setup and Use for Managing CSS Datasets.
- Visualization Basics with matplotlib.
- Bag of Words Processing and NLP Basics in Python with NLTK.

Course Website

Access: ted.ucsd.edu.

We will post lecture slides, readings and student presentations on the course website as we go. See below for details.

Grading

* 50% Class participation and reading presentations.
* 50% Lab reports and in class discussion.

Office and Lab Hours:

TA/Computer Lab: CSB 115, Mondays and Fridays 9AM-1PM (tentative).
Rona-Tas SSB 488, TTh 11-11:50AM.
Hunter: Wednesday after class or by appointment.

The Computer Lab

Location: CSB 115 (Cognitive Science Building).
Access Code: TBD

Enrolled students have all been assigned a course account (e.g. username like so290s**) that you can log into using your UCSD email account as password.

Account lookup tool:	https://sdacs.ucsd.edu/~icc/index.php
Password setup page:	http://acms.ucsd.edu/students/gpasswd.html
Host Machines:	so290s**@ieng6.ucsd.edu (standard server does not allow UI access remotely) so290s**@ieng6-240 through so290s**@ieng6-254 (use these remotely for UI access with VCNgnome)
Remote Access UI:	http://acms.ucsd.edu/info/vncgnome.html

NOTE: Do not leave processing running on these machines as they will be killed by the administrator. The labs can be completed in the time of a normal interactive session (maybe with a coffee break to wait for something in some cases).

ACMS Contact Info for Account Troubleshooting:

Maryam Sarkhosh
maryam@acsmail.ucsd.edu
Help Desk: (858) 534-ACMS (2267)

Laboratories

Labs will be done in Python using template programs and lab instructions we provide. The labs are designed for minimal programming background and provide simple step-by-step instructions for using Python and scikit-learn to conduct CSS experiments.

Reports as PDFs are due as uploads to the course website the day prior to the following class meeting. Name the pdf file soc211_lab##_lastname_firstname.pdf. There are no reports for Lab 0 (setting up and testing account) and Lab 1 (basic Python exercises). Discuss account and tool troubleshooting with the TA in lab hours, and we will discuss basic python exercises in class.

Readings and Presentations

Students will work together in 8 groups of 3-5. Each meeting, 2-3 groups will be responsible to present one of the readings with a short powerpoint slideshow. It is up to student groups to work together outside of class to discuss the readings, collect their summary and analysis of papers and construct the slides. Groups can decide who will responsible for presenting. Since there are 2-3 presentations per group, different students can present different papers. In a few cases, papers are longer and presentation can be split if that is preferred (but please construct and submit only one presentation pdf).

Slide presentations as PDFs are due as uploads to the course website the day prior to the date of presentation. Name the pdf file soc211_group##_leadauthor_year.pdf.

Readings Presentations by Week:

Date/Group	Theme/Papers
4/8	**Big Data Issues, Web Science**
1	Boyd, D., & Crawford, K. (2012)
2	Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013)
3	Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., & Weitzner, D. (2008)
4/16	**Text Analysis, Machine Learning**
4	Domingos, P. (2012)
5	Grimmer, J., & Stewart, B. M. (2013)
4/23	**Measuring Culture, Political Movements, Language Dynamics**
6	Bail, C. A. (2014)
7	Hanna, A. (2013)
8	Jensen, J., Kaplan, E., Naidu, S., & Wilse-Samson, L. (2012)
4/30	**Sentiment, Implied Support and Opposition**
1	Thomas, M., Pang, B., & Lee, L. (2006)
2	Pang, B., Lee, L., & Vaithyanathan, S. (2002)
5/7	**Classification of Political Language**
3	Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011)
4	Purpura, S., & Hillard, D. (2006)
5	Hillard, D., Purpura, S., & Wilkerson, J. (2008)
5/14	**Clustering**
6	Grimmer, J., & King, G. (2011)
7	Hopkins, D. J., & King, G. (2010)
5/21	**Topic Models, Latent Variables**
8	Blei, D. M. (2012)
TBD	DiMaggio, P., Nag, M., & Blei, D. (2013)
5/28	**Ideological Scaling**
TBD	Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008)
TBD	Laver, M., Benoit, K., & Garry, J. (2003)

Required Readings for Presentation

1. Boyd, D., & Crawford, K. (2012). Critical Questions for Big Data. Information, Communication & Society, 15(5), 662–679. doi:10.1080/1369118X.2012.678878
2. Morstatter, F., Pfeffer, J., Liu, H., & Carley, K. M. (2013). Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose (pp. 1–10). Presented at the ICWSM 2013.
3. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., & Weitzner, D. (2008). Web science. Communications of the ACM, 51(7), 60. doi:10.1145/1364782.1364798
4. Domingos, P. (2012). A few useful things to know about machine learning. Communications of the ACM, 55(10), 78. doi:10.1145/2347736.2347755
5. Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis, 21(3), 267–297. doi:10.1093/pan/mps028
6. Bail, C. A. (2014). Measuring Culture with Big Data. Theory & Society, 1–24.
7. Hanna, A. (2013). COMPUTER-AIDED CONTENT ANALYSIS OF DIGITALLY ENABLED MOVEMENTS. Mobilization: an International Quarterly, 18(4), 367–388.
8. Jensen, J., Kaplan, E., Naidu, S., & Wilse-Samson, L. (2012). Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech. Brookings Papers on Economic Activity, 2012(1), 1–81. doi:10.1353/eca.2012.0017
9. Thomas, M., Pang, B., & Lee, L. (2006). Get out the vote: Determining support or opposition from Congressional floor-debate transcripts (pp. 327–335). Presented at the Proceedings EMNLP 2006.

10. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: sentiment classification using machine learning techniques (Vol. 10). Presented at the EMNLP '02: Proceedings of the ACL-02 conference on Empirical methods in natural language processing, Association for Computational Linguistics.
11. Diermeier, D., Godbout, J.-F., Yu, B., & Kaufmann, S. (2011). Language and Ideology in Congress. British Journal of Political Science, 42(01), 31–55. doi:10.1017/S0007123411000160
12. Purpura, S., & Hillard, D. (2006). Automated Classification of Congressional Legislation (pp. 1–7). Presented at the The 7th Annual International Conference on Digital Government Research '06.
13. Hillard, D., Purpura, S., & Wilkerson, J. (2008). Computer-Assisted Topic Classification for Mixed-Methods Social Science Research. Journal of Information Technology & Politics, 4(4), 31–46. doi:10.1080/19331680801975367
14. Grimmer, J., & King, G. (2011). General purpose computer-assisted clustering and conceptualization. Proceedings of the National Academy of Sciences, 108(7), 2643–2650.
15. Hopkins, D. J., & King, G. (2010). A method of automated nonparametric content analysis for social science. American Journal of Political Science, 54(1), 229–247.
16. Blei, D. M. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77. doi:10.1145/2133806.2133826
17. DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. government arts funding. Poetics, 41(6), 570–606. doi:10.1016/j.poetic.2013.08.004
18. Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. Political Analysis, 16(4), 372–403. doi:10.1093/pan/mpn018
19. Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. American Political Science Review, 97(02), 311–331.

Supplemental Readings

1. Lohr, S. (2012, 2/11/2012). The Age of Big Data. The New York Times, pp. 1–6.
2. Simonite, T. (2021, June 13). What Facebook Knows. MIT Technology Review, 1–13. Retrieved from http://www.technologyreview.com/featuredstory/428150/what-facebook-knows/
3. Giles, J. (2012). Making the Links. Nature, 488, 448–450.
4. Anderson, C. (2008, June). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired, 16(7), 1–2. Retrieved from http://www.wired.com/science/discoveries/magazine/16-07/pb_theory
5. King, G. (2011). Ensuring the Data-Rich Future of the Social Sciences. Science, 331(6018), 717–719. doi:10.1126/science.1197448
6. Golder, S. A., & Macy, M. W. (2012). Social Science with Social Media (pp. 1–20). Presented at the American Sociological Association Footnotes.
7. Golder, S. A., & Macy, M. W. (2013). Social Media as a Research Environment. Cyberpsychology, Behavior, and Social Networking, 16(9), 627–628. doi:10.1089/cyber.2013.1525
8. Shneiderman, B. (2007). Web Science: A ProvocativeInvitation to Computer Science. Communications of the ACM, 50(6), 25–27.
9. Latour, B. (7AD, April 6). Beware, your imagination leaves digital traces. The Times Higher Literary Supplement, pp. 1–3.
10. Gray, J. (2009). Jim Gray on eScience:A Transformed Scientific Method. In T. Hey, S. Tansley, & K. Tolle (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery (pp. xvii–xxxi). Microsoft Research.
11. Savage, M., & Burrows, R. (2009). Some Further Reflections on the Coming Crisis of Empirical Sociology. Sociology, 43(4), 762–772. doi:10.1177/0038038509105420
12. Savage, M., & Burrows, R. (2007). The Coming Crisis of Empirical Sociology. Sociology, 41(5), 885–899. doi:10.1177/0038038507080443
13. Crompton, R. (2008). Forty Years of Sociology: Some Comments. Sociology, 42(6), 1218–1227. doi:10.1177/0038038508096942
14. Levallois, C., Steinmetz, S., & Wouters, P. (2012). Sloppy Data Floods or Precise Social Science Methodologies? Dilemmas in the Transition to Data-Intensive Research in Sociology and Economics. In Experimenting in the Humanities and the Social Sciences (pp. 1–32). MIT Press.
15. Zhao, S. (2006). The Internet and the Transformation of the Reality of Everyday Life: Toward a New Analytic Stance in Sociology. Sociological Inquiry, 76(4), 458–474.
16. Roberts, C. W. (2000). A Conceptual Frameworkfor Quantitative Text Analysis. Quality & Quantity, 34, 259–274.
17. Yu, B., Kaufmann, S., & Diermeier, D. (2008). Classifying Party Affiliation from Political Speech. Journal of Information Technology & Politics, 5(1), 33–48. doi:10.1080/19331680802149608

18. Chung, C. K., & Pennebaker, J. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. Journal of Research in Personality, 42(1), 96–132. doi:10.1016/j.jrp.2007.04.006
19. Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., et al. (2008). Top 10 algorithms in data mining. Knowledge and Information Systems, 14(1), 1–37. doi:10.1007/s10115-007-0114-2
20. Aggarwal, C. C., & Zhai, C. (2012). A Survey of Text Clustering Algoirthms. In Mining Text Data (pp. 77–127). Springer.
21. Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical Pattern Recognition: A Review. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22, 4–37.
22. Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34(1), 1–47.
23. Sagi, E., Diermeier, D., & Kaufmann, S. (2013). Identifying Issue Frames in Text. PLoS ONE, 8(7), e69185. doi:10.1371/journal.pone.0069185.s002
24. Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. Nature, 489(7415), 295–298. doi:10.1038/nature11421
25. Evans, M. S. (2014). A Computational Approach to Qualitative Analysis in Large Textual Datasets. PLoS ONE, 9(2), e87908. doi:10.1371/journal.pone.0087908.g009
26. Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National Academy of Sciences, 101(Suppl 1), 5228–5235.
27. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3, 993–1022.
28. Conover, M., Ratkiewicz, J., Francisco, M., Gonçalves, B., Menczer, F., & Flammini, A. (2011). Political Polarization on Twitter. Presented at the ICWSM.
29. Hornik, K., Geinerer, I., Kober, M., & Buchta, C. (2012). Spherical k-Means Clustering. Journal of Statistical Software, 50, 1–22.
30. Oboler, A., Welsh, K., & Cruz, L. (2012, July). The danger of big data: Social media as computational social science. First Monday. Retrieved March 10, 2014, from
31. Mohr, J., & Bogdanov, P. (2013). Topic models: What they are and why they matter. Poetics, 41(6), 545.
32. Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature, 401(6755), 788–791.