|  |  |
| --- | --- |
| **DATA MINING**  ***Updated April 27, 2020*** | |
| Course: GSR – 79038  Hunter College  Spring 2020  Monday- 5:35PM-7:25PM  Classroom: HW1631 | Instructor: Nga Than  Office: HW1636  Office Hours: make a 30-minute appointment at <https://calendly.com/ngathan>  Email: [nthan@gradcenter.cuny.edu](mailto:nthan@gradcenter.cuny.edu) |

**COURSE DISCRIPTION**

A tremendous amount of data is now being generated and collected through digital technology such as mobile phones, websites, and other electronic devices. How can social scientists take advantage of these digital trace data to answer pressing social questions? This winter course introduces students to basic concepts and methods in computational social science. It aims to show students new ways of thinking about data, conducting social analysis, and examining social phenomena through the lens of computational social science. Students will learn techniques such as web scraping, conducting automated text analyses, and incorporating machine learning into their research projects. By the end of the course, students will be able to conduct research using R programming language to analyze big data.

**LEARNING OUTCOMES**

* Familiarity with R Programming Language
* Learn what digital trace data is, and how to use API to extract digital trace data
* Analyze big data sets using random forest, support vector machine
* Use text data from social media to analyze social phenomena
* Familiarity with or structural topic modeling and sentiment analysis
* Formulating sociological questions and addressing them on the basis of evidence.
* Research and writing about a topic of particular interest to the student. Research could be an individual project or a group project

**REQUIRED TEXTBOOKS:**

Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.

Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.

***Optional***

Salganik, M. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

***Extra Resources***

Class Github page: <https://github.com/ngathan/DataMiningS2020>

<https://onboarding.datacamp.com/>

<https://stackoverflow.com/>

**COURSE REQUIREMENTS**

**Grades:**

|  |  |  |
| --- | --- | --- |
| **Item** | **Date** | **Weight (%)** |
| Problem Sets (10) | Mondays | 30.0 |
| Take-home Midterm | March 16 | 20.0 |
| Research Project | May 22 | 40.0 |
| Final Reflection | May 22 | 10.0 |

The grading system in this class is as follows:

|  |  |
| --- | --- |
| A | 93-100 |
| A- | 90-92 |
| B+ | 87-89 |
| B | 83-86 |
| B- | 80-82 |
| C+ | 76-79 |
| C | 74-75 |
| C- | 70-73 |
| D | 60-69 |
| F | <=59 |

***Attendance/class participation***

Regular attendance is one of the most important parameters to successful completion of the course requirements. Even though I will not take attendance, I expect that you read required readings before class, and participate in classroom discussions.

***Problem Sets:***

There are 10 problem sets that cover the class materials. Each problem set has 5-10 questions from the class exercises. Each problem set is worth 3 points. Completion of these problem sets ensures that you have a grasp of the weekly readings, in-class discussions, and lab exercises. The weekly problem sets are due at the beginning of each class on Mondays.

***Take-home Midterm Exam***

Midterm exam is a take-home exam, which you can do from the period of March 09 to March 16. You have to do this exam on your own, you can consult resources from the textbook, and online resources to solve real life problems (ethical considerations in data science, and predictions).

***Final Research Project***

The final research project would be a student-lead research project where you can form a team to formulate a research question to analyze a social problem. You can either form a team or conduct this research on your own. The final product would be a 20-page original research paper that answers the questions why the problem is important, where you get your data, and what technique you use to analyze the problem.

We will use our lab sections in the class to discuss what project you want to carry out, and how to formulate a well-research question, and what technique to use for your particular research project.

Grading of research papers. To receive an “A,” a paper must be excellent in the following three areas.

1. Research and demonstrated knowledge. The paper should cite a good range of quality scholarly sources and demonstrate depth of knowledge based on working with these sources.
2. Organization and argument. The paper should address the assigned topic, present a well-supported argument, and draw appropriate conclusions.
3. Writing and documentation. The paper should be clearly and correctly written, and footnotes and bibliography should be complete and correct.
4. Have reproducible R code

To receive a grade in the “B” range, a paper must be good in at least three of the four areas.

***Final Reflection***

If you choose to do a group research project, you should write a final reflection essay to discuss your team work process such as what role you assumed in the project, and how did you come to that role, and what have you learned from the process of researching, coding, and writing a final paper with others.

If you choose to do the research on your own, you should write a final reflection essay about the different challenges that you face while doing the research, what kind of technical problems did you encounter, how did you troubleshoot various technical issues, what have you learned from presenting data.

The final reflection should be maximum ***three*** pages.

**STATEMENT ON ACADEMIC HONESTY**

As a member of this class, you are part of both the Hunter College community and a larger community of social scientists. You have joined two communities that are devoted to sharing and creating knowledge. This process depends, however, on the intellectual honor and correct scholarly techniques of everyone involved. Always acknowledge the sources of ideas and information. Borrowing even a sentence or two from someone else’s work without proper citation is plagiarism. As you conduct your research, be scrupulous in noting and attributing your sources. The reader of your paper should be able to retrace the steps that you took in carrying out your research. We will discuss this further in class, but by all means, see me if you have questions.

**PLEASE READ CAREFULLY!**

**Laptop:** You should always carry your own laptop to the class for lab work. Since we use an open-source software, it is best for you to use it on your own laptop.

**Make-up exam** will be given only in cases of severe illness or crisis. If you miss the midterm exam, contact me as soon as possible before the test to discuss your reasons for postponement.

**Extra Resources**

Please get to know the research librarians at Hunter College, and they are extremely nice, and helpful.

If you need help about how to social computing, please visit The Social Sciences Computing Lab (Level B1 of Hunter East).

If you have any question regarding to your requirement for your master’s program, please contact Nanda Kalimootoo (Room West Building, 1622).

If you have any question pertaining to the course materials, assignments, please contact me at [nthan@gradcenter.cuny.edu.](mailto:nthan@gradcenter.cuny.edu.)

**COURSE SCHEDULE**

**Part I - Introduction & Ethical Issues in Data Science**

Week 1 (Jan - 27): Introduction

* Lab: Setting up Slack, R & Rstudio, Github repo of class, class webpage

Week 2 (Feb -03): Ethics in Data Science

* Problem set 1 (getting started)
* Lab: Introduction to Data Camp, Basics of R

**II. R Programming Language & Machine Learning 101**

Week 3 (Feb – 10): Introduction to R Programming Language

* Problem Set 2 (due)
* Lab 1: Objects and functions in R, Vectors and Data Types, Data Structure

*Readings:*

Molina, M., & Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology*.

Week 4 (Feb – 17): No class – Presidents Day

Week 5 (Feb – 24): Data Cleaning in R

* Problem Set 3 (due)
  + Introduction to R on Data Camp (4 hours of video, and exercises)
* Lab: Packages, Importing Data

*Readings:*

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, *59*(10), 1-23.

Week 6 (March – 02): Logistics Regression and Ordinary Least Square Regression

* Problem set 4 (due)
  + Importing and Cleaning Data in R on DataCamp (4 hours of videos, and exercises)
* Lab: Group work using logistics regression and OLS regression

############ Instructional Recess due to Coronavirus Outbreak March 12-March 19 ####

Week 7 (March – 23): Introduction to Machine Learning & Decision Trees

* Problem set 5
  + Practice Multiple and Logistic Regression on Data Camp (4 hours of videos and exercises)
* Lab: Decision Trees

############ Recalibration Period due to Coronavirus Outbreak March 25-April 2 ####

Week 8 (April - 6): Random Forests and Q & A for the Midterm

* Problem set 6: Decision Tree (Data Camp)
* Random Forest
* *Take home mid-term exam*

*Readings:*

Rodriguez, M. Y., DePanfilis, D., & Lanier, P. (2019). Bridging the gap: Social work insights for ethical algorithmic decision-making in human services. IBM Journal of Research and Development, 63(4/5), 8-1.

**III. Text Mining (digital trace data, Twitter API, Text processing, topic modeling)**

Week 9 (April 13): Digital Trace Data

* *Mid-term due*
* Lab: Twitter API exercises

*Readings:*

Nelson, L. K., Burk, D., Knudsen, M., & McCall, L. (2018). The future of coding: A comparison of hand-coding and three types of computer-assisted text analysis methods. *Sociological Methods & Research*, 0049124118769114.

Week 10 (April 20): Sentiment Analysis and Data Cleaning

* Problem set 6 (Random Forests - Data Camp)
* Basic Text Analysis in R
* Lab: Text Data Cleaning & Sentiment Analysis
* Screencast-o-Matic: Create a screencast of your final presentation

*Readings:*

Flores, R. D. (2017). Do anti-immigrant laws shape public sentiment? A study of Arizona’s SB 1070 using Twitter data. American Journal of Sociology, 123(2), 333-384.

Week 11 (April 27): Structural Topic Modeling

* Problem set 7 & 8
  + Create a Twitter Dataset
  + Wrangling text in R & Visualizing Text in R & Sentiment Analysis in R (Data Camp)
* Lab: Structural Topic Modeling

Readings

Roberts, M. E., Stewart, B. M., & Tingley, D. (2014). stm: R package for structural topic models. *Journal of Statistical Software*, *10*(2), 1-40.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder‐Luis, J., Gadarian, S. K., ... & Rand, D. G. (2014). Structural topic models for open‐ended survey responses. *American Journal of Political Science*, *58*(4), 1064-1082.

Fischer-Preßler, D., Schwemmer, C., & Fischbach, K. (2019). Collective sense-making in times of crisis: Connecting terror management theory with Twitter user reactions to the Berlin terrorist attack. *Computers in Human Behavior*, *100*, 138-151.

**IV. Final Project**

Week 15 (May 4): Presentations

Week 16 (May 11): Presentations & Wrap-up

**Extra Bibliography**

**Statistics/Machine Learning Theory**

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, pp. 3-7). New York: springer. (Free E-Book is available online.)

Attewell, P., Monaghan, D., & Kwong, D. (2015). *Data mining for the social sciences: An introduction*. Univ of California Press.

**Applications of Machine Learning/ Data Mining in Social Sciences**

Davidson, T. (2019). Black-Box Models and Sociological Explanations: Predicting High School Grade Point Average Using Neural Networks. *Socius*, *5*, 2378023118817702.

Nelson, L. K. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 0049124117729703.

Salganik, M. J., Lundberg, I., Kindel, A. T., & McLanahan, S. (2019). Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration. *Proceedings of the National Academy of Sciences of the United States of America*. <https://doi.org/10.1073/pnas.1915006117>