~~PS 1702: The Analysis of Political Variables~~>
PS 1702: An Introduction to Coding and Computational Social Science (i2C2S2)
==========
Class number: 30890
Term: Fall 2017=2181
Professor: Michael Colaresi, mcolaresi@pitt.edu (mailto:mcolaresi@pitt.edu), 4619 Posvar Hall
Room and Time: CL337, Tuesday and Thursday 1pm-2:15pm
Office Hours: Monday 12pm-1pm, Wednesday 12pm-1pm

Summary

"Big data", "analytics", "data science", "computational science": these are all words used to describe sets of tools that help sift and summarize massive volumes of information that are particularly important for understanding social relations today. This class is meant to be a ~~gentle~~ introduction to the opportunities and challenges with digesting, collecting and creating digitally available political and social information such as text, measures and social media connections.

We begin by going back to basics, exploring the reasons and ways we use and misuse data. We then turn our focus to flexible computational tools for data collection and visualization and how they can provide unique help in answering important questions such as what causes war and violence, who represses human rights, and what parties are likely to win elections. By the end of the class, students will be exposed to coding and computer languages that are often used in data analytics in industry, government and academia.

Goals

There are three general goals for students in this course. The first is to help students see the usefulness of dynamic, interactive data analytics for social science in particular. Understanding social relations is a complex task, and computers and code can be very important aids in tackling some of these challenges.

The second goal is to educate students on how to be skeptical consumers and users of quantitative information. There are many opportunities to either mangle or mistake messages in data visualizations. We will study clear and poor examples of communicating information to an audience, and discuss the rules that can help guide the creation of useful summaries of potentially complicated data. Having Big data or wielding random forest algorithms does not mean that inferences are clear, easy or uncontroversial.

The final goal of the class is to begin to empower students to use computers and code to collect their own data, particularly from the web, and produce clear and informative visualizations. Students will be exposed to introductory lesson in R, for munging and plotting data, and Python for webscraping. Along the way, students will learn some command line tools (bash) and be introduced to html and markdown.

Each of the goals of the course are reinforced with assignments, along with a quiz and exam. These are discussed below. We will use four books along with several online (and free!) resources.

Pages (required)

The class has four required books, only one is moderately expensive (but it can be found used online) and the other two are available as e-books.

1. Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie or Die. 2013. by Eric Siegel. Wiley. First Edition. (a bit cheaper by e-book) *PA* for short (if you had the read the book, you would already know that though).
2. The Visual Display of Quantitative Information. 2001. by Edward Tufte. Graphics Press. Second Edition. (not available as e-book) *VDoQI*
3. R for Data Science. 2016. by Hadley Wickham and Garrett Grolemund. First Edition. *RDSci*
4. Python for Beginners (http://pythonforbeginners.com). (Ok, you got me, this is not a book. But it is a web resource full of great pages of information. Plus it is free. It is required that you access this as we will draw regular readings and practice from it) When we study a concept in class (like lists) you can use this to find more information. *py4b*

Readings are *due to be read by the first class of each week*.

More Pages (not required)

While the following books are not required, I am hoping some of you will get a kick out of coding and want to hop further down the rabbit hole. Others might actually be more advanced and be in need of further information to challenge themselves as the semester goes on. These are some books for you to take the next loops. There are many, many, online resources however. So you might try searching around first.

Optional 0. Web Scraping with Python: Collecting Data from the Modern Web. 2015. by Ryan Mitchell. Oreilly. First Edition. (Much cheaper by e-book)

Optional 1. Python for Data Analysis: Data Wrangling with Pandas, NumPy and IPython. by Wes McKinney. First Edition. 2012

Optional 2. R Cookbook. by Paul Teetor. First Edition.

Optional 3. Machine Learning in Python. by Michael Bowles. First Edition.

Optional 4. Learning the bash Shell: Unix Shell Programming. by Cameron Newham. 3rd Edition. (pretty dated, but still the best place to start)

Optional 5. ggplot2: Elegant Graphics for Data Analysis. 2016. by Hadley Wickham. Second Edition. *ggplot2* (The first edition is totally different and obsolete. Lesson: Code moves pretty fast. If you don't stop and look around once and awhile. You could miss it.)

Pixels

In a class about coding and computational social science, you will need a computer that has some specific software on it. You should be able to do everything we do in class using modern Mac, Linux and even Windows operating systems. If you are a Windows user, please follow these steps (https://msdn.microsoft.com/en-us/commandline/wsl/install_guide) or these steps (https://msdn.microsoft.com/en-us/commandline/wsl/about) to install bash or the Windows subsystem for Linux. If you have questions, ask me. In the course, we will be using free, open-source software. It will be necessary to have a laptop in class as we begin to code.

- Python: We will use the free installation by Anaconda (http://continuum.io/downloads). Python is a rather easy to learn computer language that can do quite a bit. We will mainly use it for practicing some good coding practices and then scraping the web and collecting data.

- R: This is a free program, available here (https://cran.r-project.org). Install that first (just R, this is the language compiler). Then, uyou should use RStudio (https://www.rstudio.com) to run your R code, as it is very helpful. You want the free desktop version. R is a more fiddly language than Python generally, but has very easy, intuitive add-on package (also free) especially for cleaning and plotting data. We will mainly be munging and plotting data in R. Make sure you install the tidyverse packages, see here (https://github.com/tidyverse/tidyverse). Follow the instructions for the "installation", these commands would be typed in RStudio.
- A text editor. Sublime Text, BBEdit, emacs, vim, TextWranger, are all good options for text editors to write code in (Sublime Text is written in Python, btw). You do not want to write code in Word or the mis-named Mac TextEdit (which does not save to plain text by default). Really.

Assignments(<-,=)

There are four assignments, one quiz and a final exam. You will also receive credit for in-class engagement. I reserve the right the right to make changes as the semester proceeds.

- Assignments: These are due on the class noted here unless a deal has been worked out with the professor at least 48 hours before the due date (Late assignments are discussed below). Each assignment is worth 10 percent of your grade (so 40 percent total).
    - Essay (Due September 28). You are to find an example of data analytics, related to the social sciences, that inspires you. You can not pick one that I have talked about in class prior to the assignment being due. Write 1500 words that answers the questions: a) What are they trying to do?, b) How are they doing it?, c) What is new about their approach, compared to previous related attempts?, d) Why does it inspire you?. Look around the web for campaigns, advertising, research, or business applications. Ask your professors from other classes. Be prepared to discuss your examples in class. This needs to be written in Markdown (which is a really easy text formatting language– this syllabus is written in it). A very simple guide can be found here (http://daringfireball.net/projects/markdown/)
    - Picture Puzzle and 1000 Words assignment (Due November 9). I will give you a set of data visualizations, with some descriptions of them. You will find the problems with the visualizations, write about their flaws, and why the elements

are problematic. In addition, you will find a data visualization related to social science and write 1000 words analyzing its effectiveness using the concepts we have learned in class and in the readings. What could be done better?

    - 2 Coding assignments (A and B) (Due September 21 and Due November 30). After introducing a computer language (eg Python and R), I will provide you with some tasks to perform. You will turn in your code electronically (a file often referred to as a script that has the commands that will be run, we will talk about it). We will then run the code and see how you do.

- Quiz: We will have one in-class Quiz on November 16. This will cover the material up to this point in the course and will be worth 15 percent of your final grade.

- Final Exam: The Final Examination, December 11 2pm-3:50pm with be in the form of a hackathon. In the exam, I will provide you with a dataset and a set of questions to answer with that data. You will use your computer to look at the data, then you will write out a plan of how to visualize information that helps answer one of the questions (there will be a choice). Finally, you will have the chance to try and produce the visualization. This will be worth 20 percent of your final grade. The final exam is on Due from Due in the regular classroom.

- Engagement: Are you trying? Do I have proof that you did the readings? The remaining 25 percent of your grade will be allotted based on the effort I see you exerting in the class. Are you attempting to think about data and visualizations, even if this is not something you are used to? Are you trying to expand your coding skills, even if you already have some experience working with data? Asking and answering questions in class, going to office hours, bringing examples of analytics that you read about to class or emailing them to the Professor or TA, all count as engagement. It is important to note that you can't answer a question in class if you are not in class. Thus attendance is part, but only part of engagement.

Policies

- Class attendance: I will take attendance. As discussed above, this is only one part of engagement. If you miss too many classes (being unconscious for class or observationally equivalent to being unconscious is the same as missing it) this will impact your grade.
- Missing quizzes or exams: If you know that you are going to have to miss a quiz or an exam for a good reason, let me known at least 48 hours before the quiz or exam, so

that a make-up can be scheduled. If no make-up is scheduled, and you miss the quiz or exam, you will get a zero. I do not offer make-up quizzes or exams unless you bring a doctors note explaining why you were medically unable to attend that day.

- Attention: What we will be doing, will be hard for many people. Chatting in class about other topics is distracting and disrespectful. Please do not do it. It counts as anti-engagement.
- Information: Will use courseweb (http://courseweb.pitt.edu).
- Grading scale:
  - 4.0 (92-100 percent)
  - 3.5 (87-91.9 percent)
  - 3.0 (80-86.9 percent)
  - 2.5 (76-79.9 percent)
  - 2.0 (70-75.9 percent)
  - 1.5 (65-69.9 percent)
  - 1.0 (59.5-64.9 percent)
  - 0.0 (<59.5)

Class Schedule

```
Begin Progam
```

Week0<-c("Tuesday August 29", "Thursday August 31")

- Readings and Installation: None. Get the books. Read the syllabus carefully. If you want to be ahead, start reading the PA book. You can download RStudio and play with it too.
  - Get Python up and running:
    a. Install Anaconda (which gets you Python), see here (http://docs.continuum.io/anaconda/install).
    b. Look at the sheet sheet here (http://conda.pydata.org/docs/using/cheatsheet.html).
    c. Take the test slither, here (http://conda.pydata.org/docs/test-drive.html).
  - Get R up and running:
    a. Get RStudio (which gets you R), see here (https://www.rstudio.com/products/rstudio/download/).

b. Get the Data Visualization Cheat Sheet here (https://www.rstudio.com/wp-content/uploads/2015/08/ggplot2-cheatsheet.pdf) and the Data Wrangling sheet sheet here (https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf).
c. You can try some things out on here (http://tryr.codeschool.com/).
d. Check out how to make a scatter plot in ggplot2 (a package for R) here (http://docs.ggplot2.org/0.9.3/geom_point.html)

Week1<-c("Tuesday September 5", "Thursday September 7")

- What is Data Analytics and Why do we need it?
- Readings and Codings:
  - PA: Introduction and Chapter 1
  - "Computational Social Science" 2009. by David Lazer, et al. available here (http://gking.harvard.edu/files/LazPenAda09.pdf).
  - "What is code?" by Paul Ford. available here (http://www.bloomberg.com/graphics/2015-paul-ford-what-is-code/)
  - Do the learnpython Hello, World tutorial here (http://www.learnpython.org) and the first py4b module (basics).

Week2<-c("Tuesday September 12", "Thursday September 14")

- Prediction, Privacy and What's for Free: Should you be worried about Big Data?
- Readings and Codings:
  - PA: Chapter 2
  - Obama's Not-So-Big Data. by John Sides and Lynn Vavrek, January 14, 2014. Available here (http://www.psmag.com/books-and-culture/obamas-big-data-inconclusive-results-political-campaigns-72687).
  - "Big Data is Opening Doors, but Maybe Too Many" by Steve Lohr, March 23, 2013. Available here (http://www.nytimes.com/2013/03/24/technology/big-data-and-a-renewed-debate-over-privacy.html)
  - Do the learnpython variables and types AND the lists tutorial here (http://www.learnpython.org), and the second py4b module (http://www.pythonforbeginners.com/lists/python-lists-cheat-sheet).

Week3<-c(" Tuesday September 19","Thursday September 21, Coding Assignment A Due")

- Deep Data? Sentiment Analysis Examples and Observing the Unobservable
- Readings and Codings:
  - PA: Chapter 3
  - Do the learnpython basic operators tutorial here (http://www.learnpython.org) and the third, py4b module on dictionaries (http://www.pythonforbeginners.com/dictionary/how-to-use-dictionaries-in-python).

Week4<- c("Tuesday September 26", "Thursday September 28, Essay Assignment, Due!")

- Machine Learning and Teamwork between People and Computers (We hope).
- Readings and Codings:
  - PA: Chapter 4, 5
  - "How Machine Learning Works", Economist. May 13, 2015. Available here (http://www.economist.com/blogs/economist-explains/2015/05/economist-explains-14).
  - The Go Champion, The Grandmaster, and Me. by Ken Jennings. Available here (http://www.slate.com/articles/technology/technology/2016/03/google_s_alphago_defeated_go_champion_lee_sedol_ken_jennings_explains_what.html)

Week5<-c("Tuesday October 3", "Thursday October 5")

- Text Analytics Beyond Sentiment
- Readings and Codings:
  - PA: Chapter 6
  - Monsters, Men and Topic-Modeling, New York Times, May 29, 2011 Online. Available here (http://opinionator.blogs.nytimes.com/2011/05/29/of-monsters-men-and-topic-modeling/)
  - Do the learnpython string formatting and do the basic string operations tutorial here (http://www.learnpython.org) and the BeautifulSoup first module (http://www.pythonforbeginners.com/beautifulsoup/python-beautifulsoup-basic)

and second module (http://www.pythonforbeginners.com/beautifulsoup/python-beautifulsoup) on py4b.

Week6<-c("Tuesday October 10, No Class", "Thursday October 12")

- Influence and Causal Inference
- Readings and Codings:
  - PA: Chapter 7
  - No! Formal Theory, Causal Inference, and Big Data Are Not Contradictory Trends in Political Science. PS. January 2015. Available here (http://scholar.harvard.edu/files/msen/files/big-data.pdf).
  - Do the learnpython tutorial for conditions and loops here (http://www.learnpython.org)

Week7<-c("Tuesday October 17", "Thursday October 19")

- From Thinking to Doing, Part 1: Get Some Data
- Readings and Codings:
  - RDSci, Chapter 1

Week8<-c("Tuesday October 24", "Thursday October 26")

- From Thinking to Doing, Part 2: Data Munging
- Readings and Codings:
- RDSci, Chapter 2

Week9<-c("Tuesday October 31", "Thursday November 2")

- From Doing to Seeing: Types of Data Visualization and the Rules for Excellence
- Readings and Codings:
  - VDoQI, Chapter 1

- RDSci, Chapter 3

Week10<-c("Tuesday November 7", "Thursday November 9, Picture Puzzle Assignment and 100 Words, Due!")

- Not So Excellent Adventures in Data Visualization
- Readings and Codings:
  - VDoQI: Chapter 2 and 3
  - RDSci, Chapter 4 and 5

Week11<-c("Tuesday November 14", "Thursday November 16, Quiz!""")

- Elements of Good (and Bad) Visualizations Continued
- Readings and Codings:
  - VDoQI: Chapters 4 and 5 (can mostly skim, get the general ideas)
  - ggplot2, Chapter 6

Week12<-c("Tuesday November 21", "Thursday November 23, No class, something about Turkeys")

- Applying the Lessons from Pages to Pixels
- Readings and Codings:
  - VDoQI: Chapters 6 and 7 (can mostly skim, get the general ideas)
  - RDSci, Chapters 7 and 8

Week13<-c("Tuesday November 28", "Thursday November 30, Coding Assignment B Due")

- From Seeing to Predicting: Modeling and Visualizing hits and misses
- Readings and Codings:
  - VDoQI, Chapter 8
  - RDSci, Chapter 9

Week14<-c("Tuesday December 5", "Thursday December 7")

- App-lying What We Know: Vis is It
- Readings and Codings:
  - VDoQI, Chapter 9.
  - Look over tutorial for shiny, available here (http://shiny.rstudio.com/tutorial/) You do not need to do the whole thing. Just get the idea of what is going on.

FINAL EXAM WEEK, Final Exam is Monday December 11, 2-3:50pm.

Did you accomplish the goals of the course?

```
End Program
```

Special Assistance

If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and the Office of Disability Resources and Services (DRS), 140 William Pitt Union, 412-648-7890, drsrecep@pitt.edu (mailto:drsrecep@pitt.edu), 412-228-5347 for P3 ASL users, as early as possible in the term. DRS will verify your disability and determine reasonable accommodations for this course.

Plagiarism and Cheating

Take time to read the information on "Academic Integrity" and be sure that you understand your responsibilities under the guidelines set out for The Dietrich School of Arts and Sciences, which are spelled out in full here (http://www.as.pitt.edu/fac/policies/academic-integrity). We will also look at this in class during the term.