ST445 Managing and Visualizing Data

Week 1, MT 2017

Kenneth Benoit (mailto:kbenoit@lse.ac.uk) (Methodology) and Milan Vojnovic (mailto:m.vojnovic@lse.ac.uk) (Statistics)

Christine Yuen (L.T.Yuen@lse.ac.uk), TA

Plan today

- Administration and logisitics (DONE)
- A brief history of data and the origins of databases
- Information versus data
- How big is big data?
- Data types and storage units
- git and Github
- Markdown in brief
- Lab preview

Why take this course?

- You have to
- It provides "data science literacy"
- You will learn
  - basic data types and structures
  - the use of git and GitHub (http://github.com)
  - how to clean, organize, and reshape data
  - how to create and use databases
  - how to scrape data from the Internet
  - how to work with APIs
  - data visualization, including principles and practicals

Course Outline

Week	Topic
1	Introduction to Data
2	The shape of data
3	Creating and managing databases
4	Using data from the Internet
5	Working with APIs
6	*Reading Week*
7	Exploratory data analysis
8	Exploratory data analysis (cont'd)
9	Model evaluation
10	Dimensionality reduction
11	Graph data visualization

Prerequisites and Software

- Introductory course — no prerequisites

- Software
  - Python and R (Anaconda distributions) for basic work
  - SQLite (though Anaconda)
  - Jupyter notebooks for writing code and working with data
  - Github to share course documents and assignments

- Mirrors similar tool usage and learning in MY470

Readings

- Mixed set of readings, very specific to each week.
  - For instance, Week 1 (https://lse-st445.github.io/#week-1-introduction-to-data)'s readings
  - Often available electronically, otherwise, available for purchase from Amazon (often in Kindle versions)

- Often linked to Internet sources
  - Some books are available online and in print, and the online version may be more recent

Course Meetings

- Ten two-hour lectures: Tuesday 10:00–12:00 in CLM.2.02
- Ten 1.5-hour classes ("labs")
  - Thursdays 13:00–14:30 in TW2.4.02
- No lecture/class in Week 6
- Office hours
  - Ken: Mondays 16:00-17:00, Thursdays 11-12:00
  - Milan: TBC

Assessment

- 4 problem sets will be assessed (40%)
- Other problem sets will be marked with feedback, but not form part of the final grade

- Project (60%)
  - Work with a dataset to produce a series of visualizations
  - You may use either Python or R

Project data will be provided

Examples:

- a dataset of human-annotated political texts, via https://manifesto-project.wzb.eu (https://manifesto-project.wzb.eu)
- a dataset of Tweets on Brexit, about 15 million from June 2016 (Ken to provide)
- the Hansard corpus of speeches from the UK parliament (Ken to provide)
- UK government data from https://data.gov.uk (https://data.gov.uk)
- Yelp academic dataset https://www.yelp.com/dataset (https://www.yelp.com/dataset)
- UK Policing dataset https://data.police.uk/data/ (https://data.police.uk/data/)

Collaboration

- All assignments are individual unless we instruct you otherwise
- For individual assignments:
    - You can discuss solutions with peers
    - However, you are not allowed to copy-paste code – you need to write the code yourself
- You can use online resources but always give credit in comments if you borrow code/solutions
- We may very well assign teams for the project (still being discussed)

A note on tools

- Lab computers will be provided, but you will probably want to use your own
- In this and in MY470, we will be encouraging you to install Anaconda locally, and use Jupyter from your local machine's server

- However, you may also use Jupyter from the HPC system
  - Address: https://fabiancloud.lse.ac.uk/jupyter (https://fabiancloud.lse.ac.uk/jupyter)
  - You first need to email fabian@lse.ac.uk (mailto:fabian@lse.ac.uk) to request an account
  - Instructions may be found on the Fabian/HPC Moodle page (https://moodle.lse.ac.uk/course/view.php?name=fabian)

- As with MY470, we will use GitHub Classroom for the course, so you will need to send us your GitHub username so we can sign you up for the **lse-st445** organization (http://github.com/lse-st445)