Department of Social and Political Sciences

Computer Programming for the Social Sciences

This two day workshop will teach beginner level, practical computer programming skills for use in social science research. There will be a focus on the automatic collection of data (from websites, twitter feeds, etc.) and its processing into clean, formatted datasets ready for analysis. Researchers will leave the course with an installed programming environment and a variety of code snippets for use in their own projects.

4th and 5th of June 2012
The Emeroteca, Badia Fiesolana

Register with Adele Battistini
adele.battistini@eui.eu

Organized by Jonathan Bright
jonathan.bright@eui.eu

Sponsored by Professor Fabrizio Bernardi

10 credits

For more information see:
http://www.eui.eu/SeminarsAndEvents/Index.aspx?eventid=77143

Overview

The rise of what is often known as "big data" presents a huge opportunity for social scientists. The amount of information available to researchers is growing exponentially: a recent study estimated that in 2010 the quantity of digital data in the world exceeded one trillion gigabytes. Much of this information is available for free for use in research: for example, many governments have started to move towards "open data", opening up their own archives and releasing huge quantities of previously unseen data.

This flood of data creates a wealth of opportunities both to test and refine existing theories and to create new ones about the way social life works. However, while it might be available for the first time, much of this information does not come prepared and ready to plug into a statistical analysis package, and is often not quite as open and easy to use as its proponents imagine. To fully take advantage of the opportunities provided by big data, social scientists will increasingly require the ability to "program": to instruct their computer to understand this information, to capture, analyse and format it, and to put it to use.

This course is intended as an introduction to this type of practical computer programming, geared towards data collection, and created specifically with the needs of social scientists in mind. **The workshop is open to all students at the EUI and is specifically aimed at beginners with no previous programming experience.** The aim will be to convince even those who think of themselves as computer "illiterate" that basic programming ability is both a vital part of a researcher's skill set, and something that can be relatively easily learnt.

Outcomes

Those participating in the course will:

- Acquire a basic knowledge of computer programming, and the functioning of elementary programming language statements such as "if…else" statements and "for loops"
- Learn how to "mine" the wealth of data that is available on the internet in both structured and unstructured formats
- Get an insight into available tools for harvesting data from public information services such as Google and Twitter
- Learn about existing open data repositories on political systems in the US and the EU, and how to access them
- Learn how to output clean, formatted datasets for easy use with statistical software packages
- Leave with a fully functioning programming environment installed on their personal laptop, ready for application to their own research projects

The language which will be used is **Python**, which is often considered one of the simplest language to pick up yet contains within it a lot of power and flexibility. Where possible some translations of the concepts into **STATA and R** will also be offered.

Format & Requirements

The course will be structured around a series of workshop sessions. In each session, a new programming concept will be introduced, and then students will be encouraged to try it out with help from the course instructor. A variety of practical examples will be deployed, using data from government archives, Twitter, Wikipedia, etc., with an emphasis on creating re-usable code which can easily be deployed in future research projects.

This is a practical course, hence attendees should if possible bring a laptop (if you are unable to do so please contact the course instructor at: jonathan.bright@eui.eu), and should also install the Python programming language which will be used throughout the workshop (see http://www.python.org/getit/). Please install **Python version 2.7.3** (please complete this step even if you own a Mac which already contains a version of Python on it, as this version may well be out of date).

Participation in all the required sessions on the course will be sufficient to obtain the 10 credits on offer.

Programme

Day 1 – Monday June 4th 2012 – Emeroteca

*Before coming to the workshop, please install Python **version 2.7.3.** See: http://www.python.org/getit/. If you have any problems please come along between 9.00 and 9.30.*

09.30 – 11.00 || Session 1: Introduction to computer programming

This session will set the context for the rest of the workshop by examining some basic questions about the subject. What is computer programming, and why should social scientists be familiar with basic programming techniques? We will then move on to introducing the specific language being learnt during the workshop (Python), and look at some of the fundamentals of programming: how to create and run a script in python, how to assign variables, etc.

11.00 – 11.30 || Coffee break

11.30 – 13.00 || Session 2: Basic computer programming

This session examines elementary programming techniques which are fundamental to the use of computing for the social sciences. Simple if…else statements, loops, different types of data structure and file handling will all be discussed. These techniques will be applied towards the manipulation and reformatting of a simple dataset, and the basic analysis of a text file.

13.00 – 14.30 || Lunch

14.30 – 16.00 || Session 3: Web Scraping I: Using APIs

One of the advantages of popular scripting languages such as Python is that many special "interfaces" to a huge variety of data sources are available for use; hence many data collection tasks can be achieved without complex programming. This session discusses the general principles behind such interfaces, which are known as APIs. These concepts will be then put into practice using some existing APIs to access publicly available data generated by Google, Twitter and Facebook, as well as on political systems in the US and the EU.

16.00 – 16.30 || Coffee break

16.30 – 18.00 || Review session

This is an optional session, and its content will depend on the progress made during the day. It can be used to have another look at concepts covered during the day or to ask the lab instructor any questions.

Day 2 – Tuesday June 5th 2012 – Emeroteca

9.30 – 11.00 || Session 4: Web Scraping II: Semi-structured and unstructured data

Even if APIs and structured data are growing in number, the majority of information on the web is still largely "unstructured", which means in practice that it is difficult for computers to read and interpret in a fully flexible way. This session discusses how to access such data, looking at interaction with HTML (the structure of an ordinary web page) and XML (which is used by many webpages for communicating new content).

11.00 – 11.30 || Coffee break

11.30 – 13.00 || Session 5: Further resources: text processing, network analysis geocoding, etc.

This session will introduce a range of further python modules which may be of use in certain types of social science project, such as ones which tackle text processing and content analysis, network analysis, and geocoding. The aim is not to offer an introduction into the theoretical or statistical bases of these types of techniques, merely to show how they can be (simply) executed in Python, thus building familiarity with the language and showcasing the range of tools on offer.

13.00 – 14.30 || Lunch

14.30 – 16.00 || Session 6: Large Project Handling (putting it all together)

This session integrates the previous five sessions, discussing the management of large data projects in the context of social science research. Topics covered include data storage, continuous sampling periods (such as gathering tweets during an election campaign), error handling, the management of code bases and the writing of reusable code.

16.00 – 16.30 || Coffee

16.30 – 18.00 || Review session

This is an optional session, and its content will depend on the progress made during the day. It can be used to have another look at concepts covered during the day or to ask the lab instructor any questions.