



| | | | | | |
|--------------------------------|--|--------------------------|--|--|---------|
| Course Code & Title | LISS346 Advanced Text Processing | | | | |
| Convenor(s) | Dr Anda Drasovean , King's College London | | | | |
| Institution | King's College London | Department | School of Education, Communication & Society | | |
| Academic Year | 2018-19 | Term | Autumn | | |
| Number of Sessions | 1 | Research Platform | Digital & Archival Research Methods | Length of session(s) | 5 hours |
| Day, Date | | Start : End | | Room Location | |
| Thursday 16 November 2018 | | 1000 : 1600 | | B20, James Clerk Maxwell Building Waterloo Campus, SE1 8WA | |
| Enrolment Link: | You may be prompted to log into SkillsForge Details on when courses open for booking can be found here: https://liss-dtp.ac.uk/our-training-programme/ | | | | |

Course Description:

This course is aimed at researchers in the social sciences and humanities who are working with electronic texts and who might require knowledge of advanced text processing techniques. For instance, they might need to automatically remove all empty lines in a large text file or a batch of text files. The course will introduce participants to regular expressions (regex), a set of conventional symbols that can be combined to define patterns. These patterns can then be used to manipulate character strings/text, and to perform complex searches, additions, and substitutions. A typical application of regular expressions might involve, for example, standardising dates in different formats (e.g. changing all dates formatted as 'mm-dd-yyyy' or 'mm/dd/yy' to 'dd.mm.yy'), deleting unwanted whitespace characters, or finding and replacing the alternative spellings of a word with a standard one. After learning the basics of regular expressions, we will move on to manipulating multiple text files with the help of regex and an advanced text editor (e.g. Notepad++). This will be a hands-on session with a Q&A at the end to address specific questions you may have.

Course Outline:

1- Introduction to regular expressions

In this first part of the workshop, participants will be introduced to some key regex categories and concepts. These include characters, meta-characters, character classes, quantifiers, anchors, ranges, groups, and operations. Each concept will be explained through practical examples and short exercises, using an online tool called Regexpal.

2- Using regular expressions in Notepad++

The second part of the workshop will be dedicated to learning how to manipulate text files using regex and the Notepad++ text editor. Using the Notepad's "Find & Replace" functions, participants will learn how to develop complex regex searches, deletions, and substitutions. We will consolidate these skills and notions by learning to manipulate text and metadata across multiple text files and bring these to a standard format (thus obtaining a data set that can be analysed with specialised software – e.g. Nvivo, corpus analysis tools).



London Interdisciplinary Social Science Doctoral Training Partnership

Advanced Research Methods in Social Sciences

3- Q&A session

In the last part of the session, we'll address specific questions and challenges arising from the participants' own research.

Reading List: N/A

Eligibility:

Students must be registered PhD students at either King's College London, Imperial or Queen Mary University of London.

Pre-course preparation: None required

Number of students: 15