An Introduction in the Digital humanities: Data Analysis for the Humanity Scholars

Alexei Kouprianov*

Digital humanities is an umbrella term for a number of fields of inquiry dealing with the application of computers in the humanity scholarship. Traditional areas of concentration include historical data analysis, corpus linguistics, text mining, and online publication of sources. The course will include several modules dealing with (1) data analysis and data visualisation including elements of social network analysis, computer cartography, and computer text analysis. (2) data pre-processing, including cleaning the data and preparing them for the use in data analysis environments and text markup for research purposes. The course is focused on both studying examples of Digital humanities scholarship and acquiring specific skills. By the end of the course, the students will be expected to understand principles of historical data analysis and possess a sufficient command of tools for data pre-processing, analysis, and visualisation (basics of R, Perl and Regular Expressions) for implementation of individual research projects and mediation between humanity scholars and computer scientists in Digital humanities projects. The practical study of analysis will be based on both training datasets and the real-life historical datasets. The students will learn how to use the R environment for data analysis and simple Perl scripts and Regular Expressions for data collection and pre-processing. The emphasis of the course is placed on the development of practical skills.

The choice of the software products used in this course is based on the following principles: they must be widely used by the analysts, multy-functional and extensible, and they must be cross-platform, free of charge and, preferably, open-source. R and Perl as well as few other applications that will be used for more specific purposes meet these requirements. It is assumed that students are familiar with basic computer skills and can handle office applications, basic tasks of file management like navigating through the tree of directiories, locating files on the computer, copying them, etc., and software installation.

It is not that easy to recommend a single textbook that covers all necessary topics. The bibliography includes thus rather diverse sources, which should be combined for the better understanding of the matter. Some of them are typical computer science textbooks, which are structured rather as reference books, and are not intended for reading from the beginning to the end. Anscombe (1973) and Tukey (1977) provide an introduction in the philosophy and methodology of exploratory analysis (even though the latter book is rather technical), Leek (2015) accompanies them as a more recent manual. Kobacoff (2015) may serve as the main source on R, accompanied with lecture notes assembled by myself

* Department of Sociology, National Research University Higher School of Economics (St. Petersburg).

(Kouprianov, 2018, pt. 1 and 2). Field, et al. (2012) is a great advanced manual in the use of statistics in R. For a deeper understanding of how R works, I recommend Burns (2011) and Murray (2012). For those willing to go beyond the basic data transformation and visualisation techniques I recommend Wikham and Grolemund (2017). Since we need just tiny bits of Perl, no special literature is recommended. Friedl (2006) is a standard manual in Regular Expressions written in such a way that even persons with no extensive computer training can understand it. As it usually happens with computer science, in the study of R, Perl, and Regular Expressions, on-line tutorials are invaluable, and I strongly recommend to use them at every occasion.

The practical implementation of computer skills in historical scholarship is duly covered in the three manuals available on-line (Schreibman, Siemens, & Unsworth (Eds.) 2004; Cohen & Rosenzweig 2006; Shawn, Milligan, & Weingart 2015).

1 Course contents

1.1 Introduction

Visualization and its role in data analysis. Anscombe's quartet. Standard tasks of data analysis, their graphical and analytic implementation. Frequency distributions (histograms, barplots, and descriptive statistics). Bivarate models (four-filed classification of bivariate interactions, scatter-plots, multiple boxplots, and structured barplots, correlations, regresion, t-test and analysis of variance, χ^2). Non-parametric analogues of the widespread parametric methods. Why and where non-parametric methods are important. A special case of bivariate model: temporal dynamics. Less trivial visualisations. Visualising interaction of three and more variables. Sankey diagrams, maps, networks, and animated graphs.

1.2 R basics

Command-line interface. Possible responses o the interpreter, ¿ and + prompts. Commands history. Interrupting calculations when something goes wrong (Ctrl+C, Escape). Objects in R. Vectors, matrices, data frames and lists. Data types: numeric, character, factor. Some constants: TRUE, FALSE, NA and NULL. Arithmetics in R, assignment operator ¡-. Functions in R. Posible values for function arguments: constants, objects, products of other functions. Creating vectors: c(), rep(), and seq(). The default order of arguments and argument names. Some elementary math functions: log(), log10(), sqrt(). Help function: help(). Creating more complex objects: data.frame() and list() functions. Adding rows and columns, rbind(), cbind(), and their allies. Data type transformations with as.character(), as.factor(), and as.numeric(). The dimensions of objects: length(), dim(), nrow(), and ncol(). The preview functions: str(), head(), tail(), and summary(). Addressing elements of vectors, data frames and lists. Addressing by number and addressing by name. The names(), colnames(), and rownames() function, and their use for previewing and assigning names. File management in R: getwd(), setwd(), and dir(). Absolute and relative paths. Reading and saving data: read.table() and write.table(). Saving

scripts: savehistory() and loadhistory(). The use of scripts in batch mmode: source().

1.3 Descriptive statistics and data transformation

Extracting summary stats: summary(), min(), max(), mean(), median(), IQR(), sd(), fivenum(), table(). Ho to handle the NA values. Subsetting: subset(). Simple and complex conditions, Boolean operators: & (AND) and — (OR). Factor levels and subsetting, droplevels(). Loops and creation of lists. The while(){} and for(){} loops. Loops and data aggregation (including loops vs. apply() function family).

1.4 Basic R graphics

The simpliest plot: a histogram: hist(). Textual elements of the plot: main, xlab, ylab. Rough adjustment of axes (xlim and ylim). Histohgram bins, breaks argument. Numeric output of the hist() function. The connection between hist() and plot(). Colours: numbers, names, RGB-codes (#RRGGBB) and rgb() function. Black and white plots, dashing (angle and density). Border and background fill. The use of plot() for printing bar- and boxplots. Specialised barplot() and boxplot() functions and their peculiarities. The trouble with axis labels in barplot(), the use of strwidth() for setting the margin widths. Saving graph to a file. General principles of working woth basic R graphical system. Plotting devices, turning them on and off with R functions. Why it is absolutely necessary to appropriately close the plotting devices, the dev.off() function. Plotting graphics to files: troubleshooting. Raster graphics in R: the png() function. The use of variable texts in plots and variable filenames when saving them. Creation of arrays of illustrations. The paste() function and loops. Scatter plots: plot() and its arguments. The type argument. The size and shape of data points. What if the data points are overlapping (2D histograms). Adding elements to the plot: lines(), points(), text(), and abline(). The legend() function. Working with axes: the axis() function. The preparation of illustrations of on-screen presentations and for academic press. The issue of pixel size and resolution. General parameters of the plotting device, the par() function. Juxtaposition and superposition of graphs. Graphical primitives in R: segments(), arrows(), rect(), polygon(). Vector graphics in R: pdf(), postscript(), and svg().

1.5 Some advanced R chapters

Basics of network analysis in R. The libraries sna and network, network graphs and extraction of network metrics. Basics of R cartography, R as a GIS analytic tool. The libraries maps, maptools, rgeos, sp, and rgdal. Text analysis in R. From frequency lists and wordclouds to more complex models.

1.6 Implementation of basic analytic procedures in R

An overview of bivariate models. Causality and formal connection. Different classifications of variables and scales. Null-hypothesis, p-values, false positives and false negatives. The case of two quntitative variables: correlation and

linear regression: cor(), lm(), and reading their output. The case of an 'independent' qualitative and 'dependent' quantitative variables: t.test() and aov(). Post-hoc analysis. The case of two qualitative variables, contingency tables and χ^2 (chisq.test()). The case of 'independent' quantitative and 'dependent' qualitative variables. Basics of nonparametric statistics in R (wilcox.test(), fisher.test(), Theil–Sen estimator: mblm()).

1.7 Data pre-processing

Basics of Regular Expressions. The use of RegEx for search and replace in text editors. A generalised scheme of data pre-processing. The use of RegEx in perl scripts. Data fetchers and data parsers. Cleaning the data (s///) and extracting the variables for data restructurisation (m//). General principles of dataset organisation and different strategies of data parsing.

1.8 Dihital humanities applications

Quantitative history and historical demography. Historical scientometrics. Historical digital cartography. Text mining. Public digital collections. Text Encoding Initiative.

2 Grading

The final grade is based on a take-home test. A student should be able to produce a standard set of graphs based on the training datasets.

The resulting minimal set of graphs should include plots of various kinds:

(1) Univariate plots: at least one histogram and one bar plot.

(2) Bivariate plots: at least one of each kind: (2.1) a scatter plot (optional: not quite sure about whether exactly this dataset provides an opportunity to build it in a reasonable way), (2.2) time series plots for (a) absolute and (b) relative numbers of faculty as a whole and by categories, (c) median ages for the faculty as a whole and (d) by categories, and (e) for the faculty continuity (Jaccard similarity), (2.4) a multiple box plot, a (2.5) scatter plot with jitter, (2.6) a mosaic plot or a structured bar plot.

(3) Custom plots: a network graph and a map.

All graphs should be composed in two versions: (1) for academic print (black and white, vector (PDF) or a raster image of appropriate size), (2) for on-screen presentation (reasonably coloured [coloured background is possible and sometimes even advisable] raster image of appropriate size). The graphs should be accompanied by a script which should include all necessary data transformations and plotting commands.

A flawless set of graphs based on training datasets will be graded with 8. The imperfections will decrease the grade according to their severity. The script should be executable provied that the dataset is in the same folder as the script, the graphics and text elements should be easily readable, the kind of a graph should correspond to the kind of data it visualises. Highest grades (9 and 10) are reserved for the students who will present meaningful and technically correct graphs based on their own data or go beyond the required standard visualisations.

The students who will not accept their final grade based on the take-home assignment, are compelled to come to the exam and pass an additional test in data transformation and plotting meaningful graphs in-class. In this case, the grade for the in-class test will comprise 0.4 of the final grade, while the grade for take-home assignment will comprise 0.6 of it.

3 Recommended literature

Anscombe, Francis John (1973) "Graphs in Statistical Analysis". American Statistician. 27(1): 17—21. https://www.jstor.org/stable/2682899

Burns, Patrick (2011) The R Inferno. http://www.burns-stat.com/pages/Tutor/R_inferno.pdf

Field, Andy, Jeremy Miles, and Zoe Field (2012) Discovering Statistics Using R. SAGE Publications.

Friedl, Jeffrey E. F. (2006) Mastering Regular Expressions: Understand Your Data and Be More Productive. O'Reilly Media, Inc.

Kabacoff, Robert I. (2015) R in Action, Data analysis and graphics with R. Second Edition. Manning publications.

Leek, Jeff (2015) The Elements of Data Analytic Style: A guide for people who want to analyze data. Leanpub.

Murrell, Paul (2012) R Graphics. 2-nd Edn. Boca Raton, London, New York: CRC Press.

Tukey, John Wesley (1977) Exploratory data analysis. Reading, Mass.: Addison-Wesley Pub. Co.

Wikham, Hadley and Garrett Grolemund (2017) R for Data Science : Import, Tidy, Transform, Visualize, and Model Data. Sebastopol, CA: O'Reilly Media.

3.1 Lecture notes

Kouprianov, A. (2018) Basics of data analysis and graphics in R. Part one: command-line interface, objects, functions, and file management

Kouprianov, A. (2018) Basics of data analysis and graphics in R. Part two: retrieving summary statistics and drawing basic graphs

3.2 Specific software products

3.2.1 For all operating systems, mandatory:

R: A Language and Environment for Statistical Computing / The R Development Core Team. https://www.r-project.org/

3.2.2 For all operating systems, recommended:

ImageMagick http://www.imagemagick.org/script/download.php

Tesseract-OCR https://github.com/tesseract-ocr/tesseract

3.2.3 For MS Windows users:

Strawberry Perl http://strawberryperl.com/

Notepad++ text editor https://notepad-plus-plus.org/

3.2.4 For GNU Linux users:

Check whether Perl is already installed, install if necessary.

A code editor is needed, consider Medit or Gedit as a beginner's option.

3.3 Online resources

Daniel J. Cohen & Roy Rosenzweig (2006) Digital History: A Guide to Gathering, Preserving, and Presenting the Past on the Web. `http://chnm.gmu.edu/digitalhistory/`

A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004. `http://www.digitalhumanities.org/companion/`

Graham, Shawn, Ian Milligan & Scott Weingart. Exploring Big Historical Data: The Historian's Macroscope `http://www.themacroscope.org/`

3.4 Datasets

Kouprianov, Alexei (2018) Two hundred and twenty three students: height, body mass, sex, smoking behaviour, math test, and academic group. A tab-delimited dataset, V.2.1, 2018.

Presidential and Legislative elections in Russia (2000-2018).

Various small datasets are also supplied on specific occasions, including real-life examples of datasets used in historical research.