Программа учебной дисциплины «Nonreactive and big data in the social sciences: methods and approaches»

Утверждена

Академическим советом ООП

Протокол №38 от «21» июня 2018 г.

Author	Mavletova A.M.
Number of credits	4
Class work (hours)	36
Individual work	116
Year	4
Discipline's format	Without using online courses

I. GOAL, RESULTS OF STUDY AND PRE-REQUISITES

The growth of Internet penetration and the possibility of collecting and analyzing big data have produced new challenges and have offered new opportunities for researchers and official statistics. Within several years nonreactive and big data has become the main trend in the social sciences. Nonreactive methods include nonparticipant observation and analysis of digital fingerprints such as likes or shares, as well as private documents such as blogs, social media profiles and comments, or public online documents such as mass media materials.

Information is produced at an unprecedented rate nowadays. Google reports that every two days we create five exabytes of data as we did from the dawn of civilization until 2003. Most data come from user-generated content: messages, photographs or video in social media (e.g., Facebook or Vkontakte). Most of the claims on big data come from IT specialists, since the collection and analysis of big data is more easily handled by computational specialists rather than social scientists. Thus, claims about the size of data (the bigger the better) can be often met in the literature. However most of big data we have is social data. People post information, tweet, retweet or share information other people post. Social scientists can apply their experience of designing social research, as well as experimental and quasi-experimental studies to use big data for drawing valid inferences.

This course will give an introduction to key quantitative approaches to the collection of non-reactive data in social sciences. The course is taught in the form of lectures, seminars, and individual work. All teaching is conducted in English. The goal of the course is to introduce the

opportunities of nonreactive and big data for social scientists and learn basic methods and tools to collect nonreactive data.

After completing the course students are expected to acquire following competences:

Know:

- different types of big data in social sciences,

- basic methods of collecting nonreactive data in social sciences,

- opportunities and limitations of applying big data in social sciences.

Are able:

- to collect online data (VK, Twitter).

Have skills:

- of reflexive assessment of applying big data in social sciences.

The course is based on the following previously covered courses:

Sociology

Methods of data collection

- Statistical data analysis

Skills from this course should be the basis for studying the following courses:

- Research seminar
- Advanced methods of data collection and data analysis

II. THE CONTENT OF THE COURSE

Topic 1. Introduction to the course. Reactive and nonreactive methods.

Reactive and nonreactive methods. Nonreactive online methods. Nonparticipant observation and analysis of "digital footprints". Big data. The typology of nonreactive data. Social media, clickstream data, tracking data. The opportunities and limitations of big data in social sciences.

Topic 2. Big data in social sciences. Different applications of big data. Causal inference.

Different approaches of applying big data in social sciences. Experimental, quasi-experimental, and non-experimental studies. Large-scale experimental studies and causal inference in social sciences. How can social scientists use big data to design large-scale experiments for yielding causal estimates that previously have not been observed by surveys or other methods of data collection? The experiment conducted by King et al. Quasi-experimental

studies: the effect of social media on attitudes, social and political behavior. Non-experimental studies: social media as an indicator of attitudes and behavior.

Topic 3. Traps in big data. Sources of bias. Ethical concerns.

Traps in big data. Sources of bias. The case of google flu. Critical lessons we should learn. Transparency. Replicability. Are observed patterns robust? The importance of algorithms. Are algorithms robust? The importance of studying the evolution of socio-technical system. The transparency of platforms with their data collection and data analysis strategies. Ethical concerns. What kind of data we can collect and what kind of experiments conduct?

Topic 5. Introduction to network analysis in R.

Introduction to network analysis: basic definitions, centrality measures, different approaches.

Topic 6. Introduction to webscraping in R.

Introduction to webscraping in R. Collecting online data. Collecting unstructured and structured data via R. Scraping web data from APIs.

Topic 7. Collecting Twitter data.

OAuth as a nice way to authenticate with a server, and as a result, to collect data online. Consumer key, consumer secret, request token, access token. Packages in R for an interface to the OAuth specification: ROAuth, RCurl.

Using application to collect Twitter data. Streaming API and REST API for collecting data in Twitter. The difference between two options. R packages for collecting data via REST API: twitteR, netdemR, rtweet. R packages for collecting data via streaming API: streamR. Problems one might have with collecting data via streaming API and via REST API. The preferred option to collect tweets.

Topic 8. Collecting data in Vkontakte. Collecting data in Facebook.

Using application to collect *Vkontakte* and *Facebook* data.

III. ASSESSMENT AND GRADING POLICY

Each student should perform 3 out of 4 home assignments.

Type of assessment	Form of assessment	Parameters
Home	Home assignment 1.	R code and answers to the questions in R

assignments	Introduction to network analysis.	Markdown and pfd formats.
	Home assignment 2. Web Scraping in R.	R code and answers to the questions in R Markdown and pfd formats.
	Home assignment 3. Data collection in Twitter.	R code and answers to the questions in R Markdown and pfd formats.
	Home assignment 4. Data collection in VK/Facebook.	R code and answers to the questions in R Markdown and pfd formats.

Students a required to submit their home assignments within required date set by the course instructor. Home assignments sent after the deadline will not be evaluated. The answer with R code and answers to the questions should be sent in two formats: R Markdown and pfd format. Each student should perform 3 out of 4 home assignments. Home assignments are graded based on:

- Proper R code
- Answers to the questions:
  - depth of understanding of the problem,
  - depth of understanding of possible answers to the research question;
  - ability to suggest a solution to the research question;
  - consistency of the argument;
  - clear structure of the answer.

All home assignments should be done individually.

All assessments are graded from 1 (fail) to 10 (excellent). Final grade is calculated from the grades for home assignments (HA_i, class attendance (G_{class}) and participation (G_{part}). The final grade will be based only on accumulative evaluation:

$$G_{final} = 0.1*G_{class} + 0.1*G_{part} + 0.8*(HA_1 + HA_2 + HA_3)$$

There is *no exam* in the course. The final grade is based on the grade the student would get based on the seminars and home assignments.

IV. EXAMPLES OF ASSESSMENT TOOLS

Examples of tasks from home assignments:

Create edge data with 50 nodes - name these nodes as you want. Let the edge be equal 1 if there is an edge between the nodes, and 0 if there is no edge between the nodes.

Now you should write and run your R code and comment the results you have:

- What kind of network is it? How many nodes do you have and how many edges?

- Plot the network with three different layouts and comment. Try to improve the plots as much as possible.

- Calculate all centrality measures that are possible. Comment on the results.

V. Resources

V.1. Compulsory literature:

- Lazer, D., & Radford, J. Data ex Machina: Introduction to big data // *Annual Review of Sociology*, 2017, *43*(1), 7.1-7.21 https://www.annualreviews.org/doi/pdf/10.1146/annurev-soc-060116-053457

Golder, S. A., and Macy, M. W. Digital footprints: Opportunities and challenges for online social research // *Annual Review of Sociology*, 2014, 40(1), 129-152 https://www.annualreviews.org/doi/pdf/10.1146/annurev-soc-071913-043145

V.2. Additional literature:

- King, G., Pan, J., & Roberts, M. E. Reverse-engineering censorship in China: Randomized experimentation and participant observation // *Science*, 2014, 345 (6199) http://science.sciencemag.org/content/345/6199/1251722/tab-pdf

- Lazer D., Kennedy R., King G., and Vespignani A. (2014). The parable of Google flu: Traps in big data analysis // *Science*, 343. P. 1203-1205 http://science.sciencemag.org/content/343/6176/1203.full

- Ruths D, and Pfeffer J. (2015). Social media for large studies of behavior // *Science*, 346 (6213). P. 1063-1064 http://science.sciencemag.org/content/346/6213/1063/tab-pdf

V.3. Software:

№ п/п	Title	Access
1.	Microsoft Windows 7 Professional RUS Microsoft Windows 10 Microsoft Windows 8.1 Professional RUS	*From internal network of HSE (contract)*
2.	Microsoft Office Professional Plus 2010	*From internal network of HSE*

		(contract)
3.	R, RStudio. Packages for R	*Open-source license*

V.4. Professional data bases, information systmes, internet resources (online educational resources):

№ п/п	**Title**	**Access**
	Professional data bases	
1.	European Social Surveys databases	Open access. URL: https://www.europeansocialsurvey.org/

V.5. Material and technical base

Classrooms for lectures on discipline provide use and demonstration of thematic illustrations, appropriate discipline program consisting of:

- PC with Internet access (operational system, office software, anti-virus software, R, RStudio);

- Multimedia projector with remote control.

Classrooms on discipline are equipped with access to electronic information and educational HSE resources.