Computational Methods for Text Analysis

Course Syllabus

<table>
<tr><td>Title of the course</td><td colspan="3">Computational Methods for Text Analysis</td></tr>
<tr><td>Title of the Academic Programme</td><td colspan="3">Sociology and Social Informatics</td></tr>
<tr><td>Type of the course</td><td colspan="3">elective</td></tr>
<tr><td>Prerequisites</td><td colspan="3">Students are assumed to be familiar with the traditional content analysis, and have basic knowledge of statistics. Some background knowledge in linguistics will be helpful, but is not critical. R programming environment will be used for practical excercises, so basic knowledge of R is desirable. Reading any introductory book/tutorial on R (and doing the exercises it suggests) is a viable option when no formal training in R is available.</td></tr>
<tr><td>ECTS workload</td><td colspan="3">6</td></tr>
<tr><td rowspan="2">Total indicative study hours</td><td>Directed Study</td><td>Self-directed study</td><td>Total</td></tr>
<tr><td>56</td><td>172</td><td>228</td></tr>
<tr><td>Course Overview</td><td colspan="3">For social science research, written text provide essential data for studying ideology and political discourse, conflict, sentiment and political affiliation, among many other things. With a growing availability of larger collections of text in digital form it is tempting to scale the research up in terms of the population studied (e.g. “all social media users of a town”), time spans (e.g. “all of the Post-Soviet history”), and geographical scope (e.g. “all educational migration in Russia”). Computational methods for text analysis promise to aid at the scale where traditional content analysis is not feasible.

The goal of the course is to provide basic understanding on how to properly use collections of texts as quantitative evidence, and to make this knowledge practical. During the course we will cover basic word statistics, various exploratory methods, supervised and unsupervised modeling of text phenomena.</td></tr>
<tr><td>Intended Learning Outcomes (ILO)</td><td colspan="3">1) Understanding possibilities of the automated text analysis as well as its pitfalls and important caveats about applying statistical tests to language data.

2) Understanding multidimenional representation of lexical meaning and the role of the dimensionality reduction.

3) Being able to apply computational methods of text analysis (e.g. analysis of word frequency and co-occurrence, document classification, topic modeling) to collections of texts.

4) Being able to apply word embedding and clustering methods to downstream tasks, such as sentiment analysis, ideological scaling etc.

5) Being able to adequately interpret and report the results of</td></tr>
</table>

	computational text analysis in research papers.
Teaching and Learning Methods	Theoretical side of the course is centered around reading and discussing a sample of recent papers as case studies. Practical analysis of small to medium-sized text collections is used as a tool to familiarize students with the practical aspects of the methodology of computational text analysis. A set of specially prepared scripts in the R programming language are used as a sample for learning both coding practices and methods of analysis. Students are required to reproduce the analysis presented in the sample script with some modifications.

Content and Structure of the Course

№	Topic / Course Chapter	Total	Directed Study		Self-directed Study
			Lectures	Tutorials	
1	Style — Document classification	57		14	43
2	Content — Topic modeling	57		14	43
3	Sentiment — Sentiment analysis	57		14	43
4	Structure — Entities extraction	57		14	43
Total study hours		228		56	172

Indicative Assessment Methods and Strategy	Students are required to read all papers offered as case studies in the course. Upon reading each paper each student should either hand in a written summary of the paper or giva an in-class presentation on one of the aspects of the paper (a) theory and background; b) data; c) methodology; d) results). Practical analysis skill are assessed during in-class tutorials using sample scripts and homework analysis reports ("lab work"). A written report should be handed in for each lab work. General assessment includes a written mid-term test and a final group project. For a final project students are required to select a topic and data of their interest and to perform a whole cycle of data collection, analysis and presentation using methods of computational text analysis. A written report and an in-class presentation are required as a result for each project. All written work and oral presentations are assessed on a 10-point scale. The *final grade* for the course is 0.7 *course participation* (=0.3* paper summaries/presentations + 0.2 * in-class participation + 0.2 * homework + 0.3 mid-term test) + 0.3 * *final project*. If a student has a decent excuse for missing a mid-term test or any of the deadlines for written work, these could be submitted later after negotiation with the instructor.
Readings / Indicative Learning Resources	<u>Mandatory</u> 1. Bamman D., Eisenstein J., Schnoebelen T. Gender identity and lexical variation in social media // Journal of Sociolinguistics. 2014. T. 18, No 2. C. 135—160 2. Jockers M. L., Mimno D. Significant themes in 19th-century literature // Poetics. 2013. Vol. 41, No 6. C. 750—769 3. Narrative framing of consumer sentiment in online restaurant reviews / D. Jurafsky [et al.] // First Monday. 2014. Vol. 19, No 4 4. Predicting the rise and fall of scientific topics from trends in their

	rhetorical framing / V. Prabhakaran [et al.] // Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Vol. 1. 2016. C. 1170—1180. Optional 1. "How old do you think I am?"A study of language and age in Twitter / D.-P. Nguyen [и др.] // Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media. AAAI Press, 2013. C. 439—448		
Indicative Self- Study Strategies	**Type**	**+/–**	**Hours**
	Reading for seminars / tutorials (lecture materials, mandatory and optional resources)	+	40
	Assignments for seminars / tutorials / labs	+	80
	Project work	+	50
	Preparation for the mid-term test	+	12
Academic Support for the Course	Academic support for the course is provided via LMS, where students can find: guidelines and recommendations for doing the course; guidelines and recommendations for self-study; samples of assessment materials		
Facilities, Equipment and Software	All practical text analysis is done using R statistical environment.		
Course Instructor	Dr. Kirill Maslinsky, associate professor at the Sociology department		

Intended Learning Outcomes (ILO) Delivering

Programme ILO(s)	**Course ILO(s)**	**Teaching and Learning Methods for delivering ILO(s)**	**Indicative Assessment Methods of Delivered ILO(s)**
PC-6 Ability to independently formulate goals and set specific objectives for scientific research in various areas of sociology and solve them using modern research methods	A student is able to correctly formulate research objective and to implement a methodologically sound solution in case of lab work or individual project	Formulating and solving research tasks on text data analysis in the course of lab works and individual researc projects	Lab works, individual research projects
PC-7 Ability to participate in the preparation and drawing up of scientific/technical documentation and scientific reports	A student is able to fully and consistently report on all stages of the data analysis in the lab work or individual project final report	Writing final reports on the results of lab works and individual projects	Lab works, individual research projects
PC-11 Ability to plan and carry out project works in the fields of	A student is able to correctly define an	Lectures, reading assigned papers, practical	Summarizing assigned papers, in-class tests

public opinion research and organization of work of marketing services providers	applicable methodology for the assigned text analysis task	assignments	
PC-13 Ability to use the methods of collecting, processing, and interpreting comprehensive social information to solve organizational and managerial tasks, including those outside the immediate scope of activities	Carries out tasks on data analysis in the R environment fully and correctly.	Practical work with text datasets in the R environment, lab words, individual projects	In-class text analysis tasks, lab works, individual projects

Types of Assessment	**Forms of Assessment**	**Modules**			
		1	**2**	**3**	**4**
Formative Assessment	Paper summary/Presentation	*	*		
	In-class participation/tutorials	*	*		
	Homework analysis ("lab work")	*	*		
	Mid-term test	*			
Summative Assessment	Final project		*		

Assessment Criteria

In-class participation/tutorials and paper summaries

Grades	**Assessment Criteria**
«Excellent» (8-10)	A critical analysis which demonstrates original thinking and shows strong evidence of preparatory research and broad background knowledge.
«Good» (6-7)	Shows strong evidence of preparatory research and broad background knowledge. Excellent oral expression.
«Satisfactory» (4-5)	Satisfactory overall, showing a fair knowledge of the topic, a reasonable standard of expression. Some hesitation in answering follow-up questions and/or gives incomplete or partly irrelevant answers.
«Fail» (0-2)	Limited evidence of relevant knowledge and an attempt to address the topic. Unable to offer relevant information or opinion in answer to follow-up questions.

Written Assignments (Homework analysis report, Mid-term Test/Quiz)

Sample test questions:

1. Name quantitative text features that were used in stylometry for authorship detection, in the order of their historical appearence.
2. Three data set are given: (---+++++++, +-+-+-++++,++--++--++++), arrange them in the order of growing entropy.
3. What quantitative feature characterizing texts is necessary for text classification task?

Sample task for homework analysis (lab work):

A collection of literary fiction is given as data. Some of the texts are written by male authors, and the others — by female authors. A task is to build a topic model for the text collection to discover what topics significantly more or less often in the texts by male and female authors and what topics are gender-neutral. Statistical significance is to be evaluated according to the method suggested in the Jockers, Mimno 2013 paper.

Grades	Assessment Criteria
«Excellent» (8-10)	Has a clear argument, which addresses the topic and responds effectively to all aspects of the task. Fully satisfies all the requirements of the task; rare minor errors occur;
«Good» (6-7)	Responds to most aspects of the topic with a clear, explicit argument. Covers the requirements of the task; may produce occasional errors.
«Satisfactory» (4-5)	Generally addresses the task; the format may be inappropriate in places; display little evidence of (depending on the assignment): independent thought and critical judgement include a partial superficial coverage of the key issues, lack critical analysis, may make frequent errors.
«Fail» (0-2)	Fails to demonstrate any appropriate knowledge.

Project Work

Sample project topics

- A system for recommendation of relevant emoji to chat participants
- Analysis of political agenda of State Duma of Russian Federation based on session transcripts
- Language usage on the service “Yandex.District” in Saint-Petersburg

Grades	Assessment Criteria
«Excellent» (8-10)	A well-structured, analytical presentation of project work. Shows strong evidence and broad background knowledge. In a group presentation all members contribute equally and each contribution builds on the previous one clearly; Answers to follow-up questions reveal a good range and depth of knowledge beyond that covered in the presentation and show confidence in discussion.
«Good» (6-7)	Clearly organized analysis, showing evidence of a good overall knowledge of the topic. The presenter of the project work highlights key points and responds to follow up questions appropriately. In group presentations there is evidence that the group has met to discuss the topic and is presenting the results of that discussion, in an order previously agreed.
«Satisfactory» (4-5)	Takes a very basic approach to the topic, using broadly appropriate material but lacking focus. The presentation of project work is largely unstructured, and some points are

	irrelevant to the topic. Knowledge of the topic is limited and there may be evidence of basic misunderstanding. In a group presentation, most of the work is done by one or two students and the individual contributions do not add up.
«Fail» (0-2)	Fails to demonstrate any appropriate knowledge.

Recommendations for students about organization of self-study

Self-study is organized in order to:

- Systemize theoretical knowledge received at lectures;
- Extending theoretical knowledge;
- Learn how to use legal, regulatory, referential information and professional literature;
- Development of cognitive and soft skills: creativity and self-sufficiency;
- Enhancing critical thinking and personal development skills;
- Development of research skills;
- Obtaining skills of efficient independent professional activities.

Self-study, which is not included into a course syllabus, but aimed at extending knowledge about the subject, is up to the student's own initiative. A teacher recommends relevant resources for self-study, defines relevant methods for self-study and demonstrates students' past experiences. Tasks for self-study and its content can vary depending on individual characteristics of a student. Self-study can be arranged individually or in groups both offline and online depending on the objectives, topics and difficulty degree. Assessment of self-study is made in the framework of teaching load for seminars or tests.

In order to show the outcomes of self-study it is recommended:

- Make a plan for 3-5 presentation which will include topic, how the self-study was organized, main conclusions and suggestions and its rationale and importance.
- Supply the presentation with illustrations. It should be defined by an actual task of the teacher.

Recommendations for essay

An essay is a written self-study on a topic offered by the teacher or by the student him/herself approved by teacher. The topic for essay includes development of skills for critical thinking and written argumentation of ideas. An essay should include clear statement of a research problem; include an analysis of the problem by using concepts and analytical tools within the subject that generalize the point of view of the author.

Essay structure:

1. *Introduction and formulation of a research question.*

2. *Body of the essay* and theoretical foundation of selected problem and argumentation of a research question.

3. *Conclusion* and argumentative summary about the research question and possibilities for further use or development.

Special conditions for organization of learning process for students with special needs

The following types of comprehension of learning information (including e-learning and distance learning) can be offered to students with disabilities (by their written request) in accordance with their individual psychophysical characteristics:

6) *for persons with vision disorders:* a printed text in enlarged font; an electronic document; audios (transferring of learning materials into the audio); an individual advising with an assistance of a sign language interpreter; individual assignments and advising.

7) *for persons with hearing disorders: a* printed text; an electronic document; video materials with subtitles; an individual advising with an assistance of a sign language interpreter; individual assignments and advising.
8) *for persons with muscle-skeleton disorders: a* printed text; an electronic document; audios; individual assignments and advising.