TEMPLATE

Course Syllabus

<table>
<tr><td>Title of the course</td><td colspan="3">Computational Text Analysis</td></tr>
<tr><td>Title of the Academic Programme</td><td colspan="3">MA programme in Comparative Politics of Eurasia, Business and Politics in Modern Asia</td></tr>
<tr><td>Type of the course</td><td colspan="3">Elective</td></tr>
<tr><td>Prerequisites</td><td colspan="3">Students should have solid knowledge of political theory and basic knowledge of standard statistics and quantitative approach to political science research.</td></tr>
<tr><td>ECTS workload</td><td colspan="3">5</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>40</td><td>150</td><td>190</td></tr>
<tr><td>Course Overview</td><td colspan="3"><h2>Course rationale</h2> Our society is leaving more and more “digital traces” that are being accumulated at an unprecedented scale. Some of these data are publicly available and produce immediate societal effects, others have to be mined and processed before they yield some meaning, still they are available for the analysis by scholars. The process of creation / emergence of these traces is often a process of individual or mass communication, or of digitally mediated problem solving (such as online purchases, search, or rating). This process is not just a mirror or a derivation of some offline social reality, it is a new type of social processes, and a large portion of these processes are political in nature. Therefore, all these traces, either stored or evolving in real time, can be the subject of study in political science. And they demand new methods of research. Much of this data is textual, and another large portion is closely linked to the texts, and mostly they are too large to be processed manually. <h2>Course goals</h2> The aim of this course is to give students ready-to-use instruments allowing to analyze relatively large text and related data for the purposes of political science. The students will learn the types of research tasks that may be solved with text data, approaches to data preparation, analysis and interpretation, including text classification, clustering, topic modeling and other techniques. The students will also get acquainted with examples of such research by doing reading assignments. The course will use only interface software that demands no scripting skills. Neither will the students have to understand all mathematical details of algorithms they will learn to apply, however, they will have to learn their limitations. </td></tr>
<tr><td>Intended Learning Outcomes (ILO)</td><td colspan="3">$GPLO_1$: Able to conduct professional communication in Russian and/or English in a multicultural environment with the use of different communication technologies
PLO_1: Able to use relevant research results in political science and adjacent sciences, to develop applications of political science for solving practical tasks</td></tr>
</table>

	PLO_4: Able to analyze empirical data with the use of modern qualitative and quantitative methods and software PLO_6: Able to develop a design for academic and applied research, including collective one, with the use of modern political science methodology
Teaching and Learning Methods	Lectures: introduce students into methods of text analysis in political science and explain how to implement them in Orange. Tutorials: in-class individual trainings with Orange where student practice skills of text analysis for political science. Group discussion: students read a research paper implementing one of the methods studied and discuss its advantages and limitations in class. Individual home training: students are given tasks that they may perform, on their choice, to better master skills obtained during in-class practice. Team projects: students learn to set research tasks that involve automated text analysis for political science, to choose and implement relevant methods and procedures, to summarize and present results and to coordinate their work in a team.

Content and Structure of the Course

№.	**Topic**	**Total**	**Directed Study**		**Self-directed Study**
			Lectures	**Tutorials**	
1	Introduction to computational text analysis in political science.	1	1	0	0
2	Preparing texts for analysis.	23	2	1	20
3	Unsupervised machine learning: clustering	26	2	4	20
4	Unsupervised machine learning: topic modeling	26	2	4	20
5	Supervised machine learning: classification	26	2	4	20
6	Supervised machine learning: sentiment analysis	26	2	4	20
7	Text labeling for supervised methods	25	2	3	20
8	Designing, discussing and performing political science projects with automated text analysis	35	2	3	30
9	Introduction to Orange and tools overview	2	1	1	0
Total study hours		190	16	24	150

Indicative Assessment Methods and Strategy	Gfinal = 0,28Gcw1 + 0,28Gcw2 + 0,24Gpres + 0,20exam CW1 – class work 1. Individual in-class practical work on clustering and topic modeling for political science purposes, in Orange. CW2 – class work 2. Individual in-class practical work on classification

	and sentiment analysis for political science purposes, in Orange. Gpres – in-class presentation of a group project, in PPT. Should contain: project purpose, dataset description, results obtained in Orange and their interpretation. Dataset and calculation outputs should be sent beforehand. Exam: students can choose one of the tasks: (a) a four-page home essay based on group work results or on a similar individual project (b) an in-class task either on classification or clustering similar to CW1 or CW2. Home essay should be submitted 1 day prior to the official examination date. Students may be exempt of the exam on their request in which case their exam grade is considered to be equal to the average of other grades.
Readings / Indicative Learning Resources	Mandatory reading 1. *From Text to Political Positions: Text analysis across disciplines*, edited by Bertie Kaal, et al., John Benjamins Publishing Company, 2014. Part 1. 2. Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. 3. Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A. and Parnet, O. (2016). A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates. *Journal of Communication*, 66: 1007–1031. 4. Barberá et al. (2019). Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review* . DOI: https://proxylibrary.hse.ru:2120/10.1017/S0003055419000352 Published online by Cambridge University Press: 12 July 2019. Other mandatory learning resources 1. Orange data mining software download page: https://orange.biolab.si/ 2. Orange tutorial on YouTube: https://www.youtube.com/channel/UClKKWBe2SCAEyv7ZNGhIe4g Optional • Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: reproducible and agile production of political data. *American Political Science Review*, 110(2), 278-295. • Gayo-Avello, D. A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review.* December 2013 vol. 31 no. 6, 649-679. • González-Bailón, S., and Paltoglou, G. 2015. Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources, *Annals of the American Academy of Political and*

<table>
<tr><td></td><td>
Social Science, 659(1), 95-107.

Laver M., Benoit K., and John Garry. Extracting policy positions from political texts using words as data. American Political Science Review, 97(2):311–331, 2003.
Sayre, B., L. Bode, D. Shan, D. Wilcox, and C. Shah. 2010. “Agenda Setting in a Digital Age: Tracking Attention 8 in Social Media, Online News, and Conventional News.” Policy and Internet 2 (2): 7–32.
van Attenvedt W., Kleinnijenhuis J., Ruigrok N. Parsing, Semantic Networks, and Political Authority Using Syntactic Analysis to Extract Semantic Relations from Dutch Newspaper Articles. Political Analysis (2008) 16 (4): 428-446
Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. Political Communication, 29(2), 205-231.

Other optional resources
Kaggle – recommended source of datasets https://www.kaggle.com/datasets
</td></tr>
<tr><td>Indicative Self- Study Strategies</td><td>
<table>
<tr><th>Type</th><th>+/–</th><th>Hours</th></tr>
<tr><td>Reading for seminars / tutorials (lecture materials, mandatory and optional resources)</td><td>+</td><td>20</td></tr>
<tr><td>Assignments for seminars / tutorials / labs</td><td>+</td><td>80</td></tr>
<tr><td>E-learning / distance learning (MOOC / LMS)</td><td>-</td><td></td></tr>
<tr><td>Fieldwork</td><td>-</td><td></td></tr>
<tr><td>Project work</td><td>+</td><td>30</td></tr>
<tr><td>Other (please specify)</td><td>-</td><td></td></tr>
<tr><td>Preparation for the exam</td><td>+</td><td>20</td></tr>
</table>
</td></tr>
<tr><td>Academic Support for the Course</td><td>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.</td></tr>
<tr><td>Facilities, Equipment and Software</td><td>(If required) Computer class and Orange free software pre-installed.</td></tr>
<tr><td>Course Instructor</td><td>Elena Koltsova ekoltsova@hse.ru</td></tr>
</table>

Annex 1

Course Content

Introduction to computational text analysis in political science. Why modern political science needs automated text analysis. What types of tasks may be solved with such method. Advantages and limitations of automated approach to text. What is machine learning and how it is related to text analysis. What is natural language processing. This topic includes reading: Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297. Forms of work: lectures and group discussion. Forms of control: control: a question on the listed reading is included in Gcw1
Preparing texts for analysis. Text formats: plain text, vector representation, dissimilarity matrix. Text cleaning. Lemmatization and stemming. Stop words and approaches to their deletion. Transformation to vector form and types of frequencies: absolute frequencies, tf-idf and their advantages. Types of distances between texts: cosine similarity, Hamming, etc, their advantages and limitations. Dimentionality reduction and feature selection. Forms of work: lecture and tutorial with Orange. Forms of control: this element must be implemented in practice in Gcw1, 2 and Gpres.
Unsupervised machine learning: clustering Unsupervised learning. What tasks it is suited for, its advantages and limitations. Clustering and the problem of ground truth. Quality metrics for cluster analysis. Flat clustering, K-means. Types of distances between clusters (not to be confused with distances between texts). Instability and approaches to algorithm initialization. How to choose the number of clusters. Hierarchical clustering and when it is better. Where to cut the hierarchy. Difficulties of clustering texts and other high-dimentional data. How to interpret clusters and working with outputs. Clustering in Orange: tutorial with a simple dataset. Forms of work: lecture and tutorial with Orange. Forms of control: this element must be implemented in practice in Gcw1 and may be chosen by students for Gpres and for Exam.
Unsupervised machine learning: topic modeling Topic modeling as bi-clustering of texts and words. Its advantages and limitations. Research examples. Topic modeling output, its labeling and interpretation. Junk, “glued”, wide and narrow topics. Choice of the topic number with research intuition and with metrics. The problem of measuring topic modeling quality. Perplexity, coherence and entropy as metrics of quality. The problem of stability and approaches to solving it: comparing solutions and

choosing stable topics; maximizing stability. The meaning and influence of hyperparameters. This topic indludes reading: Barberá et al. (2019). Who Leads? Who Follows? Measuring Issue Attention and Agenda Setting by Legislators and the Mass Public Using Social Media Data. *American Political Science Review* .
DOI: https://proxylibrary.hse.ru:2120/10.1017/S0003055419000352 Published online by Cambridge University Press: 12 July 2019.

Forms of work: lecture, group discussion and tutorial with Orange.
Forms of control: this element must be implemented in practice in Gcw1 and may be chosen by students for Gpres and for Exam.

Supervised machine learning: classification

Preparing data for classification. Approaches to feature engineering: manual and automatic approaches. N-grams, emoticons, linguistic rules and text meta-data as features. Feature weighting. Main algorithms: Naïve Bayes, SVM, Logistic Regression, neural networks. Choosing SVM parameters. Classification quality measures: accuracy, precision, recall, F-measure; overall measures and class-specific measures. Their relative importance for different tasks. The problem of class imbalance. Performing classification in Orange.

Forms of work: lecture and tutorial with Orange.
Forms of control: this element must be implemented in practice in Gcw2 and may be chosen by students for Gpres and for Exam.

Supervised machine learning: sentiment analysis

What is sentiment? Sentiment, emotion and opinion. Machine learning and dictionary approaches in sentiment analysis. Approaches to creating dictionaries, their labeling and testing. Examples of existing dictionaries. Advantages and limitations of narrow and wide dictionaries. Creating linguistic rules. The importance of grammar and syntax for sentiment analysis. Difficulties of sentiment analysis of political texts.
This topic includes reading:
Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A. and Parnet, O. (2016). A Bad Workman Blames His Tweets: The Consequences of Citizens' Uncivil Twitter Use When Interacting With Party Candidates. *Journal of Communication*, 66: 1007–1031.

Forms of work: lecture, group discussion and tutorial with Orange.
Forms of control: this element must be implemented in practice in Gcw2 and may be chosen by students for Gpres and for Exam.

Text labeling for supervised methods

The problem of labelled corpora. Examples of labelled corpora for different tasks. Natural mark-up. Mark-up by assessors and the problem of ground truth. Crowdsourcing platforms, their advantages and limitations. Examples. Assessor work quality, criteria and methods of control and stimulation. Assessor disagreement, its causes and approaches dealing with it. Agreement metrics: simple share, Cohen's Kappa, Krippendorf's Alpha, their advantages and limitations. Labeling practice: assessor training, pilot labeling and discussion.

Forms of work: lecture, in-class practice and group discussion. Forms of control: a question on this element is included in Gcw2.
Designing, discussing and performing political science projects with automated text analysis. This topic is first learned through reading examples of research that implement some of the studied methods to political science tasks. It then includes performing a home team task that is then presented and discussed in class. Forms of work: lectures, in-class group discussions, home team work. Forms of control: this topic is controlled by Gpres.
Introduction to Orange and tools overview Introduction to Orange is a topic that goes through all other topics as their practical extension. It is controlled through Gcw1&2. It includes the following subtopics. Opening Orange and importing data. Organizing workflows. Orange functions overview. Data visualization. Exporting results. Text preprocessing with Orange. Clustering with Orange. Classification with Orange. Tools overview is given in the end of the course and is not controlled. It includes a lecture with an overview of clustering and classification functionalities given in R, Python and available GUI software.

Intended Learning Outcomes (ILO) Delivering

Course ILO(s)	Teaching and Learning Methods for delivering ILO(s)	Indicative Assessment Methods of Delivered ILO(s)
$GPLO_1$: Able to conduct professional communication in Russian and/or English in a multicultural environment with the use of different communication technologies	Interactive lectures, discussions of papers	Presentation of team work; exam in the form of an essay
PLO_1: Able to use relevant research results in political science and adjacent sciences, to develop applications of political	Interactive lectures, discussions of results obtained from data during tutorials on Orange	CW1 & CW2 – two in-class individual practical tasks; presentation of team work

science for solving practical tasks		
PLO_4: Able to analyze empirical data with the use of modern qualitative and quantitative methods and software	Interactive lectures, practical work in Orange software for automated political text analysis	CW1 & CW2 – two in-class individual practical tasks
PLO_6: Able to develop a design for academic and applied research, including collective one, with the use of modern political science methodology	Discussion of papers as examples of research, development and implementation of a small team research project	Group work presentation

Annex 2

Assessment Criteria *(please use it as an example only and insert the activities you use within your course. The assessment forms and criteria should align with real teaching of the course)*

Individual practical in-class work

Grades	Assessment Criteria
«Excellent» (8-10)	To obtain 9, all tasks are understood correctly in terms of political science. All datsets are preprocessed accordingly. All Orange workflows are implemented completely and correctly. The final results are correct mathematically and correctly interpreted in terms of political science. All test questions, if any, are answered correctly. Some elements may be imperfect to obtain 8. 10 is given when the task exceeds expectations (distinction).
«Good» (6-7)	All tasks are understood correctly in terms of political science, but one of the further elements is missing or is incorrect. Two less important elements may be missing to obtain 6.
«Satisfactory» (4-5)	The task may be understood incorrectly, but the following workflows contain no more than one or two missing or incomplete elements. Or, the task is understood correctly, but the workflow is essentially unfinished or has a wrong result and an unsatisfactory interpretation.
«Fail» (0-3)	Any work that does not meet criteria of "satisfactory". 0 is given for a missing work or to a missing student.

Project Work

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. The dataset and the results of calculations, with the brief description of the research goal, are sent to the course instructor no less than 24 hours before the presentation. The presentation contains: Research goal, dataset description, workflow description, results, interpretation / conclusions, limitations and, if applicable, practical recommendations. A brief review of presentations of the others may be included as a criterium, upon the instructor's decision. To obtain 8, some of the elements may be imperfect. 10 is given when the task exceeds expectations (distinction).
«Good» (6-7)	Clearly organized analysis, showing evidence of a good overall knowledge of the topic. The presenters of the project work highlight key points and respond to follow up questions appropriately. 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. Some of the elements of the presentation highlighted above are missing or are essentially erroneous. The dataset and the results or calculations are sent on time. The interpretation may be to some extent shallow.
«Satisfactory» (4-5)	Takes a very basic approach to the topic. 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. The goal is set in a meaningful way, but the further data processing and analysis is either incomplete or has major flaws. OR, the goal is set incorrectrly / not in a meaningful way, but the further data prcessing and analysis shows that participants have relevant data mining skills. The dataset and the calculations are not sent on time.
«Fail» (0-3)	Fails to meet criteria of "satisfactory". The data and the calculations are missing OR the

	presentation is messy and meaningless with many major errors. 0 is given for a missing work or to a missing team.

Exam in the form of a home essay

Grades	Assessment Criteria
«Excellent» (8-10)	Home essay should be a concise description of the group project that employs excellent academic writing. The flow of thoughts and data description should be logical and coherent. It should contain research goal, dataset description, workflow description, results, interpretation / conclusions, limitations and, if applicable, practical recommendations. In addition to the team work, it should also contain some brief reference to 4-5 relevant research pieces and one additional aspect of data analysis (e.g. a test of an alternative classification model or a set of features). This element should be unique (individually done). If the team works contained errors in data analysis, those should be corrected. If the team work was graded low, excellent grading of the essay is impossible, and it is recommended either to do a completely new project or to choose another type of the final exam. Data and calculations are enclosed. Rare minor errors may occur to obtain 8. 10 is given when the task exceeds expectations (distinction).
«Good» (6-7)	Academic writing is not perfect and the work has some logical flaws. Some errors in calculations or interpretation are present, that either were not corrected after the team work presentation or appered anew.
«Satisfactory» (4-5)	Generally addresses the task. Writing is understandable, but not well structured. Some serious flaws in data processing and analysis are present. Either literature or the unique element are missing. The interpretation is shallow.
«Fail» (0-3)	Does not meet critaeria of “Satisfactory”. Data and calculations are not enclosed. The text is messy, illogical and makes little sense both in terms of writing and data analysis.

Annex 3

Recommendations for students about organization of self-study

Self-study is organized in order to:

- Systemize theoretical knowledge received at lectures;
- Extend theoretical knowledge;
- Learn how to use research and professional literature;
- Develop cognitive and soft skills: creativity and self-sufficiency;
- Enhance critical thinking and personal development skills;
- Develop of research skills;
- Develop data processing and analysis skills;
- Obtain 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.

All tasks of the course (CW1, CW2 and presentation) will demand some degree of self-study. For CW1 and CW2, students are recommended to review class work and to train the application o studied methods on some datasets of their choice. It will help them to efficiently and quickly organize Orange workflows in class. For team project work self-study is required. The team should have at least two meetings, one to set the goals and to distribute work, and the other to finalize the work, the presentation and to agree on the order of speaking. Between the meetings, the team should communicate on the work progress.

Annex 4

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:

1) *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.
2) *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.
3) *for persons with muscle-skeleton disorders: a* printed text; an electronic document; audios; individual assignments and advising.