	The 3rd International IPSA – HSE Summer School for Methods of Political & Social Research Course Syllabus
Course title:	Text mining
Instructor:	Dr. Kirill Maslinsky
ECTS / academic hours	2 ECTS / 72 academic hours: 36 contact hours, 36 self – study hours
Brief course description (up to 100 words):	For social and political sciences, written text provide essential data for studying ideology and political discourse, conflict, sentiment and political affiliation, among many other things. Computational methods for text analysis promise to aid at the scale where traditional conetnt 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. We will use R programming environment as a toolbox for text analysis.
Indicative concepts (up to 10):	• Lexical statistics. Zip's law. • NLP tasks: stemming, PoS tagging, syntactic parsing. • Statistical tests for word frequency data. Dunning's log-likelihood. • Weighting schemes. TF-IDF. • Supervised text classification (Naive Bayes and Logistic regression) • Unsupervised topic modeling with (LDA and STM). • General and domain-specific sentiment lexicons.
Workshops overview:	Day 1 — *Counting words*. Preprocessing: transforming text into data in R. Appropriate statistics for text data. Day 2 — *Comparing corpora.* Locating over- and underused words in contrasting corpora. Day 3 — *Document-level modeling.* Bag-of-words: vector space model of text. Document classification task. Day 4 — *Co-occurrence.* Distributional hypothesis. Modeling topics and discourses using co-occurrence data. Day 5 — *Dictionary methods.* Sentiment analysis using lexicons. Inducing sentiment lexicons.
Assessment techiques to receive graded certificate:	To obtain ECTS credits, participants will be required to complete a small set of analytic tasks on the text dataset, either any text corpus of your own choice or the one provided by the course instructor.
Essential readings:	• 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. • Jurafsky, D., Martin, J. H. (2008) Speech and language processing. 2nd. edition. NJ: Prentice Hall, 2008. (AND some chapters from 3d edition draft, available online: http://web.stanford.edu/~jurafsky/slp3/) • Pozzi, F. A., Fersini, E., Messina, E., & Liu, B. (2016). Sentiment Analysis in Social Networks. Morgan Kaufmann.
Contacts:	kmaslinsky@hse.ru