Government of Russian Federation

Federal State Autonomous Educational Institution of High Professional Education

"National Research University Higher School of Economics"

Syllabus for the course "Research problems in Natural Language Processing"

for postgraduate program in 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 "Mathematical Modeling, Numerical Methods, and Software Systems"

Author:
Ekaterina Chernyak, assistant professor, echernyak@hse.ru

Approved by the Academic Council of the School for Postgraduate Studies in Computer Science on October 26, 2017

Moscow - 2017

1. Scope of Use

This program establishes the minimal requirements to postgraduate students' knowledge and skills for 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", "05.13.18 Mathematical Modeling, Numerical Methods, and Software Systems" and determines the content of the course and educational techniques used in teaching the course.

The present syllabus is aimed at faculty teaching the course and postgraduate students studying 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 "Mathematical Modeling, Numerical Methods, and Software Systems".

This syllabus meets the standards required by:

- Educational standards of National Research University Higher School of Economics;
- Postgraduate educational program for 09.06.01 Computer Science and Computer Engineering.
- University curriculum of the postgraduate program for 09.06.01 Computer Science and Computer Engineering / 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 Mathematical Modeling, Numerical Methods, and Software Systems", approved in 2014.

2. Learning Objectives

The learning objective of the course "Research Problems in Natural Language Proceesing" is to provide students advanced techniques and deeper theoretical and practical knowledge in modern NLP tasks, such as:

- distributional semantics;
- topic modelling;
- sequence labelling;
- structured learning;
- text classification and clustering;
- unsupervised information extraction.

3. Main Competencies Developed after Completing the Study of This Discipline

After completing the study of the discipline the PhD student should have:

- Knowledge about such models as word embeddings, Latent Dirichlet Allocation, conditional random fields, structured SVM, convolutional neural networks, recurrent neural networks, POS-tagging and syntax parsing;
- Knowledge about ongoing developments in NLP;
- Hands-on experience with large scale NLP problems;
- Knowledge about how to design, develop and evaluate NLP programs using programming language Python.

After completing the study of the discipline the student should have developed the following competencies:

Competence	Code	Descriptors (indicators of achievement of the result)	Educative forms and methods aimed at generation and development of the competence
the ability to carry out theoretical and experimental research in the field of professional activity	ОПК-1	PhD students obtain necessary knowledge in probabilistic generative models	Assignments, additional material/reading provided
the ability to develop new research methods and apply them in research in one's professional field	ОПК-2	The PhD student is able to choose an appropriate model for real-life problems and to calibrate the hyperparameters.	Examples covered during the lectures and tutorials. Assignments.
the ability to objectively evaluate the outcomes of research and development carried out by other specialists in other scientific institutions	ОПК-4	The PhD student is able to carry out comparative testing of competing models or methods.	Examples covered during the lectures and tutorials. Assignments.
the ability to do research in transformation of information into data and knowledge, models of data and knowledge representation, methods for knowledge processing, machine learning and knowledge discovery methods, principles of building and operating software for automation of these processes	ПК-4	The PhD student is able to develop and analyze machine learning models, implement them in a programming language in large scale, and select the best model using validation techniques.	Lectures, tutorials, and assignments.

4. Place of the Discipline in the Postgraduate Program Structure

This is an elective course for 05.13.01 "Systems Analysis, Control Theory, and Information Processing", 05.13.11 "Mathematical Theory and Software for Computing Machinery, Systems, and Networks", 05.13.17 "Theoretical Foundations of Computer Science", 05.13.18 "Mathematical Modeling, Numerical Methods, and Software Systems".

Postgraduate students are expected to be already familiar with some statistical learning techniques, and have skills in analysis, linear algebra, optimization, computational complexity, and probability theory.

The following knowledge and competences are needed to study the discipline:

- A good command of the English language, both oral and written.
- A sound knowledge of probability theory, complexity theory, optimization, and linear algebra

5. Schedule for one 1 module

№	Topic	Total hours	Contact hours			Self-study
			Lectures	Seminars	Practice lessons	
1.	Introduction to NLP, basic concepts		2	2		14
2.	Text preprocessing: tokenization, POS-tagging, syntax parsing		4	4		14
3.	Topic modelling		2	2		14
4.	Distributional semantics		2	2		14
5.	Sequence labelling		2	2		14
6.	Structured learning		2	2		14
7.	Text classification and clustering		4	4		14
8.	Unsupervised information extraction		2	2		14
	Total	152	20	20		112

6. Requirements and Grading

Homework	1	Homework
Presence	1	
Exam	1	Written exam. Preparation time – 180 min.

7. Assessment

Final assessments are based on the homework and the final exam. Students have to demonstrate knowledge of the material covered during the entire course.

8. The grade formula

The exam is worth 50% of the final mark.

Final course mark *is obtained from the following formula*: Final=0.4*(homework)+ 0.1*(Presence on all lectures and seminars)+0.5*(Exam).

All grades having a fractional part greater than 0.5 are rounded up.

Table of Grade Accordance

Ten-point grading Scale	Five-point grading Scale	
1 - very bad 2 – bad 3 – no pass	Unsatisfactory - 2	**FAIL**
4 – pass 5 – highly pass	Satisfactory – 3	**PASS**
6 – good 7 – very good	Good – 4	
8 – almost excellent 9 – excellent 10 – perfect	Excellent – 5	

9. Course description.

Topic 1. Introduction to NLP, basic concepts

Basic definitions of NLP tasks and methods and basic introduction to linguistics, evaluation metrics and language recourses.

Topic 2. Text preprocessing: tokenization, POS-tagging, syntax parsing

Rule-based and machine learning-bases tokenization and POS-tagging, constituency and dependency grammars, syntax parsing.

Topic 3. Topic modelling

Vector space model and dimensionality reduction. Latent semantic indexing, latent Dirichlet allocation, dynamic topic models, hierarchical Dirichlet process, autoencoders.

Topic 4. Distributional semantics

Embedding models: positive pointwise mutual information matrix decomposition, singular value decomposition, word2vec, GloVe, StarSpace, AdaGram, etc.

Topic 5. Sequence labelling

Named entity recognition, relation and event extraction and POS-tagging as sequence labelling task. Hidden Markov model, Markov maximal entropy model, conditional random fields, reccurent neural networks.

Topic 6. Structured learning

Syntax parsing and semantic role labelling as structured learning task. Structured SVM and structured perceptron.

Topic 7. Text classification and clustering

Baseline methods for text classification: naïve Bayes, logisitic regression, fasttext, convolutional neural networks, hard attention mechanism for reccurent neural networks.

Topic 8. Unsupervised Information Extraction

OpenIE paradigm. SOV triples extraction, classification and clustering. Temporal textual data analysis.

10. Educational technologies

The following educational technologies are used in the study process:

- discussion and analysis of the results during the tutorials;
- regular assignments to test the progress of the PhD student;
- consultation time on Tuesday afternoons.

11. Final exam questions

The final exam will consist of a selection of problems equally weighted. Any kind of material is allowed for the exam. Each question will focus on a particular topic presented during the lectures.

The questions consist in exercises on any topic seen during the lectures. To be prepared for the final exam, PhD students must be able to answer questions from the topics covered during the lecture.

12. Reading and Materials

Literature:

1. Manning, Christopher D., and Hinrich Schütze. Foundations of statistical natural language processing. Vol. 999. Cambridge: MIT press, 1999.
2. Martin, James H., and Daniel Jurafsky. "Speech and language processing." International Edition 710 (2000): 25.
3. Goldberg, Yoav. "Neural Network Methods for Natural Language Processing." Synthesis Lectures on Human Language Technologies 10, no. 1 (2017): 1-309.

Literature for self-study:

1. Cohen, Shay. "Bayesian analysis in natural language processing." Synthesis Lectures on Human Language Technologies 9, no. 2 (2016): 1-274.

13. Equipment.

The course requires a computer room, laptop and a projector.