

**Federal State Autonomous Educational Institution of Higher Education "Moscow
Institute of Physics and Technology
(National Research University)"**

APPROVED
**Head of the Phystech School of
Applied Mathematics and
Informatics**
A.M. Raygorodskiy

Work program of the course (training module)

course: Natural Language Processing/Обработка естественного языка
major: Applied Mathematics and Informatics
specialization: Modern State of Artificial Intelligence/Современные методы искусственного интеллекта
“Pusk” Online and Supplementary Education Centre
Chair of Machine Learning and Digital Humanities
term: 2
qualification: Master

Semester, form of interim assessment: 3 (fall) - Exam

Academic hours: 75 AH in total, including:

lectures: 45 AH.

seminars: 30 AH.

laboratory practical: 0 AH.

Independent work: 75 AH.

Exam preparation: 30 AH.

In total: 180 AH, credits in total: 4

Authors of the program:

R.G. Neychev, senior professor

A.M. Raygorodskiy, doctor of physics and mathematical sciences, associate professor, главный научный сотрудник

The program was discussed at the Chair of Machine Learning and Digital Humanities 05.03.2021

Annotation

State of the art approaches in different domains of Artificial Intelligence are based on Deep Learning techniques (e.g. in Computer Vision, Natural Language Processing, Reinforcement Learning, etc.) Deep neural architectures show great potential and promise even better results, so now is definitely the time to explore this field.

In this course we will start from the basics and rapidly dive into the latest results in Natural Language Processing, focusing on novel approaches and applied techniques. This course tends to develop both practical skills and theoretical background to provide the students thorough theoretical knowledge and ability to work on their own on the NLP projects.

1. Study objective

Purpose of the course

- Get familiar with classical and novel techniques in the NLP domain
- Get hands on experience in solving Natural Language Processing problems
- Develop skills of applying NLP models to real data

Tasks of the course

- Natural Language Processing problem statement and ability to develop the general pipeline of the solution
- Choose relevant approach and model for particular problem
- Essential experience with PyTorch framework and Python

2. List of the planned results of the course (training module), correlated with the planned results of the mastering the educational program

Mastering the discipline is aimed at the formation of the following competencies:

Code and the name of the competence	Competency indicators
Gen.Pro.C-1 Address current challenges in fundamental and applied mathematics	Gen.Pro.C-1.2 Consolidate and critically assess professional experience and research findings

3. List of the planned results of the course (training module)

As a result of studying the course the student should:

know:

- statement of tasks of morphological, syntactic analysis;
- methods for solving these problems.

be able to:

- to formulate the tasks of classification of texts, sentences or their elements to highlight structured information;
- implement a suitable text classification algorithm;
- to solve the problem of highlighting keywords and determining the sentiment.

master:

- the main software systems for highlighting hidden topics and reducing the dimension of vector models.

4. Content of the course (training module), structured by topics (sections), indicating the number of allocated academic hours and types of training sessions

4.1. The sections of the course (training module) and the complexity of the types of training sessions

№	Topic (section) of the course	Types of training sessions, including independent work			
		Lectures	Seminars	Laboratory practical	Independent work
1	Text vectorization classical approaches: BoW, TF-IDF.	9	6		15
2	Exploding in deep neural networks	9	6		15
3	Beam search	9	6		15
4	Attention in encoder-decoder architecture.	9	6		15
5	GPT family overview.	9	6		15
AH in total		45	30		75
Exam preparation		30 AH.			
Total complexity		180 AH., credits in total 4			

4.2. Content of the course (training module), structured by topics (sections)

Semester: 3 (Fall)

1. Text vectorization classical approaches: BoW, TF-IDF.

Text collocations Word embeddings; word2vec and GLoVe Language models

2. Exploding in deep neural networks

Convolutional neural networks in NLP. CNN for text processing Machine translation and Neural Machine Translation.

3. Beam search

Measuring quality of generated text. BLEU/Perplexity scores. Attention mechanism. Self-attention mechanism.

4. Attention in encoder-decoder architecture.

Transformer architecture overview. Pre-training in NLP. Contextual embeddings. ELMo. BERT overview.

5. GPT family overview.

Question answering and knowledge based systems. Bi-directional attention flow (BiDAF)
Sentiment analysis POS-tagging, dependency parsing Topic modeling (PLSA, LDA) RL
techniques in NLP. Self-critical sequence training.

5. Description of the material and technical facilities that are necessary for the implementation of the educational process of the course (training module)

A standard classroom.

6. List of the main and additional literature, that is necessary for the course (training module) mastering

Main literature

1. Машинное обучение [Текст]/Х. Бринк, Дж. Ричардс, М. Феверолф, Real-World Machine Learning, -СПб., Питер, 2017
2. Python и машинное обучение [Текст] = Python Machine Learning : крайне необходимое издание по новейшей предсказательной аналитике для более глубокого понимания методологии машинного обучения / С. Рашка; пер. с англ. А. В. Логунова .— М. : ДМК Пресс, 2017 .— 418 с.: ил. - Предм. указ.: с. 408-417. - 200 экз. - ISBN 978-5-97060-409-0 (в пер.) .— Полный текст (Доступ из сети МФТИ / Удаленный доступ).

Additional literature

1. Математические основы машинного обучения и прогнозирования [Текст] / В. В. Вьюгин ; Моск. физ.-техн. ин-т (гос. ун-т), Лаб. структурных методов анализа данных в предсказательном моделировании (ПреМоЛаб), Ин-т проблем передачи информации им. А. А. Харкевича РАН - М.МЦНМО,2013

7. List of web resources that are necessary for the course (training module) mastering

<http://dm.fizteh.ru/>

8. List of information technologies used for implementation of the educational process, including a list of software and information reference systems (if necessary)

Multimedia technologies can be employed during lectures and practical lessons, including presentations.

9. Guidelines for students to master the course

A student studying a discipline must, on the one hand, master the general conceptual apparatus, and on the other hand, must learn to apply theoretical knowledge in practice.

As a result of studying the discipline, the student must know the basic definitions, concepts, axioms.

Successful mastering of the course requires intense independent work of the student. The course program provides the minimum required time for a student to work on a topic. Independent work includes:

- reading and taking notes of the recommended literature;
- study of educational material (for educational and scientific literature), preparation of answers to questions intended for independent study, proof of individual statements, properties;
- preparation for differential credit.

Guidance and control over the student's independent work is carried out in the form of individual consultations.

It is important to achieve an understanding of the studied material, and not its mechanical memorization. If you find it difficult to study certain topics, questions, you should seek advice from the lecturer.

Assessment funds for course (training module)

major: Applied Mathematics and Informatics
specialization: Modern State of Artificial Intelligence/Современные методы искусственного интеллекта
“Pusk” Online and Supplementary Education Centre
Chair of Machine Learning and Digital Humanities
term: 2
qualification: Master

Semester, form of interim assessment: 3 (fall) - Exam

Authors:

R.G. Neychev, senior professor

A.M. Raygorodskiy, doctor of physics and mathematical sciences, associate professor, главный научный сотрудник

1. Competencies formed during the process of studying the course

Code and the name of the competence	Competency indicators
Gen.Pro.C-1 Address current challenges in fundamental and applied mathematics	Gen.Pro.C-1.2 Consolidate and critically assess professional experience and research findings

2. Competency assessment indicators

As a result of studying the course the student should:

know:

- statement of tasks of morphological, syntactic analysis;
- methods for solving these problems.

be able to:

- to formulate the tasks of classification of texts, sentences or their elements to highlight structured information;
- implement a suitable text classification algorithm;
- to solve the problem of highlighting keywords and determining the sentiment.

master:

- the main software systems for highlighting hidden topics and reducing the dimension of vector models.

3. List of typical control tasks used to evaluate knowledge and skills

1. Word representations: basic approaches (BoW, TF-IDF).
2. Word embeddings (word2vec: linearity, skip-gram, negative sampling, key ideas)
3. RNN in text processing. Context, memory
4. CNN in text processing. Relations to the n-gram approach.
5. Attention mechanism.
6. Self-attention mechanism.
7. Contextualized embeddings.
8. Transformer architecture: encoder and decoder structure main details.
9. BERT architecture. Main ideas (masking, pre-training on many problems)
10. Machine translation metrics, quality functions
11. Pre-training in NLP problems
12. Exposure bias in language generation
13. Question answering systems: key concepts
14. Self-critical sequence training approach

4. Evaluation criteria

Questions for the exam

1. Prove that if m , n are two coprime integers of different parity, then the numbers $m^2 - n^2$ and $2mn$ are also coprime.
2. Write and prove the general formula for the number of different representations of a given integer n as the sum of two squares. (Representatives that are not obtained from each other by changing signs and the order of the terms are considered different.)
3. Based on the obtained formula, derive the lower bound for the maximum number of equal distances among the given n points on the plane using a regular rectangular lattice.
4. Build a regular pentagon using a compass and a ruler.

5. Build a regular 15-gon using a compass and a ruler.
6. You are given a single segment. It is required to construct using a compass and a ruler a segment of length x satisfying the equation
7. Based on the previous task, prove that a regular heptagon cannot be built using a compass and a ruler.
8. Prove that trisection of the angle is impossible.
9. Describe all possible combinations of the amounts of black and white balls in the ballot box, so that if two balls are randomly fished in a sample without returning, the probability of fishing two white balls is exactly 0.5.
10. Consider the relation on the sides a, b, c of the triangle, in which a triangle with vertices at the bases of the bisectors is isosceles. Assuming that the sides converging on side c of the large triangle are equal, reduce this relation to the following
11. In what follows, we consider the cube defined by the first of the three equations (refusing the requirement that a, b, c be sides of a triangle). Show that the resulting cube is indecomposable, that is, the polynomial that defines it does not factor.
12. In addition to this, show that our cube is nonsingular, that is, there is not a single point on its projectivization at which each direction is tangent (or the same thing at which all three first partial derivatives of the polynomial defining it degenerate).

Exam ticket examples

Ticket number 1

1. Write and prove the general formula for the number of different representations of a given integer n as the sum of two squares.
2. Prove that trisection of the angle is impossible.

Ticket number 2

1. Consider the relationship on the sides a, b, c of the triangle, in which a triangle with vertices at the bases of the bisectors is isosceles.
2. Describe all kinds of combinations of the numbers of black and white balls in the ballot box, so that if two balls are randomly fished in the sample without returning, the probability of fishing two white balls is exactly 0.5.

Assessment “excellent (10)” is given to a student who has displayed comprehensive, systematic and deep knowledge of the educational program material, has independently performed all the tasks stipulated by the program, has deeply studied the basic and additional literature recommended by the program, has been actively working in the classroom, and understands the basic scientific concepts on studied discipline, who showed creativity and scientific approach in understanding and presenting educational program material, whose answer is characterized by using rich and adequate terms, and by the consistent and logical presentation of the material;

Assessment “excellent (9)” is given to a student who has displayed comprehensive, systematic knowledge of the educational program material, has independently performed all the tasks provided by the program, has deeply mastered the basic literature and is familiar with the additional literature recommended by the program, has been actively working in the classroom, has shown the systematic nature of knowledge on discipline sufficient for further study, as well as the ability to amplify it on one’s own, whose answer is distinguished by the accuracy of the terms used, and the presentation of the material in it is consistent and logical;

Assessment “excellent (8)” is given to a student who has displayed complete knowledge of the educational program material, does not allow significant inaccuracies in his answer, has independently performed all the tasks stipulated by the program, studied the basic literature recommended by the program, worked actively in the classroom, showed systematic character of his knowledge of the discipline, which is sufficient for further study, as well as the ability to amplify it on his own;

Assessment “good (7)” is given to a student who has displayed a sufficiently complete knowledge of the educational program material, does not allow significant inaccuracies in the answer, has independently performed all the tasks provided by the program, studied the basic literature recommended by the program, worked actively in the classroom, showed systematic character of his knowledge of the discipline, which is sufficient for further study, as well as the ability to amplify it on his own;

Assessment “good (6)” is given to a student who has displayed a sufficiently complete knowledge of the educational program material, does not allow significant inaccuracies in his answer, has independently carried out the main tasks stipulated by the program, studied the basic literature recommended by the program, showed systematic character of his knowledge of the discipline, which is sufficient for further study;

Assessment “good (5)” is given to a student who has displayed knowledge of the basic educational program material in the amount necessary for further study and future work in the profession, who while not being sufficiently active in the classroom, has nevertheless independently carried out the main tasks stipulated by the program, mastered the basic literature recommended by the program, made some errors in their implementation and in his answer during the test, but has the necessary knowledge for correcting these errors by himself;

Assessment “satisfactory (4)” is given to a student who has discovered knowledge of the basic educational program material in the amount necessary for further study and future work in the profession, who while not being sufficiently active in the classroom, has nevertheless independently carried out the main tasks stipulated by the program, learned the main literature but allowed some errors in their implementation and in his answer during the test, but has the necessary knowledge for correcting these errors under the guidance of a teacher;

Assessment “satisfactory (3)” is given to a student who has displayed knowledge of the basic educational program material in the amount necessary for further study and future work in the profession, not showed activity in the classroom, independently fulfilled the main tasks envisaged by the program, but allowed errors in their implementation and in the answer during the test, but possessing necessary knowledge for elimination under the guidance of the teacher of the most essential errors;

Assessment “unsatisfactory (2)” is given to a student who showed gaps in knowledge or lack of knowledge on a significant part of the basic educational program material, who has not performed independently the main tasks demanded by the program, made fundamental errors in the fulfillment of the tasks stipulated by the program, who is not able to continue his studies or start professional activities without additional training in the discipline in question;

Assessment “unsatisfactory (1)” is given to a student when there is no answer (refusal to answer), or when the submitted answer does not correspond at all to the essence of the questions contained in the task.

5. Methodological materials defining the procedures for the assessment of knowledge, skills, abilities and/or experience

During examination the student are allowed to use the program of the discipline.