Ministry of Science and Higher Education of the Russian Federation
NATIONAL RESEARCH
TOMSK STATE UNIVERSITY (NR TSU)

Institute of Applied Mathematics and Computer Science

APPROVE:
Director

_____ A. V. Zamyatin

Work program of the discipline

Natural Language Processing - II

in the major of training

01.04.02 Applied mathematics and informatics

Orientation (profile) of training:

Big Data and Data Science

Form of study
full-time

Qualification
Master

Y ear of admission
2025

Code of discipline in the curriculum: B1.P.V.06

AGREED;
Head of EP
_____ A.V. Zamyatin

Chairman of the EMC
_____ S.P. Sushchenko

Tomsk – 2025

1. Purpose and planned results of mastering the discipline

The purpose of mastering the discipline is the formation of the following competencies:

- UK-1 - the ability to carry out a critical analysis of problem situations based on a systematic approach, to develop an action strategy;
- UK-3 - the ability to organize and manage the work of the team, developing a team strategy to achieve the goal;
- PC-6 - the ability to manage the receipt, storage, transmission, processing of big data.s.

The results of mastering the discipline are the following indicators of the achievement of competencies:

IUK-1.1 Identifies a problem situation, on the basis of a systematic approach, carries out its multifactorial analysis and diagnostics.

IUK-1.2 Carries out the search, selection and systematization of information to determine alternative options for strategic solutions in a problem situation.

IUK-1.3 Suggests and justifies the strategy of action, taking into account the limitations, risks and possible consequences.

IUK-3.1 Forms a teamwork strategy based on a joint discussion of goals and activities for their implementation.

IUK-3.2 Organizes the work of the team, taking into account the objective conditions (technology, external factors, limitations) and the individual capabilities of the team members.

IUK-3.3 Ensures the fulfillment of the assigned tasks based on the monitoring of team work and timely response to significant deviations.

IPK-6.1 Monitors and evaluates the performance of big data processing.

IPK-6.2 Uses methods and tools for receiving, storing, transmitting, processing big data.

IPK-6.3 Develops proposals to improve the performance of big data processing.

2. Tasks of mastering the discipline

The purpose of the discipline is to teach students advanced methods, models, tools and technologies of computer processing of texts in natural languages to give the ability to represent the processes of text analysis and synthesis in an algorithmic form.

Discipline tasks:

– obtaining theoretical knowledge and practical skills in processing natural language texts;

– knowledge of the difficulties associated with the use of existing methods for processing natural language texts;

– the ability to use the acquired knowledge on the development, adaptation and use of the latest tools for processing texts in natural languages;

– teach students to conduct analytical research and develop applications using natural language processing technologies in accordance with customer requirements.

3. The place of discipline in the structure of the educational program

Discipline refers to the part of the educational program formed by the participants of educational relations.

4. Semester of mastering and form of intermediate certification in the discipline

Third semester, credit.

5. Entrance requirements for mastering the discipline

For the successful mastering of the discipline, training outcomes are required in the following disciplines: "Fundamentals of Programming", "Algorithms and Data Structures", "Intelligent Systems", "Visualization of Multidimensional Data", "Statistical Methods of Machine Learning", "Natural Language Processing - I".

6. Implementation language

English.

7. Scope of discipline

The total labor intensity of the discipline is 3 credits, 108 hours, of which:
- lectures: 10 hours
- laboratory: 20 hours
  including practical training: 0 h.
The volume of independent work of the student is determined by the curriculum.

8. The content of the discipline, structured by topics

Topic 1. Introduction, history of the development of the discipline, tasks to be solved, approaches, methods and tools

Three main stages in the development of natural language processing technologies are revealed: dictionary, probabilistic and intelligent algorithms. The classification of tasks is given. The main methods for implementing algorithms are described: local, cloud services.

Topic 2. Pre-processing of text data

The purpose and types of text preprocessing are explained: segmentation, tokenization, lemmatization. Lemmatization and stemming are compared. The role of lemmatization in the construction of search engines is explained. indexes. Explanation of the non-determinism of segmentation and tokenization.

Topic 3. Probabilistic algorithms

The main features of probabilistic algorithms are given. Their role in modern systems is explained. Hidden Markov models, Viterbi algorithm, EM-algorithm are given as examples. To describe the EM-algorithm, the purpose of topic modeling is explained.

Topic 4. Formal grammars

Chomsky's definition of analytic formal grammars. Disclosure of their features and fundamental limitations. Examples of problems that can currently be solved using formal grammars. Explanations of the functions of the Tomita-parser utility.

Topic 5. Vector representation of words

The idea of replacing words by points in a vector space is explained. Examples of algebraic operations on words replaced are given. dots. Determining the semantic similarity of words through metrics in vector space. Methods for obtaining a vector representation. Word2vec model.

Topic 6. Model Seq2seq

Explanation of transformation of sequences through recurrent cells. Concepts of encoder and decoder. The idea of long short term memory. The idea of complementing the encoder and decoder with communication through the mechanism of attention.

Topic 7. Self-attention and Transformer

Justification of the shortcomings of the Seq2seq model. Introducing the concept of Self-attention and explaining its benefits. The purpose of the query, key, and value cells. Description of the transformer model. Main advantages. Description of the structure of the encoder and decoder of the Transformer.

Topic 8. BERT and GPT-3

Description of the possibilities of constructing new models on the transformer. Separate use of encoder and decoder. BERT model. Fine tuning idea. Model GPT-3. Application of GPT-3 in practical tasks.

9. Ongoing evaluation

The ongoing evaluation is carried out by monitoring attendance, conducting tests, tests on lecture material, performing laboratory work, and is recorded in the form of a checkpoint at least once a semester.

10. The procedure for conducting and criteria for evaluating the intermediate certification

Intermediate certification is carried out in the form of a test. The results of the test - the marks "passed", "not passed" are affixed according to the results of the delivery of practical work.

List of practical works:

Practical work No. 1. Parsing sites / using api to get text data.

Practical work No. 2. Porter's stemmer implementation.

Practical work No. 3. Using libraries for morphological analysis, solving the problem of part-of-speech marking.

Practical work No. 4. Vector representation of text, word2vec, skip-gram and CBOW models.

Practical work No. 5. Thematic modeling using the gensim library.

Practical work No. 6. Sentiment analysis of text data. Deploying the trained model on the web.

Practical work No. 7. Building a language model, generating a text.

Practical work No. 8. Generation of a caption for an image.

The final grade for the subject is set based on the results of the verification of practical work:

"passed" - the student completed all practical work, answered all questions on practical work;

"failed" - the student did not pass any practical work, did not answer questions on practical work.

During the test, the student can increase his grade by retaking the relevant practical work.

11. Educational and methodological support

a) Electronic training course on the discipline at the electronic university "Moodle" - https://moodle.tsu.ru/course/view.php?id=22124

b) Assessment materials of the ongoing evaluation and intermediate certification in the discipline.

12. List of educational literature and Internet resources

a) main literature:

− Automatic processing of texts in natural language and computational linguistics: textbook. allowance / Bolshakova E.I., Klyshinsky E.S., Lande D.V., Noskov A.A., Peskova O.V., Yagunova E.V. — M.: MIEM, 2011. — 272 p.

‑ Automatic processing of texts in natural language and data analysis: textbook. allowance / Bolshakova E.I., Vorontsov K.V., Efremova N.E., Klyshinsky E.S., Lukashevich N.V., Sapin A.S. - M.: Publishing House of the National Research University Higher School of Economics, 2017. - 269 p.

‑ Introduction to cognitive linguistics: a textbook. Ed. 2nd, revised. - Kaliningrad: BFU Publishing House. I. Kant, 2012. - 313 p.

‑ Nikolenko S., Kadurin A., Arkhangelskaya E. Deep learning. - St. Petersburg: Peter, 2018. - 480 p.: ill. - (Series "Programmer's Library").

‑ Hobson Lane, Hannes Hapke, Cole Howard Natural Language Processing in Action. - St. Petersburg: Peter, 2020. - 576 p.: ill. - (Series "For professionals").

‑ Li Deng Yang Liu Deep Learning in Natural Language Processing. ISBN 978-981-10-5209-5 https://doi.org/10.1007/978-981-10-5209-5

‑ Nikolaev I.S., Mitrenina O.V., Lando T.M. Applied and COMPUTER LINGUISTICS. URSS. 2017. 320 p. ISBN 978-5-9710-4633-2.

‑ Ian Goodfellow, Joshua Bengio, Aaron Courville. Deep learning. Second color edition, revised. M.: DMK Press, 2018. - 652 p.

⸺ François Chollet. Deep learning in Python. St. Petersburg: Piter, 2018. - 400 p.

‑ Daniel Jurafsky, James H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall, 2008. - 1044 p.

13. List of information technologies
a) licensed and freely distributed software:
‑ Microsoft Visual Studio;
- publicly available cloud technologies (Google Docs, Yandex disk, etc.).

b) information reference systems:
‑ Electronic catalog of the TSU Scientific Library – http://chamo.lib.tsu.ru/search/query?locale=ru&theme=system
‑ TSU electronic library (repository) – http://vital.lib.tsu.ru/vital/access/manager/Index

14. Logistics
Halls for lectures.

Classrooms for seminars, individual and group work, ongoing evaluation and intermediate certification.

Classrooms for independent work, equipped with computer technology and access to the Internet, to the electronic information and educational environment and to information reference systems.

15. Authors information
Pozhidaev Mikhail Sergeevich, Ph.D. tech. Sciences, Associate Professor, Department of Theoretical Foundations of Informatics.