Federal State Autonomous Educational Institution of Higher Education
Peoples' Friendship University of Russia named after Patrice Lumumba

Academy of Engineering

(name of the main educational unit (MEU) that developed the educational program of higher education)

WORKING PROGRAM OF THE DISCIPLINE

INTRODUCTION TO NATURAL LANGUAGE PROCESSING

(name of discipline/module)

Recommended for the field of study/specialty:

27.04.04 CONTROL IN TECHNICAL SYSTEMS

(code and name of the field of study/specialty)

The discipline is mastered within the framework of the implementation of the main professional educational program of higher education (EP HE):

Artificial Intelligence, Machine Learning, and Space Science

(name (profile/specialization) of the educational institution of higher education)

2026

1. THE GOAL OF MASTERING THE DISCIPLINE

The course "Introduction to Natural Language Processing" is part of the Master's program "Artificial Intelligence, Machine Learning, and Space Sciences" in the 27.04.04 "Control in Technical Systems" program and is studied in the second semester of the first year. The course is offered by the department of the partner university. It consists of eight sections and 16 topics and focuses on the fundamental methods and approaches to natural language processing, as well as the principles of assessing the quality of natural language processing methods.

The goal of mastering the course is to become familiar with the basic methods and applications of automatic natural language processing (NLP), and to acquire practical skills in working with NLP tools.

2. REQUIREMENTS FOR THE RESULTS OF MASTERING THE DISCIPLINE

Mastering the course "Introduction to Natural Language Processing" aimed at developing the following competencies (parts of competencies) in students:

Table 2.1. List of competencies developed in students while mastering the discipline (results of mastering the discipline)

Cipher	Competence	Indicators of Competency Achievement (within this discipline)
PC-1	Able to formulate goals and objectives of scientific research in the field of aerospace systems management, and select methods and means for solving professional problems	PC-1.1 Knows the methods and means of solving scientific research problems in the field of artificial intelligence systems and robotic systems; PC-1.2 Able to formulate the goals and objectives of scientific research in the professional field; PC-1.3 Proficient in techniques for formulating the goals and objectives of scientific research, and knows how to select methods and means for solving problems of professional activity;

3. PLACE OF THE DISCIPLINE IN THE STRUCTURE OF THE EDUCATIONAL EDUCATIONAL INSTITUTION

Course "Introduction to Natural Language Processing" refers to the mandatory part of block 1 “Disciplines (modules)” of the educational program of higher education.

As part of the higher education program, students also master other disciplines and/or practices that contribute to the achievement of the planned results of mastering the discipline "Introduction to Natural Language Processing".

Table 3.1. List of components of the educational program of higher education that contribute to the achievement of the planned results of mastering the discipline

Cipher	Name of competence	Previous courses/modules, practical training*	Subsequent disciplines/modules, practices*
PC-1	Able to formulate goals and objectives of scientific research in the field of aerospace systems management, and select methods and means for solving professional problems		*Artificial Neural Networks (Deep Learning)***; *Artificial Neural Networks (Deep Learning)***; *Artificial Neural Networks (Reinforcement Learning)***; Undergraduate practice / Pregraduation practice;

Cipher	Name of competence	Previous courses/modules, practical training*	Subsequent disciplines/modules, practices*

* - filled in accordance with the competency matrix and the SUP EP HE

** - elective courses/practices

4. SCOPE OF THE DISCIPLINE AND TYPES OF EDUCATIONAL WORK

The total workload of the course "Introduction to Natural Language Processing" is 4 credit units.

Table 4.1. Types of educational work by periods of mastering the educational program of higher education for full-time education.

Type of academic work	**TOTAL,**academic hours		**Semester(s)**
			2
Contact work, academic hours	*34*		*34*
Lectures (LC)	17		17
Laboratory work (LW)	17		17
Practical/seminar classes (SC)	0		0
Independent work of students, academic hours	*83*		*83*
Control (exam/test with assessment), academic hours	*27*		*27*
Total complexity of the discipline	**academic hours**	***144***	*144*
	credit	***4***	*4*

5. CONTENT OF THE DISCIPLINE

Table 5.1. Content of the discipline (module) by types of academic work

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
Section 1	Introduction	1.1	Definition, features, tasks, applications, methods.	Natural language processing is defined as an interdisciplinary field at the intersection of linguistics, computer science, and artificial intelligence. Characteristics of natural language processing include ambiguity, context dependence, and exceptions. Key tasks include tokenization, morphological analysis, syntactic parsing, sentiment analysis, information retrieval, machine translation, information extraction, and automated summarization. Application areas include search engines, voice assistants, review analysis, automatic translation, and chatbots. Methods include rules and dictionaries, statistical methods, and machine learning.	LC, LW
		1.2	A brief history of global and Soviet/Russian OEL. Course content and "ideology." Resources.	Stages of development in global natural language processing: the first works on machine translation, the period of using rules and dictionaries, the transition to probabilistic and statistical methods, the era of machine learning. The contribution of the Soviet and Russian schools of computational linguistics. Course content and philosophy: a sequential study of language levels and machine processing methods. Resources for study: textbooks, open text corpora.	LC, LW
Section 2	Morphological analysis	2.1	What does morphology do, and why is machine morphology needed? Stemming, lemmatization, part-of-speech (POS) tagging. Porter's algorithm for English.	Stamping, lemmatization, part-of-speech annotation. Porter's Algorithm for the English language. Morphological tasks: studying the structure of words, their forms, and inflection. The purpose of machine morphology: reducing words to normal form and determining grammatical characteristics. Stamping as a simplified truncation of endings. Lemmatization as a context-sensitive conversion to dictionary form. Part-of-speech annotation as determining the part of speech of a word in a sentence. Porter's Algorithm for the English language: sequential removal of endings according to rules.	LC, LW
		2.2	Lemmatization for Russian: mystem, pymorphy2, AOT. Zaliznyak's dictionary as a data source for lemmatization. How to build hypotheses for unfamiliar words. The task of HR annotation. Hidden Markov models for HR	Zaliznyak's dictionary as a data source for lemmatization. Software tools: mystem, pymorphy2, AOT. Hypothesis generation for unfamiliar words. Part-of-speech tagging task. Hidden Markov models for part-of-speech tagging. Data sources for training models.	LC, LW

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
			annotation. Data sources. The Viterbi algorithm.		
Section 3	Classical scenario of information retrieval, features of information retrieval tasks	3.1	Basic search quality metrics: precision, recall, F1. Morphology in search. Frequency properties of terms in a collection: Heaps' and Zipf's laws. Vector space model.	Morphology in search. Term frequency properties in a collection. Vector space model. Search quality metrics: precision, recall, F1. Using morphology in information retrieval to account for word forms. Heaps' and Zipf's laws describing the distribution of term frequencies. Vector space model for representing documents and queries as vectors.	LC, LW
		3.2	Term weighting: the tf.idf approach. Other NLE tasks in search. Automatic abstracting. Problem statement, application areas, examples. Abstract types. Text document abstracting methods. Abstract post-processing. Evaluation and initiatives for evaluating automatic abstracting methods. Web snippets (query-tailored abstracts): problem features, methods. Web snippet evaluation.	Other tasks of natural language processing in search. Automatic summarization. Problem statement, application areas, examples. Abstract types. Methods for summarizing text documents. Abstract post-processing. Abstract quality assessment. The tf.idf approach for assessing the importance of a term in a document relative to the entire collection. Other tasks: keyword extraction, document classification. Automatic summarization as the creation of a summary of the source text. Abstract types: indicative, informative, overview. Summarization methods: sentence extraction, abstract summarization.	LC, LW
Section 4	Language Models	4.1	Application areas and limitations. N-grams, probability estimation, available large n-gram collections. Language model evaluation, perplexity.	N-grams, probability estimation, available large n-gram collections. Language model evaluation and prospects. Application areas of language models: speech recognition, machine translation, spell checking, text generation. Limitations: difficulty accounting for long dependencies, data volume requirements. N-grams as sequences of n elements. N-grams as sequences of n elements. Probability estimation based on corpus frequencies. Available large n-gram collections. Prospects for the development of language models: neural network models, large language models.	LC, LW
		4.2	Smoothing, backoff, and interpolation. Techniques for working with web-scale language models, language models with "memory." Laplace, Good-Turing, and Kneser-Nei smoothing.	Techniques for working with web-scale language models, language models with memory. Laplace, Good-Turing, and Kneser-Nei smoothing. Smoothing as a method for estimating the probability of sequences missing from the corpus. Backtracking as a transition to shorter n-grams when data is insufficient. Interpolation as a combination of probabilities from n-grams of different lengths. Language models with memory for accounting for long dependencies.	LC, LW
Section 5	Syntactic parsing	5.1	Two formalisms for describing syntax: the	Constituent system: hierarchical grouping of words into phrases	LC, LW

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
			constituent system (constituency) and the dependency tree (dependency). Context-free grammars (CFG): possibilities and limitations. Probabilistic CFG. Probabilistic parsing algorithm.	and sentences. Dependency tree: grammatical relationships between words. Context-free grammars for describing the structure of constituents. Limitations of context-free grammars: inability to account for lexical and contextual preferences. Probabilistic context-free grammars with rule probability estimates. A probabilistic parsing algorithm for finding the most probable tree.	
		5.2	Evaluation of the analysis results. Lexicalization of the VKSG.	Lexicalization of probabilistic context-free grammars. Parser evaluation metrics. Lexicalization as the consideration of specific lexical units in grammar rules.	LC, LW
Section 6	Extracting information from text documents	6.1	Problem characteristics, data sources. Named entities (NEs), relations. Main approaches. Evaluation. Machine translation: a brief history, challenges, approaches, and applications. Rule-based translation (RBMT) and statistical machine translation (SMT).	Named entities, relations. Main approaches. Evaluation. Machine translation: a brief history, challenges, approaches, applications. Information extraction features: working with unstructured text, fact extraction. Data sources: news feeds, scientific publications, social media. Named entities: names of people, names of organizations, geographical names, dates. Relationships between entities. Main approaches to extraction: rule-based and machine learning-based. Information extraction quality assessment. Machine translation: a brief history of development. Translation challenges: polysemy, word order, idioms. Approaches to machine translation. Application areas.	LC, LW
		6.2	Data sources for statistical machine translation. Parallel corpus alignment. IBM Models 1 and 2. Phrase-based machine translation. Evaluation of MT systems: manual and automatic (BLUE)	Parallel corpus alignment. Phrase-based statistical machine translation. Machine translation system evaluation: manual and automatic. Data sources: parallel corpora of texts in two or more languages. Alignment as establishing correspondence between source and target language sentences. Phrase-based statistical machine translation. Machine translation system evaluation: manual evaluation and automatic metrics.	LC, LW
Section 7	Sentiment analysis: data sources and various problem statements, application examples	7.1	A classification-based approach. Complexities of sentiment analysis: the variety of expression forms, irony and sarcasm, the order and relationship of evaluations. Use of dictionaries of sentiment-laden words. Automatic replenishment/creation of dictionaries.	Difficulties in sentiment analysis. Using dictionaries of sentiment-laden words. Automatic dictionary replenishment. Sentiment analysis as a means of determining the emotional tone of a text. Data sources: reviews, social media posts, news comments. Problem statements: binary classification (positive/negative), three-class classification (including neutrality), rating scale. Classification-based approach with learning on labeled examples. Difficulties: variety of emotional expressions, irony and sarcasm, order and relationship of ratings. Using dictionaries of sentiment-laden words	LC, LW

Section number	Name of the discipline section	Topic Title		Topic Contents	Type of academic work*
				with positive and negative meanings. Automatic dictionary replenishment from text corpora.	
		7.2	Handling negations. Phrases vs. individual words. Aspects/attributes: automatic highlighting and manually compiled lists.	Phrases versus individual words. Aspects and attributes: automatic extraction and manual compilation of lists. Processing of negations by inverting sentiment in the presence of negating words. Phrase analysis versus individual word analysis for contextual consideration. Extraction of aspects and attributes of the object of assessment. Automatic extraction of aspects from texts and manual compilation of lists.	LC, LW
Section 8	Semantics, different approaches and definitions	8.1	Semantics, various approaches and definitions: propositional logic, semantic web, knowledge bases, domain ontologies, thesauri. Lexical semantics: homonyms, polysemantic words, synonyms, antonyms, hyponyms/hypernyms. Semantic dictionaries -- thesauri. WordNet: synsets and relations. Thesaurus-based methods for determining semantic similarity of words.	Propositional logic, semantic web, knowledge bases, domain ontologies, thesauri. Lexical semantics: homonyms, polysemantic words, synonyms, antonyms, hyponyms, and hypernyms. Various approaches to semantics. Propositional logic for formalizing meaning. The Semantic Web as a project to impart machine-readable meaning to data. Knowledge bases as structured fact repositories. Domain ontologies as formal descriptions of concepts and their relationships. Thesauri as dictionaries of semantic relations. Lexical semantics studies the meanings of words. Homonyms are spelled the same but have different meanings. Polysemantic words have several related meanings. Synonyms are similar in meaning. Antonyms are opposite in meaning. Hyponyms and hypernyms are related by the species-genus relationship.	LC, LW
		8.2	Distributional semantics: corpus-based semantic proximity. Positive Pointwise Mutual Information (PPMI). Similarity Computations Based on Phrase Structure. Neural Network-Based Vector Representation of Word Semantics.	Positive poetic mutual information. Similarity calculations based on phrase structure. Vector representation of words. Distributive semantics is based on the principle that words with similar meanings occur in similar contexts. Semantic similarity is calculated based on the analysis of large corpora. Positive poetic mutual information as a measure of the relationship between a word and its context. Similarity calculations based on phrase structure. Vector representation of words in a multidimensional feature space.	LC, LW

* - to be completed only for FULL-TIME education: LC – lectures; LW – laboratory work; SC – practical/seminar classes.

6. LOGISTIC AND TECHNICAL SUPPORT OF DISCIPLINE

Table 6.1. Material and technical support for the discipline

Audience type	Equipment of the auditorium	Specialized educational/laboratory equipment, software and materials for mastering the discipline (if necessary)
Lecture	A lecture hall equipped with specialized furniture, a whiteboard (screen), and multimedia presentation equipment.	
Computer class	A computer room for conducting classes, group and individual consultations, ongoing monitoring and midterm assessment, equipped with personal computers (in the amount of ____ units), a board (screen) and technical means for multimedia presentations.	
For independent work	A classroom for independent student work (can be used for seminars and consultations), equipped with a set of specialized furniture and computers with access to the Electronic Information System.	

* - the classroom for independent work of students MUST be indicated!

7. EDUCATIONAL, METHODOLOGICAL AND INFORMATIONAL SUPPORT OF THE DISCIPLINE

Main literature:

1. Kang Y. et al. Natural language processing (NLP) in management research: A literature review //Journal of Management Analytics. – 2020. –T. 7. – No. 2. – pp. 139-172.

2. Vajjala S. et al. Practical natural language processing: a comprehensive guide to building real-world NLP systems. – O'Reilly Media, 2020.

Further reading:

1. Cambria E., White B. Jumping NLP curves: A review of natural language processing research //IEEE Computational intelligence magazine. – 2014. –T. 9. – No. 2. – pp. 48-57.

2. Mihalcea R., Liu H., Lieberman H. NLP (natural language processing) for NLP (natural language programming) //International Conference on intelligent text processing and computational linguistics. – Berlin, Heidelberg: Springer Berlin Heidelberg, 2006. –P. 319-330.

Resources of the information and telecommunications network "Internet":

1. RUDN University Electronic Library System and third-party electronic library systems to which university students have access based on concluded agreements

- RUDN University Electronic Library System – RUDN University Electronic Library System https://mega.rudn.ru/MegaPro/Web
- Electronic Library System "University Library Online" http://www.biblioclub.ru
- EBS "Urayt" http://www.biblio-online.ru
- Electronic Library System "Student Consultant" www.studentlibrary.ru
- EBS “Knowledge” https://znanium.ru/

2. Databases and search engines

- Sage https://journals.sagepub.com/

- Springer Nature Link https://link.springer.com/
- Wiley Journal Database https://onlinelibrary.wiley.com/
- Scientometric database Lens.org https://www.lens.org

Educational and methodological materials for independent work of students in mastering a discipline/module:*

1. Lecture course on the subject "Introduction to Natural Language Processing".

* - all teaching and methodological materials for independent work of students are posted in accordance with the current procedure on the discipline page in TUIS!

DEVELOPER:

Associate Professor		Saltykova Olga Alexandrovna
Position, DEPARTMENT	*Signature*	*Surname I.O.*

HEAD OF THE DEPARTMENT:

Position of the DEPARTMENT	*Signature*	*Surname I.O.*

HEAD OF THE EP HE:

Professor		Razumny Yuri Nikolaevich
Position, DEPARTMENT	*Signature*	*Surname I.O.*