**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 training area/specialty)

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

AIML and Space Sciences / Artificial Intelligence, Machine Learning and Space Sciences

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

2025

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 direction 27.04.04 "Control in Technical Systems" and is studied in the 2nd semester of the 1st year. The course is implemented by the Department of the Partner University. The course consists of 8 sections and 16 topics and is aimed at studying the main methods and approaches to NLP, principles of assessing the quality of NLP methods.

The goal of mastering the discipline 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 TO THE RESULTS OF MASTERING THE DISCIPLINE

Mastering the course "Introduction to Natural Language Processing" is 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 Competence Achievement (within the framework of this discipline)
PC-1	Able to formulate goals and objectives of scientific research in the field of aerospace systems management, 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 EDUCATION

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, practices*	Subsequent disciplines/modules, practices*
PC-1	Able to formulate goals and objectives of scientific research in the field of aerospace systems management, 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 / Pre-graduation practice;

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

* - filled in in accordance with the competency matrix and the SUP EP HE
** - elective disciplines/practices

4. SCOPE OF THE DISCIPLINE AND TYPES OF STUDY WORK

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

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,ac.h.		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*
General complexity of the discipline	**ac.h.**	*144*	*144*
	credit.ed.	*4*	*4*

5. CONTENT OF THE DISCIPLINE

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

Section number	Name of the discipline section		Section Contents (Topics)	Type of academic work*
Section 1	Introduction	1.1	Variants of the discipline name, definition, features, objectives, applications, methods.	LC, LW
		1.2	Brief history of world and Soviet/Russian OEL. Contents and "ideology" of the course. Resources.	LC, LW
Section 2	Morphological analysis	2.1	What does morphology do, what is machine morphology for? Stemming, lemmatization, part-of-speech (POS) tagging. Porter's algorithm for English.	LC, LW
		2.2	Lemmatization for Russian: mystem, pymorphy2, AOT. Zaliznyak's dictionary as a source of data for lemmatization. How to build hypotheses for unfamiliar words. The task of HR-labeling. Hidden Markov models for HR-labeling. Data sources. Viterbi algorithm.	LC, LW
Section 3	Classic 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.	LC, LW
		3.2	Term weighting: the tf.idf approach. Other OEL tasks in search. Automatic abstracting. Problem statement, areas of application, examples. Abstract types. Text document abstracting methods. Abstract post-processing. Evaluation, initiatives for evaluating automatic abstracting methods. Web snippets (query-based abstracts): problem features, methods. Web snippet evaluation.	LC, LW
Section 4	Language Models	4.1	Applications, limitations. N-grams, probability estimation, available large n-gram collections. Evaluation of language models, perplexity.	LC, LW
		4.2	Smoothing, backoff, and interpolation. Techniques for working with web-scale language models, language models with "memory". Laplace, Good-Turing, Kneser-Ney smoothing.	LC, LW
Section 5	Syntactic parsing	5.1	Two formalisms for describing syntax: a system of constituents (constituency) and a tree of dependencies (dependency). Context-free grammars (CFG): possibilities and limitations. Probabilistic CFG. Algorithm for probabilistic syntactic analysis.	LC, LW
		5.2	Evaluation of the analysis results. Lexicalization of VKSG.	LC, LW
Section 6	Extracting information from text documents	6.1	Problem features, data sources. Named entities (NEs), relations. Main approaches. Evaluation. Machine translation: brief history, challenges, approaches, applications. Rule-based translation (RBMT) and statistical machine translation (STM).	LC, LW
		6.2	Data sources for statistical machine translation. Parallel corpus alignment. IBM Models 1, 2. Phrase MT. Evaluation of MT systems: manual, automatic (BLUE)	LC, LW
Section 7	Sentiment analysis: data sources and various problem statements, application examples	7.1	Classification-based approach. Difficulties of sentiment analysis: variety of expression forms, irony and sarcasm, order and connection of evaluations. Use of dictionaries of sentimentally	LC, LW

Section number	Name of the discipline section		Section Contents (Topics)	Type of academic work*
			colored words. Automatic replenishment/creation of dictionaries.	
		7.2	Handling negations. Phrases vs. single words. Aspects/attributes: automatic highlighting and manually compiled lists.	LC, LW
Section 8	Semantics, different approaches and definitions	8.1	Semantics, different 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. Methods for determining semantic closeness of words based on a thesaurus.	LC, LW
		8.2	Distributional semantics: corpus-based semantic proximity. Positive Pointwise Mutual Information (PPMI). Proximity Computations Based on Phrase Structure. Vector Representation of Word Semantics Based on Neural Networks	LC, LW

* - filled in 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 of the discipline

Audience type	Equipping the auditorium	Specialized educational/laboratory equipment, software and materials for mastering the discipline (if necessary)
Lecture	An auditorium for conducting lecture-type classes, equipped with a set of specialized furniture; a board (screen) and technical means for multimedia presentations.	
Computer class	A computer room for conducting classes, group and individual consultations, ongoing monitoring and midterm assessment, equipped with personal computers (14 in total), a board (screen) and technical means for multimedia presentations.	
For independent work	A classroom for independent work of students (can be used for conducting seminars and consultations), equipped with a set of specialized furniture and computers with access to the Electronic Information System.	

* - the audience 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. – pp. 319-330.

Resources of the information and telecommunications network "Internet":

1. RUDN University EBS and third-party EBSs to which university students have access on the basis of concluded agreements

    - Electronic library system of RUDN - ELS RUDN https://mega.rudn.ru/MegaPro/Web

    - Electronic library system "University library online"http://www.biblioclub.ru

    - EBS "Yurait"http://www.biblio-online.ru

    - Electronic Library System "Student Consultant" www.studentlibrary.ru

    - EBS "Znanium"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 educational 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:

Head of Department		Razumny Yuri Nikolaevich
Position, Department	*Signature*	*Surname I.O.*