

MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE  
ODESSA I.I. Mechnikov NATIONAL UNIVERSITY  
Department of mathematical support of computer systems



Vice-rector for scientific and pedagogical work

20\_\_

**WORKING PROGRAM OF EDUCATIONAL COURSE**

***БЕ6 "Methods of natural language text processing"***

(name of academic discipline)

Level of higher education Second (master's)

Field of knowledge 12 – Information technologies

Specialty 126 – Information systems and technologies  
(code and name of specialty(s))

Educational and professional program Information systems and technologies  
(name of OPP/ONP)

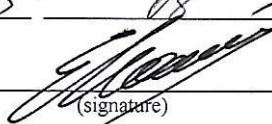
The working program of the academic course "Methods of natural language text processing"— Odesa: ONU, 2022. – 8 p.

Developer:

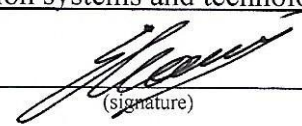
Penko V.G., Ph.D. (Tech.), associate professor of the Department of MSCS

The work program was approved at the meeting of the Department of Mathematical Support of Computer Systems

Protocol No. 1 from " 25 " 08 2022 year


Head of the department  ( Eugene MALAKHOV )  
(signature) (First Name Surname)

Agreed with the guarantor of the EPP "Information systems and technologies"

 ( Eugene MALAKHOV )  
(signature) (First Name Surname)

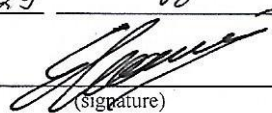
Approved by the educational and methodical commission (EMC) for IT specialties of the FMPhIT

Protocol No. 1 from " 31 " 08 2022 year

Head of EMC  ( Alla RACHYNSKA )  
(signature) (First Name Surname)

Reviewed and approved at the meeting of the department \_\_\_\_\_

Protocol No. 1 from " 29 " 08 2023 year

Head of Department  ( \_\_\_\_\_ )  
(signature) (First Name Surname)

Reviewed and approved at the meeting of the department \_\_\_\_\_

Protocol No. \_\_\_\_ from " \_\_\_\_ " \_\_\_\_\_ 20\_\_ year

Head of Department \_\_\_\_\_ ( \_\_\_\_\_ )  
(signature) (First Name Surname)

## 1. Course description

Name of indicators	Field of knowledge, direction of training, educational and qualification level	Characteristics of the academic discipline	
		<i>full-time education</i>	<i>external form of education</i>
The total number of:  credits - 4  hours - 120  content modules - 2	Branch of knowledge <u>12 - Information technologies</u> (code and name)	<i>Mandatory</i>	
	Specialty <u>126 – Information systems and technologies</u>	<b><i>Year of preparation:</i></b>	
		1st	
		<b><i>Semester</i></b>	
		2nd	
	Level of higher education: <u>Second (master's)</u>	<b><i>Lectures</i></b>	
		16 hours	6 hours
		<b><i>Practical, seminar</i></b>	
		<b><i>Laboratory</i></b>	
		18 hours	6 hours
		<b><i>Independent work</i></b>	
		86 hours	108 hours
		<b><i>Individual tasks:</i></b>	
		Final control form: exam	

\* - in the presence

## 2. The purpose and tasks of the educational course

**The purpose** of the course is a study of the main modern approaches to solving basic tasks of text processing in natural language and the practical application of the Python language and libraries to solve these tasks.

### **Task:**

- mastering the skills of using the Python language for the development of general-purpose programs;
- familiarity with the problems of processing texts in natural language;
- application of the capabilities of specialized packages to increase the efficiency of natural language processing.

The process of studying the discipline is aimed at forming elements of the following competencies (according to the OPP "Information Systems and Technologies" from 2019):

1) general: -

2) special (professional):

*SC04. The ability to develop mathematical, informational, and computer models of objects and processes related to informatization.*

*SC05. The ability to utilize modern data analysis technologies for optimizing processes in information systems.*

*SKM03. The ability to mathematically model digital data and apply efficient algorithms for the analysis and transformation of multimedia data in modern information systems.*

*SKM07. The ability to conduct information analysis and create multi-dimensional models of subject areas.*

### **Program learning outcomes:**

*LOO9. Develop and use data repositories, and perform data analysis to support decision-making.*

*LOOM5. Present research results, conduct discussions and publish research findings.*

*LOOM6. Develop mathematical models and software-information systems to solve current problems of multimedia information analysis and processing.*

*LOOM8. Create optimized pipelines for data preparation for subsequent storage and processing.*

**Expected learning outcomes.**As a result of studying the course, student should

**know:** *basic opportunities of Python language for developing text processing software; the main capabilities of specialized packages for processing texts in natural language; the main types of tasks related to the processing of texts in natural language; features of the corpus-oriented approach to the processing of natural language texts.*

**be able:** *develop software provision that performs basic operations with texts; use specialized Python packages to improve the efficiency of basic text processing tasks in natural language; apply several varieties of language corpora as a resource for solving text processing tasks in natural language.*

### 3. Content of the academic discipline

**Content module 1** Text processing in Python.

**Tema 1.** Computing over language is simple statistics.

Literature: [1, 2, 7].

**Tema 2.** Classification of natural language text processing tasks.

References: [4, 7].

**Tema 3.** Obtaining access to corpora of texts and lexical resources.

Literature: [1, 3].

**Tema 4.** Use of basic lexical resources.

Literature: [1, 2, 6, 8].

**Content module 2** Learning and using the main techniques of corpus linguistics.

**Tema 1.** WordNet Corpus.

References: [6, 10].

**Tema 2.** Access to text from the Web and to local text.

Literature: [1, 6].

**Tema 3.** NLP Pipeline. Implementation of individual stages.

Literature: [1, 4, 5, 6].

### 4. The structure of the academic discipline

Names of content modules and topics	Number of hours									
	Full-time					Correspondence form				
	That's all	including				That's all	Including			
		1	p	lab	W ed		1	p	lab	Wed
1	2	3	4	5	6	7	8	9	10	11
Content module 1. Text processing in Python..										
Topic 1.	14	2		2	10		3		3	13
Topic 2.	14	2		2	10					13
Topic 3.	14	2		2	10					13
Topic 4.	14	2		2	12					15
Content module 2. Learning and using the main techniques of corpus linguistics.										
Topic 1.	16	2		2	12		3		3	18
Topic 2.	18	2		4	12					18
Topic 3.	28	4		4	20					18
Hours in general	120	16		18	86	120	6		6	108

### 5. Topics of seminar classes

Seminar classes are not provided

### 6. Topics of practical classes

Practical classes are not provided

## 7. Topics of laboratory classes

No s/p	Topic name	Number Hours
1	Using Python programming environments to implement basic text processing tasks.	2
2	Installing and using the NLTK package to improve the performance of natural language text processing.	4
3	Conducting software experiments with text corpora embedded in NLTK.	4
4	Conducting software experiments on creating corpora based on local and Web-texts.	4
5	Implementation of individual stages of the NLP pipeline.	4
	<b>Total</b>	<b>18</b>

## 8. Independent work

No s/p	Topic name	Number Hours
1	Mastering additional features of the Python language.	14
2	Mastering additional features of the NLTK package.	14
3	Demonstration of applied aspects of using text corpora.	24
4	Development of own software implementations of individual stages of the NLP pipeline.	24
	<b>Total</b>	<b>86</b>

Independent work includes:

[1] – preparation for lectures and laboratory classes;

**8.1. Individual educational and research task** (course project or calculation and graphic work) is not provided

## 9. Teaching methods

Lectures using multimedia presentation material.

## 10. Control methods

During the final control, the student must answer 2 questions of the examiner from the list given in clause 11.1.

### 10.1. Evaluation criteria at the final inspection:

The examination ticket for the discipline consists of two parts: theoretical and practical. The minimum number of points counted as a positive result is 60 (on a 100-point scale). Points are distributed as follows: 60 points - theoretical part and 40 points - practical.

The theoretical part contains 2 questions, the practical part - 1 question.

For an impeccable answer to each theoretical question, the student receives - 30 points. At the same time, the answer is considered flawless if the student fully disclosed the essence of the question, presented it consistently and logically, gave examples,

illustrated the answer with the necessary and sufficient number of records, graphs, formulas, schemes; made references to relevant literary sources.

For perfect performance of the task of the practical part, the student receives - 40 points. The task of the practical part of the exam is considered flawlessly completed if the correct answer is obtained, the solution is presented consistently and logically, and all the results formulated in the task are obtained.

### 11. Questions for the final control

1. Comparative analysis of different language programming environments Python.
2. How can you automatically determine keywords and phrases that characterize the style and content of the text?
3. Consider some interesting problems in natural text processing.
4. What is the role and components of the NLTK package?
5. Software methods of working with corpora (Class Text), included in the NLTK package.
6. Functions and methods for working with strings in Python.
7. Use of frequency and conditional frequency distribution.
8. Collocations and bigrams - definition, role in text processing and methods of use in NLTK.
9. A simple sequential architecture of a conversational dialogue system.
10. The Gutenberg corpus - structure and methods of accessing information.
11. Webtext corpus - structure and methods of accessing information.
12. Brown corpus - structure and ways of accessing information.
13. Reuters corpus - structure and methods of access to information.
14. Corpus of inaugural addresses - structure and methods of access to information.
15. Brief characteristics of other cases.
16. Peculiarities of working with corpora in different languages.
17. Classification of corpus texts.
18. Loading own case.
19. Corpus WordList and Comparative Wordlists - structure and methods of accessing information.
20. CMU Pronouncing Dictionary - structure and methods of accessing information.
21. The structure of the WordNet corpus.
22. Ways of using the WordNet corpus.

### 12. Distribution of points received by students

Current testing and independent work							Exam	Sum
Content module No.1			Content module No. 2					
T1	T2	T3	T4	T1	T2	T3	25	100
8	10	10	12	10	10	15		

T1, T2 ... - topics of content modules.

## Evaluation scale: national and ECTS

Total points	ECTS assessment	National scale	
90 — 100	A - "excellent"	5 "excellent"	"test"
85 - 89	B - "very good"	4 "good"	
75 - 84	C - "good"		
70 - 74	D - "satisfactory"	3 "satisfactory"	
60 - 69	E - "permissible"		
35 — 59	F - "unsatisfactory with the possibility of reassembly"	2 "unsatisfactory"	"uncountable"
0 — 34	FX – "unsatisfactory with mandatory repeat course"		

### 13. Educational and methodical support

Synopsis of lectures in electronic format.

### 14. Recommended Books

#### 14.1. Basic literature

1. Bird S. Natural Language Processing with Python / S. Bird, E. Klein, E. Loper. - O'Reilly, 2009. - 514p.
2. Hobson Lane, Howard Cole, Hannes Max Hapke Natural Language Processing in Action: Understanding, analyzing, and generating text with Python: Manning Shelter Island, 2019. - 576p.
3. A. Vasiliev Programming in Python Learning book – Bohdan, 2019. - 514 p.
4. N. Indurkha, FJDamerau Handbook of Natural Language Processing: A Chapman & Hall Book/CRC Machine Learning & Pattern Recognition Series, 2010. - 676p.

#### 14.2. Auxiliary literature

5. CD Manning, H. Schutze Foundations of Statistical Natural Language Processing The MIT Press Cambridge, Massachusetts London, England, 1999. - 717 p.
6. D. Jurafsky, JH Martin Speech and Language Processing - Prentice Hall, Englewood Cliffs, New Jersey 2008. - 975 p.

### 15. Electronic information resources

7. Natural Language Toolkit - Access Mode: <https://www.nltk.org/>
8. Natural Language Processing with Python - Access mode: <https://www.nltk.org/book/>
9. Project Gutenberg - Access mode: <https://www.gutenberg.org/>
10. WordNet: A Lexical Database for English - Access mode: <https://wordnet.princeton.edu/>