МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ОДЕСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМЕНІ І.І. МЕЧНИКОВА
Кафедра математичного забезпечення комп'ютерних систем

"ЗАТВЕРДЖУЮ"
Проректор з науково-педагогічної роботи
_____
"30" 08 2024 р.

РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

ВЛ02 «Методи обробки текстів природної мови»
(назва навчальної дисципліни)

Рівень вищої освіти *Перший (бакалаврський)*

Галузь знань *12 – Інформаційні технології*

Спеціальність *123 – Комп'ютерна інженерія*
(код і назва спеціальності (тей)

Освітньо-професійна програма *Комп'ютерна інженерія*
(назва ОПП)

ОНУ – 2024

Робоча програма навчальної дисципліни
«Методи обробки текстів природної мови»: – Одеса: ОНУ, 2024. – 104с.

Розробник:
Пенко В.Г., к.т.н., доцент кафедри МЗКС

Робоча програма затверджена на засіданні кафедри математичного забезпечення комп'ютерних систем

Протокол № 1 _ від. "_28___" __серпня_____2024 р.

Завідувач кафедри _____ (_Євгеній МАЛАХОВ _)
(підпис) (Ім'я ПРІЗВИЩЕ)

Погоджено із гарантом ОПП _«Комп'ютерна інженерія»_

_____ (_Людмила ВОЛОЩУК_)
(підпис) (Ім'я ПРІЗВИЩЕ)

Схвалено навчально-методичною комісією (НМК) _з ІТ факультету МФІТ_

Протокол №1 від. "30 "серпня 2024 р.

Голова НМК _____ (_Лариса МАРТИНОВИЧ_)
(підпис) (Ім'я ПРІЗВИЩЕ)

Переглянуто та затверджено на засіданні кафедри _____

Протокол № ___ від. "____" _____20__ р.

Завідувач кафедри _____ (_____)
(підпис) (Ім'я ПРІЗВИЩЕ)

Переглянуто та затверджено на засіданні кафедри _____

Протокол № ___ від. "____" _____20__ р.

Завідувач кафедри _____ (_____)
(підпис) (Ім'я ПРІЗВИЩЕ)

1. **Опис навчальної дисципліни**

<table>
<tr><td rowspan="2">Найменування показників</td><td rowspan="2">Галузь знань, напрям підготовки, освітньо-кваліфікаційний рівень</td><td colspan="2">Характеристика навчальної дисципліни</td></tr>
<tr><td>денна форма навчання</td><td>заочна форма навчання</td></tr>
<tr><td rowspan="15">Загальна кількість:

кредитів – 3

годин – 90

змістових модулів – 2

ІНДЗ* – _____
(вид завдання)</td><td>Галузь знань
<u>12 – Інформаційні технології</u>
(шифр і назва)</td><td colspan="2">Обов’язкова</td></tr>
<tr><td rowspan="4">Спеціальність
<u>123 – Комп’ютерна інженерія</u>
(шифр і назва)</td><td colspan="2">Рік підготовки:</td></tr>
<tr><td>3</td><td>4</td></tr>
<tr><td colspan="2">Семестр</td></tr>
<tr><td>2 (6)</td><td>1 (7)</td></tr>
<tr><td rowspan="11">Рівень вищої освіти:
<u>Перший (бакалаврський)</u></td><td colspan="2">Лекції</td></tr>
<tr><td>18 год.</td><td>6 год</td></tr>
<tr><td colspan="2">Практичні, семінарські</td></tr>
<tr><td></td><td></td></tr>
<tr><td colspan="2">Лабораторні</td></tr>
<tr><td>16 год.</td><td>6 год</td></tr>
<tr><td colspan="2">Самостійна робота</td></tr>
<tr><td>56 год.</td><td>78 год</td></tr>
<tr><td colspan="2">Індивідуальні завдання:</td></tr>
<tr><td colspan="2">Вид контролю: залік</td></tr>
</table>

* – за наявності

2. Мета та завдання навчальної дисципліни

Метою курсу є вивчення практичних засобів застосування мови та бібліотек Python для вирішення базових завдання обробки текстів на природній мові.

Завдання:

– освоєння навичок застосування мови Python для розробки програм загального призначення;

– знайомство з проблематикою обробки текстів на природній мові;

– застосування можливостей спеціалізованих пакетів для підвищення ефективності обробки природної мови.

Процес вивчення дисципліни спрямований на формування елементів наступних **компетентностей** (згідно ОПП «Інформаційні системи та технології» від **2024** р.):

а) загальних

Z2. Здатність вчитися і оволодівати сучасними знаннями.

Z7. Вміння виявляти, ставити та вирішувати проблеми.

Z12. Здатність застосовувати базові знання з фундаментальної та прикладної математики в професійній діяльності.

б) спеціальних (фахових):

P2. Здатність використовувати сучасні методи і мови програмування для розроблення алгоритмічного та програмного забезпечення.

P17.Здатність застосовувати закономірності випадкових явищ, ймовірнісно-статистичні методи, основи теорії чисельних методів та сучасні методи дискретної математики для аналізу і синтезу складних систем, методи кількісної оцінки інформації і створення коригуючих кодів при розв'язанні прикладних і наукових завдань в області комп'ютерної інженерії.

P19. Здатність використовувати декларативну парадигму програмування та мови, підходи, методи і технології штучного інтелекту, технології інженерії знань, інструментальні засоби підтримки інтелектуальних систем, розробляти та застосовувати моделі представлення знань, стратегії логічного виведення.

Програмні результати навчання:

N2. Мати навички проведення експериментів, збирання даних та моделювання в комп'ютерних системах.

N3. Знати новітні технології в галузі комп'ютерної інженерії.

NM3. Вміти застосовувати закономірності випадкових явищ, ймовірнісно-статистичні методи, основи теорії чисельних методів та сучасні методи дискретної математики для аналізу і синтезу складних систем, методи кількісної оцінки інформації і створення коригуючих кодів при розв'язанні прикладних і наукових завдань в області комп'ютерної інженерії.
N7. Вміти розв'язувати задачі аналізу та синтезу засобів, характерних для спеціальності.
NM4 Розробляти та застосовувати моделі представлення знань, стратегії логічного виведення, технологій інженерії знань, технологій і інструментальних засобів побудови інтелектуальних систем і систем штучного інтелекту.

Очікувані результати навчання. У результаті вивчення навчальної дисципліни студент повинен
знати*: основні причини, що ускладнюють розробку програмного забезпечення, основні моделі життєвого циклу програмного забезпечення, принципи об'єктно-орієнтованого проектування програмного забезпечення.*
вміти*: застосовувати патерни об'єктно-орієнтованого проектування при розробці програм; здійснювати розробку у відповідності з методикою розробки через тестування; здійснювати рефакторинг програмного коду; використовувати системи контролю версій програмного продукту.*

3. Зміст навчальної дисципліни

Змістовний модуль 1 Обробка текстів в Python.
Тема 1. Обчислення над мовою - проста статистика.
Література: [1, 2, 6].
Тема 2. Класифікація завдань обробки текстів природної мови.
Література: [4, 6].
Тема 3. Отримання доступу до корпусів текстів і лексичних ресурсів.
Література: [1, 3].
Тема 4. Використання основних лексичних ресурсів.
Література: [1, 2, 6, 8].

Змістовний модуль 2 Вивчення та використання основних прийомів корпусної лінгвістики.

Тема 1. Корпус WordNet.
Література: [6, 10].
Тема 2. Доступ до тексту з Web і до локального тексту.
Література: [1, 6].
Тема 3. Трубопровід NLP. Реалізація окремих етапів.
Література: [1, 4, 7, 9].

4. Структура навчальної дисципліни

Назви змістових модулів і тем	Кількість годин									
	Денна форма					Заочна форма				
	Усього	у тому числі				Усього	у тому числі			
		л	п	лаб	ср		л	п	лаб	ср
1	2	3	4	5	6	7	8	9	10	11
Змістовний модуль 1. Обробка текстів в Python..										
Тема 1.	12	2		2	8	11.5	0.5		1.0	10
Тема 2	12	2		2	8	11.5	0.5		1.0	10
Тема 3.	12	2		2	8	12.0	1.0		1.0	10
Тема 4.	12	2		2	8	12.0	1.0		1.0	10
Змістовний модуль 2. Вивчення та використання основних прийомів корпусної лінгвістики.										
Тема 1.	12	2		2	8	11.5	1.0		0.5	10
Тема 2.	14	4		2	8	11.5	1.0		0.5	10
Тема 3.	16	4		4	8	20.0	1.0		1.0	18
Всього годин	**90**	18		**16**	**56**	**90**	**6**		**6**	**78**

Форма контролю: **КО** –контрольне опитування (поточне)
КР – контрольна робота

5. Теми семінарських занять

Семінарські заняття не передбачені

6. Теми практичних занять

Практичні заняття не передбачені

7. Теми лабораторних занять

№ з/п	Назва теми	Кількість годин	
		денне	заочне
1	Використання середовищ програмування Python для реалізації базових завдань обробки текстів.	2	0,5
2	Встановлення та використання пакету NLTK для підвищення ефективності обробки текстів на природній мові.	3	0,5
3	Проведення програмних експериментів з корпусами текстів, що вбудовані в NLTK.	4	2

4	Проведення програмних експериментів зі створення корпусів на основі локальних та Web-текстів.	3	1
5	Реалізація окремих етапів трубопроводу NLP.	4	2
	Разом	**16**	**6**

8. Самостійна робота

№ з/п	Назва теми	Кількість годин	
		денне	заочне
1	Опанування додаткових можливостей мови Python.	14	15
2	Опанування додаткових можливостей пакету NLTK.	14	15
3	Демонстрація прикладних аспектів використання корпусів текстів.	14	20
4	Розробка власних програмних реалізацій окремих етапів трубопроводу NLP.	14	28
	Разом	**56**	**78**

До самостійної роботи відноситься:
[1] – підготовка до лекцій та лабораторних занять;

8.1. Індивідуальне навчально-дослідне завдання

Курсовий проект або розрахунково-графічна робота не передбачено.

9. Методи навчання

Лекції з використанням мультимедійного презентаційного матеріалу.

10. Методи контролю

Підсумковий контроль відбувається шляхом урахування балів, які були накопичені на протязі поточного виконання теоретичних та лабораторних завдань по темам. Кількість балів по кожній темі може бути збільшено за рахунок відповідей на 2 запитання з переліку, наведеному у п. 11.1.

10.1. Критерії оцінювання на підсумковому контролі:

1. Відповідь повинна бути повною і короткою. Вона не повинна мати в собі матеріал, що не відноситься до сутті питання.
2. Чітко формулювати твердження, вправно застосовувати необхідні формули і знання основних питань програми.
3. Відповіді, що мають помилкові твердження оцінюються виходячи з близькості відповіді до правильної.
4. Пропуски в обґрунтуванні тверджень враховуються і це призводить до зменшення кількості балів.
5. Малі недоліки, неточності при викладенні матеріалу, зменшують кількість балів.

6. Незнання і нерозуміння основної ідеї теоретичного питання або задачі призводить до зняття до 90 % балів.
7. Якщо відповідь на питання відсутня то виставляється нуль балів.

8. Питання до підсумкового контролю

1. Порівняльний аналіз різних середовищ програмування на мові Python.
2. Як можна автоматично визначати ключові слова і фрази, які характеризують стиль і зміст тексту?
3. Розгляньте деякі цікаві проблем в обробці природних текстів.
4. Яка роль і складові частини пакета NLTK?
5. Програмні методи роботи з корпусами (клас Text), що входять в пакет NLTK.
6. Функції та методі работі з рядками в Python.
7. Використання частотного і умовного частотного розподілу.
8. Колокації і біграми - визначення, роль в обробці текстів і способи використання в NLTK.
9. Проста послідовна архітектура розмовної діалогової системи.
10. Корпус Gutenberg - структура і способи доступу до інформації.
11. Корпус Webtext - структура і способи доступу до інформації.
12. Корпус Brown - структура і способи доступу до інформації.
13. Корпус Reuters - структура і способи доступу до інформації.
14. Корпус інаугураційних звернень - структура і способи доступу до інформації.
15. Коротка характеристика інших корпусів.
16. Особливості роботи з корпусами на різних мовах.
17. Класифікація корпусів текстів.
18. Завантаження власного корпусу.
19. Корпуси WordList і Comparative Wordlists - структура і способи доступу до інформації.
20. Словник вимов CMU Pronouncing Dictionary - структура і способи доступу до інформації.
21. Структура корпуса WordNet.
22. Способи використання корпусу WordNet.

9. Розподіл балів, які отримують студенти

Поточне тестування та самостійна робота							Сума
Змістовий модуль №1				Змістовий модуль № 2			
Т1	Т2	Т3	Т4	Т1	Т2	Т3	
15	15	15	15	15	15	10	100

Т1, Т2 ... – теми змістових модулів.

Шкала оцінювання: національна та ECTS

<table>
<tr><td>Загальна сума балів</td><td>Оцінка ECTS</td><td colspan="2">Національна шкала</td></tr>
<tr><td>90 — 100</td><td>A – «відмінно»</td><td>5 «відмінно»</td><td rowspan="5">«зараховано»</td></tr>
<tr><td>85 — 89</td><td>B – «дуже добре»</td><td rowspan="2">4 «добре»</td></tr>
<tr><td>75 — 84</td><td>C – «добре»</td></tr>
<tr><td>70 — 74</td><td>D – «задовільно»</td><td rowspan="2">3 «задовільно»</td></tr>
<tr><td>60 — 69</td><td>E – «допустимо»</td></tr>
<tr><td>35 — 59</td><td>F – «незадовільно з можливістю повторного складання»</td><td rowspan="2">2 «незадовільно»</td><td rowspan="2">«не зараховано»</td></tr>
<tr><td>0 — 34</td><td>FX – «незадовільно з обов’язковим повторним курсом»</td></tr>
</table>

10. Навчально-методичне забезпечення

Конспект лекцій в електронному форматі.

11. Рекомендована література

11.1. Основна література

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. А. Васильев Програмування мовою Python Навчальна книга – Богдан, 2019. - 514 с.
4. N. Indurkhya, F.J.Damerau Handbook of Natural Language Processing: A Chapman & Hall Book/CRC Machine Learning & Pattern Recognition Series , 2010. - 676 p.

11.2. Допоміжна література

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

12. Електронні інформаційні ресурси

7. Natural Language Toolkit – Режим доступу: https://www.nltk.org/
8. Natural Language Processing with Python – Режим доступу: https://www.nltk.org/book/
9. Project Gutenberg – Режим доступу: https://www.gutenberg.org/
10. WordNet:A Lexical Database for English – Режим доступу: https://wordnet.princeton.edu/