МІЖНАРОДНИЙ НАУКОВО-НАВЧАЛЬНИЙ ЦЕНТР ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА СИСТЕМ НАН УКРАЇНИ ТА МОН УКРАЇНИ

ЗАТВЕРДЖЕНО
Директор Міжнародного науково-навчального центру інформаційних технологій та систем НАН та МОН України
_____
Олександр ВОЛКОВ
«23» ___квітня___ 2024 р.

РОБОЧА НАВЧАЛЬНА ПРОГРАМА

З дисципліни «ІНТЕЛЕКТУАЛЬНІ МЕТОДИ ОБРОБЛЕННЯ ТЕКСТОВОЇ ІНФОРМАЦІЇ»

Рівень вищої освіти третій
Ступінь вищої освіти доктор філософії
Галузь знань12 – інформаційні технології
Спеціальність122 – комп'ютерні науки

Шифр ДВА 07 Дисципліна за вибором аспіранта

Форма навчання_____денна_____Курс _2___Семестр _4____
Всього годин /кредитів ЄКТС____60 /2,0 за навчальним планом

- лекції (Л) _____10____
- семінарські заняття (СЗ) __6_____
- практичні заняття (ПЗ) _____
- індивідуально-консультативна робота (ІКР) ____4_____
- самостійна робота студентів (СРС) 40____
- підсумковий контроль дисципліни – залік

- **м. Київ**

Укладач(і) робочої навчальної програми:
доктор фізико-математичних наук, професор, завідувач відділу інтелектуалізації інформаційних технологій

_____ Олександр МАРЧЕНКО
(підпис)
e-mail: rozenkrans17@gmail.com

Робочу програму погоджено з гарантом освітньо-наукової програми
Гарант освітньої програми _____/ Володимир СТЕПАШКО
(підпис)

Затверджено Вченою радою Міжнародного науково-навчального центру інформаційних технологій та систем НАН та МОН України
Протокол № 3 від 23.04.2024 р.

Вчений секретар Вченої ради _____ Микола КОМАР

Ухвалено Науково-методичною радою Міжнародного науково-навчального центру інформаційних технологій та систем НАН та МОН України_
Протокол № 2 від 15.04.2024 р.

Голова Науково-методичної ради _____ Людмила КОЗАК

Ухвалено Радою молодих вчених Міжнародного науково-навчального центру інформаційних технологій та систем НАН та МОН України
Протокол № 4 від 15.04.2024 р.

Голова Ради молодих вчених _____ Дмитро ВОЛОШЕНЮК

Введено в дію наказом директора Міжнародного науково-навчального центру інформаційних технологій та систем НАН та МОН України № 57 від 23.04.2024 р.

1.ЗАГАЛЬНІ ВІДОМОСТІ

Найменування показників	Характеристика дисципліни за денною формою навчання
Вид дисципліни	**Дисципліна за вибором аспіранта**
Мова викладання, навчання та оцінювання	Українська
Загальний обсяг кредитів / годин	2/60
Курс	2
Семестр	4
Кількість змістових модулів з розподілом:	2
Обсяг кредитів	2
Обсяг годин, в тому числі:	30
Лекції	10
Семінарські заняття	6
Самостійна робота	40
Форма підсумкового контролю	диференційований залік

Дисципліна «Інтелектуальні методи оброблення текстової інформації» належить до переліку дисциплін циклу професійної підготовки аспіранта (за вибором аспіранта). Вона забезпечує важливий аспект професійного світогляду аспіранта та спрямована на формування вміння розробляти та використовувати в наукових дослідженнях сучасні інформаційні технології оброблення природної мови, застосовувати моделі та методи інтелектуальної обробки текстів як основного універсального носія інформації в сучасному світі.

2. МЕТА І ЗАВДАННЯ ДИСЦИПЛІНИ

Метою дисципліни «Інтелектуальні методи оброблення текстової інформації» є навчити аспіранта формулювати та розв'язувати завдання аналізу, класифікації та інтерпретації текстів природною мовою, використовувати моделі та методи обробки природної мови для створення інтелектуальних інформаційних технологій оброблення інформації, представленої у текстовому вигляді, шукати власні шляхи розв'язування завдань, ефективно працювати з інформацією, створювати нові знання шляхом проведення оригінальних теоретичних та експериментальних досліджень.

Основними завданнями є: 1) ознайомлення з основними напрямами та методами оброблення природної мови (текстів природною мовою); 2) ознайомлення з принципами та підходами до моделювання структур та явищ природної мови, розробки лінгвістичних баз знань онтологічного типу, формування та оброблення текстових корпусів як основного джерела для здобування лінгвістичних знань в контексті машинного навчання.

Інтегральна компетентність

Здатність продукувати нові ідеї, розв'язувати комплексні проблеми у сфері комп'ютерних наук, застосовувати методологію наукової та педагогічної діяльності, а також проводити власне наукове дослідження, результати якого мають наукову новизну, теоретичне та практичне значення.

Загальні компетентності:

ЗК01. Здатність до абстрактного мислення, аналізу та синтезу.

ЗК02. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.

ЗК03. Здатність працювати в міжнародному контексті.

Спеціальні (фахові) компетентності:

СК02. Здатність застосовувати сучасні методології, методи та інструменти експериментальних і теоретичних досліджень у сфері комп'ютерних наук, сучасні цифрові технології, бази даних та інші електронні ресурси у науковій та освітній діяльності.

СК03. Здатність виявляти, ставити та розв'язувати дослідницькі науково-прикладні завдання та/або вирішувати проблеми в сфері комп'ютерних наук, оцінювати та забезпечувати якість виконуваних досліджень.

СК05. Здатність здійснювати науково-педагогічну діяльність у вищій освіті у сфері комп'ютерних наук.

СК06. Здатність аналізувати та оцінювати сучасний стан і тенденції розвитку комп'ютерних наук та інформаційних технологій.

СК10. Здатність до проведення наукових досліджень з інтелектуального оброблення, аналізу та інтерпретації інформації про об'єкти різної природи.

Програмні результати навчання

РН01. Мати передові концептуальні та методологічні знання з комп'ютерних наук і на межі предметних галузей, а також дослідницькі навички, достатні для проведення наукових і прикладних досліджень на рівні останніх світових досягнень з відповідного напряму, отримання нових знань та/або здійснення інновацій.

РН02. Вільно презентувати та обговорювати з фахівцями і нефахівцями результати досліджень, наукові та прикладні проблеми комп'ютерних наук державною та іноземною мовами, оприлюднювати результати досліджень у наукових публікаціях у провідних міжнародних наукових виданнях.

РН06. Застосовувати сучасні інструменти і технології пошуку, оброблення та аналізу інформації, зокрема, статистичні методи аналізу даних великого обсягу та/або складної структури, спеціалізовані бази даних та інформаційні системи.

РН08. Визначати актуальні наукові та практичні проблеми у сфері комп'ютерних наук, глибоко розуміти загальні принципи та методи комп'ютерних наук, а також методологію наукових досліджень, застосувати їх у власних дослідженнях у сфері комп'ютерних наук та у викладацькій практиці.

РН12. Здійснювати інтелектуальний аналіз електронних масивів даних для розв'язання конкретних практичних завдань, зокрема побудови нейронних мереж, комп'ютерних систем автоматичного керування, розв'язання задач штучного інтелекту, створення систем інтелектуального керування динамічними об'єктами у реальному часі.

РН15.Проводити інтелектуальний аналіз об'єктів різних природи за різними видами первинної інформації (зображення, складні сигнали, тексти, електронні медичні записи, відео та аудіо записи).

РН16. Застосовувати методи побудови систем штучного інтелекту, визначати механізми використовування знань про предметну область для виконання прикладних завдань на основі інтелектуальних інформаційних систем різної спрямованості.

РН 17. Розробляти комп'ютерні системи оброблення та аналізу інформації різного виду (цифрової, текстової, зображень, відеоряду, сигналів тощо).

3. СТРУКТУРА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Номер лекції	Назва лекції	Кількість годин		
		Лекції	Семінарські /практичні заняття	Самостійна робота
ЗМ1: Моделі та методи обробки природної мови(1 кредит)				
1	Оброблення природної мови. Основні моделі представлення структур та процесів природної мови.	2		
2	Методи та моделі машинного навчання в задачах оброблення природної мови/текстів природною мовою.	2	3	10
3	Знання. Бази знань. Онтології. Побудова БЗ.	2		10
ЗМ2: Прикладні задачі оброблення текстів природною мовою (1 кредит)				
4	Розпізнавання сутностей тексту. Кореферентний аналіз текстів.	2	2	10
5	Задача генерування текстів	2	1	10
	ВСЬОГО	**10**	**6**	**40**

4. ПРОГРАМНИЙ МАТЕРІАЛ

Змістовий модуль 1. Моделі та методи обробки природної мови

Тема 1. Оброблення природної мови. Основні моделі представлення структур та процесів природної мови.

В даній темі представлено матеріал про основні задачі, моделі та методи прикладної комп'ютерної лінгвістики. Описуються основні структури для представлення об'єктів, властивостей, відношень та процесів природної мови на морфологічно-лексичному, синтаксичному, семантичному та прагматичному рівнях аналізу.

Надано опис архітектури системи послідовного аналізу тексту природною мовою з описом входів і виходів блоків кожного рівня. Описуються деякі модифікації загальної архітектури під конкретну специфіку окремих прикладних задач.

Тема 2. Методи та моделі машинного навчання в задачах оброблення природної мови/текстів природною мовою.

В темі надано короткий опис основних моделей та методів машинного навчання застосованих для обробки природномовних текстів: лінгвістичні марківські моделі, приховані марківські моделі, умовні випадкові поля, модель максимальної ентропії, метод опорних векторів, нейронні моделі різних архітектур.

Дається порівняльний аналіз двох напрямів моделювання – генеративного та дискримінативного, із обговоренням їх основних переваг та недоліків. Окрема увага

приділяється методикам вимірювання ефективності методів аналізу та класифікації текстів.

Тема 3. Знання. Бази знань. Онтології. Побудова БЗ

Дається основні визначення поняттю «знання». Представлено основні моделі подання знань: семантичні мережі, фреймові моделі, онтології. Розглядаються сучасні системи баз знань.

Також розглянуто основні підходи до розробки та розбудови великих загальних і спеціалізованих баз знань онтологічного типу. Представлено методи автоматизації здобування нових знань для автоматичного заповнення та розширення БЗ.

Тема 4. Розпізнавання сутностей тексту. Кореферентний аналіз текстів.

Представлено дві класичні прикладні задачі комп'ютерної лінгвістики - розпізнавання сутностей тексту та кореферентний аналіз текстів. Перша задача призначена для виявлення та класифікації згадувань у тексті сутностей певних класів (наприклад людина, організація, місцевість, дата і т.д.). Друга – для встановлення кореферентних зв'язків між тими елементами тексту, що посилаються на одну сутність незалежно від типу згадування – прямого чи анафоричного (наприклад, за допомогою займенника).

Розглянуто основні моделі та методи розв'язання даних задач. Разом ці методи дають змогу виконувати пошук по текстах всієї релевантної інформації по певному запиту стосовно, наприклад, деякої особи чи організації з виявленням їх властивостей, відношень, подій та процесів, в яких вони беруть участь.

Тема 5. Задача генерування текстів

Розглядається класична задача комп'ютерної лінгвістики – синтез текстів. В темі представлено моделі для подання речень та моделі для подання текстів. Розглядається цілий клас споріднених задач: генерування тексту на задану тему, генерування машинного перекладу, генерування парафразу (переказу оригінального тексту своїми словами), генерування реферату (короткого переказу тексту).

Представлено низку моделей та методів для генерування текстів – структурних та нейронних (зокрема моделі deep learning). Окрема увага приділена проблематиці породження зв'язних та цілісних текстів та методам вимірювання зв'язності та цілісності текстів.

5. САМОСТІЙНА РОБОТА

Самостійна робота охоплює:

1) підготовку до семінарських занять,

2) опрацювання наукової літератури,

3) підготовку до заліку.

№ п/п	Зміст самостійної роботи	Обсяг СР (годин)
1.	Підготовка до семінарських занять	10
2.	Опрацювання наукової літератури	20
3.	Підготовка до заліку	10
Усього за навчальною дисципліною		40

6. РЕЙТИНГОВА СИСТЕМА ОЦІНЮВАННЯ РЕЗУЛЬТАТІВ НАВЧАННЯ

Контроль знань аспірантів здійснюється на підставі Положення про організацію та проведення поточного і підсумкового/семестрового контролю результатів навчання здобувачів вищої освіти третього (освітньо-наукового) рівня.

Контроль знань аспірантів складається з двох складників: поточного і підсумкового/семестрового контролю результатів навчання здобувачів вищої освіти третього (освітньо-наукового) рівня. Кожний складник оцінюється за стобальною системою .

Загальна оцінка результатів за дисципліною (ЗО) розраховують:

$$ЗО = к1*ПоК + к2*ПідК,$$

де к1, к2 - коефіцієнти переведення балів поточного (ПоК) та підсумкового контролю (ПідК) відповідно; к1 = 0,4, к2 =0,6.

Максимальна кількість балів у поточному контролі встановлюється таким чином:

Форми навчальної діяльності	Максимальна сумарна оцінка в балах
усне опитування	65
активна робота на заняттях	35
Всього	100

Порядок перерахунку рейтингових показників нормованої 100-бальної шкали оцінювання в національну шкалу та шкалу ECTS

<table>
<tr><th rowspan="2">За 100-бальною шкалою</th><th colspan="2">За національною шкалою</th><th rowspan="2">За шкалою ECTS</th></tr>
<tr><th>Іспит</th><th>Залік</th></tr>
<tr><td>91 - 100</td><td>відмінно</td><td rowspan="5">зараховано</td><td>A (відмінно)</td></tr>
<tr><td>81 - 90</td><td rowspan="2">добре</td><td>B (дуже добре)</td></tr>
<tr><td>71 - 80</td><td>C (добре)</td></tr>
<tr><td>66 - 70</td><td rowspan="2">задовільно</td><td>D (задовільно)</td></tr>
<tr><td>60 - 65</td><td>E (достатньо)</td></tr>
<tr><td>40 - 59</td><td rowspan="2">незадовільно</td><td rowspan="2">не зараховано</td><td>FX (незадовільно – з можливістю повторного складання)</td></tr>
<tr><td>1 - 39</td><td>F (неприйнятно – з обов’язковим повторним навчанням</td></tr>
</table>

7. ОРІЄНТОВНИЙ ПЕРЕЛІК ЗАПИТАНЬ НА ЗАЛІК

1. Основні задачі комп’ютерної лінгвістики.
2. Архітектура комп’ютерної лінгвістичної системи.
3. Моделі представлення лексики.
4. Моделі представлення синтаксису.
5. Моделі представлення семантики.
6. Модель представлення тексту. Модель представлення дискурсу.
7. Діалогова система.

8. Лінгвістичні моделі машинного навчення
9. Модель Байєса
10. Логістична регресія
11. Максимальна ентропія
12. Метод опорних векторів
13. Приховані марківські моделі
14. Умовні випадкові поля
15. Нейромережеві архітектури NLP-систем
16. Deep-learning в NLP задачах
17. Знання. Визначення. Основні моделі представлення
18. Семантичні мережі
19. Фреймові моделі
20. Онтології
21. Модальні рольові відношення
22. Методи автоматизації побудови та розширення баз знань
23. Методи автоматизації локалізації БЗ до іншої мови
24. Формальний концептуальний аналіз
25. Розпізнавання сутностей тексту
26. Кореферентний аналіз текстів
27. Задача генерування текстів
28. Трансформерні архітектури у генеруванні текстів
29. Проблема зв'язності та цілісності текстів
30. Методи та метрики для вимірювання ефективності розв'язання задач комп'ютерної лінгвістики.

8. ПОЛІТИКА ДОБРОЧЕСНОСТІ

Виконання навчальних завдань має відповідати вимогам Кодексу академічної доброчесності Міжнародного науково-навчального центру інформаційних технологій та систем НАН та МОН України, затвердженого вченою радою Міжнародного центру 20.січня 2022 року, протокол № 1.

9. РЕКОМЕНДОВАНА ЛІТЕРАТУРА

Основна

1. D Jurafsky, JH Martin Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. - Prentice Hall, 2020
2. Rish, Irina. (2001). «An empirical study of the naive Bayes classifier». IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence.
3. Карташов М. В. Імовірність, процеси, статистика — Київ, ВПЦ Київський університет, 2007.
4. Hosmer, David W., Stanley Lemeshow (2000). Applied Logistic Regression, 2nd ed.. New York; Chichester,

5. Harremoës P. and Topsøe F., 2001, Maximum Entropy Fundamentals, Entropy, 3(3), 191—226.
6. Nello Cristianini, John Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. — Cambridge University Press, 2000.
7. Lawrence Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77 (2): 257–286.
8. Lafferty, J., McCallum, A., Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proc. 18th International Conf. on Machine Learning. Morgan Kaufmann. с. 282–289.

Допоміжна література

1.Patrick Henry Winston. Artificial Intelligence. 1992. ISBN 0-201-53377-4.

2. Nils J. Nilsson Principles of Artificial Intelligence. 1980. ISBN 0-934613-10-9