

# НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ ІНСТИТУТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА СИСТЕМ

## Силабус ДВА 14 Інтелектуальне оброблення текстової інформації

### Реквізити навчальної дисципліни

Рівень вищої освіти Третій (освітньо-науковий)

Галузь знань	F – інформаційні технології
Спеціальність	F3 – комп'ютерні науки
Освітньо-наукова програма	Інтелектуальні методи та засоби комп'ютерних наук
Статус дисципліни	Вибіркова
Форма навчання	Очна(денна)
Рік підготовки, семестр	2-й рік/4-й семестр
Обсяг дисципліни	2 кредити/60 годин
Семестровий контроль/ контрольні заходи	Диференційований залік
Розклад занять	1 година аудиторних занять/тиждень,
Мова викладання	Українська
Інформація про керівника курсу / викладачів	Лектор: Марченко Олександр Олександрович; доктор фізико-математичних наук, професор; завідувач відділу інтелектуалізації інформаційних технологій Контактна інформація (e-mail rozenkrans17@gmail.com )
Розміщення курсу	<a href="https://aspirant.irtc.org.ua/silabus/">https://aspirant.irtc.org.ua/silabus/</a>

### ХАРАКТЕРИСТИКА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

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.

### МЕТА, ЗАВДАННЯ, ПРИЗНАЧЕННЯ НАВЧАЛЬНОЇ 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, представлені 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 напрямками 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.

Пререквізити - попередні вимоги до навчання за освітнім компонентом:

Знання, вміння, навички, якими повинен володіти здобувач, щоб приступити до вивчення дисципліни: знання з першого та другого рівня вищої освіти: знання та вміння за освітнім компонентом ЗП-2 «аспірантські студії з комп'ютерних наук», а також: методи та моделі машинного навчання, основні поняття щодо баз знань, онтологій.

Постреквізити: Вивчення дисципліни забезпечить виконання завдань дисертаційного дослідження відповідного напрямку, одержання та осмислення одержаних результатів для написання наукових статей, підготовки та захисту дисертації. Наявність можливості подальшого навчання та дослідження для підготовки та захисту дисертації доктора наук.

### **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.

### **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 напрямку, 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.

РН13. Проводити інтелектуальний аналіз складних об'єктів за різними видами первинної

інформації (зображення, складні сигнали, тексти, електронні медичні записи, відео та аудіо записи).

## ПЕРЕДУМОВИ ВИВЧЕННЯ ДИСЦИПЛІНИ ДЛЯ ФОРМУВАННЯ ПРОГРАМНИХ РЕЗУЛЬТАТІВ НАВЧАННЯ ТА КОМПЕТЕНТНОСТЕЙ

Ефективність засвоєння змісту дисципліни «Інтелектуальні методи оброблення текстової інформації» значно підвищиться, якщо здобувач вищої освіти попередньо опанував матеріалом таких дисциплін як: «Штучний інтелект», «Машинне навчання», «Методи математичної статистики».

### СТРУКТУРА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ.

Назви змістових модулів і тем	Кількість годин			
	Усього	У тому числі		
		Лекції	Семінарські заняття	Самостійна робота
<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>
<i>Змістовий модуль 1. Моделі та методи обробки природної мови (1 кредит)</i>				
<b>Тема 1.</b> Оброблення природної мови. Основні моделі представлення структур та процесів природної мови.	2	2		
<b>Тема 2.</b> Методи та моделі машинного навчання в задачах	15	2	3	10

оброблення природної мови/текстів природною мовою.				
<b>Тема 3.</b> Знання. Базизнань. Онтології. Побудова БЗ.	12	2		10
<i>Усього годин за змістовим модулем 1</i>	29	6	3	20
<i>Змістовий модуль 2.</i> Прикладні задачі оброблення текстів природною мовою (1 кредит)				
<b>Тема 4.</b> Розпізнавання сутностей тексту. Кореферентний аналіз текстів .	14	2	2	10
<b>Тема 5.</b> Задача генерування текстів	13	2	1	10
<b>РАЗОМ:</b>	56	10	6	40

### ТЕХНОЛОГІЧНА КАРТКА ОПАНУВАННЯ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

№ з/п/	Теми та форми занять (год.)	Зміст занять і навчальних завдань	Форми контролю
<b>Змістовий модуль 1.</b> Моделі та методи обробки природної мови(1 кредит)			
<b>Тема 1.</b> Оброблення природної мови. Основні моделі представлення структур та процесів природної мови.			
1	<i>Лекція</i> Оброблення природної мови. Основні моделі представлення структур та процесів природної мови.  (2 год.)	В даній лекції 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.	
<b>Тема 2.</b> Методи та моделі машинного навчання в задачах оброблення природної мови/текстів природною мовою			

2	<i>Лекція</i> Методи та моделі машинного навчання в задачах оброблення природної мови/текстів природною мовою  (2 год.)	В лекції 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.	
3	<i>Семінарське заняття.</i> Методи та моделі машинного навчання в задачах оброблення природної мови/текстів природною мовою. (3 год.)	Моделі та методи машинного навчання для обробки природномовних текстів: лінгвістичні марківські моделі, приховані марківські моделі, умовні випадкові поля, модель максимальної ентропії, метод опорних векторів, нейронні моделі різних архітектур.	Усне опитування
4	<i>Самостійна робота</i> (10 год.)	Моделі та методи машинного навчання для обробки природномовних текстів: лінгвістичні марківські моделі, приховані марківські моделі, умовні випадкові поля, модель максимальної ентропії, метод опорних векторів, нейронні моделі різних архітектур.	Усне опитування
<b>Тема 3.</b> Знання. Бази знань. Онтології. Побудова БЗ.			
5	<i>Лекція</i> Знання. Бази знань. Онтології. Побудова БЗ. (2 год.)	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.	
6	<i>Самостійна робота</i> (10 год.)	Опрацювання наукової літератури (Бази знань. Онтології. Побудова БЗ.) Формальний концептуальний аналіз	Усне опитування
<b>Змістовий модуль 2.</b> Прикладні задачі оброблення текстів природною мовою 1 кредит)			
<b>Тема 4.</b> Розпізнавання сутностей тексту. Корелативний аналіз текстів .			
7	<i>Лекція</i> Розпізнавання сутностей тексту. Корелативний аналіз текстів . (2 год.)	В лекції представлено дві класичні прикладні задачі комп'ютерної лінгвістики - розпізнавання сутностей тексту та корелативний аналіз текстів.	

8	<i>Семінарське заняття.</i> Розпізнавання сутностей тексту. Корелативний аналіз текстів . (2 год.)	Розглянуто основні моделі та методи розв'язання даних задач. Разом ці методи дають змогу виконувати смисловий 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.	Усне опитування
9	<i>Самостійна робота (10 год.)</i>	Програмна реалізація вивчених моделей та методів машинного навчання для розв'язання задач розпізнавання сутностей тексту та корелативного аналізу текстів .	Перевірка та опитування
<b>Тема 5. Задача генерування текстів</b>			
10	<i>Лекція</i> Задача генерування текстів (2 год.)	В лекції розглядається класична задача комп'ютерної лінгвістики – синтез текстів. Представлено 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)	
11	<i>Семінарське заняття.</i> Задача генерування текстів . (1 год.)	Моделі для подання речень та моделі для подання текстів.	Усне опитування
12	<i>Самостійна робота (10 год.)</i>	Дослідження проблематики породження зв'язних та цілісних текстів та методів вимірювання зв'язності та цілісності текстів. Підготовка до заліку	Перевірка та опитування

Обов'язкове індивідуальне завдання		
№№ з/п	1. Методи розв'язання займенникової анафори 2. Методи Машинного перекладу 3. Методи Сентимент аналізу 4. Діалогові системи 5. Системи типу «Питання-Відповідь»	реферат

### КОНТРОЛЬ І 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:

$$ЗО = k_1 * \text{ПоК} + k_2 * \text{ПідК},$$

де  $k_1$ ,  $k_2$  - коефіцієнти переведення балів поточного (ПоК) та підсумкового контролю (ПідК) відповідно;  $k_1 = 0,4$ ,  $k_2 = 0,6$ .

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

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

### ПОЛІТИКА НАВЧАЛЬНОГО КУРСУ

#### Політика щодо академічної доброчесності

Дотримання академічної доброчесності здобувачами передбачає, зокрема:

- самостійне виконання навчальних завдань, завдань поточного та підсумкового контролю результатів навчання (для осіб з особливими освітніми потребами ця вимога застосовується з урахуванням їхніх індивідуальних потреб і можливостей);

- посилання на джерела інформації у разі використання ідей, розробок, тверджень, відомостей інших дослідників;
- дотримання норм законодавства про авторське право і суміжні права;
- надання достовірної інформації про результати власної (наукової, творчої) діяльності, використанні методики досліджень і джерела інформації.

#### ***Політика щодо відвідування занять та поведінки на заняттях***

Відвідування занять є обов'язковим компонентом навчання. За об'єктивних причин (наприклад, хвороба, міжнародне стажування тощо) навчання може відбуватись в он-лайн формі за погодженням із викладачем навчальної дисципліни та затвердженням директора Інституту.

#### ***Політика щодо правил поведінки на заняттях***

Здобувачі вищої освіти третього рівня беруть активну участь у всіх заняттях: обговорюють проблемні ситуації, запропоновані викладачем на лекціях; активно включаються і за потреби ініціюють спільну (групову роботу) під час семінарських занять; Спілкування учасників освітнього процесу (викладач, здобувачі) відбувається на засадах партнерських стосунків, взаємодопомоги, толерантності та поваги до особистості кожного, спрямованості на здобуття істинного наукового знання.

#### ***Політика щодо термінів виконання завдань і перескладання***

Здобувачі вищої освіти третього рівня повинні виконувати всі навчальні завдання вчасно, відповідно до робочої навчальної програми, за невчасне виконання знижується бальна оцінка. Графіки перескладання формують викладачі відповідних дисциплін.

## **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.
9. Marchenko O. O., Radyvonenko O. S., Ignatova T. S., Titarchuk P. V., Zhelezniakov D. V. Improving Text Generation Through Introducing Coherence Metrics. *Cybernetics and Systems Analysis*; 2020; №: 1; стор.: 13-21;
10. Marchenko Oleksandr, Isoieva Mariam. Automatic Generation of Coherent Natural Language Text. Flexible Query Answering Systems. Lecture Notes in Computer Science; 2023; p: 79-92
11. Skurzhashkyi O. H., Marchenko O. O., Anisimov A. V. Specialized Pre-Training of Neural Networks on Synthetic Data for Improving Paraphrase Generation. *Cybernetics and Systems Analysis*; 2024; №: 2; pp 167-174

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

