НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ
ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ

ЗАТВЕРДЖУЮ
Директор ІПРІ НАН України
академік НАН України
_____В.В. Петров
«16» 11 2021 р.

ОСНОВИ ГЛИБИННОГО АНАЛІЗУ ДАНИХ І ТЕКСТУ (TEXT/ DATA MINING)

(назва навчальної дисципліни)

РОБОЧА ПРОГРАМА
кредитного модуля

ГАЛУЗЬ ЗНАНЬ 12 «Інформаційні технології»

СПЕЦІАЛЬНІСТЬ 122 «Комп'ютерні науки»

СПЕЦІАЛІЗАЦІЯ «Інформаційні технології»

Ухвалено Вченою радою ІПРІ НАН України
(протокол від «16» 11 2021 р. №11)

Київ
ІПРІ НАН України
2021

Робоча програма кредитного модуля Основи теорії інформаційного пошуку складена відповідно до програми навчальної дисципліни Основи теорії інформаційного пошуку.

Розробник робочої програми:

Завідувач відділу, професор., д.т.н. Ланде Дмитро Володимирович
(посада, науковий ступінь, вчене звання, прізвище, ім'я, по батькові)

(підпис)

1. Опис кредитного модуля

<table>
<tr><th>Рівень ВО, спеціальність, освітня програма, форма навчання</th><th>Загальні показники</th><th>Характеристика кредитного модуля</th></tr>
<tr><td>Рівень ВО
третій (доктор філософії)</td><td>Основи глибинного аналізу даних і тексту (Text/ Data mining)</td><td>Лекції
_16_ год.</td></tr>
<tr><td>Спеціальність
122 Комп’ютерні науки</td><td>Цикл
професійної підготовки</td><td>Практичні (семінарські)
_14_ год.</td></tr>
<tr><td rowspan="3">Освітньо-наукової програми спеціальності 122 Комп’ютерні науки спеціалізації Інформаційні технології</td><td rowspan="2">Статус кредитного модуля обов’язковий</td><td>Лабораторні роботи
-</td></tr>
<tr><td>Самостійна робота
_30__год.,</td></tr>
<tr><td>Семестр __4__</td><td>Індивідуальне завдання
Не передбачено</td></tr>
<tr><td>Форма навчання:
денна</td><td>Кількість
кредитів (годин):
2 (60)</td><td>Вид та форма семестрового контролю:
залік</td></tr>
</table>

Предметом вивчення навчальної дисципліни «Основи глибинного аналізу даних і тексту (Text/ Data mining)» є питання, що відносяться до інформаційної структури Web-простору, повнотекстовим інформаційно-пошуковим системам, їх алгоритмічному та лінгвістичному забезпеченню, можливостям ранжирування, аналітичного узагальнення результатів пошуку, загальним закономірностям сучасного інформаційного простору.

Дисципліна складається з двох змістовних модулів:

1. Теоретико-методологічні основи глибинного аналізу даних і тексту (Text/ Data mining)
2. Застосування методів глибинного аналізу даних і тексту (Text/ Data mining).

Вивчення дисципліни спирається на знання, отримані за програмою попередніх років навчання за спеціальністю 122 Комп’ютерні науки. Аспіранти мають досвід у таких областях математики, як теорія множин, вища алгебра, диференціальні рівняння та теорія ймовірностей, а також здатні оволодіти необхідними навичками для проведення практичних занять: мовами програмування Perl, Python програмуванням у веб-середовищі, прикладними програмами статистичної обробки даних.

2. Мета та завдання кредитного модуля

*Метою викладання дисципліни “*Основи глибинного аналізу даних і тексту (Text/ Data mining)” є формування у слухачів наукового світогляду, цілісного уявлення про методологію наукового дослідження та навичок практичного

застосування конкретних методів Text/ Data mining у професійній діяльності, формування системи теоретичних і практичних знань в галузі інформаційних технологій, зокрема, при роботі з інформаційно-пошуковими системами, серед яких сьогодні важлива роль відводиться пошуковими системами в мережі Інтернет.

2.1. Відповідно до вимог Освітньо-наукової програми третього (доктор філософії) рівня вищої освіти спеціальності 122 Комп'ютерні науки аспіранти після засвоєння навчальної дисципліни мають продемонструвати такі результати навчання:

Здатності:

- працювати з основними мережевими інформаційно-пошуковими системами з питань глибинного аналізу даних і тексту (Text/ Data mining);
- застосовувати сучасні інформаційні технології у різних видах професійної діяльності (ІК-2);
- знаходити, обробляти й аналізувати необхідну інформацію для рішення проблем й прийняття рішень (ІК-3);
- проводити теоретичні й експериментальні дослідження, математичне й комп'ютерне моделювання в галузі знань "Інформаційні технології" (ПК-3);
- застосовувати високопродуктивні технології розподілених систем та паралельних обчислень при вирішенні науково-практичних задач (ПК-5с);
- проектування та програмної реалізації методів інформаційного пошуку в інформаційних середовищах різноманітного призначення, систем управління бізнес-процесами, мереж Інтернету речей, сервіс-орієнтованих середовищ та систем високопродуктивних кластерних обчислень;
- вирішувати масштабні обчислювальні задачі у розподілених інтелектуальних середовищах та контролювати хід обчислень за допомогою спеціалізованого програмного забезпечення;
- вибирати адекватні методи глибинного аналізу даних і тексту (Text/ Data mining) для вирішення конкретних задач прогнозування, керування, класифікації та інтелектуального аналізу даних.

2.2. Основні завдання кредитного модуля.

Згідно з вимогами програми навчальної дисципліни аспіранти після засвоєння кредитного модуля мають продемонструвати такі результати навчання:

знання :

- концептуальних засад і технологій глибинного аналізу даних і тексту (Text/ Data mining);
- сучасних інформаційно-комунікаційних технологій, програмно-апаратних засобів проведення наукових досліджень;
- можливостей використання інформаційно-пошукових технологій для автоматизації експерименту, обробки даних, оформлення результатів досліджень;
- технологій інтелектуальних обчислень та аналізу даних;
- комп'ютерного моделювання та обробки даних, паралельних обчислень з використанням спеціалізованих програмних засобів;
- методів розподіленого моделювання складних об'єктів і систем в обчислювальному середовищі, застосувань технологій штучного інтелекту в

розподілених обчисленнях, базових алгоритмів глибинного аналізу даних і тексту (Text/ Data mining);

- з проектування математичного, лінгвістичного, інформаційного і програмного забезпечення інформаційних систем, з розроблення інформаційних систем, комплексів та мереж.

уміння:

- застосовувати знаннєво-орієнтовані мережеві інформаційні системи при вирішенні наукових та прикладних задач, пов'язаних з аналізом, моделюванням, прогнозуванням та управлінням економічних та соціальних процесів суспільства;
- використовувати інформаційно-комунікаційні технології, універсальні та спеціалізовані програмно-апаратні засоби;
- здійснювати автоматизацію експерименту, статистичну обробку даних, оформлення результатів досліджень засобами інформаційних технологій;
- застосовувати методи розподіленого моделювання складних об'єктів і систем, технології штучного інтелекту в розподілених обчисленнях, інтелектуальні обчислення, проектувати та програмно реалізовувати методи комп'ютерної обробки надвеликих за обсягом даних;
- проектувати математичне, лінгвістичне, інформаційне і програмне забезпечення інформаційних систем, розробляти інформаційні системи, комплекси та мережі
- використовувати сучасні технології Text/ Data mining та відповідні інфраструктури програмних рішень;
- використовувати методи глибинного аналізу даних і тексту (Text/ Data mining) для вирішення практичних задач.

3. Структура кредитного модуля

Назви розділів і тем	Кількість годин				
	Всього	у тому числі			
		Лекції	Практичні	Лабораторні	СРС
1	2	3	4	5	6
Розділ 1. Теоретико-методологічні основи глибинного аналізу даних і тексту (Text/ Data mining)					
Тема 1. Введення в дисципліну Text/ Data mining	7	2	1	-	4
Тема 2. Методи розпізнавання образів	8	2	2	-	4
Тема 3. Методи підтримки прийняття рішень	8	2	2	-	4
Разом за розділом 1	23	6	5	-	12
Розділ 2. Застосування методів глибинного аналізу даних і тексту (Text/ Data mining)					
Тема 1. Основи м'яких обчислень.	6	2	1	-	3
Тема 2. Кореляційний та вейвлет-аналіз.	8	2	2	-	4
Тема 3 Фрактальний і мультифрактальний аналіз.	8	2	2	-	4
Тема 4. Основи теорія складних	8	2	2	-	4

1	2	3	4	5	6
мереж					
Тема 5. Мережевий аналіз даних і візуалізація.	7	2	2	-	3
Разом за розділом 2	37	10	9	-	18
Всього годин	60	16	14	-	30

4. Лекційні заняття

№ з/п	Назва теми лекції та перелік основних питань
1	**Введення в дисципліну Text/Data Mining.** Поняття інтелектуального аналізу даних. Поняття "інформація" і "знання", співставлення и порівняння цих понять. Сфери застосування інтелектуального аналізу даних. Класифікації методів Text / Data Mining.
2	**Методи розпізнавання образів.** Методи дисперсійного аналізу (МНК, ЛДА, КДА). Поняття дисперсіного і регресивного аналізу. Метод найменших квадратів. . Лінійний дискримінантний аналіз. Квадратичний дискримінантний аналіз. Статистичні методи розпізнавання образів. Евристичні методи розпізнавання образів.
3	**Методи підтримки прийняття рішень.** Основи теорії нечіткої логіки. Нечіткі відношення та нечітке логічне виведення. Аналіз нечітких експертних заключень. Прийняття рішень в нечітких умовах. Основні поняття, визначення та сфера застосувань методу аналізу ієрархій. Побудова ієрархічної структури проблеми. Математичні методи прогнозування.
4	**Основи м'яких обчислень.** Методи м'яких обчислень. Багатоагентні системи. Концепція колективного розуму. Введення в еволюційні обчислення. Основні принципи. Генетичний алгоритм. Призначення та формальна постановка задачі. Цикл генетичного алгоритму. Програмна реалізація генетичного алгоритму. Концепція колективного розуму або інтелекту зграї. Принципи побудови мурашиного алгоритму.
5	**Кореляційний та вейвлет-аналіз.** Формалізм кореляційного аналізу. Автокореляційна функція. Кореляційний аналіз самоподібних процесів. Визначення понять вейвлет, вейвлет-перетворення. Побудова вейвлет-спектограм. Інструментарій у середовищі системи MATLAB.
6	**Фрактальний і мультифрактальний аналіз.** Основні поняття і визначення фрактального аналізу. Фрактальна розмірність, абстрактні фрактали. Фрактальна розмірність числових рядів. Обчислення коефіцієнту Херста. Метод DFA. Обчислення скейлингового коефіцієнта. Концепція мультифрактального формалізму. Модель множини Кантора. Поняття мультифрактального спектру та його інтерпретація. Алгоритми розрахунку і візуалізації мультифрактального спектру.
7	**Основи прикладної теорії складних мереж.** Основні показники мереж та вузлів. Близькі та далекі зв'язки. Концепція малих світів. Перколяція. Структура веб-простору. Ранжирування результатів

	пошуку з урахуванням структури мережі. Алгоритми HITS і PageRank.
8	**Мережевий аналіз даних і візуалізація.** Візуалізація багатовимірних даних, карти Кохонена. Карти Кохонена як нейронні мережі. Алгоритм карти Кохонена. Візуалізація карти Кохонена. Поняття графів видимості. Граф горизонтальної видимості. Застосування методів мережевого аналізу числових рядів.

5. Семінарські заняття

Основні завдання циклу практичних занять полягають у набутті аспірантами практичних навичок з використання програмного забезпечення інформаційно-пошукових систем.

№ з/п	Назва теми заняття
1	Генерування псевдовипадкових чисел із заданим розподілом. Основи техніки програмування у середовищі MATLAB.
2	Перевірка статистичних гіпотез. Вивчення можливостей пакету Statistical Toolbox у середовищі MATLAB.
3	Методи дисперсійного аналізу. Рішення задачі класифікації методом ЛДА.
4	Еволюційний алгоритм. Генетичний алгоритм.
5	Технологія моделювання за допомогою клітинних автоматів. Модель дифузії інформації.
6	Кореляційний аналіз. Розрахунок автокореляційної функції. Фрактальний аналіз. Розрахунок показника Херста.
7	Мережевий аналіз даних і візуалізація. 3D-візуалізація результатів аналізу числових рядів методом DFA. Вейвлет-аналіз у середовищі MATLAB.

5. Практичні заняття

Практичних занять не передбачено.

6. Лабораторні заняття

Лабораторних занять не передбачено.

7. Самостійна робота

№ з/п	Назви тем і питань, що виноситься на самостійне опрацювання та посилання на навчальну літературу	Кількість годин СРС
1	**Введення в дисципліну Text/ Data mining.** Поняття "інформація" і "знання", співставлення и порівняння цих понять. Сфери застосування інтелектуального аналізу даних.	4
2	**Методи розпізнавання образів.** Методи дисперсійного аналізу (МНК, ЛДА, КДА). Статистичні методи розпізнавання образів. Евристичні методи розпізнавання	4

	образів.	
3	**Методи підтримки прийняття рішень.** Основи теорії нечіткої логіки. Аналіз нечітких експертних заключень. Методу аналізу ієрархій. Методи прогнозування.	4
4	**Основи м'яких обчислень.** Методи м'яких обчислень. Багатоагентні системи Еволюційні обчислення. Генетичний алгоритм. Принципи побудови мурашиного алгоритму.	3
5	**Кореляційний та вейвлет- аналіз.** Автокореляційна функція. Узагальнене перетворення Фурьє. Вейвлет-перетворення.	4
6	**Фрактальний і мультифрактальний аналіз.** Фрактальна розмірність, абстрактні фрактали. Обчислення коефіцієнту Херста. Метод DFA. Поняття мультифрактального спектру та його інтерпретація.	4
7	**Основи прикладної теорія складних мереж.** Концепція малих світів.. Структура веб-простору. Ранжирування результатів пошуку з урахуванням структури мережі. Алгоритми HITS і PageRank.	4
8	**Мережевий аналіз даних і візуалізація.** Карти Кохонена. Граф горизонтальної видимості.	3

8. Індивідуальні завдання

Індивідуальних завдань не передбачено.

9. Контрольні роботи

Передбачається одна модульна контрольна робота, метою якої є перевірка та закріплення набутих аспірантами знань. Варіант контрольної роботи містить два теоретичні питання.

10. Рейтингова система оцінювання результатів навчання

Оцінка з дисципліни виставляється за багатобальною системою, з подальшим перерахуванням у 4-бальну.

2. Максимальна кількість балів з дисципліни дорівнює 100.

3. Нарахування балів по окремих видах робіт:

Рейтинг аспіранта з кредитного модуля складається з балів, що він отримав за:

1) виконання практичних робіт;
2) написання контрольної роботи (МКР);

Система рейтингових (вагових) балів та критерії оцінювання

1.Виконання практичних робіт

Оцінюються 8 робіт, передбачених робочою програмою. Максимальний ваговий бал $r_{ЛР}$ =64

Сума вагових балів практичних робіт:

N	Назва практичної роботи	Максимальний ваговий бал
1	Генерування псевдовипадкових чисел із заданим	9

	розподілом. Основи техніки програмування у середовищі MATLAB.	
2	Перевірка статистичних гіпотез. Вивчення можливостей пакету Statistical Toolbox у середовищі MATLAB.	9
3	Методи дисперсійного аналізу. Рішення задачі класифікації методом ЛДА.	9
4	Еволюційний алгоритм. Генетичний алгоритм.	9
5	Технологія моделювання за допомогою клітинних автоматів. Модель дифузії інформації.	9
6	Кореляційний аналіз. Розрахунок автокореляційної функції. Фрактальний аналіз. Розрахунок показника Херста.	9
7	Мережевий аналіз даних і візуалізація. 3D-візуалізація результатів аналізу числових рядів методом DFA. Вейвлет-аналіз у середовищі MATLAB.	9
Разом		63

Оцінювання практичних робіт:

–якщо робота виконана невчасно знімається 10-30% від максимальної кількості балів (кількість процентів залежить від терміну запізнення);

–якщо робота виконана не самостійно та простежується не індивідуальне виконання то знімається 50% від максимальної кількості балів;

–якщо в програмі не витримані основні правила створення програмних продуктів (модульність, дружній інтерфейс, наявність коментарів та т.п.) знімається 5%.

2.Модульний контроль

На одному з лекційних занять проводиться модульна контрольна робота: Максимальний ваговий бал $r_{МКР} = 11$.

Оцінювання модульної контрольної роботи виконується наступним чином:

–якщо на всі питання дані повні та чітко аргументовані відповіді, контрольна виконана охайно, з дотримання основних правил, то виставляється 9 - 11 балів;

–якщо методика виконання запропонованого завдання розроблена вірно, але допущені непринципові помилки у теоретичному описі або розрахунках, то виставляється 6 - 8 балів;

–від 3 до 5 балів нараховується, якщо методика виконання завдання розроблена в основному вірно, але допущені деякі з наступних помилок: помилки у представлені вихідних даних, не обґрунтовані теоретичні рішення, помилки у методиці розрахунків;

–нижче 3 балів нараховується, якщо завдання не виконане або допущені грубі помилки.

3. Залік

Залік відбувається у письмовій формі. Максимальна оцінка за залік складає $r_{ЕК} = 25$ балів.

Розрахунок шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру складає:

R=64 +11+25 = 100 балів

Таким чином, рейтингова шкала з кредитного модуля складає 100 балів.

Умови допуску до заліку: зарахування всіх лабораторних робіт, а також стартовий рейтинг r ≥ 40 балів.

Для отримання аспірантом відповідних оцінок (ECTS та традиційних) його рейтингова оцінка ***R*** переводиться згідно таблиці:

Шкала оцінювання:

1

За 100 – бальною шкалою	*За національною шкалою*
90 – 100	*Зараховано*
85 – 89	
75 – 84	
65 – 74	
60 – 64	
1 – 59	*не зараховано*

11. Методичні рекомендації

Для кращого засвоєння матеріалу дисципліни рекомендується використовувати на лекціях мультимедійні засоби навчання, які дозволяють інтенсифікувати навчальний процес, стимулювати розвиток мислення та уяви аспірантів, збільшувати обсяг навчального матеріалу для творчого засвоєння і використання його аспірантами, викликати зацікавленість та позитивне ставлення до навчання.

Методика побудована таким чином, що матеріал майже кожної лекції закріплюється виконанням завдання комп'ютерного практикуму. Завдання аспіранти отримують заздалегідь і на аудиторному занятті під керівництвом викладача виправляють помилки в разі їх наявності та відповідають на запитання щодо програмної реалізації та теоретичних засад роботи.

Якість самостійної роботи перевіряється на заняттях комп'ютерного практикуму.

12. Рекомендована література

12.1. Базова:

1. *Ситюк В.Є.* Прогнозування. Моделі. Методи. Алгоритми: Навчальний посібник. – К.: «Маклаут», 2008. – 364 с.
2. *Шумейко А.А., Сотник С.Л.* Интеллектуальный анализ данных (введение в Data Mining). – Днепропетровск: Белая Е.А., 2012. – 212 с.
3. *Ланде Д.В., Фурашев В.М.* Основи інформаційного і соціально-правового моделювання: монографія. – К.: ТОВ "ПанТот", 2012. – 144 с.
4. *В.П. Горбулін, О.Г. Додонов, Д.В. Ланде.* Інформаційні операції та безпека суспільства: загрози, протидія, моделювання: монографія. – К.: Інтертехнологія, 2009. – 164 с.
5. *Дебок Г.* Анализ финансовых данных с помощью самоорганизующихся карт / Г. Дебок, Т. Кохонен. − М.: Альпина, 2001.− 317 с.

6. *Саати Т. Л.* Принятие решений. Метод анализа иерархий. – М.: Радио и связь, 1989. – 316 с.
7. *Ландэ Д.В., Снарский А.А., Безсуднов И.В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. – М.: Либроком (Editorial URSS), 2009. – 264 с.
8. *Кетко Ю. Л., Кетков А. Ю., Шульц М. М*. MATLAB 7. Программирование численных методов. – С-Пб.: БХВ-Петербург, 2005. – 742 с.

12.2. Допоміжна:

9. *Прикладная статистика: Классификация и снижение размерности:* Справ. изд. / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. Под ред. С.А. Айвазяна, – М.: Финансы и статистика, 1989. – 607 с.
10. *Рассел С., Норвиг П.* Искусственный интеллект. Современный подход. – М.: Вильямс, – 2006. – 1408 с.
11. *Фогель Л., Оуенс А., Уолш М.* Искусственный интеллект и эволюционное моделирование. – М.: Мир, 1969. – 230 с.
12. *Koza J.R.* Genetic Programming: On the Programming of Computers by means of Natural Selection. – Cambridge MA, MIT Press, 1992.
13. *Ивахненко А.Г.* Долгосрочное прогнозирование и управление сложными системами. – К.: Техніка, 1975. – 312 с.
14. *Заде Л.* Понятие лингвистической переменной и ее применение к принятию приближенных решений. – М.: Мир, 1976. – 167 с.
15. *Нечеткие множества в моделях управления и искусственного интеллекта* / Под. ред. Д.А.Поспелова. – М.: Наука, 1986. – 312 с.
16. *Люгер Ф.Дж*. Искусственный интеллект. Стратегии и методы решения сложных систем. – М.: Вильямс, – 2003. – 864 с.
17. *Дуброва Т.А., Архипова М.Ю.* Статистические методы прогнозирования в экономике. Учебно-методический комплекс. – М.: Изд. Центр ЕАОИ, 2008. – 136 с.
18. *Семченок М.С., Семченок Н. М.* Система MATLAB. Часть 1.: Учебное пособие. – СПб.: изд. СПбГУКиТ, 2004. – 140 с.

КАТАЛОГ ПРОГРАМ ДЛЯ ПЕОМ

- ОС Windows 7
- Система MATLAB
- Система MathCad
- Текстовий редактор Notepad.
- Microsoft Office.