

НАЦІОНАЛЬНА АКАДЕМІЯ НАУК УКРАЇНИ  
ІНСТИТУТ ПРОБЛЕМ РЕЄСТРАЦІЇ ІНФОРМАЦІЇ

ЗАТВЕРДЖУЮ

Директор ІПРІ НАН України  
академік НАН України

В.В. Петров

«    » \_\_\_\_\_ 2019 р.



**КОНЦЕПТУАЛЬНІ ЗАСАДИ І ТЕХНОЛОГІЇ РОБОТИ  
З ВЕЛИКИМИ ДАНИМИ (BIG DATA)**

(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 «29» 08 2019 р. № 9)

Київ  
ІПРІ НАН України  
2019

Робоча програма кредитного модуля Концептуальні засади і технології роботи з великими даними (Big Data) (назва кредитного модуля)  
складена відповідно до програми навчальної дисципліни Концептуальні засади і технології роботи з великими даними (Big Data)

Розробник робочої програми:

старший науковий співробітник, к.т.н. Андрійчук Олег Валентинович  
(посада, науковий ступінь, вчене звання, прізвище, ім'я, по батькові)

  
(підпис)

## 1. Опис кредитного модуля

Рівень ВО, спеціальність, освітня програма, форма навчання	Загальні показники	Характеристика кредитного модуля
Рівень ВО третій (доктор філософії)	Оброблення надвеликих масивів даних	Лекції <u>16</u> год.
Спеціальність 122 Комп'ютерні науки	Цикл професійної підготовки	Практичні (семінарські) <u>14</u> год.
Освітньо-наукової програми спеціальності 122 «Комп'ютерні науки» галузі знань 12 «Інформаційні технології»	Статус кредитного модуля обов'язковий	Лабораторні роботи -
		Самостійна робота <u>60</u> год.,
	Семестр <u>4</u>	Індивідуальне завдання <i>Не передбачено</i>
Форма навчання: денна	Кількість кредитів (годин): 2 (90)	Вид та форма семестрового контролю: <i>залік</i>

Предмет навчальної дисципліни – серія підходів, інструментів і методів обробки структурованих і неструктурованих різноманітних даних великих розмірів для отримання результатів, які легко сприймаються людиною.

Дисципліна складається з одного кредитного модуля.

Вивчення дисципліни спирається на знання, отримані за програмою попередніх років навчання за спеціальністю 122 Комп'ютерні науки. Аспіранти мають досвід у імперативному, об'єктно-орієнтованому і функціональному програмуванні.

## 2. Мета та завдання кредитного модуля

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:

- проводити критичний аналіз, оцінку і синтез нових та складних ідей (СК-1);
- переосмислювати наявне та створювати нове цілісне знання та/або професійну практику і розв'язувати значущі соціальні, наукові, культурні, етичні та інші проблеми (СК-2);
- знаходити, обробляти й аналізувати необхідну інформацію для рішення проблем й прийняття рішень (ІК-3);
- використовувати сучасні методи і технології наукової комунікації на українській ті іноземній мовах (ІК-4);
- самостійно виконувати науково-дослідну діяльність у галузі знань "Інформаційні технології". з використанням сучасних теорій, методів та інформаційно-комунікаційних технологій (ПК-1);

- адаптувати і узагальнювати результати сучасних досліджень в галузі знань "Інформаційні технології" для вирішення наукових і практичних проблем (ПК-2);
- проводити теоретичні й експериментальні дослідження, математичне й комп'ютерне моделювання в галузі знань "Інформаційні технології" (ПК-3);
- узагальнювати результати науково-технічної діяльності, готувати науково-технічні публікації за результатами виконаних досліджень (ПК-4);
- аналізувати проблеми і визначати головні тенденції розвитку сучасних інформаційних технологій (ПК-1с);
- застосовувати інноваційні технології в сфері інформаційної безпеки (ПК-2с);
- обґрунтовано обирати математичний апарат для вирішення задач в галузі інформаційних технологій (ПК-3с);
- розробляти абстрактні моделі предметних областей із використанням різних парадигм програмування (ПК-4с);
- застосовувати високопродуктивні технології розподілених систем та паралельних обчислень при вирішенні науково-практичних задач (ПК-5с);
- аналізувати і обирати методи вирішення задач обчислювального (штучного) інтелекту із застосуванням нейромережних технологій, експертних систем та систем підтримки прийняття рішень (ПК-6с).

## 2.2. Основні завдання кредитного модуля.

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

### **знання :**

- системні знання в галузі природничих наук;
- міждисциплінарні знання в широкому контексті наукової та інноваційної діяльності;
- сучасних інформаційно-комунікаційних технологій, програмно-апаратних засобів проведення наукових досліджень;
- державної та іноземної мови, сучасної термінології у предметній області для наукового спілкування;
- фундаментальні природничо-наукові й математичні знання та теорії про будову і властивості матерії, фізичні процеси, на яких ґрунтується професійна та наукова діяльність у галузі інформаційних технологій;
- методології наукових досліджень у предметній області;
- сучасних методів планування та постановки експериментів;
- методів одержання, обробки та зберігання наукової інформації з використанням інформаційних технологій;
- можливостей використання інформаційних технологій для автоматизації експерименту, обробки даних, оформлення результатів досліджень;
- світових тенденції розвитку інноваційних та інформаційних технологій;
- сучасного стану інновацій у сфері інформаційної безпеки в провідних країнах світу;
- досвіду інформаційного забезпечення протидії інформаційним операціям в Україні та інших державах;

- теоретичних основ аналізу проблем наукових досліджень в галузі інформаційних технологій спеціального призначення;
- технологій інтелектуальних обчислень та аналізу даних.

#### уміння:

- самостійно добирати напрямки та методологію наукових досліджень, проводити критичний аналіз, узагальнювати нові знання та складні науково-технічні ідеї та рішення;
- самостійно визначати засоби розв'язання наукових задач з адаптацією до об'єкта та предмета досліджень;
- самостійно добирати напрямки та методологію наукових досліджень, проводити критичний аналіз, узагальнювати нові знання та складні науково-технічні ідеї та рішення;
- самостійно визначати засоби розв'язання наукових задач з адаптацією до об'єкта та предмета досліджень;
- аналізувати та планувати напрямки пошуку та добору науково-технічної інформації;
- готувати науково-технічні публікації;
- використовувати засоби збору та оброблення науково-технічної інформації для складання звітів та документації;
- проводити експерименти, аналізувати та узагальнювати їх результати;
- вибирати ефективні засоби інформаційних технологій для використання в науковій діяльності;
- використовувати інформаційні технології для підготовки традиційних та електронних наукових публікацій та презентацій;
- здійснювати автоматизацію експерименту, статистичну обробку даних, оформлення результатів досліджень засобами інформаційних технологій;
- здійснювати формалізацію поставлених задач, вибирати адекватні моделі та методи дослідження.

### 3. Структура кредитного модуля

Назви розділів і тем	Кількість годин				
	Всього	у тому числі			
		Лекції	Практичні	Лабораторні	СРС
1	2	3	4	5	6
<b><i>Розділ 1. Концепція “Великих даних”. Програмні засоби роботи з “Великими даними”</i></b>					
Тема 1. Поняття “великих даних”	4	1	-	-	3
Тема 2. Базові програмні засоби роботи з надвеликими масивами даних.	19	4	2	-	13
Тема 3. Програмна платформа розподілених даних Spark.	10	2	2	-	6
Разом за розділом 1	33	7	4	-	22
<b><i>Розділ 2. Машинне навчання на “Великих даних”</i></b>					
Тема 1. Основи машинного навчання	8	1	-	-	7
Тема 2. Навчання з учителем (Supervised Learning).	18	3	5	-	10

1	2	3	4	5	6
Тема 3. Навчання без вчителя (Unsupervised Learning).	13	2	3	-	8
Тема 4. Рекомендаційні системи.	12	2	2	-	8
Тема 5. Перспективні напрямки розвитку програмних засобів обробки надвеликих даних і машинного навчання.	6	1	-	-	5
Разом за розділом 2	57	9	10	-	38
<b>Всього годин</b>	90	16	14	-	60

#### 4. Лекційні заняття

№ з/п	Назва теми лекції та перелік основних питань
1	<p><b>Поняття великих даних. Програмний засіб Hadoop</b></p> <p>Основні аспекти та складові елементи трактування поняття “Великі дані”. Сфери застосування надвеликих масивів даних. Характеристики “великих даних”. Проблема масштабування. Базові компоненти аналізу даних. Поняття розподіленої файлової системи. Програмні моделі “великих даних”. Hadoop екосистема. Концепція MapReduce. Шаблони підсумовування (summarization). Шаблони фільтрації (filtering). Шаблони організації даних. Шаблони об’єднань (join). Меташаблони. Шаблони вводу/виведення.</p>
2	<p><b>Обробка текстової інформації та проектування ознак засобами Spark</b></p> <p>Передумови створення. Основні концепції та архітектура Spark. Призначення та використання SparkContext. Програмування з RDD. Використання RDD з парами ключ / значення. Завантаження та збереження даних. Partitioning та Shuffling. Структуровані та неструктуровані дані. Огляд Spark SQL. Бібліотеки Spark SQL. Запит із використанням Spark SQL. Додавання схеми до RDDs. RDDs як Relations. Використання Spark SQL для аналізу даних. Машинне навчання, його історичний розвиток і сучасний стан. Підходи до визначення поняття машинного навчання. Типи машинного навчання. Основні класи задач. Огляд алгоритмів машинного навчання. Приклади прикладних задач, які використовують методи машинного навчання. Етапи розробки моделі машинного навчання. Огляд бібліотеки Spark Mllib. Основи попередньої обробки даних. Нормалізація. Масштабування. Стандартизація числових даних. Токенізація текстових даних. TF-IDF. Word2Vec/ Spark Mllib типи даних. Transformer/ Estimator/ Pipeline API бібліотеки Mllib.</p>
3	<p><b>Навчання з учителем. Дерева рішень. Класифікатор Баєса. Метод опорних векторів</b></p> <p>Постановка задачі навчання з учителем. Основні класи задач. Задача регресії Лінійна регресія. Градієнтний спуск. Метод нормальних рівнянь. Задача класифікації. Логістична регресія. Бінарні дерева рішень. Навчання дерев рішень. Вхідні параметри про створенні дерев рішень та рандомних лісів. Ефект від налаштування специфічних параметрів дерев рішень. Bagging (Bootstrap Aggregation). Random Forests та Gradient-Boosted Trees. Generative Probabilistic моделі. Принцип максимуму апостеріорної вірогідності. Оцінювання щільності розподілу: три основних підходи. Наївний байесовський класифікатор. Метод опорних векторів. Оптимальна</p>

	розділювальна гіперплоща. Поняття зазора між класами (margin). Випадки лінійної роздільності та її відсутності. Поняття опорних векторів. Функція ядра (kernel functions).
4	<b>Нейронні мережі. Основні поняття</b> Біологічний нейрон, модель Маккаллока-Питтса як лінійний класифікатор. Функції активації. Проблема повноти. Завдання виключного або. Типи нейронних мереж. Класи задач. Сфери застосування. Навчання нейронних мереж: алгоритм зворотнього поширення похибки, проблеми навчання нейронних мереж, типи оптимізаторів, швидкі методи стохастичного градієнта, метод випадкових відключень нейронів (Dropout), інтерпретації Dropout.
5	<b>Навчання без учителя. Зниження розмірності</b> Постановка задачі навчання без учителя. Задача зниження розмірності. Сингулярний розклад матриці. Метод головних компонент. T-SNE алгоритм.
6	<b>Методи кластеризації. Карти самоорганізації.</b> Задача кластеризації. Типи методів кластеризації. Алгоритм k-середніх. Змішені моделі. Максимізація очікування. Ієрархічна кластеризація. Застосування методів кластеризації. Основні концепції. Конкурентне навчання. Архітектура. Топологія. Алгоритм навчання. Підходи до визначення відстаней. Застосування.
7	<b>Базові підходи до створення рекомендаційних систем</b> Види й області застосування рекомендаційних систем. Огляд підходів до створення рекомендаційних систем. Алгоритм SVD. Вимірювання якості рекомендацій. Задача колаборативної фільтрації. Реалізація рекомендаційної системи засобами Spark: memory-based collaborative filtering, model-based collaborative filtering, архітектура рекомендаційної системи, екстракція даних, навчання моделі, застосування отриманого програмного рішення.
8	<b>Перспективні напрямки розвитку програмних засобів обробки надвеликих даних і машинного навчання. Програмні рішення обробки надвеликих даних в задачах машинного навчання від лідерів ринку (Amazon, Google, Facebook)</b> Сучасні напрямки в машинному навчанні. Automated machine learning. Generative Adversarial Networks. Спеціалізоване апаратне забезпечення. Cloud Object Storage. Платформи для машинного навчання та роботи з надвеликими даними. Машинне навчання як сервіс. Машинне навчання на AWS. Azure Machine Learning Packages. Google Cloud ML Engine. IBM Data Science Experience. FBLearner.

## 5. Практичні заняття

Практичних занять не передбачено.

## 5. Семінарські заняття

Основні завдання циклу семінарських занять полягають у формуванні вмінь та навичок у аспірантів самостійно систематизувати і поглиблювати знання концептуальних засад та технологій роботи з великими даними.

№ з/п	Назва теми заняття
1	Використання шаблонів проектування MapReduce
2	Затосування SparkSQL, робота з даними з використанням Dataframes та DataSet. Обробка текстової інформації та проектування ознак засобами Spark
3	Задачі класифікації та регресії з використанням бібліотеки машинного навчання MLlib
4	Дерева рішень та ансамблеві алгоритми навчання
5	Алгоритми кластеризації
6	Алгоритми зниження розмірності
7	Створення рекомендаційних систем засобами Spark MLlib

## 6. Лабораторні заняття

Лабораторних занять не передбачено.

## 7. Самостійна робота

№ з/п	Назви тем і питань, що виносяться на самостійне опрацювання та посилання на навчальну літературу	Кількість годин СРС
1	<b>Поняття “великих даних”</b> Враливі технології великих даних на сучасне програмне забезпечення [11]. Програмна реалізація шаблонів проектування Mapreduce [1]	6
2	<b>Базові програмні засоби роботи з надвеликими масивами даних.</b> Програмна екосистема Hadoop. Практичне застосування Hadoop в задачах Data Science. Огляд HiveQL, Pig, Yarn, Hbase. Аналітика даних з використанням Spark. [2, 4, 5, 7]	15
3	<b>Програмна платформа розподілених даних Spark</b> <b>Модель паралельних обчислень Spark.</b> Spark Job Scheduling. DataFrame API. Представлення даних в DataFrames and Datasets. Core Spark Joins. Ефективні трансформації. [5]	7
4	<b>Основи машинного навчання</b> Теоретичні засади машинного навчання [9]. Кодування та підготовка даних Mllib. Масштабування та вибір характеристик. MLlib Model Training. Оцінка моделі машинного навчання. [5]	4
5	<b>Навчання з учителем (Supervised Learning).</b> Метод стохастичного градієнта SG. Логістична регресія. Принцип максимуму правдоподібності і логарифмічна функція втрат. Метод стохастичного градієнта для логарифмічної функції втрат. Математичні основи методу опорних векторів.	10



	Завдання квадратичного програмування і двоїста задача. Побудова ядер для методу опорних векторів. Критерії якості класифікації: чутливість і специфічність, ROC-крива і AUC, точність і повнота, AUC-PR. [6, 9, 13, 16]	
6	<b>Навчання без вчителя (Unsupervised Learning).</b> Постановка завдання кластеризації. Постановка завдання Semisupervised Learning, приклади застосування. Алгоритм k-середніх і EM-алгоритм для поділу гауссовської суміші. Агломеративної кластеризація, Алгоритм Ланса-Вільямса та його окремі випадки. Алгоритм побудови дендрограми. Визначення числа кластерів. Сингулярний розклад в задачі зниження розмірності. Alternating Least Squares з використанням Spark [6, 9, 13, 14, 16]	7
7	<b>Рекомендаційні системи.</b> Завдання колаборативної фільтрації, транзакційні дані і матриця суб'єкти-об'єкти. Collaborative Filtering and Recommendation. [5, 3, 9, 13]	7
8	<b>Перспективні напрямки розвитку програмних засобів обробки надвеликих даних і машинного навчання.</b> Ризики, пов'язані з застосуванням надвеликих даних. Проблема конфіденційності. Датифікація. [11]	4

### 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

1. Оцінка з дисципліни виставляється за 100 бальною системою.
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) участь у семінарських заняттях;
- 2) написання контрольної роботи (МКР);

### Система рейтингових (вагових) балів та критерії оцінювання

#### 1. Участь у семінарських заняттях

Оцінюються участь у 7-ми семінарських заняттях, передбачених робочою програмою. Максимальний ваговий бал  $r_c = 64$

Сума вагових балів семінарських занять:

№	Назва теми семінарського заняття	Максимальний ваговий бал
1	Використання шаблонів проектування MapReduce	8
2	Затосування SparkSQL, робота з даними з використанням Dataframes та DataSet. Обробка текстової інформації та проектування ознак засобами Spark	16
3	Задачі класифікації та регресії з використанням бібліотеки машинного навчання MLlib	8
4	Дерева рішень та ансамблеві алгоритми навчання	8
5	Алгоритми кластеризації	8
6	Алгоритми зниження розмірності	8
7	Створення рекомендаційних систем засобами Spark MLlib	8
Разом		64

Критерії оцінювання роботи на семінарському занятті:

а) послідовний виклад матеріалу з висновками в кінці відповіді, знання фактичного матеріалу, володіння понятійним апаратом, уміння творчо розв'язувати завдання – 90-100% балів;

б) послідовний виклад матеріалу, знання фактичного матеріалу, володіння категоріальним апаратом, допускається 1-2 неточності у використанні понятійного матеріалу – 75-89% балів,

в) зміст матеріалу викладено частково, з недотриманням в окремих випадках послідовності, аспірант частково володіє знаннями фактичного матеріалу, при поясненні понять допущено помилки, суть питання в основному розкрита, не зважаючи на зазначені вище упущення – 60-74% балів;

Доповнення та участь у дискусії на семінарському занятті оцінюється у 1 бал.

#### 2. Модульний контроль

На одному з лекційних занять проводиться модульна контрольна робота: Максимальний ваговий бал  $r_{МКР} = 16$ .

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 15 - 16 балів;

– якщо методика виконання запропонованого завдання розроблена вірно, але допущені непринципові помилки у теоретичному описі або розрахунках, то виставляється 13 - 14 балів;

– від 10 до 12 балів нараховується, якщо методика виконання завдання розроблена в основному вірно, але допущені деякі з наступних помилок: помилки у представленні вихідних даних, не обґрунтовані теоретичні рішення, помилки у методиці розрахунків;

– нижче 10 балів нараховується, якщо завдання не виконане або допущені грубі помилки.

### 3. Залік

Залік відбувається у письмовій формі. Максимальна оцінка за залік складає  $r_3 = 20$  балів.

#### ***Розрахунок шкали (R) рейтингу:***

Сума вагових балів контрольних заходів протягом семестру складає:

$$R = 64 + 16 + 20 = 100 \text{ балів}$$

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

Умови допуску до заліку: стартовий рейтинг  $r \geq 48$  балів.

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

#### Шкала оцінювання:

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

## 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 участю в семінарських заняттях. Завдання аспіранти отримують заздалегідь і на аудиторному занятті під керівництвом викладача виправляють помилки в разі їх наявності та відповідають на запитання щодо теоретичних та технологічних засад.

Якість самостійної роботи перевіряється на семінарських заняттях.

## 12. Рекомендована література

### 12.1. Базова

1. Miner, Donald, and Adam Shook. Mapreduce Design Patterns. Beijing: O'Reilly, 2012.
2. Lam, Chuck. Hadoop in Action. Greenwich, Conn: Manning Publications, 2011.
3. Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell Learning Spark Lightning-Fast Big Data Analysis: O'Reilly Media, 2015, 276 p.
4. Holden Karau, Rachel Warren High Performance Spark: Best Practices for Scaling and Optimizing Apache Spark : O'Reilly Media, 2017, 358 p.
5. Sandy Ryza, Uri Laserson, Josh Wills, Sean Owen Advanced Analytics with Spark, 2nd Edition Patterns for Learning from Data at Scale, O'Reilly Media, 2017, 280p.
6. Stephen Marsland Machine Learning: An Algorithmic Perspective, Chapman & Hall/CRC, 2009m 390 p.
7. Додонов А.Г., Ландэ Д.В., Цыганок В.В., Андрійчук О.В., Каденко С.В., Грайворонская А.Н., Распознавание информационных операций, К.: ООО «Инжиниринг», 2017. – 282 с.

### 12.2. Допоміжна

8. Holmes, Alex. Hadoop in Practice. Shelter Island, NY: Manning, 2012.
9. Alpaydin, Ethem. Introduction to Machine Learning., 2014.
10. Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. Cambridge, MA: MIT Press, 2012.
11. Shalev-Shwartz, Shai, and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms., 2014.
12. Tom White Hadoop: The Definitive Guide, 4th Edition O'Reilly Media. 2015 – 756p.
13. Witten, I H, and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman, 2005.
14. Cherkassky, Vladimir S, and Filip Mulier. Learning from Data: Concepts, Theory, and Methods. Hoboken, N.J: IEEE Press, 2007.

15. Marsland, Stephen. Machine Learning: An Algorithmic Perspective. Boca Raton: CRC Press, 2009. Print.
16. Harrington, Peter. Machine Learning in Action. Shelter Island, NY: Manning Publications, 2012.
17. Виктор Маер-Шенбергер, Кеннет Кукьер Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим М.: «Манн, Иванов и Фербер», 2013, 240 с. ISBN 978-5-91657-936-9

### **13. Інформаційні ресурси**

1. Welcome to Apache™ Hadoop®: <http://hadoop.apache.org/>
2. Apache Spark™ is a unified analytics engine for large-scale data processing: <https://spark.apache.org/>
3. Machine Learning Library (MLlib) Guide: <https://spark.apache.org/docs/latest/ml-guide.html>