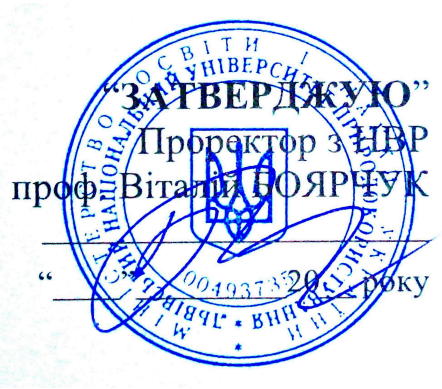
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Львівський національний університет природокористування

Факультет механіки, енергетики та інформаційних технологій

Кафедра інформаційних технологій



РОБОЧА ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

ПІДГОТОВКА ТА ОБРОБКА ВЕЛИКИХ ДАНИХ

спеціальність 122 «Комп'ютерні науки»
перший (бакалаврський) рівень вищої освіти

Львів 2024 р.

Робоча програма із дисципліни «Підготовка та обробка великих даних» для здобувачів першого (бакалаврського) рівня вищої освіти ОП «Комп'ютерні науки» спеціальності 122 «Комп'ютерні науки»

Розробник: Тригуба А.М., д.т.н., професор

Робочу програму схвалено на засіданні кафедри **Інформаційних технологій**
Протокол № 1 від 12 серпня 2024 року

Завідувач кафедри **Інформаційних технологій**

_____ (Тригуба А.М.)
(підпис) (прізвище та ініціали)

Робочу програму схвалено на засіданні методичної комісії факультету механіки, енергетики та інформаційних технологій Протокол № 1 від 29 серпня 2024 року.

Голова методичної комісії факультету механіки, енергетики та інформаційних технологій.

_____ (Ковалишин С.Й.)
(підпис) (прізвище та ініціали)

1. Опис навчальної дисципліни

Галузь знань, спеціальність, освітня програма, рівень вищої освіти,

Галузь знань 12 – інформаційні технології
(шифр і назва)
Спеціальності 122 «Комп'ютерні науки»
Освітня програма «Комп'ютерні науки»
(шифр і назва)
Рівень вищої освіти: перший (бакалаврський)

Характеристика навчальної дисципліни:
Вибіркова
Кількість кредитів 4
Загальна кількість годин – 120
Індивідуальне науково-дослідне завдання –
(назва
Вид контролю: іспит
Тижневих аудиторних годин для денної форми навчання – 4
Співвідношення кількості годин аудиторних занять до самостійної і індивідуальної роботи становить (%):
для денної форми навчання – 66,7%
для заочної форми навчання – 15,4%

1. Програма навчальної дисципліни

Тема 1. Аналіз та обробка великих даних.
1.1. Актуальність даних в сучасному світі.
1.2. Поняття великих даних.
1.3. Характеристики великих даних.
1.4. Методи аналізу та обробки великих даних.
Тема 2. Методи та засоби інтеграції даних.
1.1. Актуальність задачі обробки різнотипних даних.
1.2. Визначення поняття «консолідація» та його зв'язок з процесами обробки різнотипних даних.
1.3. Методи обробки даних з різних джерел.
1.4. Стандарти обробки різнотипних даних. Засоби інтеграції даних з різнотипних джерел.
Тема 3. Методи та засоби забезпечення якості даних.
3.1. Формулювання причин забезпечення якості консолідованих даних.
3.2. Стандарти забезпечення якості даних. Моделі оцінювання якості великих даних.
Тема 4. Відновлення відсутніх даних.
4.1. Аналіз методів та засобів відновлення відсутніх даних в наборах великих даних.
4.2. Модель великих даних для задачі відновлення відсутніх даних.
4.3. Метод відновлення відсутніх даних з використанням продукційних залежностей і правил асоціації.
4.4. Оцінка складності алгоритмів відновлення відсутніх даних.
Тема 5. Бази даних та сховища даних.
5.1. Особливості баз даних NoSQL.

5.2. Організація та принципи функціонування MongoDB, PostgreSQL.
5.3. Дослідження швидкодії роботи різних баз даних.
5.4. Обробка запитів на локальній машині.
5.5. Обробка запитів на машині на сервісах Amazon.

Тема 6. Розподілені платформи.

6.1. Інфраструктура розподілених обчислень Hadoop.
6.2. Інфраструктура розподілених обчислень Spark.
6.3. Інфраструктура розподілених обчислень Dryad.
6.4. Інфраструктура розподілених обчислень Message Passing Interface.

Тема 7. Інструментальні засоби.

7.1. Вимоги до програмних систем.
7.2. Система Rapid Miner. Система Orange. Система Weka. Система Knime. Система Alteryx.
7.3. Мови програмування: Java, Scala, R, Python. Порівняльний аналіз мов програмування.
7.4. Бібліотеки мови програмування Python.

Тема 8. Застосування глибоких нейронних мереж для аналізу та обробки великих даних.

8.1. Особливості архітектури глибоких нейронних мереж.
8.2. Навчання глибоких нейронних мереж.
8.3. Застосування глибоких нейронних мереж для аналізу та обробки великих даних.

Тема 9. Методи аналізу великих даних на основі глибоких нейронних мереж.

9.1. Класифікація мережевих пакетів на основі глибоких нейронних мереж.
9.2. Розпізнавання об’єктів на зображеннях супутникових знімків.
9.3. Розпізнавання об’єктів на зображеннях текстових документів.

Тема 10. Методи та засоби підвищення ефективності підготовки та обробки великих даних.

10.1. Створення та функціонування глибоких нейронних мереж на основі еволюційного підходу.
10.2. Підвищення швидкодії нейронної мережі.
10.3. Підвищення швидкодії прийняття рішень на основі нечіткої логіки.

Тема 11. Побудова інформаційної технології підготовки та обробки великих даних.

11.1. Структурна модель інформаційної технології підготовки та обробки великих даних.
11.2. Особливості системи підготовки та обробки великих даних BigDDL.
11.3. Архітектура системи BigDDL. Приклад застосування.
11.4. Оцінка ефективності алгоритмів підготовки та обробки великих даних.

Тема 12. Когнітивний аналіз даних.

12.1. Характеристика технологій когнітивного аналізу даних.
12.2. Огляд проекту DeepQA.
12.3. Когнітивна система типу IBM Watson.
12.4. Функції та можливості системи IBM Watson.

3. Структура навчальної дисципліни

Назви тем	Кількість годин											
	денна форма						заочна форма					
	усього	у тому числі					усього	у тому числі				
		л	п	лаб.	інд.	с. р.		л	п	лаб.	інд.	с. р.
1	2	3	4	5	6	7	8	9	10	11	12	13
	Рік підготовки 4 Семестр 8						Рік підготовки 4 Семестр 8					
Тема 1	8	2	–	2	–	4	8	1	–	–	–	7
Тема 2	7	2	–	2	–	3	7	–	–	1	–	6
Тема 3	8	2	–	2	–	4	8	1	–	–	–	7
Тема 4	7	2	–	2	–	3	7	–	–	1	–	6
Тема 5	8	2	–	2	–	4	8	1	–	1	–	6
Тема 6	7	2	–	2	–	3	7	–	–	1	–	6
Тема 7	8	2	–	2	–	4	8	1	–	1	–	6
Тема 8	7	2	–	2	–	3	7	–	–	1	–	6
Тема 9	7	2	–	2	–	3	7	1	–	–	–	6
Тема 10	8	2	–	2	–	4	8	1	–	1	–	6
Тема 11	7	2	–	2	–	3	7	1	–	–	–	6
Тема 12	8	2	–	2	–	4	8	1	–	1	–	6
Іспит	30	–	–	–	–	30	30	–	–	–	–	30
Разом за семестр	120	24	0	24	0	72	120	8	0	8	0	104
Індивідуальні завдання												
–	–	–	–	–	–	–	–	–	–	–	–	–
Усього годин	120	24	–	24	–	72	120	8	–	8	–	104

4. Перелік лабораторних занять

№ з/п	Назва теми	Кількість годин
1	Методи підготовки та обробки даних.	4
2	Алгоритми кластеризації великих даних.	2
3	Розподілена платформа Apache Hadoop.	4
4	Розподілена платформа Apache Spark.	2
5	Налаштування та запуск кластера Hadoop, YARN та Spark.	4
6	Реалізація методів аналізу та обробки даних на Python.	2
7	Використання нейронних мереж для підготовки та обробки даних.	4
8	Когнітивна система IBM Watson.	2
Разом		24

5. Теми, питання та завдання, винесені на самостійне вивчення

№ п/п	Назва теми
1.	Основні характеристики великих даних.
2.	Роль великих даних в різних галузях.
3.	Консолідація даних.
4.	Візуалізація даних, Gaphi.
5.	Основні конструкції мови R, консолідація даних, візуалізація.
6.	HDFS - основи організації.
7.	Архітектура Hadoop.
8.	Виконання Map/Reduce.
9.	Виконання програм в Hadoop.
10.	Основи YARN.
11.	Аналітика потокових даних в платформі Storm.
12.	Архітектура Apache Spark.
13.	Організація даних в Apache Spark.
14.	Обробка даних в GraphX
15.	Алгоритми класифікації.
16.	Нейронні мережі як реалізація алгоритмів машинного навчання
17.	Інтелектуальні алгоритми.

6. Індивідуальні завдання

Комплексні індивідуальні завдання (КІЗ) з дисципліни «Підготовка та обробка великих даних» виконуються самостійно кожним студентом і охоплюють усі основні теми лабораторних робіт дисципліни. КІЗ оформляється у відповідності з встановленими вимогами. Виконання КІЗ є одним із обов'язкових складових модулів залікового кредиту. КІЗ пов'язане з підготовкою та обробкою великих даних. Підготовка та обробка даних виконується у відповідності до наборів даних офіційного сайту Державної служби статистики України (ukrstat.gov.ua) (розділ «Статистична інформація»):

Варіанти КІЗ з дисципліни «Підготовка та обробка великих даних»:

1. Будівництво – http://www.ukrstat.gov.ua/operativ/menu/menu_u/bud.htm
2. Внутрішня торгівля – http://www.ukrstat.gov.ua/operativ/menu/menu_u/spr.htm
3. Державні фінанси, податки та публічний сектор – http://www.ukrstat.gov.ua/operativ/menu/menu_u/ekon/publ_u.htm
4. Діяльність підприємств – http://www.ukrstat.gov.ua/operativ/menu/menu_u/sze.htm
5. Доходи та умови життя – http://www.ukrstat.gov.ua/operativ/menu/menu_u/virdg.htm
6. Енергетика – http://www.ukrstat.gov.ua/operativ/menu/menu_u/energ.htm
7. Зовнішньоекономічна діяльність – http://www.ukrstat.gov.ua/operativ/menu/menu_u/zed.htm
8. Інформаційне суспільство – http://www.ukrstat.gov.ua/operativ/menu/menu_u/zv.htm
9. Капітальні інвестиції – http://www.ukrstat.gov.ua/operativ/menu/menu_u/ioz.htm
10. Комплексна статистика – http://www.ukrstat.gov.ua/operativ/menu/menu_u/mp.htm
11. Культура – http://www.ukrstat.gov.ua/operativ/menu/menu_u/cult.htm
12. Макроекономічна статистика – http://www.ukrstat.gov.ua/operativ/menu/menu_u/tda.htm

13. Навколишнє середовище – http://www.ukrstat.gov.ua/operativ/menu/menu_u/ns.htm
14. Населені пункти та житло – http://www.ukrstat.gov.ua/operativ/menu/menu_u/if.htm
15. Наука, технології та інновації – http://www.ukrstat.gov.ua/operativ/menu/menu_u/ni.htm
16. Національні рахунки – http://www.ukrstat.gov.ua/operativ/menu/menu_u/nac_r.htm
17. Освіта – http://www.ukrstat.gov.ua/operativ/menu/menu_u/osv.htm.
18. Основні засоби – http://www.ukrstat.gov.ua/operativ/menu/menu_u/voz.htm
19. Охорона здоров'я – http://www.ukrstat.gov.ua/operativ/menu/menu_u/oz.htm
20. Правосуддя та злочинність – http://www.ukrstat.gov.ua/operativ/menu/menu_u/ppr.htm
21. Промисловість – http://www.ukrstat.gov.ua/operativ/menu/menu_u/prom.htm
22. Реєстр статистичних одиниць – http://www.ukrstat.gov.ua/operativ/operativ2013/kap/kap.htm
23. Ринок праці – http://www.ukrstat.gov.ua/operativ/menu/menu_u/dem/r_pr.htm
24. Сільське, лісове та рибне господарство – http://www.ukrstat.gov.ua/operativ/menu/menu_u/cg.htm
25. Соціальний захист – http://www.ukrstat.gov.ua/operativ/menu/menu_u/sz.htm
26. Транспорт – http://www.ukrstat.gov.ua/operativ/menu/menu_u/tr.htm
27. Туризм – http://www.ukrstat.gov.ua/operativ/menu/menu_u/tur.htm
28. Ціни – http://www.ukrstat.gov.ua/operativ/menu/menu_u/cit.htm.

7. Методи навчання

1. Словесні методи (розповідь, пояснення, бесіда, лекція.)
2. Наочні методи
- ілюстрація (презентації, таблиці, моделі, муляжі, малюнки тощо),
- демонстрування засобу демонстрування: навчальна телепередача або кіно-відеофільм чи його фрагмент; діюча модель, дослід; експеримент, спостереження та досліди в практичних умовах тощо,

3. Практичні методи: лабораторні та самостійні роботи.

8. Методи контролю

1. Усне опитування: фронтальне, індивідуальне.
2. Письмова аудиторна та позааудиторна перевірка: рішення задач із інтелектуального аналізу даних, контрольні роботи.
3. Практична перевірка: виконання практичних робіт, рішення ситуаційних завдань.
4. Стандартизований контроль: тести.
Види контролю: Поточний контроль, проміжна та семестрова атестація.

Політика оцінювання

Політика щодо дедлайнів та перескладання: Роботи, які здаються із порушенням термінів без поважних причин, оцінюються на нижчу оцінку (75% від можливої максимальної кількості балів). Перескладання проміжних модулів відбувається за наявності поважних причин (наприклад, лікарняний).
Політика щодо академічної доброчесності: Списування під час тестування, виконання контрольних робіт або підсумкового контролю заборонені (в т.ч. із використанням мобільних девайсів та генеративного інтелекту). Мобільні пристрої дозволяється технічно використовувати лише під час он-лайн тестування та підготовки до виконання завдань.

Політика щодо відвідування: Відвідування занять є обов'язковим компонентом оцінювання. За об'єктивних причин (наприклад, хвороба, працевлаштування, міжнародне стажування) навчання може відбуватись в он-лайн формі за погодженням із керівником курсу.

9. Результати навчання

Формування програмних компетентностей

Індекс в матриці ОПП	Програмні компоненти
ЗК6	Здатність вчитися й оволодівати сучасними знаннями.
ЗК7	Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
СК2	Здатність до виявлення статистичних закономірностей недетермінованих явищ, застосування методів обчислювального інтелекту, зокрема статистичної, нейромережевої та нечіткої обробки даних, методів машинного навчання та генетичного програмування тощо.
СК3	Здатність до логічного мислення, побудови логічних висновків, використання формальних мов і моделей алгоритмічних обчислень, проектування, розроблення й аналізу алгоритмів, оцінювання їх ефективності та складності, розв'язності та нерозв'язності алгоритмічних проблем для адекватного моделювання предметних областей і створення програмних та інформаційних систем.
СК11	Здатність до інтелектуального аналізу даних на основі методів обчислювального інтелекту включно з великими та погано структурованими даними, їхньої оперативної обробки та візуалізації результатів аналізу в процесі розв'язування прикладних задач.
ПРН3	Використовувати знання закономірностей випадкових явищ, їх властивостей та операцій над ними, моделей випадкових процесів та сучасних програмних середовищ для розв'язування задач статистичної обробки даних і побудови прогнозних моделей.
ПРН5	Проектувати, розробляти та аналізувати алгоритми розв'язання обчислювальних та логічних задач, оцінювати ефективність та складність алгоритмів на основі застосування формальних моделей алгоритмів та обчислюваних функцій.
ПРН12	Застосовувати методи та алгоритми обчислювального інтелекту та інтелектуального аналізу даних в задачах класифікації, прогнозування, кластерного аналізу, пошуку асоціативних правил з використанням програмних інструментів підтримки багатовимірного аналізу даних на основі технологій DataMining, TextMining, WebMining.

10. Розподіл балів, які отримують студенти

Остаточна оцінка за курс розраховується наступним чином: поточний контроль оцінюється в 50 балів, та складається із двох модулів по 25 балів кожен. В суму балів кожного модуля входять бали за підготовку, виконання та захисту 8 лабораторних робіт по 5 бали за кожну роботу (8 х 5 = 40) та 10 балів за індивідуальну роботу, яка оцінюється під час її захисту (співбесіда із викладачем).

Поточне тестування та самостійна робота (разом 50 балів)		Самостійна робота	Підсумковий контроль	Сума
Модуль 1 (20 балів)	**Модуль 2** (20 балів)		екзамен	
Л1- Л4	Л5- Л8			
4 х 5 =**20**	4 х 5 =**20**	**10**	**50**	**100**

Л1, Л2 ... Л8 – лабораторні роботи; СР – самостійна робота.

11. Методичне забезпечення

Підручники і навчальні посібники; інструктивно-методичні матеріали до лабораторних занять; індивідуальні навчально-дослідні завдання; контрольні роботи; текстові та електронні варіанти тестів для поточного контролю, методичні матеріали для організації самостійної роботи студентів, виконання індивідуальних завдань.

12. Рекомендована література

Базова

1. Томас Єрл, Ваджид Хаттак, Пол Булер. Основи Big Data: Концепції, алгоритми та технології / Пер.з англ. Анатолія Гладуна; За наук. ред. Олексія Найди. Дніпро: «Баланс Бізнес Букс», 2018. 320 с.
2. Кучеров Д.П. Методи аналізу великих даних «Big Data». Київ. 2020. 237 с.
3. Ланде, Д. В. Оброблення надвеликих масивів даних (Big Data) [Електронний ресурс] : навчальний посібник для використання у навчальному процесі з підготовки фахівців другого (магістерського) рівня вищої освіти зі спеціальності 122 «Комп'ютерні науки» / Д. В. Ланде, І.Ю. Субач, А. Я. Гладун ; КПІ ім. Ігоря Сікорського. Електронні текстові дані. Київ : КПІ ім. Ігоря Сікорського, 2021. 168 с.
4. Навчальний посібник з дисципліни "Технології Big Data" для студентів спеціальності 123 "Комп'ютерна інженерія" / Таран В.І., Гордієнко Ю.Г., Стіренко С.Г. Київ: КПІ, 2022. 56 с.

Допоміжна

5. Технології оброблення великих даних: конспект лекцій з дисципліни «Технології оброблення великих даних» [Електронний ресурс] : навч. посіб. / Л.М. Олещенко; КПІ ім. Ігоря Сікорського. Електронні текстові дані. Київ: КПІ ім. Ігоря Сікорського, 2021. 227 с.
6. Zgurovsky M.Z., Zaychenko Y.P. Big Data: Conceptual Analysis and Applications. Springer, 2020. 298 p.
7. Wiktorski Tomasz. Data-intensive Systems: Principles and Fundamentals using Hadoop and Spark. Springer, 2019. 105 p.
8. Wang C., Shakhovska N., Sachenko A., Komar M. A New Approach for Missing Data Imputation in Big Data Interface. Information Technology and Control. 2020. Vol. 49. No 4. P. 541-555.
9. Комплект методичних посібників виданих кафедрою, конспект лекцій.

Інформаційні ресурси в Інтернеті

10. The latest in machine learning. Papers With Code [Електронний ресурс]. Електрон. дан. Режим доступу: World Wide Web. URL: https://paperswithcode.com/

11. Платформа для змагань з аналітики та передбачувального моделювання. [Електронний ресурс]. Режим доступу: https://www.kaggle.com/
12. Портал відкритих даних України. [Електронний ресурс]. Режим доступу: https://data.gov.ua/
13. Shaw J. Why “Big Data” Is a Big Deal [Електронний ресурс]. Режим доступу: http://harvardmag.com/pdf/2014/03- pdfs/0314-HarvardMag.pdf
14. Schutt P. What is Big Data? [Електронний ресурс]. Режим доступу: https://blogs.oracle.com/bigdata/big-data-andanalytic-top-10-trends-for-2014/
15. Відкритий посібник з відкритих даних [Електронний ресурс]. Режим доступу: https://socialdata.org.ua/manual/
16. Big Data Та Блокчейн – Прорив В Області Аналізу Даних [Електронний ресурс]. Режим доступу: https://busines.in.ua/big-data-ta-blokchejn-proryv-v-oblasti-analizu-danyh/
17. Weka Machine learning software to solve data mining problems [Електронний ресурс]. Режим доступу: https://sourceforge.net/projects/weka/?source=typ_redirect
18. Books Ngram Viewer [Електронний ресурс]. Режим доступу: https://books.google.com/ngrams
19. Мова програмування R [Електронний ресурс]. Режим доступу: https://cran.r-project.org
20. Середовище для розробки програм на R – R Studio [Електронний ресурс]. Режим доступу: http://www.r-studio.com
21. Manyika James and others. Big data: The next frontier for innovation, competition, and productivity [Електронний ресурс]. Режим доступу: http://www.mckinsey.com/business-functions/businesstechnology/our-insights/big-data-the-next-frontier-for-innovation
22. IBM Analytics [Електронний ресурс]. Режим доступу: http://www.ibm.com/analytics/us/en/technology/hadoop/hadoop-trials.html
23. IBM Cloud [Електронний ресурс]. Режим доступу: https://www.ibm.com/cloud-computing/bluemix/?lnk=hp_trials_uauk
24. IBM Bluemix Promo Code - 6 Month Trial [Електронний ресурс]. Режим доступу: https://ibm.onthehub.com/WebStore/OfferingDetails.aspx?o=bb3528b7-2b63-e611-9420-b8ca3a5db7a1
25. Hadoop: Built for big data, insights, and innovation [Електронний ресурс]. Режим досупу: http://www.ibm.com/analytics/us/en/technology/hadoop/
26. IBM BigInsights [Електронний ресурс]. Режим доступу: http://www.ibm.com/analytics/us/en/technology/biginsights/