МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Львівський національний університет імені Івана Франка
Механіко-математичний факультет
Кафедра математичної статистики і диференціальних рівнянь

Затверджено
На засіданні
кафедри математичної статистики і диференціальних рівнянь
механіко-математичного факультету
Львівського національного університету імені Івана Франка
(протокол № 1 від 22.06.2023 р.)

Завідувач кафедри: Олег БУГРІЙ

Силабус з навчальної дисципліни
“ Аналіз великих даних ”,
що викладається в межах ОПП “ Статистичний аналіз даних ”
першого (бакалаврського) рівня вищої освіти для здобувачів з спеціальності 112 - Статистика

Львів 2023 р.

Назва дисципліни	Аналіз великих даних
Адреса викладання дисципліни	Головний корпус ЛНУ ім. І. Франка м. Львів, вул. Університетська 1, 79000
Факультет та кафедра, за якою закріплена дисципліна	Механіко-математичний факультет Кафедра математичної статистики і диференціальних рівнянь
Галузь знань, шифр та назва спеціальності	11 - Математика та статистика 112 - Статистика
Викладачі дисципліни	Холявка О.Т., кандидат фізико-математичних наук, асистент кафедри математичної статистики і диференціальних рівнянь
Контактна інформація викладачів	oksana.kholiavka@lnu.edu.ua, https://new.mmf.lnu.edu.ua/employee/kholyavka_o_t
Консультації з питань навчання по дисципліні відбуваються	Консультації в день проведення лекцій/практичних занять (за попередньою домовленістю). Головний корпус ЛНУ ім. І. Франка, каб. 267. м. Львів, вул. Університетська, 1
Сторінка курсу	https://new.mmf.lnu.edu.ua/course/aveld-112-bak23
Інформація про дисципліну	Дисципліна "Аналіз великих даних" є вибірковою дисципліною зі спеціальності 112 – Статистика для освітньої програми "Статистичний аналіз даних", яка викладається в 8-му семестрі в обсязі 5-х кредитів (за Європейською Кредитно-Трансферною Системою ECTS).
Коротка анотація дисципліни	Курс розроблено для ознайомлення студентів із теоретичними основами та можливостями практичного застосування методів аналізу великих даних для дослідження процесів та систем різного призначення.
Мета та цілі дисципліни	*Мета:* ознайомити студентів з підходами, що використовуються при аналізі великих даних. *Цілі:* навчити студентів правильно структурувати код для паралелізації; масштабувати обчислення; використовувати бібліотеки для розподілених обчислень Dask та Spark.
Література для вивчення дисципліни	1) Ivan Marin, Ankit Shukla, Sarang VK. *Big Data Analysis with Python.* Packt Publishing, 2019. 2) J.T. Wolohan. *Mastering Large Datasets with Python.* Manning, 2019. 3) Jonathan Rioux. *Data Analysis with Python and PySpark.* Manning, 2022. 4) Bastiaan Sjardin, Luca Massaron, Alberto Boschetti. *Large Scale Machine Learning with Python.* Packt Publishing, 2016. 5) Jesse C. Daniel. *Data Science with Python and Dask.* Manning, 2019. 6) Hassan A. Karimi. *Big data: techniques and technologies in geoinformatics.* CRC Press, Taylor & Francis Group, Boca Raton, 2014. 7) Jimmy Lin and Chris Dyer. *Data-Intensive Text Processing with*

	MapReduce. Morgan & Claypool, 2010.
Обсяг курсу	Загальний обсяг: 150 годин. Аудиторних занять: 78 год., з них 39 год. лекційних та 39 год. лабораторних занять. Самостійної роботи: 72 год.
Очікувані результати навчання	У результаті вивчення даного курсу студент буде: **знати**: методи аналізу великих даних **вміти**: аналізувати дані
Ключові слова	Великі дані, Pandas, MapReduce, Dask, Spark
Формат курсу	Очний
Теми	Див. **Схема курсу**
Підсумковий контроль, форма	Залік
Пререквізити	Для вивчення даного курсу студенту потрібні базові знання з: - Програмування - Інформатики
Навчальні методи та техніки, які будуть використовуватися під час викладання курсу	Інформаційні методи (лекція, бесіда, ілюстрація, демонстрація); дедуктивні методи на основі узагальнень; евристичні методи (проблемна лекція); інтерактивні методи (дискусія)
Необхідне обладнання	Для проведення лекційних занять: комп'ютер (мінімальні характеристики: процесор Intel Core i3, 4ГБ оперативної пам'яті), доступ до мережі Internet, засоби мультимедіа (в т.ч. проектор). Для проведення практичних/лабораторних занять: комп'ютер (мінімальні характеристики: процесор Intel Core i3, 4ГБ оперативної пам'яті), доступ до мережі Internet. Необхідне програмне забезпечення включає в себе ОС Windows 10, програмні додатки (MS Teams).
Критерії оцінювання (окремо для кожного виду навчальної діяльності)	Оцінювання проводиться за 100-бальною шкалою. Бали нараховуються за наступним співвідношенням: • Змістовий модуль 1: 42% семестрової оцінки за контрольну роботу та виконання домашніх завдань, максимальна кількість балів 42. • Змістовий модуль 2: 36% семестрової оцінки за контрольну роботу та виконання домашніх завдань, максимальна кількість балів 36. • Контрольне тестування: 22% семестрової оцінки, максимальна кількість балів 22. Підсумкова максимальна кількість балів 100. **Академічна доброчесність**: Очікується, що роботи студентів будуть оригінальними дослідженнями чи міркуваннями. Списування та втручання в роботу інших студентів становлять, але не обмежують, приклади можливої академічної недоброчесності. Виявлення ознак академічної недоброчесності в написанні завдань є підставою для її незарахуванння викладачем, незалежно від масштабів плагіату чи обману. Жодні форми порушення академічної доброчесності не толеруються.

	Відвідання занять є важливою складовою навчання. Очікується, що всі студенти відвідають усі лекції та практичні/лабораторні зайняття курсу. Студенти повинні інформувати викладача про неможливість відвідати заняття. У будь-якому випадку студенти зобов'язані дотримуватися термінів визначених для виконання всіх видів робіт, передбачених курсом. **Література.** Уся література, яку студенти не зможуть знайти самостійно, буде надана викладачем виключно в освітніх цілях без права її передачі третім особам. Студенти заохочуються до використання також й іншої літератури та джерел, яких немає серед рекомендованих. **Політика виставлення балів.** Враховуються бали, набрані при поточному контролі та бали підсумкового тестування. При цьому обов'язково враховуються присутність на заняттях та активність студента під час практичного заняття; недопустимість пропусків та запізнень на заняття; користування мобільним телефоном, планшетом чи іншими мобільними пристроями під час заняття в цілях не пов'язаних з навчанням; списування та плагіат; несвоєчасне виконання поставленого завдання і т. ін. **Оцінювання лабораторних робіт** (змістовий модуль 1 містить 7 лабораторних робіт, змістовий модуль 2 містить 6 лабораторних робіт, загалом 13 лабораторних робіт, максимальна кількість балів: 78) відбувається шляхом оцінки роботи студента під час проведення практичної роботи в аудиторії (0-2 балів за одну роботу) та захисту написаної студентом вдома практичної роботи (0-4 балів за одну роботу). Бали оцінювання аудиторного виконання лабораторних робіт нараховуються за наступним співвідношенням: 2 – студент в повному обсязі володіє навчальним матеріалом, має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання; 1.5 – студент достатньо розуміє розглянутий матеріал та принципи написаного ним коду програми, присутні неточності та незначні помилки у відповідях на запитання по темі, код програми функціонує відповідно до завдання; 1 – студент не досить добре розуміє розглянутий матеріал та написаний ним код програми, вагається та надає неточні/не конкретні відповіді на запитання по темі, код програми функціонує з помірними недоліками; 0.5 – студент погано розуміє розглянутий матеріал та написаний ним код програми, студент в більшості надає помилкові відповіді на питання по темі, код програми не функціонує належним чином; 0 - студент зовсім не засвоїв розглянутий матеріал, написаний ним код програми не відповідає темі/не функціонує взагалі. Бали оцінювання домашнього завершення виконання практичних робіт та наданого звіту нараховуються за наступним співвідношенням: 4 – звіт цілком і повністю відображає індивідуальне завдання студента, містить правильні висновки, ілюстрований (за потреби) відповідними графіками і таблицями які правильно відображають суть виконаного завдання, студент має повне розуміння розглянутої теми, надає правильні відповіді на запитання по темі, код програми функціонує відповідно до завдання; 3 – звіт в достатній мірі відображає індивідуальне завдання студента, містить допустимі висновки, ілюстрований (за потреби) відповідними

	графіками і таблицями які частково відображають суть виконаного завдання, студент достатньо розуміє принципи написаного ним коду програми, присутні неточності та незначні помилки у відповідях на запитання по темі, код програми функціонує відповідно до завдання; 2 – звіт містить загальні формулювання завдання, висновки нечіткі, необхідні ілюстрації чи таблиці відсутні, студент не досить добре розуміє розглянутий матеріал та представлений код програми, надає неточні/не конкретні відповіді на запитання по темі, код програми функціонує з помірними недоліками; 1 – звіт не містить формулювання завдання, висновки необґрунтовані чи неповні, необхідні ілюстрації чи таблиці відсутні, студент погано розуміє розглянутий матеріал та представлений код програми, студент в більшості надає помилкові відповіді на питання по темі, код програми не функціонує належним чином; 0 – звіт відсутній/не відповідає темі, студент зовсім не засвоїв розглянутий матеріал, написаний ним код програми не відповідає темі/не функціонує взагалі. **Оцінювання контрольного тестування** (завдання з тематики кожного змістового модуля) відбувається шляхом оцінки письмових відповідей студента на поставлені запитання. Відсотки нарахування балів оцінювання відповіді на кожне запитання нараховуються за наступним співвідношенням: 75-100% – тема відтворюється в повному обсязі, правильно, обґрунтовано, логічно; 50-75% – відтворюється значна частина розглянутої теми, проте присутні неточності та/або невідповідності; 25-50% – виявлено множинні неточності та невідповідності, пояснення відсутні чи частково помилкові; 0-25% – тему майже не розкрито, кількість викладеного матеріалу не відповідає загальним нормам обраного виду роботи. **Критерії оцінювання результатів неформальної освіти:** Нарахування балів відбувається за публікацію студентом тез доповідей на конференціях, наукових статей, за участь студента у діяльності наукових гуртків, семінарів, круглих столів, конкурсів, участь у заходах неформальної освіти, за отримання сертифікатів про проходження навчання на різних освітніх платформах (Coursera, Prometheus тощо), курсах провідних IT компаній за тематикою навчальної дисципліни. Кількість балів визначається відсотком покриття результатів відповідної активності до вимог результатів навчання з навчальної дисципліни
Питання для контрольного тестування	Типи даних, оператори, вказівники, керування пам’яттю, синтаксис функцій.
Опитування	Анкету-оцінку з метою оцінювання якості курсу буде надано по завершенню курсу.

Схема курсу “Аналіз великих даних”
для студентів спеціальності 112 - Статистика

Тижні	Лекційний курс		Практичні заняття		К-сть год СР	Літе-ратура
	Назва теми	К-сть год	Назва теми	К-сть год		
1	2	3	4	5	6	7
1	Вступ у аналіз великих даних. Означення великих даних. Основні підходи до збору та аналізу великих даних.	3	Налаштування середовища розробки. Установка необхідних пакетів для Python.	3	2	[1], [2], [4], Сайт курсу
2	Обробка великих даних у Pandas. Фільтрування. Агрегація та групування. Ввід-вивід у Pandas. Візуалізація.	3	Практичні завдання на обробку тестових csv-файлів у Pandas. Експорт та візуалізація даних.	3	6	[1], [2], [4], Сайт курсу
3	Паралелізація обчислень. Поняття про паралелізацію. Порядок і стан в паралелізації.	3	Задачі на модифікацію коду із використанням паралельних алгоритмів. Практичні завдання на побудову графів даних, стягнених з Вікіпедії.	3	6	[2], [4], Сайт курсу
4	Масштабування обчислень. Різні підходи до масштабування. Out-of-core learning. Обробка потокових даних.	3	Вправи на завантаження та обробку великих даних у потоковому форматі.	3	6	[2], [4], Сайт курсу
5	MapReduce. Функції map та reduce. Допоміжні функції. Ланцюжки функцій. Підходи до пришвидшення коду, що використовує MapReduce.	3	Обробка інтернет-даних із використанням MapReduce. Групування слів, статистичні обчислення із використанням MapReduce.	3	6	[2], Сайт курсу
6	Ліниві обчислення. Ітератори в Пайтон. Функції map, range, filter, zip.	3	Написання власних ітераторів у Пайтон. Практичні завдання на використання функцій filter та zip.	3	6	[2], Сайт курсу
7	Масштабування глибокого навчання. Навчання без нагляду із використанням theanets. Масштабування класифікаційних та	3	Практичні завдання на навчання без нагляду, класифікацію та регресію із використанням бібліотек H2O та theanets.	3	6	[4], Сайт курсу

	регресивних дерев.					
8	Аналіз великих даних із Dask. Опис бібліотеки. Орієнтовані ациклічні графи. Робота із структуризованими даними із використанням датафрейм. Масштабування NumPy і Pandas.	3	Вправи на обробку датафреймів у Dask. Дескриптивна статистика та аналіз даних у Dask.	3	6	[5], Сайт курсу
9	Вступ у Spark. Структури даних в Spark. RDD. DataFrame. Створення схем для датафреймів у Spark.	3	Налаштування PySpark на локальній машині. Запуск тестових завдань локально.	3	5	[1], [3], Сайт курсу
10	Аналіз табличних даних із PySpark. Використання SparkReader для читання даних із csv-файлів. Дослідження структури даних. Денормалізація даних.	3	Практичні завдання на роботу із даними із використанням модуля pyspark.sql.	3	6	[3], Сайт курсу
11	Робота з датафреймами в PySpark. Маніпуляція даних у датафреймах. Об'єднання та сумування даних. Конверсія датафреймів у формат Pandas.	3	Вправи на маніпуляцію даними у датафреймах Spark. Підрахунок кількості слів у тексті за заданими характеристиками.	3	6	[2], [3], Сайт курсу
12	Робота з відсутніми значеннями у Spark. Кореляційний аналіз. Пошук та обробка відсутніх значень. Опис кореляції між змінними. Створення кореляційної матриці.	3	Вправи на видалення слів з тексту. Обробка неповних записів у датафреймі.	3	6	[1], [3], Сайт курсу
13	Підсумкове заняття.	3	Перевірка знань та вмінь студентів.	3	5	[1]-[7], Сайт курсу
	Разом	**39**		**39**	**72**	
	Викладач: **Холявка О. Т.**		Викладач: **Холявка О.Т.**			