Національний технічний університет України
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Кафедра інженерії
програмного забезпечення
в енергетиці (ІПЗЕ)

ОБРОБЛЕННЯ НАДВЕЛИКИХ МАСИВІВ ДАНИХ
Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	**Другий (магістерський)**
Галузь знань	12 Інформаційні технології
Спеціальність	121 Інженерія програмного забезпечення
Освітня програма	Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці
Статус дисципліни	Вибіркова
Форма навчання	заочна
Рік підготовки, семестр	1 курс, весняний
Обсяг дисципліни	4 кред/120 год.(лекцій 10 год., практ. 6 год.)
Семестровий контроль/ контрольні заходи	Екзамен, мкр
Розклад занять	http://rozklad.kpi.ua/
Мова викладання	Українська/Англійська/Німецька / Французька
Інформація про керівника курсу / викладачів	Лектор: д.т.н., Федорова Наталія Володимирівна, Natasha_f@ukr.net, telegram, viber, Zoom session Практичні: д.т.н., Федорова Наталія Володимирівна, Natasha_f@ukr.net, telegram, viber, Zoom session
Розміщення курсу	Кампус

Програма навчальної дисципліни

1. Опис навчальної дисципліни, її мета, предмет вивчання та результати навчання

Метою дисципліни "Оброблення надвеликих масивів даних" є набуття знань та практичних навичок використання методів та алгоритмів обробки великих даних для вирішення комплексних задач аналізу надвеликих масивів даних.

Предметом дисципліни "Обробка надвеликих масивів даних" є серія підходів, інструментів і методів обробки структурованих і неструктурованих різноманітних даних великих розмірів для отримання результатів, які легко сприймаються людиною.

Основні завдання кредитного модуля.

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

Загальні компетенції:

- здатність проводити дослідження на відповідному рівні (ЗК 3);
- здатність генерувати нові ідеї (креативність) (ЗК 5).

Фахові компетентності:

- здатність проєктувати та розробляти програмну систему з використанням методів інтелектуального аналізу даних (ФК-11);
- здатність проєктувати та розробляти програмне забезпечення для роботи в хмарі (ФК-14).
- здатність проектувати та програмно реалізовувати метод комп'ютерної обробки надвеликих за обсягом даних в інформаційних середовищах різноманітного призначення, систем управління бізнес-процесами, мереж Інтернету речей, сервіс-орієнтованих середовищ та систем високопродуктивних кластерних обчислень;

- здатність вирішувати масштабні обчислювальні задач у розподілених інтелектуальних середовищах та контролювати хід обчислень за допомогою спеціалізованого програмного забезпечення;

- здатність обирати адекватних методи машинного навчання, включаючи методи глибокого навчання, та використовувати їх для налаштування нейронних мереж для вирішення конкретних задач прогнозування, керування, класифікації та інтелектуального аналізу даних.

Згідно з вимогами ОПП/ОНП Інженерія програмного забезпечення інтелектуальних кібер-фізичних систем в енергетиці, студенти після засвоєння навчальної дисципліни мають продемонструвати такі результати навчання.

Програмні результати навчання:

- збирати, аналізувати, оцінювати необхідну для розв'язання наукових і прикладних задач інформацію, використовуючи науково-технічну літературу, бази даних та інші джерела (ПРН 17);

- вміти проектувати та розробляти програмні системи з використанням методів інтелектуального аналізу даних (ПРН 19);

- вміти проектувати та розробляти програмне забезпечення для роботи в хмарі (ПРН 22).

- використовувати сучасні технологій обробки надвеликих масивів даних за допомогою інфраструктури програмних рішень Spark;

- використовувати методи машинного навчання для вирішення практичних задач.

2. Пререквізити та постреквізити дисципліни (місце в структурно-логічній схемі навчання за відповідною освітньою програмою)

Дисципліна містить чотири кредити.

Вивчення дисципліни спирається на знання, отримані за програмою попередніх років навчання за спеціальністю 121 Інженерія програмного забезпечення. Студенти мають досвід у імперативному, об'єктно-орієнтованому і функціональному програмуванні.

Викладений матеріал може бути інструментальною основою для підготовки магістерських дисертацій.

Міждисциплінарні зв'язки забезпечуються дисциплінами: «Аналітика обробки даних в сенсорних мережах», «BigData в енергетиці» «Математичні методи моделювання систем з розподіленими параметрами», «Розробка застосунків Інтернету речей та сенсорних мереж в енергетиці».

Зміст навчальної дисципліни

Розділ. Концепція "Великих даних". Основні концепції програмної платформи розподілених даних Spark. Машинне навчання з використанням бібліотеки Spark MLlib

Тема 1. Поняття "великих даних".

Основні аспекти та складові елементи трактування поняття "Великі дані". Сфери застосування надвеликих масивів даних. Характеристики "великих даних". Проблема масштабування. Базові компоненти аналізу даних.

Тема 2. Базові програмні засоби роботи з надвеликими масивами даних.

Поняття розподіленої файлової системи. Програмні моделі "великих даних". Hadoop екосистема. Концепція MapReduce. Шаблони проектування з підходом MapReduce.

Тема 3. Програмна платформа розподілених даних Spark.

Основні концепції та архітектура Spark. Програмування з RDD. Використання RDD з парами ключ / значення. Завантаження і збереження даних. Spark SQL, DataFrames, Datasets.

Тема 4. Основи машинного навчання.

Машинне навчання, його історичний розвиток і сучасний стан. Проблема навчання. Приклади прикладних задач, які використовують методи машинного навчання. Огляд можливостей бібліотеки Spark Mllib.

Тема 5. Перспективні напрямки розвитку програмних засобів обробки надвеликих даних і машинного навчання.

Можливості глибинного навчання з допомогою Spark. Інтеграція з Tensorflow. Прграмні рішення та сервісм для обробки надвеликих даних в галузі машинного навчання віл Google, Amazon, Facebook та інших лідерів ринку.

3. Навчальні матеріали та ресурси

Базова література

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.

Додаткова література

7. Holmes, Alex. Hadoop in Practice. Shelter Island, NY: Manning, 2012.
8. Alpaydin, Ethem. Introduction to Machine Learning., 2014.
9. Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of Machine Learning. Cambridge, MA: MIT Press, 2012.
10. Shalev-Shwartz, Shai, and Shai Ben-David. Understanding Machine Learning: From Theory to Algorithms., 2014.
11. Tom White Hadoop: The Definitive Guide, 4th Edition O'Reilly Media. 2015 –756p.
12. Witten, I H, and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques. Amsterdam: Morgan Kaufman, 2005.
13. Cherkassky, Vladimir S, and Filip Mulier. Learning from Data: Concepts, Theory, and Methods. Hoboken, N.J: IEEE Press, 2007.
14. Marsland, Stephen. Machine Learning: An Algorithmic Perspective. Boca Raton: CRC Press, 2009. Print.
15. Harrington, Peter. Machine Learning in Action. Shelter Island, NY: Manning Publications, 2012.

Навчальний контент

4. Методика опанування навчальної дисципліни (освітнього компонента)

Лекційні заняття

№ з/п	Назва теми лекції та перелік основних питань
1	Поняття великих даних. Програмний засіб Hadoop 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.
2	Шаблони проектування MapReduce Шаблони підсумовування (summarization). Шаблони фільтрації (filtering). Шаблони організації даних. Шаблони об'єднань (join). Меташаблони. Шаблони вводу/виведення.

3	Основи Spark 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.
4	Spark SQL. DataFrames. DataSet. 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.
5	Основи машинного навчання. Spark Mllib 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.

Лабораторні заняття

Основні завдання циклу лабораторних занять полягають у набутті студентами практичних навичок з використання спеціалізованого програмного забезпечення обробки надвеликих масивів даних та реалізації алгоритмів машинного навчання

№ з/п	Назва теми заняття
1	Використання шаблонів проектування MapReduce в системах енергоменеджменту
2	Застосування SparkSQL, робота з даними з використанням Dataframes та DataSet в системах збирання даних з лічильників
3	Обробка текстової інформації та проектування ознак засобами Spark для систем енергоменеджменту

5. Самостійна робота студента/аспіранта

№ з/п	Назви тем і питань, що виноситься на самостійне опрацювання та посилання на навчальну літературу
1	Поняття “великих даних” Технології великих даних на сучасне програмне забезпечення [11]. Програмна реалізація шаблонів проектування Mapreduce [1].
2	Базові програмні засоби роботи з надвеликими масивами даних. Програмна екосистема Hadoop. Практичне застосування Hadoop в задачах Data Science. Огляд HiveQL, Pig, Yarn, Hbase. Аналітика даних з використанням Spark. [2, 4, 5, 7].
3	Програмна платформа розподілених даних Spark Модель паралельних обчислень Spark. Spark Job Scheduling. DataFrame API. Представлення даних в DataFrames and Datasets. Core Spark Joins. Ефективні трансформації [5].
4	Основи машинного навчання Теоретичні засади машинного навчання [9]. Кодування та підготовка даних Mllib. Масштабування та вибір характеристик. MLlib Model Training. Оцінка моделі машинного навчання [5].
5	Навчання з учителем (Supervised Learning). Метод стохастичного градієнта SG. Логістична регресія. Принцип максимуму правдоподібності і логарифмічна функція втрат. Метод стохастичного градієнта для логарифмічною функції втрат. Математичні основи методу опорних векторів. 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, 15].
6	Навчання без вчителя (Unsupervised Learning). Постановка завдання кластеризації. Постановка завдання Semisupervised Learning, приклади застосування. Алгоритм k-середніх і EM-алгоритм для поділу Гаусовскої суміші. 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, 15].
7	Рекомендаційні системи. Завдання колаборативної фільтрації, транзакційні дані і матриця суб'єкти-об'єкти. Collaborative Filtering and Recommendation [5, 3, 9, 13].
8	Перспективні напрямки розвитку програмних засобів обробки надвеликих даних і машинного навчання. Ризики, пов'язані з застосуванням надвеликих даних. Проблема конфіденційності. Датифікація [11].

Політика та контроль

6. Політика навчальної дисципліни (освітнього компонента)

Застосовуються стратегії активного і колективного навчання, які визначаються наступними методами і технологіями:

1) методи проблемного навчання (проблемний виклад, частково-пошуковий (евристична бесіда) і дослідницький метод);

2) особистісно-орієнтовані (розвиваючі) технології, засновані на активних формах і методах навчання («мозковий штурм», «аналіз ситуацій» дискусія, експрес-конференція);

3) інформаційно-комунікаційні технології, що забезпечують проблемно-дослідницький характер процесу навчання та активізацію самостійної роботи студентів (електронні презентації для лекційних занять, використання аудіо-, відео-підтримки навчальних занять).

4) лекційні та лабораторні заняття відносяться до аудиторних занять. Відвідування аудиторних занять є обов'язковим;

5) правила поведінки на заняттях: активність, підготовка коротких доповідей чи текстів, відключення телефонів, використання засобів зв'язку для пошуку інформації на гугл-диску викладача чи в інтернеті тощо;

6) правила захисту лабораторних робіт. На лабораторних заняттях студенти під керівництвом викладача вивчають методику експериментальних досліджень. На кожній лабораторній роботі студенти оформляють звіт у письмовому вигляді. До звіту заноситься перебіг досліду, його результати і даються пояснення отриманих результатів з урахуванням похибок експерименту.

7) індивідуальні завдання з дисципліни (реферати, розрахункової, графічні, тощо) видаються студентам в терміни, передбачені вищим навчальним закладом. Індивідуальні завдання виконуються студентом самостійно при консультуванні викладачем. Допускаються випадки виконання комплексної тематики кількома студентами.

8) правила призначення заохочувальних балів: своєчасне виконання та здача лабораторних, індивідуальних завдань, нестандартний підхід до вирішення певного завдання;

правила призначення штрафних балів: несвоєчасне виконання лабораторних та індивідуальних завдань, а також користування допоміжними засобами (наприклад, мобільний телефон, конспект лекцій) під час виконання контрольної роботи.

9) політика дедлайнів та перескладань: невчасно виконані та здані лабораторні роботи оцінюються нижчою оцінкою (-10-15% від загальної підсумкової оцінки).

10) політика щодо академічної доброчесності: письмові роботи можуть перевірятися на наявність плагіату і допускаються до захисту із коректними текстовими запозиченнями не більше 40%. Списування під час контрольних робіт та екзаменів заборонені.

11) інші вимоги, що не суперечать законодавству України та нормативним документам Університету:

- політика щодо відвідування: відвідування занять є обов'язковим компонентом оцінювання, за яке нараховуються бали. За об'єктивних причин (підтверджених документально) дозволяється перескладання пропущених тем курсу.
- політика щодо виконання завдань: позитивно оцінюється відповідальність, старанність, креативність, фундаментальність.

7. Види контролю та рейтингова система оцінювання результатів навчання (РСО)

1. 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 студента з кредитного модуля складається з балів, що він отримав за:

- виконання практичних робіт;
- написання контрольної роботи (МКР).

Система рейтингових (вагових) балів та критерії оцінювання

1.Виконання практичних робіт

Оцінюється 3 роботи, передбачених робочою програмою. Максимальний ваговий бал rЛР =60

Сума вагових балів практичних робіт:

N л. р.	Назва практичної роботи	Максимальний ваговий бал
1	Використання шаблонів проектування MapReduce в системах енергоменеджменту	20
2	Застосування SparkSQL, робота з даними з використанням Dataframes та DataSet в системах збирання даних з лічильників	20
3	Обробка текстової інформації та проектування ознак засобами Spark для систем енергоменеджменту	20
Разом		60

Оцінювання лабораторних робіт:

–якщо робота виконана невчасно знімається 10-15% 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 rМКР = 10.

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 - 10 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 обгрунтовані теоретичні рішення, помилки у методиці розрахунків;

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

3. Екзамен

Екзамен відбувається у письмовій формі. Максимальна оцінка за екзамен складає rЕК = 30 балів.

Умови позитивної проміжної атестації

Для отримання „зараховано" з першої проміжної атестації студент повинен мати не менше, ніж 12 балів (за умови, що за 8 тижнів згідно з календарним планом контрольних заходів студент повинен отримати 24 бали).

Для отримання „зараховано" з другої проміжної атестації студент повинен мати не менше, ніж 40 балів (за умови, що за 14 тижнів згідно з календарним планом контрольних заходів студент повинен отримати 76 балів).

Розрахунок шкали (R) рейтингу:

Сума вагових балів контрольних заходів протягом семестру складає:

$$R=60+10+30=100 \text{ балів}$$

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

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

Для отримання студентом відповідних оцінок (ECTS та традиційних) його рейтингова оцінка RD переводиться згідно таблиці:

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Кількість балів	*Оцінка*
100-95	Відмінно
94-85	Дуже добре
84-75	Добре
74-65	Задовільно
64-60	Достатньо
Менше 60	Незадовільно
Не виконані умови допуску	Не допущено

8. Додаткова інформація з дисципліни (освітнього компонента)

Методичні рекомендації

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

Методика побудована таким чином, що матеріал майже кожної лекції закріплюється виконанням завдання комп'ютерного практикуму. Завдання студенти 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.

Перелік питань, які виносяться на семестровий контроль:

1. Поняття великих даних. Програмний засіб Hadoop
2. Методи і техніка аналізу великих даних
3. Дайте визначення MapReduce
4. Дайте визначення Apache Spark
5. Архітектура розподіленого додатку Spark. Ком'ютерний кластер. Вузол
6. Основні концепції Spark. RDD та граф перетворень
7. Основні етапи обробки даних

8. Загрузка даних із зовнішнього сховища
9. Зміна розміщення даних та кількості партицій
10. Як відбувається обчислення над даними в Spark
11. Розгалуження та ітеративні обчислення
12. Shuffle механізм
13. Управління памяттю в Apache Spark
14. DataFrame API та Spark SQL. Датафрейми
15. Робота з DataFrame API в рамках побудови сценаріїв обробки даних на Spark
16. Використання функцій користувачів(UDF)
17. Функції користувачів агрегації в Spark
18. Створення, налаштування та запуск Spark проекту. Налаштування оточення
19. Створення нового проекту в Spark
20. Основи машинного навчання. Spark Mllib
21. Загальні відомості про класифікацію та логістичну регресію
22. Створення додатку машинного навчання Apache Spark MLlib. Створення вхідного кадру даних
23. Створення моделі логістичної регресії
24. Навчання моделі логістичної регресії
25. Створення візуального представлення прогнозу
26. Обробка текстової інформації та проектування ознак засобами Spark
27. Для аналітики великих даних використовуються компоненти: • MLLib – машинне навчання, • GraphX – робота з графами, • Spark-SQL – інтерфейс, • SparkStreaming – потокова аналітика. Дайте визначення цих компонент
28. Мова програмування Scala. Призначення.
29. Алгоритм “дерево рішень”.
30. Застосування випадкових лісів для класифікації даних
31. Як працюють ансамблі. Стекінг. Беггінг. Бустінг
32. Алгоритми кластеризації. Задача кластеризації. Типи методів кластеризації.
33. Ієрархічна кластеризація. Застосування методів кластеризації.
34. Алгоритми зниження розмірності
35. Карти самоорганізації. Основні концепції.
36. Конкурентне навчання. Архітектура. Топологія. Алгоритм навчання. Підходи до визначення відстаней. Застосування.
37. Створення рекомендаційних систем засобами Spark MLlib
38. Перспективні напрямки розвитку програмних засобів обробки надвеликих даних і машинного навчання.
39. Сучасні напрямки в машинному навчанні. Automated machine learning. Generative Adversarial Networks. Спеціалізоване апаратне забезпечення. Cloud Object Storage.
40. Програмні рішення обробки надвеликих даних в задачах машинного навчання від лідерів ринку (Amazon, Google, Facebook). 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.
41. Назвати основні причини, за якими сервіси застосовують системи рекомендацій.
42. Назвати суттєві. характеристики рекомендаційних систем.
43. Навести приклади явного і неявного збору даних про користувача.
44. Навести приклади популярних ресурсів, де спостерігається вплив рекомендаційної системи на запити користувача.
45. На які фактори спирається система рекомендацій YouTube?
46. Назвати цінності, які покладає на своїх користувачів Facebook.
47. Які фактори суттєво впливають на формування стрічки новин Facebook для конкретного користувача?
48. Назвати формати взаємодії користувача з оточенням та їх вагомість для ранжирування.

49. За яким принципом працює алгоритм Facebook для рекомендації друзів?
50. Які фактори суттєво впливають на формування стрічки новин Instagram для конкретного користувача?

Робочу програму навчальної дисципліни (силабус): «Оброблення надвеликих масивів даних»:

Складено професором кафедри ІПЗЕ, д.т.н., доц. Федоровою Наталією Володимирівною

Ухвалено кафедрою ІПЗЕ (протокол № 34 від 10.05.2024 р.)

Погоджено Методичною комісією факультету[1] (протокол № 9 від 31.05.2024 р.)

---

[1]Методичною радою університету – для загальноуніверситетських дисциплін.