



МЕТОДИ ДОБУВАННЯ ДАНИХ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти *Третій (доктор філософії)*

Галузь знань	<i>12 Інформаційні технології</i>
Спеціальність	<i>123 Комп'ютерна інженерія</i>
Освітня програма	<i>Комп'ютерна інженерія</i>
Статус дисципліни	<i>Вибіркова</i>
Форма навчання	<i>очна(денна)</i>
Рік підготовки, семестр	<i>2 курс, осінній семестр</i>
Обсяг дисципліни	<i>5 кредити 150 годин</i>
Семестровий контроль/ контрольні заходи	<i>Залік</i>
Розклад занять	<i>Практичні – 36 годин Самостійна -114</i>
Мова викладання	<i>Українська</i>
Інформація про керівника курсу / викладачів	Лектор: д.т.н, професор, Новотарський Михайло Анатолійович novotar@gmail.com Лабораторні: д.т.н, професор, Новотарський Михайло Анатолійович novotar@gmail.com
Розміщення курсу	

Програма навчальної дисципліни

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

Дисципліна “Методи добування даних” спрямована на вивчення основних методів, які дозволяють дослідити та визначити приховані функціональні залежності в великих наборах даних. В рамках вивчення даної дисципліни також розглядаються основні технологічні рішення у сфері науки про дані (Data Science), які забезпечують пошук прихованих невідомих раніше корисних та доступних для інтерпретації знань, необхідних для використання при прийнятті рішень у різних сферах діяльності.

Вивчення даної дисципліни майбутніми науковцями дозволить їм набути важливих компетенцій, оскільки, в зв'язку з накопиченням значних обсягів даних у різних сферах діяльності людини, виникає широкий спектр задач з добування корисних даних, на основі яких отримують нові знання, що не можуть бути одержані у інший спосіб.

Метою вивчення дисципліни «Методи добування даних» є підготовка фахівців, здатних розв'язувати комплексні проблеми в галузі дослідницько-інноваційної діяльності у сфері представлення та опрацювання даних для отримання прихованої в даних корисної інформації шляхом вивчення теоретичних та практичних положень побудови моделей прихованих в наборах даних функціональних залежностей та використання результатів аналізу даних для уточнення

наукових висновків та формування прогнозів щодо майбутніх станів об'єктів досліджень, що передбачає глибоке переосмислення наявних та створення нових цілісних знань.

Предметом дисципліни є:

- методи та способи представлення та опрацювання даних для отримання прихованої в даних корисної інформації;
- методи побудови та дослідження математичних моделей опису функціональних залежностей між параметрами даних та шуканою інформацією;
- технології адаптивних та інтелектуальних обчислень при добуванні даних для розроблення прогнозів щодо досліджуваних об'єктів та процесів.

Основні результати навчання

Здобувачі наукового ступеня доктора філософії після засвоєння навчальної дисципліни мають продемонструвати такі **компетентності**.

1. *Загальні компетентності*: здатність до абстрактного мислення, аналізу і синтезу; здатність до пошуку, оброблення та аналізу інформації з різних джерел щодо методів та технологій добування даних; здатність формування системного наукового світогляду у сфері отримання прихованої в даних корисної інформації; здатність набуття універсальних навичок усної та письмової презентації власного наукового дослідження за результатами аналізу даних, застосування сучасних інформаційних технологій у науковій діяльності.

2. *Спеціальні компетентності*: здатність ефективно застосовувати основні методи добування даних, основні статистичні методи навчання, кластерні методи машинного навчання та методи навчання без учителя при проведенні наукових досліджень; здатність інтегрувати знання з різних дисциплін, застосовувати системний підхід при проведенні досліджень; здатність аргументувати вибір методу розв'язання наукової задачі, критично оцінювати отримані результати та захищати прийняті рішення щодо вибору методів та способів обробки даних.

За результатами вивчення навчальної дисципліни «Методи добування даних» мають бути отримані такі **знання**.

1. Мати передові концептуальні та методологічні знання у сфері аналізу великих обсягів даних для отримання нових знань, а також дослідницькі навички, достатні для проведення наукових і прикладних досліджень в галузі добування даних на рівні останніх світових досягнень з комп'ютерної інженерії, IT-інфраструктур, інформаційних технологій.

2. Знати сучасні методи проведення досліджень у сфері представлення та опрацювання даних для отримання прихованої в даних корисної інформації шляхом вивчення теоретичних та практичних положень побудови моделей прихованих в наборах даних функціональних залежностей та використання результатів аналізу даних для уточнення наукових висновків та передбачення майбутніх станів об'єктів досліджень.

3. Знати і розуміти наукові і математичні положення, що лежать в основі методів опрацювання даних для отримання прихованої корисної інформації, методів побудови та дослідження математичних моделей та технологій адаптивних та інтелектуальних обчислень при добуванні даних.

Уміння, які мають бути отримані у рамках вивчення навчальної дисципліни «Методи добування даних».

1. Вміти ефективно здійснювати пошук та критичний аналіз інформації з різних джерел щодо методів та технологій добування даних.

2. Вміти розв'язувати задачі синтезу та аналізу об'єктів дослідження при опрацюванні даних для отримання прихованої корисної інформації.

3. Вміти створювати та реалізовувати математичні моделі опису функціональних залежностей, що приховані у даних;

4. Вміти застосовувати технології адаптивних та інтелектуальних обчислень при добуванні даних для передбачення майбутніх станів досліджуваних об'єктів та процесів.

Здобувачі наукового ступеня також мають бути **здатні**.

1. Застосовувати прикладні бібліотеки та програмні системи, які використовуються при машинному аналізі даних.

2. Володіти методами та технологіями програмування з використанням прикладних бібліотек та програмних систем, призначених для машинного аналізу великих наборів даних.

Таке поєднання загальних та спеціальних компетентностей, теоретичних та практичних знань, умінь та здатностей сприяє підвищенню науково-практичного рівня здобувачів наукового ступеня доктора філософії задля здійснення ними ефективних наукових досліджень.

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

Для успішного оволодіння дисципліною необхідні знання:

- основ математичного аналізу, теорії функцій та математичної статистики;
- основ функціонування операційних систем;
- основ програмування мовою Python.

Відповідно до освітньої програми необхідно попередньо оволодіти знаннями з дисциплін: “Програмування”, “Об’єктно-орієнтоване програмування”, “Системне програмування”, “Структури даних та алгоритми”, “Інженерія програмного забезпечення”, “Алгоритми та методи обчислень”, “Дискретна математика”.

Компетентності, знання та вміння, отримані в рамках вивчення даної дисципліни, можуть бути застосовані для отримання обґрунтованих результатів досліджень та підвищення наукового рівня дисертаційних робіт.

3. Зміст навчальної дисципліни

Розділ 1. Загальне поняття про методи аналізу та добування даних

Тема 1.1. Огляд методів та засобів.

Тема 1.2. Статистичні обмеження для аналізу даних

Тема 1.3. Основні інструменти та базові теоретичні положення

Тема 1.4. Ефект Метью

Розділ 2. Пошук подібних об'єктів

Тема 2.1. Застосування методу пошуку найближчих сусідів

Тема 2.2. Шинглінг документів

Тема 2.3. Збереження схожості скорочених наборів

Тема 2.4. Локально-чутливе хешування для документів

Розділ 3. Міри відстані

Тема 3.1. Визначення міри відстані

Тема 3.2. Відстань Евкліда

Тема 3.3. Відстань Жакарта

Тема 3.4. Косинусна відстань

Тема 3.5. Відстань редагування

Тема 3.6. Відстань Хеммінга

Розділ 4. Локально-чутливі функції

Тема 4.1. Теорія локально-чутливих функцій
Тема 4.2. LSH сімейства для інших мір відстані

Розділ 5. Застосування локально-чутливого хешування

Тема 5.1. Приклади розпізнавання об'єктів
Тема 5.2. Методи для високих ступенів подібності

Розділ 6. Аналіз потоків даних

Тема 6.1. Модель поточкових даних
Тема 6.2. Вибір даних у потоці
Тема 6.3. Фільтрація потоків
Тема 6.4. Підрахування різних елементів у потоці

Розділ 7. Оцінка моментів

Тема 7.1. Визначення моментів та практичні приклади
Тема 7.2. Підрахунки у вікні
Тема 7.3. Згасання вікон

Розділ 8. Аналіз посилань

Тема 8.1. Класифікація сторінок. PageRank
Тема 8.2. Ефективне обчислення PageRank

Розділ 9. Покращення при обчисленні Page Rank

Тема 9.1. Тематично чутливий PageRank
Тема 9.2. «Лінковий» спам
Тема 9.3. Хаби та авторитети

Розділ 10. Набори елементів, які часто зустрічаються (часті набори елементів)

Тема 10.1. Модель ринкова корзина
Тема 10.2. Ринкові корзини і алгоритм A-Priori

Розділ 11. Обробка великих наборів даних у основній пам'яті

Тема 11.1. Основні алгоритми обробки даних в пам'яті
Тема 11.2. Алгоритми з обмеженням проходів
Тема 11.3. Підрахунок частих елементів у потоці

Розділ 12. Основні принципи роботи NumPy, SciPy, Matplotlib

Тема 12.1. Принципи роботи з NumPy
Тема 12.2. Робота з SciPy
Тема 12.3. Використання бібліотеки Matplotlib

Розділ 13. Основи бібліотеки Pandas

Тема 13.1. Структури даних Pandas
Тема 13.2. Робота з елементами Series
Тема 13.3. Робота з елементами DataFrame
Тема 13.4. Доступ до даних у структурах Pandas
Тема 13.5. Використання атрибутів для доступу до даних в Pandas

Розділ 14. Отримання випадкового набору з структур Pandas

Тема 14.1. Робота з випадковим набором даних

Тема 14.2. Додавання елементів у структури

Тема 14.3. Індексція з використанням логічних виразів

Тема 14.4. Використання isin для роботи з даними в Pandas

Тема 14.5. Робота з пропусками в даних

Розділ 15. Робота Pandas у середовищі Jupyter Notebook інтегратора Anaconda

Тема 15.1. Установка програмного забезпечення

Тема 15.2. Запуск Jupyter Notebook. Приклади роботи

Тема 15.3. Основні елементи інтерфейсу

Тема 15.4. Робота з бібліотекою Pandas

4. Навчальні матеріали та ресурси

Базова:

1. Han J., Kamber M. Data Mining Concepts and Techniques. – Morgan Kaufmann, 2011. –626 p.
2. Larose D.T. Data Mining Methods and Models. – NY: Wiley-IEEE Press, 2006. –344 p.
3. Rajaraman A., Ullman D.J. Mining of Massive Datasets. – Cambridge: Cambridge University Press, 2011. – 326 p.
4. Lawrence K.D., Kudyba S., Klimberg R.K. Data Mining Methods and Applications. – Auerbach Publications, 2008. –366 p.
5. Гладун А.Я., Рогушина Ю.Ф. Data Mining: пошук знань в даних. – К.: ТОВ «ВД « АДЕФ-Україна», 2016. –452 с.

Додаткова:

6. Runkowski L., Jaworski M., Duda P. Stream Data Mining: Algorithms and Their Probabilistic Properties – Springer, 2020. – 504 p.
7. Olson D.L., Delen D. Advanced Data Mining Techniques. – Springer, 2008. – 192 p.
8. Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques. – Morgan Kaufmann, 2005. – 560 p.
9. Nisbet R., Elder J., Miner G. Handbook of Statistical Analysis. Data Minig Applications. –Amsterdam: Elsevier, 2009 – 864 p.
10. Fu X., Wang L. Data Mining with Computational Intelligence. – Springer, 2005. – 287 p.
11. Черняк О.І., Захарченко П.В. Інтелектуальний аналіз даних. – К.: Знання, 2014. – 599 с.

Навчальний контент

12. Методика опанування навчальної дисципліни (освітнього компонента)

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Практичні роботи	СРС
1. Загальне поняття про методи аналізу та добування даних Тема 1.1. Огляд методів та засобів. Статистичне моделювання. Машинне навчання. Обчислювальні підходи до моделювання. Спосіб	8		4	4

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Практичні роботи	СРС
узагальнення даних. Виділення ознак. Тема 1.2. . Статистичні обмеження для аналізу даних. Загальна інформативність. Принцип Бонферроні. Приклад застосування принципу Бонферроні. Тема 1.3. Основні інструменти та базові теоретичні положення. Важливість слів у документах. Хеш-функції. Індекси. Зовнішня пам'ять. Основа натуральних логарифмів. Степенева залежність. Тема 1.4. Ефект Метью. Висновки за розділом.				
2. Пошук подібних об'єктів Тема 2.1. Застосування методу пошуку найближчих сусідів. Тема 2.2. Шинглінг документів. Тема 2.3. Збереження схожості скорочених наборів. Тема 2.4. Локально-чутливе хешування для документів.	8		2	6
3. Міри відстані Тема 3.1. Визначення міри відстані. Тема 3.2. Відстань Евкліда. Тема 3.3. Відстань Жакарта. Тема 3.4. Косинусна відстань. Тема 3.5. Відстань редагування. Тема 3.6. Відстань Хеммінга.	8		2	6
4. Локально-чутливі функції Тема 4.1. Теорія локально-чутливих функцій. Тема 4.2. LSH сімейства для інших мір відстані.	8		2	6
5. Застосування локально-чутливого хешування Тема 5.1. Приклади розпізнавання об'єктів. Тема 5.2. Методи для високих ступенів подібності.	8		2	6
6. Аналіз потоків даних Тема 6.1. Модель поточкових даних. Тема 6.2. Вибір даних у потоці. Тема 6.3. Фільтрація потоків. Тема 6.4. Підрахування різних елементів у потоці.	8		2	6
7. Оцінка моментів Тема 7.1. Визначення моментів та практичні приклади. Тема 7.2. Підрахунки у вікні. Тема 7.3. Згасання вікон.	8		2	6

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Практичні роботи	СРС
8. Аналіз посилань Тема 8.1. Класифікація стрінок. PageRank. Тема 8.2. Ефективне обчислення PageRank.	8		2	6
9. Покращення при обчисленні Page Rank Тема 9.1. Тематично чутливий PageRank. Тема 9.2. «Лінковий» спам. Тема 9.3. Хаби та авторитети.	8		2	6
10. Набори елементів, які часто зустрічаються (часті набори елементів) Тема 10.1. Модель ринкова корзина. Тема 10.2. Ринкові корзини і алгоритм A-Priori.	8		2	6
11. Обробка великих наборів даних у основній пам'яті Тема 11.1. Основні алгоритми обробки даних в пам'яті. Тема 11.2. Алгоритми з обмеженням проходів. Тема 11.3. Підрахунок частих елементів у потоці.	10		2	8
12. Основні принципи роботи NumPy, SciPy, Matplotlib Тема 12.1. Принципи роботи з NumPy. Тема 12.2. Робота з SciPy. Тема 12.3. Використання бібліотеки Matplotlib.	10		2	8
13. Основи бібліотеки Pandas Тема 13.1. Структури даних Pandas. Тема 13.2. Робота з елементами Series. Тема 13.3. Робота з елементами DataFrame. Тема 13.4. Доступ до даних у структурах Pandas. Тема 13.5. Використання атрибутів для доступу до даних в Pandas.	12		2	10
14. Отримання випадкового набору з структур Pandas Тема 14.1. Робота з випадковим набором даних. Тема 14.2. Додавання елементів у структури. Тема 14.3. Індексція з використанням логічних виразів. Тема 14.4. Використання isin для роботи з даними в Pandas. Тема 14.5. Робота з пропусками в даних.	14		2	12
15. Робота Pandas у середовищі Jupyter Notebook інтегратора Anaconda Тема 15.1. Установка програмного забезпечення.	16		6	10

Назви розділів, тем	Кількість годин			
	Всього	У тому числі		
		Лекції	Практичні роботи	СРС
Тема 15.2. Запуск Jupyter Notebook. Приклади роботи. Тема 15.3. Основні елементи інтерфейсу. Тема 15.4. Робота з бібліотекою Pandas.				
Залік	8		0	8
Всього в семестрі:	150	-	36	114

13. Самостійна робота аспіранта

Метою проведення циклу лабораторних робіт є набуття студентами необхідних практичних навичок використання методів та способів представлення та опрацювання даних для отримання прихованої інформації, методів дослідження математичних моделей опису шуканої інформації, технології добування даних для розроблення прогнозів щодо досліджуваних об'єктів та процесів.

№ з/п	Назва самостійної роботи	Кількість ауд. годин
1	СР № 1. Створення зошита в середовищі Jupyter notebook.	2
2	СР № 2. Практичне застосування атрибутів та функцій модуля NumPy.	2
3	СР № 3. Вивчення та практичне застосування бібліотеки Matplotlib.	2
4	СР № 4. Вивчення та практичне застосування бібліотеки Pandas.	2
5	СР № 5. Первинна обробка набору даних "Yelp Dataset JSON" для тематичного дослідження.	2
6	СР № 6. Фільтрування великого набору даних фрагментами	2
7	СР № 7. Збереження структури DataFrame у форматі csv	2
8	СР № 8. Практичний аналіз текстових представлень великих обсягів даних	2
9	СР № 9. Візуалізація одержаних даних та підготовка презентації за результатами досліджень	2

14. Політика навчальної дисципліни (освітнього компонента)

Під час занять з навчальної дисципліни «Методи добування даних» аспіранти повинні дотримуватись певних дисциплінарних правил:

- забороняється запізнюватись на заняття;
- при вході викладача, на знак привітання, особи, які навчаються в КПІ ім. Ігоря Сікорського повинні встати;
- не допускаються сторонні розмови або інший шум, що заважає проведенню занять;
- виходити з аудиторії під час заняття допускається лише з дозволу викладача.
- не допускається користування мобільними телефонами та іншими технічними засобами без дозволу викладача.

Лабораторні роботи здаються особисто з попередньою перевіркою теоретичних знань, які необхідні для виконання лабораторної роботи. Перевірка практичних результатів включає перевірку коду та виконання тестових завдань.

В процесі навчання викладач має право нарахувати до 5 заохочувальних балів за дострокове виконання лабораторної роботи, за проявлений творчий підхід при виконанні індивідуального завдання або за активну участь у обговоренні питань, що пов'язані з тематикою лекції або практичного заняття.

За виконання та здачу лабораторної роботи після зазначеного дедлайну, за значну кількість пропущених занять, або за порушення правил поведінки на заняттях викладач може призначити до 5 штрафних балів.

При проведенні контрольних заходів та при виконанні лабораторних робіт аспіранти повинні дотримуватися правил академічної доброчесності. При виявленні значного відсотку списування або плагіату викладач може відмовити у прийнятті даної роботи та вимагати доброчесного виконання навчального плану.

15. Види контролю та рейтингова система оцінювання результатів навчання (PCO)

Види контролю з навчальної дисципліни «Методи добування даних» включають:

Поточний контроль: тестування закритими тестами.

Календарний контроль: провадиться двічі на семестр як моніторинг поточного стану виконання вимог силабусу.

Семестровий контроль: залік

Умови допуску до семестрового контролю: семестровий рейтинг більше 59 балів.

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

16. Додаткова інформація з дисципліни (освітнього компонента)

Перелік питань, які виносяться на семестровий контроль.

1. Поняття про методи аналізу даних.
2. Статистичні обмеження аналізу даних. Принцип Бонферроні.
3. Базові теоретичні положення та інструменти аналізу даних. Ефект Метью.
4. Метод пошуку найближчих сусідів та його застосування.
5. Збереження схожості скорочених наборів. Шинглінг документів.
6. Локально-чутливе хешування для документів.
7. Принципи формування міри відстані, відстані Евкліда, Жакарта, Хемінга. Косинусна відстань та відстань редагування.
8. Базові поняття теорії локально-чутливих функцій.
9. Сімейства локально-чутливих функцій для різних мір відстані.
10. Поняття про потокові дані. Базова модель поточкових даних.
11. Способи вибору даних в потоці.
12. Методи фільтрації потоків.
13. Способи організації підрахунку різних елементів у потоці.
14. Способи визначення моментів та практичні приклади.
15. Поняття згасання вікон та підрахунку у вікні.
16. PageRank, як інструмент для класифікації сторінок.
17. Способи ефективного обчислення для PageRank.
18. Тематично чутливий PageRank.
20. «Лінковий» спам, хаби та авторитети.
21. Модель аналізу на основі задачі про ринкову корзину
22. Використання алгоритму A-Priori для задач про ринкову корзину.
23. Методи обробки великих наборів даних у основній пам'яті.
24. Основні алгоритми обробки даних в пам'яті, алгоритм обмеження проходів.
25. Основні бібліотеки на мові Python для аналізу даних: модулі NumPy та SciPy, бібліотека Matplotlib.
26. Основні поняття про бібліотеку Pandas, структури даних Pandas.
27. Робота в Pandas з елементами Series та DataFrame.
28. Способи доступу до даних у структурах бібліотеки Pandas.
29. Використання атрибутів для доступу до даних в Pandas.
30. Отримання випадкового набору даних з структур бібліотеки Pandas.
31. Додавання елементів та індексація з використанням логічних виразів при роботі зі структурами бібліотеки Pandas.
32. Використання isin для роботи з даними в Pandas, робота з пропусками даних.
33. Робота з бібліотекою Pandas у середовищі Jupyter Notebook інтегратора Anaconda.
34. Порядок та основні етапи установки програмного забезпечення для вирішення задач аналізу та добування даних.
35. Реалізація практичної задачі добування даних на основі завдання, заданого викладачем.

Умова зарахування додаткових балів.

В рамках вивчення навчальної дисципліни «Методи добування» допускається зарахування балів, одержаних в результаті дистанційних курсів на платформі “Coursera”, за умови попереднього погодження програми даного курсу з викладачем та за умови отримання офіційного сертифікату.

Робочу програму навчальної дисципліни (силабус):

Складено , д.т.н, професор, Новотарський Михайло Анатолійович

Ухвалено кафедрою обчислювальної техніки (протокол № 10 від 27 05 2020 р.)

Погоджено Методичною комісією ФІОТ (протокол № 10 від 18 06 2020 р.)