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

Кафедра штучного інтелекту

ОБРОБКА НАДВЕЛИКИХ МАСИВІВ ДАНИХ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	***Другий (магістерський)***
Галузь знань	*12 Інформаційні технології*
Спеціальність	*124 Системний аналіз*
Освітня програма	*Системний аналіз і управління*
Статус дисципліни (код)	*Нормативна*
Форма навчання	*очна(денна)*
Рік підготовки, семестр	*1 курс, осінній семестр*
Обсяг дисципліни	*4 кредити ЄКТС (36 год. лекцій, 18 год. лабораторних, 66 год. СРС)*
Семестровий контроль/ контрольні заходи	*Залік, МКР*
Розклад занять	*https://schedule.kpi.ua* *2 год. лекційних та 1 год. лабораторних робіт на тиждень*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лектор: *к.ф.-м.н., доцент, Пишнограєв Іван Олександрович, pyshnograiev@wdc.org.ua* Лабораторні: *к.ф.-м.н., доцент, Пишнограєв Іван Олександрович*
Розміщення курсу	Google classroom https://classroom.google.com/c/NjgxMjU5ODk5NTAz?cjc=wwwa7yc

Програма навчальної дисципліни

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

Дисципліна є нормативною в освітній програмі. Вивчення навчальної дисципліни націлено на формування, розвиток та закріплення у здобувачів таких загальних та фахових компетентностей:
ЗК 01 Здатність до абстрактного мислення, аналізу та синтезу,
ЗК 03 Здатність до пошуку, оброблення та аналізу інформації з різних джерел,
ФК 06 Здатність застосовувати теорію і методи Data Science для здійснення інтелектуального аналізу даних з метою виявлення нових властивостей та генерації нових знань про складні системи,
ФК 07 Здатність управляти робочими процесами у сфері інформаційних технологій, які є складними, непередбачуваними та потребують нових стратегічних підходів,

Внаслідок вивчення курсу студент повинен бути здатний продемонструвати такий програмний результат навчання ОПП:
ПРН 2 Будувати та досліджувати моделі складних систем і процесів застосовуючи методи системного аналізу, математичного, комп'ютерного та інформаційного моделювання,
ПРН 6 Застосовувати методи машинного навчання та інтелектуального аналізу даних, математичний апарат нечіткої логіки, теорії ігор та розподіленого штучного інтелекту для розв'язання складних задач системного аналізу,
ПРН 8 Здійснювати ідентифікацію та оцінювання параметрів математичних моделей об'єктів керування,
ПРН 12 Знати законодавчі акти щодо забезпечення захисту інтелектуальної власності, вимоги до дотримання установлених вимог при оформленні заявок з патентів на винаходи; дотримуватися академічної доброчесності.

*У кінці вивчення курсу студент повинен **знати**:*

- *особливості роботи з великими даними;*
- *методи обробки та аналізу надвеликих даних;*
- *засоби обробки, зберігання та аналізу великих масивів даних;*

***вміти**:*

- *аналізувати великі масиви даних;*
- *створювати та модифікувати алгоритми роботи з великими даними;*
- *налагоджувати засоби обробки та зберігання великих масивів даних.*

.

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

Дисципліна базується на знаннях та навичках суміжних дисциплін, що вивчаються на попередньому освітньому рівні. Дана дисципліна передує ОК ПО 07 «Теорія управління і прогнозування в складних системах. Курсова робота» та може являтись однією з головних складових магістерської дисертації.

Зміст навчальної дисципліни:

Розділ 1. Методи інтелектуального аналізу даних

Тема 1.1. Вступ до великих даних.

1. *Загальні відомості про великі дані;*
2. *Основні виклики великих даних;*
3. *Що таке великі дані;*
4. *Основні особливості та огляд методів роботи з великими даними.*

Тема 1.2. Огляд методів математичної статистики та Data Mining.

1. *Кореляційно-регресійний аналіз;*
2. *Перевірка даних на помилки, заповнення пропусків;*
3. *Нормалізація даних, шкали;*
4. *Методи прогнозування та передбачення.*

Тема 1.3. Візуалізація великих даних.

1. *Основні проблеми;*
2. *Типи візуалізацій даних;*
3. *Приклади успішних представлень;*
4. *Основні напрями досліджень;*
5. *Системи для представлення даних.*

Тема 1.4. Огляд особливостей методів Machine Learning та інших методів штучного інтелекту.

1. *Задачі кластеризації великих даних;*
2. *Задачі класифікації великих даних;*
3. *Зниження розмірності простору даних;*
4. *Нейронні мережі та великі дані;*
5. *Приклади проєктів.*

Тема 1.5. Використання мови R для аналізу великих даних.

1. *Огляд основних можливостей роботи з великими даними;*
2. *Огляд функцій і бібліотек Data Mining;*
3. *Огляд функцій і бібліотек візуалізації даних;*
4. *Огляд функцій і бібліотек Machine Learning.*

Розділ 2. Методи аналізу надвеликих масивів даних

Тема 2.1. Hadoop та MapReduce.

1. *Огляд технологій збереження великих даних та їх обробки;*
2. *Огляд технологій Hadoop та MapReduce;*
3. *NoSQL бази даних, їх особливості, переваги та недоліки;*
4. *Мікросередовище Hadoop, супутні інструменти.*

Тема 2.2. R і Hadoop

1. *Огляд бібліотек і можливостей;*
2. *Паралельні обчислення.*

Тема 2.3. Обробка слабкоструктурованих даних.

1. *Що таке слабкоструктуровані дані, їх особливості;*

2. Огляд інструментів для роботи.

Тема 2.4. Теорія ігор та великі дані.

1. Використання методів теорії ігор для обробки великих даних;

2. Приклади задач теорії ігор, що використовують великі дані.

Тема 2.5. Обробка потокових даних.

1. Що таке потокові дані, їх особливості;

2. Огляд інструментів для роботи.

Тема 2.6. Робота з текстовими даними та соціальними мережами.

1. Особливості; роботи з текстовими даними та соціальними мережами;

2. Огляд інструментів для роботи.

3. Навчальні матеріали та ресурси

Базова:

1. *Згуровский, М. З., Zgurovsky, M., Згуровский, М. З., Згуровський Михайло Захарович, Згуровський, М. З., Zgurovsky, M. Z., . Zaychenko, Y. (2020). Big Data: Conceptual Analysis and Applications. Cham, Switzerland: Springer. https://link.springer.com/book/10.1007/978-3-030-14298-8*
2. *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 (1 файл: 6,95 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2021. – 168 с. https://ela.kpi.ua/handle/123456789/46129*
3. *SpringerLink (Online service), Zomaya, A. Y., & Sakr, S. (2017). Handbook of Big Data Technologies (1st ed. 2017.). Cham: Springer International Publishing. https://link.springer.com/book/10.1007/978-3-319-49340-4*
4. *Staudacher, J., & Anwander, J. (2019). Using the R package CoopGame for the analysis , solution and visualization of cooperative games with transferable utility. https://cran.r-project.org/web/packages/CoopGame/vignettes/UsingCoopGame.pdf*

Допоміжна:

4. *IoT Fundamentals: Big Data & Analytics // Електронний ресурс. Режим доступу: https://www.netacad.com/courses/iot/big-data-analytics*
5. *R програмування // Електронний ресурс. Режим доступу: https://coderlessons.com/tutorials/mashinnoe-obuchenie/r-programmirovanie/r- programmirovanie*
6. *Virtualization Technology // Електронний ресурс. Режим доступу: https://www.sciencedirect.com/topics/computer-science/virtualization-technology*
7. *Apache Hadoop // Електронний ресурс. Режим доступу:: http://hadoop.apache.org/*
8. *Apache Spark // Електронний ресурс. Режим доступу: https://spark.apache.org/*
9. *Аналіз даних в Spark-кластері за допомогою пакета dplyr // Електронний ресурс. Режим доступу: https://r-analytics.blogspot.com/2020/03/spark-dplyr.html*
10. *Таран, В. І. Технології Big Data. Практикум [Електронний ресурс] : навчальний посібник для здобувачів ступеня магістра за освітньою програмою «Комп'ютерні системи та мережі» спеціальності 123 Комп'ютерна інженерія / В. І. Таран, Ю. Г. Гордієнко, С. Г. Стіренко ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 2,27 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2022. – 56 с. https://ela.kpi.ua/bitstream/123456789/50014/1/BigData_praktykum.pdf*
11. *Олещенко, Л. М. Технології оброблення великих даних. Конспект лекцій [Електронний ресурс] : навчальний посібник для студентів спеціальності 121 «Інженерія програмного забезпечення» (освітня програма «Інженерія програмного забезпечення мультимедійних та інформаційно-пошукових систем») / Л. М. Олещенко ; КПІ ім. Ігоря Сікорського. – Електронні текстові дані (1 файл: 5,55 Мбайт). – Київ : КПІ ім. Ігоря Сікорського, 2021. – 227 с. https://ela.kpi.ua/handle/123456789/42206*
12. *Giacomo Bonanno. GAME THEORY. 2nd Edition. CreateSpace Independent Publishing Platform. 2018. 592 p. http://faculty.econ.ucdavis.edu/faculty/bonanno/PDF/GT_book.pdf*
13. *Georgios Chalkiadakis, Edith Elkind, and Michael Wooldridge. Computational Aspects of Cooperative Game Theory.2012. 150 p. DOI 10.2200/S00355ED1V01Y201107AIM016*

14. *Луцків, Андрій Мирославович. Паралельні та розподілені обчислення : підручник для студентів вищих навчальних закладів / А. Луцків, С. Лупенко, В. Пасічник. – Львів : Видавництво "Магнолія 2006", 2021. – 565 с. Замовити в Бібліотеці КПІ: https://opac.kpi.ua/F/?func=direct&doc_number=000636469&local_base=KPI01*

Навчальний контент

4. Методика опанування навчальної дисципліни (освітнього компонента)

Назви розділів і тем	Кількість годин				
	Всього	у тому числі			
		Лекції	Практ.	Лаб. роб.	СРС
1	2	3	4	5	6
Розділ 1. *Методи інтелектуального аналізу даних*					
Тема 1. Вступ до великих даних	8	2	-	2	4
Тема 2. Огляд методів математичної статистики та Data Mining.	10	4	-	2	4
Тема 3. Візуалізація даних.	6	2	-	-	4
Тема 4. Огляд методів Machine Learning та інших методів штучного інтелекту.	8	4	-	-	4
Тема 5. Використання мови R для аналізу даних.	10	2	-	4	4
Разом за розділом 1	42	14	-	8	20
Розділ 2. *Методи аналізу надвеликих масивів даних*					
Тема 1. Hadoop та MapReduce.	8	2	-	2	4
Тема 2. R і Hadoop.	8	2	-	2	4
Тема 3. Обробка слабкоструктурованих даних.	9	4	-	2	3
Тема 4. Теорія ігор та великі дані.	4	2	-	-	2
Тема 5. Обробка потокових даних	9	4	-	2	3
Тема 6. Робота з текстовими даними та соціальними мережами.	10	4	-	2	4
Модульна контрольна робота	18	2	-	-	16
Разом за розділом 2	66	20	-	10	36
Залік	12	2	-	-	10
Всього годин	120	36	-	18	66

Перелік лекційних занять

№ з/п	Назва теми лекції та перелік основних питань (перелік дидактичних засобів, посилання на літературу та завдання на СРС)
1	**Вступ до великих даних.** *Лекція №1* 1. Загальні відомості про великі дані;

	2. Основні виклики великих даних; 3. Що таке великі дані; 4. Основні особливості та огляд методів роботи з великими даними.
2	**Огляд методів математичної статистики та Data Mining.** *Лекція №2,3* 1. Кореляційно-регресійний аналіз; 2. Перевірка даних на помилки, заповнення пропусків; 3. Нормалізація даних, шкали; 4. Методи прогнозування та передбачення.
3	**Візуалізація даних.** *Лекція №4* 1. Основні проблеми; 2. Типи візуалізацій даних; 3. Приклади успішних представлень; 4. Основні напрями досліджень; 5. Системи для представлення даних.
4	**Огляд методів Machine Learning.** *Лекція №5,6* 1. Що таке Machine Learning; 2. Задачі кластеризації; 3. Задачі класифікації; 4. Зниження розмірності простору даних; 5. Нейронні мережі; 6. Приклади проектів.
5	**Використання мови R для аналізу даних.** *Лекція №7* 1. Огляд основних можливостей; 2. Огляд функцій і бібліотек Data Mining; 3. Огляд функцій і бібліотек візуалізації даних; 4. Огляд функцій і бібліотек Machine Learning.
6	**Hadoop та MapReduce.** *Лекція №8* 1. Огляд технологій збереження великих даних та їх обробки; 2. Огляд технологій Hadoop та MapReduce; 3. NoSQL бази даних, їх особливості, переваги та недоліки; 4. Мікросередовище Hadoop, супутні інструменти.
7	**R і Hadoop.** *Лекція №9* 1. Огляд бібліотек і можливостей; 2. Паралельні обчислення.
8	**Обробка слабкоструктурованих даних.** *Лекція №10,11* 1. Що таке слабкоструктуровані дані, їх особливості; 2. Огляд інструментів для роботи.
	Теорія ігор та великі дані. *Лекція №12* 1. Використання методів теорії ігор для обробки великих даних; 2. Приклади задач теорії ігор, що використовують великі дані.
9	**Обробка потокових даних.**

	Лекція №13,14 1. Що таке потокові дані, їх особливості; 2. Огляд інструментів для роботи.
10	**Робота з текстовими даними та соціальними мережами.** *Лекція №15,16* 1. Особливості; роботи з текстовими даними та соціальними мережами; 2. Огляд інструментів для роботи;

Перелік лабораторних робіт

Ціллю є формування умінь та навичок практичного використання певних оболонок, програм з використанням ПЕОМ для розв'язування задач пов'язаних з обробкою великих даних.

№ з/п	Назва комп'ютерного практикуму	Кількість ауд. годин
1	Data Mining за допомогою базових інструментів аналізу даних.	4
2	Machine Learning засобами мови R.	4
3	Установка та налагодження програмного середовища для обробки великих даних. Виконання паралельних обчислень засобами R.	6
4	Спеціальні алгоритми обробки великих даних.	4

5. Самостійна робота студента

Індивідуальні завдання складаються з підготовки до лабораторних робіт, модульної контрольної роботи та опрацюванні лекційного матеріалу. Розподіл СРС за темами представлено в п. 4.

Політика та контроль

6. Політика навчальної дисципліни (освітнього компонента)

Усі роботи студенти мають прикріплювати в особистому кабінеті гугл-класу. Дедлайни кожного завдання позначені в щотижневих завданнях у гугл-класі. Роботи мають бути виконані з дотриманням академічної доброчесності. Політика та принципи академічної доброчесності, етична поведінка студентів визначені у Кодексі честі https://kpi.ua/code. Лектор може запропонувати студентам пройти онлайн-курси на платформі Coursera. Також сертифікати цих курсів можуть бути частково зараховані згідно до Положення.
Тематика лабораторних робіт спрямована на поглиблення засвоєного матеріалу лекцій.

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

*Семестровий контроль: **залік**.*
Семестровий рейтинг студента з дисципліни виставляється лектором та складається з балів, що він отримує за:

~виконання модульної контрольної;
~виконання 4 лабораторних робіт.

Критерії нарахування балів за семестр:

1) Модульна контрольна робота оцінюється у 20 балів.
2) Кожна з лабораторних робіт оцінюється в 20 балів.

За кожний тиждень запізнення з поданням роботи на перевірку нараховується штрафний – 2 бали.
Критерії нарахування балів за контрольні заходи:

- *"відмінно": 95 -100% - здобувач виявив всебічні, систематичні та глибокі знання навчального матеріалу з дисципліни; продемонстрував уміння вільно виконувати всі завдання, передбачені програмою; засвоїв основну та додаткову літературу; проявив творчі здібності в розумінні, в логічному, чіткому, стислому та ясному трактуванні навчального матеріалу; засвоїв взаємозв'язок основних понять дисципліни, їх значення для подальшої професійної діяльності*
- *"дуже добре": 85-94% - здобувач виявив систематичні знання навчального матеріалу з дисципліни вище середнього рівня; продемонстрував уміння добре виконувати всі завдання,*

передбачені програмою, допустивши незначні помилки; засвоїв основну та додаткову літературу; засвоїв взаємозв'язок основних понять дисципліни, їх значення для подальшої професійної діяльності

- *"добре": 75-84% - здобувач виявив загалом добрі знання навчального матеріалу при виконанні передбачених програмою завдань, але припустив ряд помітних помилок; засвоїв основну літературу; показав систематичний характер знань з дисципліни; здатний до їх самостійного використання та поповнення в процесі подальшої навчальної роботи і професійної діяльності*
- *"задовільно": 65-74% - здобувач виявив знання основного навчального матеріалу з дисципліни в обсязі, необхідному для подальшого навчання та майбутньої професійної діяльності; ознайомився з основною літературою; впорався з виконанням завдань, передбачених програмою, але припустив значну кількість помилок або недоліків на запитання при співбесіді, тестуванні та при виконанні завдань тощо, принципові з яких може усунути самостійно*
- *"достатньо": 60-64% - здобувач виявив знання основного навчального матеріалу з дисципліни в мінімальному обсязі, необхідному для подальшого навчання та майбутньої професійної діяльності; ; ознайомився з основною літературою; в основному виконав завдання, передбачені програмою, але припустив помилки у відповіді на запитання при співбесідах, тестуванні та при виконанні завдань тощо, які він може усунути лише під керівництвом та за допомогою викладача*
- *"незадовільно": 30-54% - здобувач мав значні прогалини в знаннях основного навчального матеріалу; допускав принципові помилки при виконанні передбачених програмою завдань, але спроможній самостійно доопрацювати програмний матеріал і підготуватися для перездачі дисципліни*
- *"незадовільно": 0-29% - здобувач не мав знань зі значної частини навчального матеріалу з дисципліни; допускав принципові помилки при виконанні більшості передбачених програмою завдань або не виконував ці завдання*

Умовою першої атестації є поточний рейтинг не менше 30% запланованих балів за семестр. Умова другої атестації ~ поточний рейтинг не менше 50% запланованих балів.

Необхідною умовою допуску до заліку є зарахування всіх комп'ютерних практикумів та 40 балів семестрового рейтингу. Студенти, які мають менше 40 балів або не здали комп'ютерні практикуми, не допускаються до здачі заліку. Залік виставляється за результатом семестрового рейтингу, рейтинг повідомляється на передостанньому занятті.
Якщо студент не набрав 60 балів або хоче покращити свою оцінку, можливо написати залікову роботу. На залікову роботу виноситься 100 балів, бали за семестр в цьому випадку анулюються. Залік проводиться у вигляді письмової роботи, в якій три теоретичних питання та два практичних. Кожне завдання оцінюється в 20 балів за такими критеріями:

- *«відмінно», повна відповідь, не менше 90% потрібної інформації, що виконана згідно з вимогами до рівня «умінь», (повне, безпомилкове розв'язування завдання) – 18-20 балів;*
- *«добре», достатньо повна відповідь, не менше 75% потрібної інформації, що виконана згідно з вимогами до рівня «умінь або є незначні неточності (повне розв'язування завдання з незначними неточностями) – 15-17 балів;*
- *«задовільно», неповна відповідь, не менше 60% потрібної інформації, що виконана згідно з вимогами до «стереотипного» рівня та деякі помилки (завдання виконане з певними недоліками) – 12-14 балів;*
- *«незадовільно», відповідь не відповідає умовам до «задовільно» – 0 балів.*

Сума рейтингових балів, отриманих студентом протягом семестру, переводиться до підсумкової оцінки згідно з таблицею.

Таблиця відповідності рейтингових балів оцінкам за університетською шкалою:

Бали:	*Оцінка*
100…95	Відмінно
94…85	Дуже добре
84…75	Добре
74…65	Задовільно

64…60	Достатньо
Менше 60	Незадовільно
не зараховано лабораторні роботи або менше 36	Не допущено

8. Додаткова інформація з дисципліни (освітнього компонента)

Теоретичні питання:

1. Поняття Big Data. Головні особливості та виклики.
2. Помилки в даних. Їх ідентифікація. Заповнення пропусків.
3. Нормалізація даних. Особливості, способи.
4. Кореляційно-регресійний аналіз.
5. Кластеризація і класифікація, їх особливості.
6. Зниження розмірності даних.
7. Методи проведення кластеризації і класифікації.
8. Тестування гіпотез. Гіпотези про розподіл даних, значення математичного сподівання, квадратичного відхилення. Різниця між генеральною сукупністю та вибіркою.
9. Нейронні мережі.
10. Візуалізація даних, її особливості. Стандартні елементи, графіки і їх типи, ефективність графіків.
11. Hadoop. Map Reduce.
12. Реляційні та NoSQL бази даних. Основні відмінності.
13. Потокова обробка даних.
14. Обробка слабкоструктурованих даних.
15. Методи теорії ігор для обробки великих даних.

Робочу програму навчальної дисципліни (силабус):

Складено доцент, к.ф.-м.н., доцент *Пишограєв Іван Олександрович*

Ухвалено кафедрою ШІ (протокол № 14 від 11.05.2024)

Погоджено Методичною комісією НН ІПСА (протокол № 10 від 24.06.2024)