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

КАФЕДРА ШТУЧНОГО ІНТЕЛЕКТУ

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	*122 Комп'ютерні науки*
Освітня програма	*Системи і методи штучного інтелекту*
Статус дисципліни (код)	*Нормативна*
Форма навчання	*очна(денна)/дистанційна/змішана*
Рік підготовки, семестр	*1 курс, осінній семестр*
Обсяг дисципліни	*4 кредити ЄКТС*
Семестровий контроль/ контрольні заходи	*Екзамен*
Розклад занять	*rozklad.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/u/1/c/MjAwMTM4NDc5NjM5

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 02 Здатність застосовувати знання у практичних ситуаціях,
ЗК 05 Здатність вчитися й оволодівати сучасними знаннями,
СК 01 Усвідомлення теоретичних засад комп'ютерних наук,
СК 03 Здатність використовувати математичні методи для аналізу формалізованих моделей предметної області,
СК 04 Здатність збирати і аналізувати дані (включно з великими), для забезпечення якості прийняття проєктних рішень,
СК 05 Здатність розробляти, описувати, аналізувати та оптимізувати архітектурні рішення інформаційних та комп'ютерних систем різного призначення,
СК 06 Здатність застосовувати існуючі і розробляти нові алгоритми розв'язування задач у галузі комп'ютерних наук,
СК 07 Здатність розробляти програмне забезпечення відповідно до сформульованих вимог з урахуванням наявних ресурсів та обмежень,
СК 08 Здатність розробляти і реалізовувати проєкти зі створення програмного забезпечення, у тому числі в непередбачуваних умовах, за нечітких вимог та необхідності застосовувати нові стратегічні підходи, використовувати програмні інструменти для організації командної роботи над проєктом,
СК 09 Здатність розробляти та адмініструвати бази даних та знань,

СК 13 Здатність розробляти та застосовувати технології розподілених високопро-дуктивних обчислень, грід-технології.

Внаслідок вивчення курсу студент повинен бути здатний продемонструвати такий програмний результат навчання ОПП:
РН 1 Мати спеціалізовані концептуальні знання, що включають сучасні наукові здобутки у сфері комп'ютерних наук і є основою для оригінального мис- лення та проведення досліджень, критичне осмислення проблем у сфері комп'ютерних наук та на межі галузей знань,
РН 2 Мати спеціалізовані уміння/навички розв'язання проблем комп'ютерних наук, необхідні для проведення досліджень та/або провадження інновацій- ної діяльності з метою розвитку нових знань та процедур,
РН 4 Управляти робочими процесами у сфері інформаційних технологій, які є складними, непередбачуваними та потребують нових стратегічних підходів,
РН 6 Розробляти концептуальну модель інформаційної або комп'ютерної системи,
РН 7 Розробляти та застосовувати математичні методи для аналізу інформаційних моделей,
РН 8 Розробляти математичні моделі та методи аналізу даних (включно з великими),
РН 9 Розробляти алгоритмічне та програмне забезпечення для аналізу даних (включно з великими),
РН 11 Створювати нові алгоритми розв'язування задач у сфері комп'ютерних наук, оцінювати їх ефективність та обмеження на їх застосування,
РН 12 Проєктувати та супроводжувати бази даних та знань,
РН 14 Тестувати програмне забезпечення,
РН 15 Виявляти потреби потенційних замовників щодо автоматизації обробки інформації.

*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 передує дисциплінам «Інтелектуальні системи прийняття рішень» 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.
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. Обробка потокових даних.
1. *Що таке потокові дані, їх особливості;*
2. *Огляд інструментів для роботи.*

Тема 2.5. 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*

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)

У гугл-класі будуть щотижневі завдання з детальними інструкціями та необхідним матеріалом, які необхідно вчасно виконувати.

5. Самостійна робота студента

Індивідуальні завдання складаються з підготовки до комп'ютерних практикумів та опрацюванні лекційного матеріалу.

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 робіт лабораторних робіт спрямована на поглиблення засвоєного матеріалу лекцій. На заняттях комп'ютерного практикуму розв'язуються задачі та вправи по темам лекції.

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

*Семестровий контроль: **екзамен**.*
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 контрольні роботи по 10 балів).
2) Кожна з лабораторних робіт оцінюється в 10 балів.

За кожний тиждень запізнення з поданням роботи на перевірку нараховується штрафний – 1 бал.
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 60% запланованих балів.

Необхідною умовою допуску до екзамену є зарахування всіх лабораторних робіт та 36 балів семестрового рейтингу. Студенти, які мають менше 36 балів не допускаються до здачі екзамену. На екзамен виноситься 40 балів. Екзамен проводиться у вигляді письмової роботи, в якій два теоретичних питання та два практичних. Кожне завдання оцінюється в 10 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) – 9-10 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) – 8 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) – 7 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.

Робочу програму навчальної дисципліни (силабус):

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

Ухвалено кафедрою ШІ (протокол № 14 від 24.05.2023)

Погоджено Методичною комісією НН ІПСА (протокол № 4 від 16.06.2023)