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

Кафедра програмного забезпечення комп'ютерних систем

ТЕХНОЛОГІЇ ОБРОБЛЕННЯ ВЕЛИКИХ ДАНИХ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

Рівень вищої освіти	***Перший (бакалаврський)***
Галузь знань	*12 Інформаційні технології*
Спеціальність	*121 Інженерія програмного забезпечення*
Освітня програма	*Інженерія програмного забезпечення мультимедійних та інформаційно-пошукових систем*
Статус дисципліни	*Вибіркова*
Форма навчання	*Очна (денна)*
Рік підготовки, семестр	*3 рік підготовки, 5 семестр*
Обсяг дисципліни	*Лекції: 36 год., комп'ютерний практикум: 18 год., самостійна робота: 66 год.*
Семестровий контроль/ контрольні заходи	*Залік, модульна контрольна робота, календарний контроль*
Розклад занять	*Згідно розкладу на весняний семестр поточного навчального року (rozklad.kpi.ua)*
Мова викладання	*Українська*
Інформація про керівника курсу / викладачів	Лектор: *к.т.н., доцент, Олещенко Любов Михайлівна, oleshchenkoliubov@gmail.com* Комп'ютерний практикум: *к.т.н., доцент, Олещенко Любов Михайлівна, oleshchenkoliubov@gmail.com*
Розміщення курсу	Google classroom: *https://classroom.google.com/u/2/c/MjQ3NjM5MTUwODg1*

Програма навчальної дисципліни

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

Вивчення дисципліни «Технології оброблення великих даних» дозволяє сформувати у здобувачів освіти компетенції, необхідні для розв'язання практичних задач професійної та наукової діяльності, пов'язаної з підготовкою та аналізом великих даних.

***Метою** вивчення дисципліни «Технології оброблення великих даних» є формування у студентів здатностей застосовувати програмні методи та засоби обробки даних для аналізу великих даних та прийняття управлінських рішень в різних галузях науки та бізнесу.*

***Предметом** дисципліни «Технології оброблення великих даних» є загальні принципи та підходи до оброблення великих даних, можливості та технології мов програмування Python та R для оброблення, аналізу та візуалізації великих даних.*

Вивчення дисципліни «Технології оброблення великих даних» сприяє формуванню у студентів фахової компетентності (ФК) за освітньою програмою:

***ФК08** Здатність застосовувати фундаментальні і міждисциплінарні знання для успішного розв'язання завдань інженерії програмного забезпечення.*

Вивчення дисципліни «Технології оброблення великих даних» сприяє формуванню у студентів наступних програмних результатів навчання (ПРН) за освітньою програмою:

ПРН01 *Аналізувати, цілеспрямовано шукати і вибирати необхідні для вирішення професійних завдань інформаційно-довідникові ресурси і знання з урахуванням сучасних досягнень науки і техніки.*

ПРН07 *Знати і застосовувати на практиці фундаментальні концепції, парадигми і основні принципи функціонування мовних, інструментальних і обчислювальних засобів інженерії програмного забезпечення.*

ПРН18 *Знати та вміти застосовувати інформаційні технології обробки, зберігання та передачі даних.*

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

Успішному вивченню дисципліни «Технології оброблення великих даних» передує вивчення дисциплін «Програмування» та «Бази даних» навчального плану підготовки бакалаврів за спеціальністю 121 Інженерія програмного забезпечення.

Отримані при засвоєнні дисципліни «Технології оброблення великих даних» теоретичні знання та практичні уміння сприяють успішному виконанню курсових проєктів та дипломної роботи бакалаврів.

3. Зміст навчальної дисципліни

Дисципліна «Технології оброблення великих даних» передбачає вивчення таких тем:

Тема 1. Джерела та типи великих даних. Підготовка та аналіз даних в Python.

Тема 2. Можливості мови R. Архітектурні моделі Big Data.

Модульна контрольна робота.

Залік.

4. Навчальні матеріали та ресурси

Базова література:

1. Олещенко Л. М. Технології оброблення великих даних: конспект лекцій з дисципліни «Технології оброблення великих даних»: навч. посіб. для студ. спеціальності 121 «Інженерія програмного забезпечення» / Л.М. Олещенко; КПІ ім. Ігоря Сікорського. – 2021. – 225 с.

2. Олещенко Л. М. Технології оброблення великих даних: комп'ютерний практикум з дисципліни «Технології оброблення великих даних»: навч. посіб. для студ. спеціальності 121 «Інженерія програмного забезпечення» / Л.М. Олещенко; КПІ ім. Ігоря Сікорського. – 2021. – 85 с.

Додаткова література:

1. MapReduce. Tutorial. Електронний ресурс. Режим доступу: https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html

2. Programming with Databases – Python. Електронний ресурс. Режим доступу: https://swcarpentry.github.io/sql-novice-survey/10-prog/index.html

3. Pandas. Електронний ресурс. Режим доступу: https://www.w3schools.com/python/pandas/pandas_intro.asp

4. Мова програмування R. Електронний ресурс. Режим доступу: https://coderlessons.com/tutorials/mashinnoe-obuchenie/r-programmirovanie/r-programmirovanie

Матеріали знаходяться у вільному доступі в Інтернеті.

Навчальний контент

5. Методика опанування навчальної дисципліни (освітнього компонента)

№ з/п	Тип навчального заняття	Опис навчального заняття
Тема 1. Джерела та типи великих даних. Підготовка та аналіз даних в Python.		
1	*Лекція 1. Джерела великих даних. Інтернет Речей. Визначення Big Data.*	*Інтернет Речей та зростання даних. Платформа Kaggle. DrivenData. Визначення великих даних. Приклади великих даних у реальному світі. Відкриті дані. Приватність даних. Структуровані та неструктуровані дані. Хмарні та туманні обчислення. Дані в спокої та дані в русі. Інфраструктура великих даних. Розподілені дані та їх обробка.*
2	*Комп'ютерний практикум 1. Дослідження джерел відкритих даних. Завантаження датасету та збереження даних в форматі csv.*	*Завдання: дослідити джерела відкритих даних за допомогою Open Government Partnership та вебсайтів, які надають відкриті дані, можливості збереження та візуалізації даних, використовуючи вебсайти www.knoema.com та www.gapminder.org, дослідити право власності на персональні дані, коли ці дані не зберігаються локально та обмеження електронних таблиць при завантаженні даних.*
3	*Лекція 2. Розроблення програмного забезпечення для аналізу вебсайтів, які надають відкриті дані за допомогою Python Pandas. Відкриті дані, їх формати та засоби обробки.*	*Можливості інструментів аналізу даних. Роль Python в аналізі даних. Традиційна аналітика великих даних та аналітика нового покоління. Життєвий цикл аналізу даних. Відкриті дані, їх формати та засоби обробки. Вебскрепінг. Витягування, перетворення та завантаження даних.*
4	*Комп'ютерний практикум 2. Аналіз та візуалізація даних у Python.*	*Завдання: продемонструвати свої знання про життєвий цикл аналізу даних, використовуючи заданий набір даних та вказані інструменти, імпортувати пакети Python, необхідні для аналізу набору даних, використати засоби Python та Jupyter, щоб підготувати ці дані до аналізу, проаналізувати їх, побудувати графіки.*
5	*Лекція 3. Форматування даних про час та дату, читання та запис файлів в Python.*	*Форматування даних про час та дату у Python. Читання та запис файлів в Python. Взаємодія із зовнішніми додатками.*
6	*Лекція 4. Програмування Python та SQLite. Призначення утиліти csvsql.*	*Основні операції SQL. Робота Python з SQLite. Призначення утиліти csvsql. Метод execute().*

7	*Лекція 5. Процедура імпорту даних із файлів у Pandas. Імпорт даних з Інтернету за допомогою Pandas. Засоби для кореляційного аналізу в Pandas.*	*Статистичні підходи до аналітики великих даних. Використання Pandas. Імпорт даних з файлів. Імпорт даних з мережі Інтернет. Описова статистика в Pandas. Засоби для кореляційного аналізу в Pandas.*
8	*Комп'ютерний практикум 3. Кореляційний аналіз у Python.*	*Завдання: продемонструвати свої навички виконання кореляційного аналізу даних, використовуючи заданий набір даних та вказані інструменти Python для обчислення кореляції. Потрібно налаштувати набір даних, визначити, чи змінні в даному наборі даних є корельованими, використати Python для обчислення кореляції між двома наборами змінних та здійснити візуалізацію результатів дослідження.*
9	*Лекція 6. Оброблення відсутніх даних. Перетворення типів даних та маніпулювання дата фреймами.*	*Оброблення відсутніх даних. Перетворення типів даних. Маніпулювання дата фреймами у Python.*
10	*Лекція 7. Регресійний аналіз даних в Python.*	*Регресійний аналіз. Типи регресійного аналізу. Застосування регресійного аналізу для аналізу даних.*
11	*Комп'ютерний практикум 4. Побудова лінійної регресії в Python.*	*Завдання: ознайомитись з поняттями лінійної регресії та роботи з даними для прогнозування в Python, проаналізувати запропоновані дані про продажі та побудувати лінійну регресію для прогнозування річного чистого обсягу продажів на основі кількості магазинів у районі.*
12	*Лекція 8. Помилки в аналізі даних та прогнозній аналітиці. Оцінка помилок регресії засобами Python.*	*Помилки в аналізі даних та прогнозній аналітиці. Оцінка помилок регресії засобами Python. Призначення бібліотеки scikit-learn.*
13	*Лекція 9. Алгоритми класифікації даних. Застосування та проблеми класифікацій.*	*Проблеми класифікації. Алгоритми класифікації. Візуалізація класифікацій. Застосування та валідація класифікацій. Модель класифікатора дерева рішень.*
14	*Лекція 10. Модуль Pyplot. Інструмент Plotly. Типи візуалізації даних. Візуалізація аномалій. Використання бібліотек Folium та Leaflet.js для побудови карт.*	*Модуль Pyplot. Інструмент Plotly. Типи візуалізації даних. Візуалізація аномалій. Використання бібліотек Folium та Leaflet.js для побудови карт.*

Тема 2. Можливості мови R. Архітектурні моделі Big Data.		
15	*Лекція 11. Аналіз даних в R. Фактори, списки, фрейми та дії над ними.*	*Історія розвитку мови R. Можливості мови R. Об'єкти, пакети, функції. Вектори, матриці та операції над ними в R. Фактори, списки, фрейми та дії над ними.*
16	*Лекція 12. Експорт, імпорт та оброблення даних в R.*	*Експорт та імпорт даних в R. Використання R для аналізу часових рядів. Оброблення даних в R.*
17	*Лекція 13. Основні інструменти аналізу та візуалізації даних в R.*	*Функція plot () і її параметри. Управління загальними параметрами - аргументами графічних функцій. Типи графіків в R.*
18	*Комп'ютерний практикум 5. Аналіз та візуалізація даних в R.*	*Завдання: ознайомитись з можливостями мови програмування R для аналізу та візуалізації даних, використати бібліотеку R dplyr для очищення та трансформації даних та бібліотеку ggplot2 для візуалізації даних.*
19	*Лекція 14. Архітектурні моделі Big Data. Технології віртуалізації. Гіпервізори. Контейнерна технологія виконання програмного коду на сервері. SaaS, PaaS і IaaS.*	*Архітектурні моделі інженерії Big Data. Технології віртуалізації. Шари абстракції. Гіпервізори. Контейнерна технологія виконання програмного коду на сервері. Інжиніринг даних.*
20	*Лекція 15. Технології Hadoop Big Data. Розподілена обробка MapReduce. HDFS.*	*Масштабованість за допомогою великих даних. Зберігання та оброблення даних в розподілених файлових системах. Розподілені бази даних. Розподілена файлова система Hadoop (HDFS).*
21	*Лекція 16. Розподілена потокова платформа Kafka. Переваги Cassandra.*	*Проблема прийому даних. Розподілена потокова платформа Kafka. Переваги Cassandra.*
22	*Лекція 17. Платформа Apache Spark. Lambda та Kappa архітектури оброблення великих даних.*	*Проблема обчислювальної функції. Технологія Spark. Порівняння Spark та MapReduce. Spark і sparklyr для роботи з великими даними в R. Lambda - архітектура. Переваги і недоліки Lambda - архітектури. Kappa - архітектура. Переваги і недоліки Kappa-архітектури.*
23	*Комп'ютерний практикум 6. Розподілені обчислення даних з використанням Spark-кластера та мови R.*	*Завдання: встановити Spark на локальній машині, виконати розподілені обчислення для набору даних з використанням Spark-кластера та мови R.*
24	*Лекція 18. Підсумкове заняття. Залік.*	*Повторення вивченого матеріалу. Залікова контрольна робота.*

6. Самостійна робота студента

Дисципліна «Технології оброблення великих даних» ґрунтується на самостійній підготовці до аудиторних занять на теоретичні та практичні теми.

№ з/п	Назва теми, що виноситься на самостійне опрацювання	Кількість годин	Література
1	Підготовка до лекції 1	2	1, стор. 8-27.
2	Підготовка до комп'ютерного практикуму 1	2	2, стор. 5-16.
3	Підготовка до лекції 2	2	1, стор. 28-40.
4	Підготовка до комп'ютерного практикуму 2	2	2, стор. 17-23.
5	Підготовка до лекції 3	2	1, стор. 41-46.
6	Підготовка до лекції 4	2	1, стор. 47-52.
7	Підготовка до лекції 5	2	1, стор. 52-64.
8	Підготовка до комп'ютерного практикуму 3	2	2, стор. 24-31.
9	Підготовка до лекції 6	2	1, стор. 65-71.
10	Підготовка до лекції 7	2	1, стор. 72-82.
11	Підготовка до комп'ютерного практикуму 4	2	2, стор. 32-38.
12	Підготовка до лекції 8	2	1, стор. 83-92.
13	Підготовка до лекції 9	2	1, стор. 93-100.
14	Підготовка до лекції 10	2	1, стор. 101-118.
15	Підготовка до лекції 11	2	1, стор. 119-142.
16	Підготовка до лекції 12	2	1, стор. 142-157.
17	Підготовка до лекції 13	2	1, стор. 158-168.
18	Підготовка до комп'ютерного практикуму 5	2	2, стор. 33-69.
19	Підготовка до лекції 14	2	1, стор. 169-182.
20	Підготовка до лекції 15	2	1, стор. 183-189.
21	Підготовка до лекції 16	2	1, стор. 190-201.
22	Підготовка до лекції 17	2	1, стор. 201-226.
23	Підготовка до комп'ютерного практикуму 6	2	2, стор. 70-85.
24	Підготовка до модульної контрольної роботи	20	1, стор. 5-226.

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

- *Відвідування лекційних занять є обов'язковим.*
- *Відвідування занять комп'ютерного практикуму може бути епізодичним та за потреби захисту робіт комп'ютерного практикуму.*
- *Правила поведінки на заняттях: активність, повага до присутніх, відключення телефонів.*
- *Дотримання політики академічної доброчесності.*

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

Протягом семестру студенти виконують 6 комп'ютерних практикумів. Максимальна кількість балів за кожний комп'ютерний практикум: 10 балів.

Бали нараховуються за:
- якість виконання комп'ютерного практикуму: 0-4 бали;
- відповідь під час захисту комп'ютерного практикуму: 0-4 бали;
- своєчасне представлення роботи до захисту: 0-2 бали.

Критерії оцінювання якості виконання:
3-4 бали – робота виконана якісно, в повному обсязі;
1-2 бали – робота виконана якісно, в повному обсязі, але має недоліки;
0 балів – робота виконана не в повному обсязі, або містить суттєві помилки.

Критерії оцінювання відповіді:
3-4 бали – відповідь повна, добре аргументована;
1-2 бал – у відповіді є суттєві помилки;
0 балів – немає відповіді або відповідь невірна.

Критерії оцінювання своєчасності представлення роботи до захисту:
1-2 бали – робота представлена до захисту не пізніше вказаного терміну;
0 балів – робота представлена до захисту пізніше вказаного терміну.

Максимальна кількість балів за виконання та захист комп'ютерних практикумів:
10 балів × 6 комп. практ. = 60 балів.

*Завдання на **модульну контрольну роботу** складається з 8 питань – 5 теоретичних та 3 практичних. Відповідь на кожне теоретичне/практичне запитання оцінюється 5 балами.*

Критерії оцінювання кожного теоретичного/практичного запитання модульної контрольної роботи:
5 балів – відповідь вірна, повна, добре аргументована;
3-4 балів – відповідь вірна, але неповна або погано аргументована;
2 бали – у відповіді є незначні помилки;
1 бал – у відповіді є суттєві помилки;
0 балів – немає відповіді або відповідь невірна.

Максимальна кількість балів за модульну контрольну роботу:
5 балів × 5 теоретичні запитання + 5 балів × 3 практичні запитання = 40 балів.

Рейтингова шкала з дисципліни дорівнює:
$R = R_C$ = 60 балів + 40 балів = 100 балів.
За описом: $R = R_{комп.практ} + R_{МКР.}$ = 60+40 балів = 100 балів

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

На першій атестації (8-й тиждень) студент отримує «зараховано», якщо його поточний рейтинг не менше 50 % від максимальної кількості балів, яку може отримати студент до першої атестації (20 балів).

На другій атестації (14-й тиждень) студент отримує «зараховано», якщо його поточний рейтинг не менше 50 % від максимальної кількості балів, яку може отримати студент до другої атестації (30 балів).

Семестровий контроль: ***залік****.*

Умови допуску до семестрового контролю:

При семестровому рейтингу (r_C) не менше 60 % (60 балів) та зарахуванні усіх робіт комп'ютерного практикуму.

Необхідною умовою допуску до заліку є виконання і захист комп'ютерного практикуму.

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

1. Додаткова інформація з дисципліни (освітнього компонента)

Сертифікати проходження онлайн курсів «Programming Essentials in Python» та «IoT Fundamentals: Big Data & Analytics» дозволяють зарахувати студенту модульну контрольну роботу згідно отриманої загальної оцінки в Мережевій академії (Cisco Networking Academy), написання статей або участь у конференціях/ проєктах за відповідною тематикою також оцінюється як додаткові 5 балів.

Складено к.т.н., доц. Олещенко Л.М.

Ухвалено кафедрою ПЗКС (протокол №8 від 25.01.23)

Погоджено Методичною комісією факультету прикладної математики (протокол № 6 від 27.01.2023)