



ОБРОБЛЕННЯ НАДВЕЛИКИХ МАСИВІВ ДАНИХ

Робоча програма навчальної дисципліни (Силабус)

Реквізити навчальної дисципліни

| | |
|---|---|
| Рівень вищої освіти | <i>другий (магістерський)</i> |
| Галузь знань | <i>Інформаційні технології</i> |
| Спеціальність | <i>F2 Інженерія програмного забезпечення</i> |
| Освітня програма | <i>Інженерія програмного забезпечення інформаційних систем</i> |
| Статус дисципліни | <i>Нормативна</i> |
| Форма навчання | <i>очна(денна)</i> |
| Рік підготовки, семестр | <i>1-й курс, осінній семестр</i> |
| Обсяг дисципліни | <i>5 кредитів, 150 годин (32 години – Лекції, 28 годин - Лабораторні роботи, 90 годин – СРС)</i> |
| Семестровий контроль/ контрольні заходи | <i>Екзамен/тестування, МКР, захист лабораторних робіт</i> |
| Розклад занять | <i>Згідно розкладу на осінній семестр поточного навчального року https://schedule.kpi.ua/</i> |
| Мова викладання | <i>Українська</i> |
| Інформація про керівника курсу / викладачів | Лектор: доцент кафедри ІП, к.т.н., Олійник Ю.О., yurii.oliinyk-fiot@iit.kpi.ua Лабораторні роботи: доцент кафедри ІП, к.т.н., Олійник Ю.О yurii.oliinyk-fiot@iit.kpi.ua асистент кафедри ІП, доктор філософії Зарічковий О.А. alexkiras1998@gmail.com |
| Розміщення курсу | Дистанційний курс на платформі https://classroom.google.com/ Код доступу до курсу: nebvprgw |

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

Мета вивчення дисципліни – набуття ключових фахових компетентностей, теоретичних знань і практичних навичок аналізу, аргументування, прийняття рішень при розв'язанні задач та практичних проблем оброблення та аналізу надвеликих масивів даних.

Предметом вивчення дисципліни є технології, моделі, архітектура розподілених сховищ даних; процеси обробки та аналізу надвеликих масивів даних.

Завдання вивчення дисципліни:

- оволодіння основними поняттями обробки та аналізу надвеликих масивів даних;
- ознайомлення з новітніми підходами створення розподілених сховищ даних;
- набуття практичних навичок обробки та аналізу надвеликих масивів даних для вирішення задач машинного навчання та підтримки прийняття рішень.

Навчальна дисципліна покликана допомогти студенту отримати:

- вивчення сучасних концепцій та підходів до оброблення надвеликих масивів даних та створення сховищ даних;
- уміння вільно орієнтуватися на сучасному світі розподілених сховищ даних; проектувати та створювати розподілені сховища даних, застосовувати сучасні методи та технології обробки та аналізу надвеликих масивів даних.
- здатність використовувати можливості сучасних засобів та технологій обробки потоків даних.
- здатність до використання методів машинного навчання при розробці програмного забезпечення інформаційних систем.

КОМПЕТЕНТНОСТІ

Спеціальні (фахові, предметні) компетентності

- ФК13 – Здатність до аналізу, проектування та розробки нових та використання існуючих систем зберігання та обробки надвеликих масивів даних.
- ФК14 – Здатність до використання методів машинного навчання при розробці програмного забезпечення інформаційних систем.

ПРОГРАМНІ РЕЗУЛЬТАТИ НАВЧАННЯ

- ПРН20 – Знання методів машинного навчання.
- ПРН23 – Розробляти, реалізувати та застосовувати різні методи інтелектуального аналізу даних до Big Data, формулювати алгоритми обробки в парадигмі Map Reduce, обирати відповідну технологію зберігання і оброблення надвеликих даних, використовувати сучасні високонавантажені системи зберігання та оброблення великих даних.

2. Пререквізити та постреквізити дисципліни

Для успішного засвоєння дисципліни студент повинен володіти базовими знаннями з:

- Баз даних.
- Аналізу даних.
- Проектування інформаційних систем.
- Теорії алгоритмів.

Знання, отримані студентами при вивченні дисципліни, можуть використовуватись у таких дисциплінах:

- Науково-дослідна практика.
- Виконання магістерської дисертації.

3. Зміст навчальної дисципліни

| |
|--|
| Тема 1. Обробка даних та побудова розподілених сховищ в Hadoop та Hive |
| Тема 2. Архітектура Apache Spark |
| Тема 3. Графові алгоритми в Apache Spark |
| Тема 4. Інтелектуальний аналіз даних в Apache Spark |
| Тема 5. Обробка потоків даних |
| Тема 6. Обробка текстових даних |

4. Навчальні матеріали та ресурси

Основна література

1. Дистанційний курс на платформі Google Classroom.
2. 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.. (доступ за посиланням <https://ela.kpi.ua/handle/123456789/42206>)
3. 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. (доступ за посиланням <https://ela.kpi.ua/handle/123456789/46129>)
4. Graph algorithms: practical examples in Apache Spark and Neo4j [Електронний ресурс] / М. Needham, А. Е. Hodler. – O'Reilly Media, 2019. – Назва з екрана (доступ за посиланням https://www.academia.edu/41944402/Mark_Needham_and_Graph_Algorithms_Practical_Examples_in_Apache_Spark_and_Neo4j)

Додаткова література

1. Олійник, Ю. О. Оброблення надвеликих масивів даних [Електронний ресурс] : рек. до виконання курсової роботи : для здобувачів ступеня магістра за освітньою програмою «Інженерія програмного забезпечення інформаційних систем» спеціальності «Інженерія програмного забезпечення» / КПІ ім. Ігоря Сікорського ; уклад.: Ю. О. Олійник, Д. А. Гобов. – Електронні текстові дані (1 файл: 450,34 Кбайт). – Київ : КПІ ім.

- Пояснювально-ілюстративного методу Послідовна та логічно ув'язана подача матеріалу надає уявлення та знання у його логічної цілісності.
- Метод проблемного викладу надає уяву та методи отримання нових знань та фактів з використанням вже відомих фактів та тверджень.
- Інтерактивний метод під час лекційних занять використовується для встановлення діалогу з аудиторією.

Лабораторні роботи проходять з використанням наступних методів:

- 1) репродуктивного методу, завдяки якому студенти закріплюють вивчений теоретичний матеріал та навчаються використовувати його в конкретних задачах
 - 2) проблемного методу, при застосуванні якого студенти залучаються до обговорення та вирішення задач, пов'язаних з новітніми інформаційними технологіями аналітичної обробки інформації
- Самостійна робота з можливістю особистих консультацій з викладачем.

Тематичний план проведення лекційних занять

Таблиця 1. План проведення лекційних занять

| | |
|--|---|
| Лекція 1. Задачі оброблення надвеликих масивів даних | 2 |
| Лекція 2. Архітектура та принципи обробки даних в Apache Hadoop. Концепція MapReduce. | 2 |
| Лекція 3. Архітектура та принципи обробки даних в Apache Hive | 2 |
| Лекція 4. Архітектура Apache Spark | 2 |
| Лекція 5. Принципи обробки даних в Apache Spark | 2 |
| Лекція 6. Обробка графів в Apache Spark. Алгоритми пошуку шляхів | 2 |
| Лекція 7. Обробка графів в Apache Spark. Алгоритми зв'язності | 2 |
| Лекція 8. Класифікація даних в Apache Spark | 2 |
| Лекція 9. Кластеризація даних в Apache Spark | 2 |
| Лекції 10. Створення рекомендаційних систем на основі Apache Spark | 2 |
| Лекція 11. Обробка потоків даних | 2 |
| Лекції 12. Обробка текстової інформації | 2 |
| Лекції 13. Великі мовні моделі | 2 |
| Лекція 14-15. Кращі практики створення рішень з обробки надвеликих масивів даних | 4 |
| Проведення МКР | 2 |

Тематичний план проведення лабораторних робіт

Таблиця 2. План проведення лабораторних робіт

| № | Лабораторна робота | Тема | Години |
|---|--------------------|--|--------|
| 1 | ЛР 1 | Розподілена обробка даних в Apache Hadoop та Apache Hive: <ul style="list-style-type: none"> – створення обчислювального кластеру Apache Hadoop; – завантаження даних в Apache Hadoop; | 4 |

| | | |
|-------------------|---|-----------|
| | <ul style="list-style-type: none"> – завантаження даних в Apache Hive; – виконання запитів в Apache Hive; – створення партицій та бакетів в Apache Hive як виконання стратегії обробки надвеликих масивів даних; – підрахунок кількості слів в в Apache Hadoop. | |
| ЛР 2 | <p>Використання графових алгоритмів в Apache Spark:</p> <ul style="list-style-type: none"> – завантаження та підготовка даних; – створення графу GraphFrame; – пошук маршрутів між заданими країнами за допомогою Motifs або BFS; – робота з алгоритмами PageRank, Connected Components, Label Propagation і т.д. | 4 |
| ЛР 3 | <p>Класифікація та кластеризація даних в Apache Spark:</p> <ul style="list-style-type: none"> – завантаження та підготовка даних для створення та використання моделей та методів машинного навчання для інформаційних систем; – робота з методами класифікації; – робота з методами кластеризації. | 6 |
| ЛР 4 | <p>Створення рекомендацій на основі Apache Spark:</p> <ul style="list-style-type: none"> – завантаження та підготовка даних для створення моделі машинного навчання; – робота з моделлю ALS; – комбінування методів кластеризації та ALS. | 4 |
| ЛР 5 | <p>Обробка потоків даних в Apache Spark:</p> <ul style="list-style-type: none"> – завантаження та підготовка даних; – створення вхідного потоку даних; – підготовка моделі машинного навчання; – робота з запитом до вхідного потоку даних; – створення вихідного потоку даних; – робота з запитом до вихідного потоку даних. | 4 |
| ЛР 6 | <p>Обробка текстової інформації:</p> <ul style="list-style-type: none"> – завантаження та підготовка даних; – створення моделі BOW та підрахунок метрик TF-IDF; – робота з векторним представленням тексту моделі Bert; – підрахунок подібності тексту на основі різних моделей; – робота з LLM моделями. | 6 |
| В підсумку | | 28 |

6. Самостійна робота студента

Тематика самостійної роботи студентів наведено у таблиці 3.

Таблиця 3. Самостійна робота студентів

| | |
|--|-----------|
| Тема 1. Характеристики технологій обробки надвеликих масивів даних | 2 |
| Тема 2. Виконання операцій join в концепції MapReduce | 2 |
| Тема 3. Типи моделей сховищ даних | 2 |
| Тема 4. ETL процеси в технологіях обробки надвеликих масивів даних | 6 |
| Тема 5. Бібліотеки ML, MLib в Apache Spark | 2 |
| Тема 6. Графові алгоритми в Neo4j | 4 |
| Тема 7. Методи визначення аномалій для надвеликих масивів даних | 2 |
| Тема 8. Фреймворк TEZ | 2 |
| Підготовка до лекцій | 8 |
| Виконання лабораторних робіт | 28 |
| Підготовка до модульної контрольної роботи | 2 |
| Підготовка до екзаменаційної роботи по всьому матеріалу модуля. | 30 |
| ВСЬОГО | 90 |

Політика та контроль

7. Політика навчальної дисципліни (освітнього компонента)

Форми організації освітнього процесу, види навчальних занять і оцінювання результатів навчання регламентуються Положенням про організацію освітнього процесу в Національному технічному університеті України «Київському політехнічному інституті імені Ігоря Сікорського».

Політика виставлення оцінок: кожна оцінка виставляється відповідно до розроблених викладачем та заздалегідь оголошених студентам критеріїв, а також мотивується в індивідуальному порядку на вимогу студента; у випадку не виконання студентом усіх передбачених навчальним планом видів занять до екзамену він не допускається.

Політика щодо виконання лабораторних робіт:

- у випадку виявлення факту академічної недоброчесності робота не зараховується;
- для виконання та захисту лабораторних робіт встановлюються дедлайни. У разі пропуску дедлайну із поважної причини — можливе перенесення за попереднім погодженням, без поважної причини – оцінка знижується.
- кожна лабораторна робота має 3 критерії оцінювання: виконання, захист та підготовку звіту;
- лабораторна робота вважається зданою у випадку її захисту студентом на занятті;
- за одне заняття (здачу) студентом може бути здано не більше 2 (двох) лабораторних робіт;

Політика щодо перескладань:

- якщо студент не проходив або не з'явився на МКР (без поважної причини), його результат оцінюється у 0 балів; переписування модульних контрольних робіт не передбачено;
- перескладання лабораторних робіт можливо лише у визначені терміни, не пізніше ніж за тиждень до екзаменаційної сесії.

- в окремих випадках, при наявності об'єктивних причин (хвороба, наявність «повітряної тривоги», перебування студента на прифронтових територіях, академічна мобільність, тощо), що унеможливають можливість виконання студентом контрольних заходів згідно оприлюдненого графіка, терміни здачі контрольних заходів можуть бути змінені за умови попереднього узгодження з викладачем.

Визнання результатів навчання, набутих у неформальній/інформальній освіті.

Порядок визнання таких результатів регламентується Положенням про визнання результатів навчання, набутих у неформальній / інформальній освіті (<https://osvita.kpi.ua/index.php/node/179>). Можуть бути зараховані окремі змістовні модулі або теми дисципліни. В такому разі здобувач звільняється від виконання відповідних завдань, отримуючи за них максимальний бал відповідно до рейтингової системи оцінювання.

Політика академічної поведінки та доброчесності: конфліктні ситуації мають відкрито обговорюватись в академічних групах з викладачем, необхідно бути взаємно толерантним, поважати думку іншого. Плагіат та інші форми нечесної роботи неприпустимі. Всі індивідуальні завдання студент має виконати самостійно із використанням рекомендованої літератури й отриманих знань та навичок. Цитування в письмових роботах допускається тільки із відповідним посиланням на авторський текст. Недопустимі підказки і списування у ході захисту практикумів, на контрольних роботах, на іспиті.

Дотримання академічної доброчесності. Обов'язковою умовою виконання завдань з освітньої компоненти є дотримання політики та принципів академічної доброчесності (<https://kpi.ua/academic-integrity>), які, у тому числі, викладено у Кодексі честі Національного технічного університету України «Київський політехнічний інститут імені Ігоря Сікорського» (<https://kpi.ua/code>), Положенні про систему запобігання академічному плагіату в КПІ ім. Ігоря Сікорського (<https://osvita.kpi.ua/node/47>). У разі виявлення дублювання робіт, плагіату роботи здобувачі отримують нульовий рейтинг.

Політика використання штучного інтелекту. Використання штучного інтелекту (далі, ШІ) регламентується «Політикою використання штучного інтелекту для академічної діяльності в КПІ ім. Ігоря Сікорського» (<https://osvita.kpi.ua/node/1225>). Усі навчальні завдання з дисципліни мають бути результатом власної оригінальної роботи здобувача. За порушення принципів академічної доброчесності, зокрема плагіат лабораторних робіт, студент втрачає всі бали за дану лабораторну роботу.

Норми академічної етики: дисциплінованість; дотримання субординації; чесність; відповідальність; робота в аудиторії з відключеними мобільними телефонами. Повага один до одного дає можливість ефективніше досягати поставлених командних результатів. При виконанні практикумів студент може користуватися ноутбуками. Проте під час лекційних занять та обговорення завдань практикумів не слід використовувати ноутбуки, смартфони, планшети чи комп'ютери. Це відволікає викладача і студентів групи та перешкоджає навчальному процесу. Якщо ви використовуєте свій ноутбук чи телефон для аудіо - чи відеозапису, необхідно заздалегідь отримати дозвіл викладача.

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

Рейтинг студента з дисципліни складається з балів, що він отримує за:

1. Виконання та захист 6 лабораторних робіт;
2. Виконання модульної контрольної роботи (МКР);
3. Заохочувальні бали.
4. Екзаменаційна робота.

Лабораторна робота (ЛР):

Лабораторні роботи здається очно/в дистанційному форматі та оцінюються згідно таблиці 4:

Таблиця 3. Критерії оцінювання лабораторних робіт

| № л.р. | Виконання | Підготовка протоколу до л.р. | Захист | Максимальна сума балів по відповідним критеріям |
|--------|---|---------------------------------------|---|---|
| 1,6 | робота виконана без зауважень – 7 балів | Оформлений належним чином – 1 бал | робота захищена без зауважень – 2 бали | 10 |
| | достатньо повне виконання роботи з деякими похибками – 5 балів | | робота захищена, однак при захисті є зауваження – 1,5 бали | |
| | неповністю виконана роботи – 3 бали | Оформлений неналежним чином – 0 балів | робота захищена, однак на частину питань відсутні відповіді або надані часткові відповіді – 1 бал | |
| | при виконанні роботи є суттєві зауваження – 0 балів | | є суттєві зауваження при захисті роботи – 0 балів | |
| 2,4,5 | робота виконана без зауважень – 4 бали | Оформлений належним чином – 1 бал | робота захищена без зауважень – 2 бали | 7 |
| | достатньо повне виконання роботи з деякими похибками – 3 бали | | робота захищена, однак при захисті є зауваження – 1,5 бали | |
| | неповністю виконана роботи – 2 бали | Оформлений неналежним чином – 0 балів | робота захищена, однак на частину питань відсутні відповіді або надані часткові відповіді – 1 бал | |
| | при виконанні роботи є суттєві зауваження – 0 балів | | є суттєві зауваження при захисті роботи – 0 балів | |
| 3 | робота виконана без зауважень – 6 балів | Оформлений належним чином – 1 бал | робота захищена без зауважень – 2 бали | 9 |
| | достатньо повне виконання роботи з деякими похибками – 4,5 бали | | робота захищена, однак при захисті є зауваження – 1,5 бали | |
| | неповністю виконана робота – 3 бали | Оформлений неналежним чином – 0 балів | робота захищена, однак на частину питань відсутні відповіді або надані часткові відповіді – 1 бал | |
| | при виконанні роботи є суттєві зауваження – 0 балів | | є суттєві зауваження при захисті роботи – 0 балів | |

Заохочувальні бали

– за виконання творчих робіт з кредитного модуля (наприклад, участь у факультетських та інститутських олімпіадах з навчальних дисциплін, участь у конкурсах робіт, підготовка оглядів наукових праць тощо); за активну роботу на лекції (питання, доповнення, зауваження за темою лекції, коли лектор пропонує студентам задати свої питання) 1-2 бали, але в сумі не більше 6;

– презентації по СРС – від 1 до 5 балів.

– додаткові факультативні лабораторні роботи/комп'ютерні практикуми – від 1 до 7 балів.

Сумарна кількість заохочувальних балів не більше 10.

Модульна контрольна робота

Модульна контрольна робота складається з практичного завдання.

Ваговий бал МКР – 10 балів. *Критерії оцінювання кожної частини МКР:*

– “відмінно”, повна відповідь (не менше 90% потрібної інформації) – 8–10 балів;

– “добре”, достатньо повна відповідь (не менше 75% потрібної інформації), або повна відповідь з незначними помилками – 6-7 балів;

– “задовільно”, неповна відповідь (не менше 60% потрібної інформації) та незначні помилки – 4-5 балів;

– “незадовільно”, незадовільна відповідь (взагалі неправильна відповідь) – 0-3 бали.

Міжсесійна атестація

На першій атестації (8-й тиждень) студент отримує «зараховано», якщо він має два зданих практикуми/лабораторних роботи.

На другій атестації (14-й тиждень) студент отримує «зараховано», якщо він має 4 здані практикуми/лабораторних роботи та написати МКР.

РОЗПОДІЛ БАЛІВ, ЯКІ ОТРИМУЮТЬ СТУДЕНТИ З ДИСЦИПЛІНИ

| Види контролю | бали |
|--|------|
| ЛР1 «Розподілена обробка даних в Apache Hadoop та Apache Hive» | 10 |
| ЛР2 «Використання графових алгоритмів в Apache Spark» | 7 |
| ЛР3 «Класифікація та кластеризація даних в Apache Spark» | 9 |
| ЛР4 «Створення рекомендацій на основі Apache Spark» | 7 |
| ЛР5 «Обробка потоків даних в Apache Spark» | 7 |
| ЛР6 «Обробка текстової інформації» | 10 |
| МКР | 10 |

$$S=10+7+9+7+7+10+10+P=60, \text{ де } S - \text{ стартовий рейтинг,}$$

що складається з балів за лабораторні роботи студента, ДКР та суми заохочувальних балів (P)

Семестровий контроль: *Екзамен*

Сумарний бал за виконання лабораторних робіт складає 50 балів. Критерії оцінювання лабораторних робіт включають якість її виконання, захисту та відповідь на запитання.

Екзамен оцінюється 40 балами, МКР оцінюється 10 балами. Умови проведення МКР та екзамену визначаються додатково і оголошуються на консультації. МКР та екзамен не переносяться та повторно не проводяться.

Умовою допуску до екзамену є набір не менше ніж 50% від максимально можливої кількості балів за усі заходи поточного контролю впродовж семестру - тобто 30 балів (із 60) по всім видам контролю та захист не менше 4 лабораторних робіт.

Студенти, які наприкінці семестру мають стартовий рейтинг $30 \leq S < 60$ виконують екзаменаційну роботу. При цьому рейтингова оцінка з кредитного модуля складається з балів стартового рейтингу та балів за екзаменаційну роботу. Підсумкова оцінка формується за результатами оцінювання знань та навичок студента в семестрі та на екзамені за формулою: $R=S+E$, де S – стартовий рейтинг, E – бал за екзамен.

Оцінювання екзаменаційної відповіді відбувається обчисленням суми балів за правильні відповіді на питання під час тесту. Правильна відповідь на одне питання дозволяє отримати 1 бал. Тест складається з 40 питань.

Студенти, які наприкінці семестру мають стартовий рейтинг $S < 30$ до іспиту не допускаються і повинні виконувати додаткову роботу для підвищення свого рейтингу.

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 |

9. Додаткова інформація з дисципліни (освітнього компонента)

Перелік питань, які виносяться на семестровий контроль розміщений на дистанційні платформі Classroom компанії Google.

Робочу програму навчальної дисципліни (силабус):

Складено доцент кафедри ІПІ, Олійник Ю.О., ас. кафедри ІПІ Зарічковий О.А.

Ухвалено кафедрою ІПІ (протокол №2/1 від 10.10.2025)

Погоджено Методичною комісією факультету (протокол №3 від 17.10.2025)