І СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

1. Загальна інформація про навчальну дисципліну

Повна назва навчальної дисципліни	Методи аналізу великих даних
Повна офіційна назва закладу вищої освіти	Сумський державний університет
Повна назва структурного підрозділу	Факультет електроніки та інформаційних технологій. Кафедра прикладної математики та моделювання складних систем
Розробник(и)	Марченко Анна Вікторівна
Рівень вищої освіти	другий рівень вищої освіти, НРК – 7 рівень, QF-LLL – 7 рівень, FQ-EHEA – другий цикл
Семестр вивчення навчальної дисципліни	8 тижнів протягом 3-го семестру
Обсяг навчальної дисципліни	5 кредитів ЄКТС, 150 годин, з яких 48 години становить контактна робота з викладачем (24 годин лекцій, 24 години практичних робіт), 102 години становить самостійна робота
Мова викладання	українська

2. Місце навчальної дисципліни в освітній програмі

Статус дисципліни	Обов`язкова навчальна дисципліна для освітньої програми 'Прикладна математика'
Передумови для вивчення дисципліни	Системи баз даних, Алгоритми машинного навчання
Додаткові умови	Додаткові умови відсутні
Обмеження	Обмеження відсутні

3. Мета навчальної дисципліни

Мета дисципліни – набуття знань та практичних навичок використання методів видобування даних, які застосовуються для формування паралельних обчислень, хешування, потокової обробки даних тощо

4. Зміст навчальної дисципліни

Тема 1 Вступ до великих даних *Огляд курсу. Поняття обробки, здобуття даних. Хеш-функція. Індекси Знайомство з тех-нологіями Hadoop та Spark*
Тема 2 Map Reduce *Розподілені файлові системи. Map Reduce (Згортка-Відображення). Алгоритми, по-будовані на MapReduce. Модель комунікативної вартості. Вступ до PySpark. Основи машинного обчислення засобами PySpark*
Тема 3 Пошук схожих об’єктів *Додатки пошуку найближчого сусіда. Розбиття документів на шингли. Сигнатури множин із збереженням схожості. Хешування документів. Метрики. Застосування хешування з урахуванням близькості*
Тема 4 Аналіз потоків даних *Потокова модель даних. Вибірка даних з потоку. Фільтрація потоків. Підрахунок різних елементів у потоці. Оцінювання моментів*

Тема 5 Аналіз посилань *PageRank. Ефективне обчислення PageRank. Тематичний PageRank. Посилальний спам. Хаби та авторитетні сторінки*
Тема 6 Реклама в Інтернеті *Онлайнові алгоритми. Задача про паросумісність. Задача про ключові слова. Реалізація алгоритму Adwords*
Тема 7 Рекомендаційні системи *Модель рекомендаційної системи. Рекомендація на основі фільтрації вмісту. Колаборативна фільтрація. Зниження розмірності*
Тема 8 Аналіз графів соціальних мереж *Вступ до графів. Соціальні мережі як графи. Кластеризація графа соціальної мережі. Пряме знаходження спільнот. Знаходження спільнот, що перетинаються. Околиці в графах*
Тема 9 Зниження розмірності *Власні значення та власні вектори. Метод головних компонент. CUR-декомпозиція. Правильний вибір рядків та стовпців*
Тема 10 Машинне навчання на великих даних *Модель машинного навчання. Перцептрони. Метод опорних векторів. Навчання за найближчими сусідами. Порівняння методів навчання*

5. Очікувані результати навчання навчальної дисципліни

Після успішного вивчення навчальної дисципліни здобувач вищої освіти зможе:

РН1	Знати сучасні методи обробки великих даних
РН2	Визначати алгоритми обробки даних для вирішення професійних задач аналізу великих даних
РН3	Застосовувати спеціалізовані алгоритми обробки потокових даних для аналізу швидкозмінних даних, технологій пошукових систем, методи кластеризації великогабаритних даних
РН4	Поєднувати існуючі алгоритми обробки великих даних для вирішення комплексних задач аналізу складних даних

6. Роль навчальної дисципліни у досягненні програмних результатів

Програмні результати навчання, досягнення яких забезпечує навчальна дисципліна.

Для спеціальності 113 Прикладна математика:

ПР2	Уміти формалізувати задачі певної предметної галузі, формулювати їх математичну постановку та обирати раціональний метод вирішення; розв’язувати отримані задачі аналітичними та чисельними методами, оцінювати точність та достовірність отриманих результатів.
ПР17	Володіти математичними методами обробки великих наборів даних. Вміти обирати до застосовування оптимальні методи для конкретної задачі побудови моделі поведінки складної системи за існуючим набором даних та будувати графові ймовірнісні моделі для розв’язання технічних задач
ПР19	Уміти оцінити на адекватність результат обробки великих масивів даних

7. Види навчальних занять та навчальної діяльності

7.1 Види навчальних занять

Тема 1. Вступ до великих даних

Лк1 "Вступ до великих даних" (денна) *Поняття обробки, здобуття даних. Хеш-функція. Індекси Знайомство з технологіями Hadoop та Spark*
Пр1 "Вступ до мови Scala" (денна) *Вступ до мови Scala та інтерфейс командного рядка*
Тема 2. Map Reduce
Лк2 "Алгоритми Map Reduce" (денна) *Алгоритми, побудовані на MapReduce. Модель комунікативної вартості*
Лк3 "Map Reduce та новий програмний стек" (денна) *Розподілені файлові системи. Map Reduce (Згортка-Відображення)*
Пр2 "Spark - швидкий обчислювальний кластер" (денна) *Вступ до PySpark. Основи машинного обчислення засобами PySpark*
Тема 3. Пошук схожих об’єктів
Лк4 "Пошук схожих об’єктів. Частина 1" (денна) *Додатки пошуку найближчого сусіда. Розбиття документів на шингли. Сигнатури множин із збереженням схожості. Хешування документів. Метрики*
Лк5 "Пошук схожих об’єктів. Частина 2" (денна) *Застосування хешування з урахуванням близькості. Корпоративна культура та організація бізнес-процесів*
Пр3 "Spark – надбудова SQL" (денна) *Виконання запитів на створення об'єктів збереження даних та вибірки даних*
Тема 4. Аналіз потоків даних
Лк6 "Аналіз потоків даних" (денна) *Потокова модель даних. Вибірка даних з потоку. Фільтрація потоків. Підрахунок різних елементів у потоці. Оцінювання моментів*
Пр4 "Spark Streaming" *Обробка потоків даних засобами Spark*
Тема 5. Аналіз посилань
Лк7 "PageRank" (денна) *PageRank. Ефективне обчислення PageRank. Тематичний PageRank. Посилальний спам. Хаби та авторитетні сторінки*
Пр5 "Візуалізація великих даних засобами MicroStrategy. Огляд." (денна) *Вступ до бізнес-аналітики. Інформаційні панелі в режимі офлайн із розширеною візуалізацією*
Тема 6. Реклама в Інтернеті
Лк8 "Реклама в Інтернеті" (денна) *Онлайнові алгоритми. Задача про паросумісність. Задача про ключові слова. Реалізація алгоритму Adwords*
Пр6 "Візуалізація великих даних засобами MicroStrategy. Побудова простої інформаційної панелі." (денна) *Аналіз показників для побудови світового сценарію викидів вуглецю*
Тема 7. Рекомендаційні системи
Лк9 "Рекомендаційні системи" (денна) *Модель рекомендаційної системи. Рекомендація на основі фільтрації змісту. Колаборативна фільтрація. Зниження розмірності*
Пр7 "MLLib" (денна) *Машинне навчання з MLLib. Створення рекомендацій фільмів*
Пр8 "Візуалізація великих даних засобами MicroStrategy. Побудова комплексної інформаційної панелі" (денна) *Вступ до державних курсів дистанційної освіти*
Тема 8. Аналіз графів соціальних мереж

Лк10 "Аналіз графів соціальних мереж" (денна) *Вступ до графів. Соціальні мережі як графи. Кластеризація графа соціальної мережі. Пряме знаходження спільнот. Знаходження спільнот, що перетинаються. Околиці в графах*
Пр9 "Аналіз графів з GraphX" (денна) *Дослідження веб-графів і алгоритмів графів (PageRank) за допомогою GraphX*
Пр10 "Візуалізація великих даних засобами MicroStrategy. Побудова звітної інформаційної панелі" (денна) *Інтеграція додатку MicroStrategy з Salesforce.com для створення звітної інформаційної панелі*
Тема 9. Зниження розмірності
Лк11 "Зниження розмірності" (денна) *Власні значення та власні вектори. Метод головних компонент. CUR-декомпозиція. Правильний вибір рядків та стовпців*
Пр11 "Tachyon" (денна) *Розгортання надійної файлової системи в пам'яті через кластер*
Тема 10. Машинне навчання на великих даних
Лк12 "Машинне навчання на великих даних" (денна) *Модель машинного навчання. Перцептрони. Метод опорних векторів. Навчання за найближчими сусідами. Порівняння методів навчання*
Пр12 "BlinkDB" (денна) *Використання SQL зі статистичною вибіркою*

7.2 Види навчальної діяльності

НД1	Розв’язання базових практичних завдань за допомогою онлайн-технологій
НД2	Розв’язання розширених практичних завдань за допомогою онлайн-технологій
НД3	Виконання практичних завдань
НД4	Розв'язування вправ на основі теоретичного матеріалу
НД5	Перегляд відеолекцій МВОК
НД6	Підготовка карти пам’яті
НД7	Виконання індивідуальних розрахунково-аналітичних завдань

8. Методи викладання, навчання

Дисципліна передбачає навчання через:

МН1	Інтерактивні лекції
МН2	Практико-орієнтоване навчання
МН3	Проблемно-пошуковий метод

9. Методи та критерії оцінювання

9.1. Критерії оцінювання

Шкала оцінювання ECTS	Визначення	Чотирибальна національна шкала оцінювання	Рейтингова бальна шкала оцінювання

5 (відмінно)	Відмінне виконання лише з незначною кількістю помилок	A	90 ≤ RD ≤ 100
4 (добре)	Вище середнього рівня з кількома помилками	B	82 ≤ RD < 89
4 (добре)	Загалом правильна робота з певною кількістю помилок	C	74 ≤ RD < 81
3 (задовільно)	Непогано, але зі значною кількістю недоліків	D	64 ≤ RD < 73
3 (задовільно)	Виконання задовольняє мінімальні критерії	E	60 ≤ RD < 63
2 (незадовільно)	Можливе повторне складання	FX	35 ≤ RD < 59
2 (незадовільно)	Виконання задовольняє мінімальні критерії	F	0 ≤ RD < 34

9.2 Методи поточного формативного оцінювання

МФО1	Експрес-тестування
МФО2	Оцінювання практичних завдань
МФО3	Взаємооцінювання (peer assessment)
МФО4	Оцінювання вправ на основі теорії
МФО5	Оцінювання індивідуальних розрахунково-аналітичних завдань

9.3 Методи підсумкового сумативного оцінювання

МСО1	Звіт за результатами виконання базових практичних завдань на основі онлайн-технологій
МСО2	Звіт за результатами виконання розширених практичних завдань на основі онлайн-технологій
МСО3	Звіт за результатами виконання практичних завдань
МСО4	Виконання вправ на основі теоретичного матеріалу
МСО5	Виконання індивідуальних розрахунково-аналітичних завдань
МСО6	Розробка карт пам'яті
МСО7	Експрес-тестування
МСО8	Поточні контрольні роботи (проміжний модульний контроль)

Контрольні заходи:

3-й семестр		**100 балів**
МСО1. Звіт за результатами виконання базових практичних завдань на основі онлайн-технологій		**3**
	3x1	3
МСО2. Звіт за результатами виконання розширених практичних завдань на основі онлайн-технологій		**15**
	5x3	15
МСО3. Звіт за результатами виконання практичних завдань		**8**
	4x2	8
МСО4. Виконання вправ на основі теоретичного матеріалу		**8**
	4x2	8
МСО5. Виконання індивідуальних розрахунково-аналітичних завдань		**16**

		16
МСО6. Розробка карт пам'яті		**10**
	10x1	10
МСО7. Експрес-тестування		**20**
	10x2	20
МСО8. Поточні контрольні роботи (проміжний модульний контроль)		**20**
		20

Контрольні заходи в особливому випадку:

10. Ресурсне забезпечення навчальної дисципліни

10.1 Засоби навчання

ЗН1	Мультимедіа, проєкційна апаратура
ЗН2	Комп’ютери, комп’ютерні системи та мережи
ЗН3	Програмне забезпечення для підтримки дистанційного навчання

10.2 Інформаційне та навчально-методичне забезпечення

Основна література	
1	Юре Лесковец, Ананд Раджараман, Джеффри Д. Ульман Анализ боль-ших наборов данных. / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2016. – 498 с.: ил
Інформаційні ресурси в Інтернеті	
2	BigData Mini Course. – UC Berkeley
3	Online Course. Mining of Massive Datasets. Jure Leskovec, Anand Rajaraman, Jeff Ullman