СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

1. Загальна інформація про навчальну дисципліну

Повна назва навчальної дисципліни	Маніпулювання великими наборами даних засобами Scala
Повна офіційна назва закладу вищої освіти	Сумський державний університет
Повна назва структурного підрозділу	Факультет електроніки та інформаційних технологій. Кафедра інформаційних технологій
Розробник(и)	Неня Анна Вікторівна
Рівень вищої освіти	Другий рівень вищої освіти, НРК – 7 рівень, QF-LLL – 7 рівень, FQ-EHEA – другий цикл
Семестр вивчення навчальної дисципліни	16 тижнів протягом 3-го семестру
Обсяг навчальної дисципліни	Обсяг становить 5 кред. ЄКТС, 150 год. Для денної форми навчання 32 год. становить контактна робота з викладачем (16 год. лекцій, 16 год. практичних занять), 118 год. становить самостійна робота.
Мова викладання	Українська

2. Місце навчальної дисципліни в освітній програмі

Статус дисципліни	Вибіркова навчальна дисципліна для освітньо-наукової програми "Інформаційні технології проектування"
Передумови для вивчення дисципліни	Знання основ Python
Додаткові умови	Додаткові умови відсутні
Обмеження	Обмеження відсутні

3. Мета навчальної дисципліни

Отримати знання алгоритмів та навички маніпулювання великих наборів даних з використанням мови Scala

4. Зміст навчальної дисципліни

Тема 1 Вступ до мови Scala Огляд мови. Базовий синтаксис
Тема 2 Основи Scala Колекції та керуючі структури

Тема 3 Обробка розподілених даних засобами Spark Огляд розподілених наборів даних. Структура. Основні алгоритми обробки розподілених даних
Тема 4 Обробка структурованих наборів даних засобами Spark Огляд структурованих наборів даних. Алгоритми обробки структурованих наборів даних
Тема 5 Пошук схожих об’єктів Огляд алгоритмів аналізу великих наборів даних для пошуку схожих об'єктів. Розбиття документів на шингли. Сигнатури множин із збереженням схожості. Хешування документів. Метрики
Тема 6 Рекомендаційні системи Модель рекомендаційної системи. Рекомендація на основі фільтрації змісту. Колаборативна фільтрація. Зниження розмірності
Тема 7 Аналіз потоків даних Огляд алгоритмів аналізу великих наборів даних для обробки потокових даних. Потокова модель даних. Вибірка даних з потоку. Фільтрація потоків
Тема 8 Аналіз посилань веб-сторінок Огляд алгоритмів аналізу великих даних для обчислення рангу веб-сторінок. Ефективне обчислення PageRank. Тематичний PageRank.

5. Очікувані результати навчання навчальної дисципліни

Після успішного вивчення навчальної дисципліни здобувач вищої освіти зможе:

РН1	Знати синтаксис програмування мови SCALA
РН2	Розробляти програмні додатки мовою SCALA для обробки великих наборів даних
РН3	Визначати алгоритми обробки даних для вирішення професійних задач аналізу великих даних

7. Роль освітнього компонента у формуванні соціальних навичок

Загальні компетентності та соціальні навички, формування яких забезпечує навчальна дисципліна:

СН1	Здатність до пошуку, оброблення та аналізу інформації з різних джерел.
СН2	Здатність застосовувати знання у практичних ситуаціях.

8. Види навчальних занять

Тема 1. Вступ до мови Scala
Лк1 "Огляд мови Scala" (денна) Функціональність мови Scala. Маштабованість мови Scala. Scala і аналіз великих наборів даних. Spark для аналізу даних. Програмна модель Spark

Пр1 "Налаштування програмного середовища" (денна) Налаштування програмного середовища для обробки великих масивів даних з використанням IDE IntelliJ. Підключення наборів великих даних для їх подальшої обробки.
Тема 2. Основи Scala
Лк2 "Основи мови Scala" (денна) Основи мови Scala. Синтаксис. Елементи та конструкції.
Пр2 "Вступ до Scala. Основні елементи та структури мови." (денна) Знайомство з основними елементами Scala. Отримання навичок створення основних конструкцій та застосування базових операцій обробки їх значень. Отримання навичок створення структур даних та виконання основних операцій над ними.
Тема 3. Обробка розподілених даних засобами Spark
Лк3 "RDD API" (денна) Основні поняття RDD. Операції RDD. Трансформації і дії.
Пр3 "Обробка розподілених наборів даних" (денна) Знайомство із структурою розподілених наборів даних. Отримання навичок створення розподілених наборів даних, звернення до елементів розподілених наборів даних та виконання операції над ними.
Тема 4. Обробка структурованих наборів даних засобами Spark
Лк4 "Spark SQL. DataFrames/" (денна) Вступ до Spark SQL. Огляд стурктурованих наборів даних DataFrames Операції DataFrames.
Пр4 "Обробка структурованих наборів даних" (денна) Ознайомлення із структурою наборів даних DataFrame та DataSets. Отримання навичок створення структурованих наборів даних з різних джерел даних. Отримання навичок звернення до елементів структурованих наборів дани та виконання операції над ними.
Тема 5. Пошук схожих об’єктів
Лк5 "Пошук схожих об’єктів" (денна) Огляд алгоритмів пошуку найближчого сусіда. Розбиття документів на шингли. Сигнатури множин із збереженням схожості. Хешування документів. Метрики.
Пр5 "Підрахунок кількості вживаності кожного слова в текстовому файлі" (денна) Реалізація алгоритму пошуку схожих об'єктів для підрахунку кількості вживаності кожного слова в текстовому файлі
Тема 6. Рекомендаційні системи

Лк6 "Огляд рекомендаційниї систем" (денна) Модель рекомендаційної системи. Рекомендація на основі фільтрації змісту. Колаборативна фільтрація. Зниження розмірності
Пр6 "Використання методів колаборативної фільтрації засобами Spark." (денна) Ознайомлення із підходами використання методів колаборативної фільтрації для побудови рекомендаційної системи. Реалізація алгоритму роботи рекомендаційної системи на основі одного набору даних. Реалізація алгоритму роботи рекомендаційної системи на основі двох і більше наборів даних.
Тема 7. Аналіз потоків даних
Лк7 "Аналіз потоків даних" (денна) Потокова модель даних. Вибірка даних з потоку. Фільтрація потоків
Пр7 "Використання API Structured Streaming Spark для обробки потокових даних" (денна) Реалізація алгоритму аналізу потокових даних з використанням інструментів API Structured Streaming для обробки та аналізу журналу доступу до сайту
Тема 8. Аналіз посилань веб-сторінок
Лк8 "PageRank. Використання технології Page Rank." (денна) PageRank. Ефективне обчислення PageRank. Тематичний PageRank. Посилальний спам. Хаби та авторитетні сторінки
Пр8 "Аналіз графів з GraphX" (денна) Дослідження веб-графів і алгоритмів графів (PageRank) за допомогою GraphX

9. Стратегія викладання та навчання

9.1 Методи викладання та навчання

Дисципліна передбачає навчання через:

МН1	Інтерактивні лекції
МН2	Практико-орієнтоване навчання

Лекції надають матеріали щодо основних алгоритмів обробки великих наборів даних мовою Scala, зокрема (РН1). Леції доповнюються практичними роботами для опанування навичок обробки велики наборів даних основними констуркціями мови Scala (РН3, РН2)

Навички комунікації, здатність брати на себе відповідальність і працювати в критичнихумовах, вміння управляти своїм часом, розуміння важливості дедлайнів, здатність логічно і системно мислити

9.2 Види навчальної діяльності

НД1	Розв’язання завдань практикуму з обробки великих наборів даних
НД2	Підготовка карти пам’яті

НДЗ	Інтерактивні лекціЇ

10. Методи та критерії оцінювання

10.1. Критерії оцінювання

Визначення	Чотирибальна національна шкала оцінювання	Рейтингова бальна шкала оцінювання
Відмінне виконання лише з незначною кількістю помилок	5 (відмінно)	90 ≤ RD ≤ 100
Вище середнього рівня з кількома помилками	4 (добре)	82 ≤ RD < 89
Загалом правильна робота з певною кількістю помилок	4 (добре)	74 ≤ RD < 81
Непогано, але зі значною кількістю недоліків	3 (задовільно)	64 ≤ RD < 73
Виконання задовольняє мінімальним критеріям	3 (задовільно)	60 ≤ RD < 63
Можливе повторне складання	2 (незадовільно)	35 ≤ RD < 59
Необхідний повторний курс з навчальної дисципліни	2 (незадовільно)	0 ≤ RD < 34

10.2 Методи поточного формативного оцінювання

	Характеристика	**Дедлайн, тижні**	**Зворотний зв'язок**
МФО1 Експрес-тестування	Експрес-тестування призначене для проміжного оцінювання рівня засвоєння теоретичного матеріалу. Проводиться на початку наступної лекції з використанням хмарних інтерактивних технлогій. Результати тестування обговорюються протягом виконання завдання	Протягом лекційного заняття	google meet
МФО2 Виконання практичних завдань	Практичні роботи призначені для опанування практичних навичок налаштування програмного середовища та реалізацій базових алгоритмів обробки великих наборів даних. Для успішного зарахування завдання з практичної роботи неохідно виконати мініальний рівень складнощі завдання відповідно методичним вказівкам	Протягом поточного прктичного заняття	google meet

10.3 Методи підсумкового сумативного оцінювання

	Характеристика	**Дедлайн, тижні**	**Зворотний зв'язок**

МСО1 Звіт за результатами виконання завдань практикуму з обробки великих наборів даних	Звіт за результатами виконання практичних робіт повинен містити ілюстрацію основних результатів виконання завдань відповідно методичних вказівок. Для успішного зарахування необхідно виконати мінімальний рівень завдання та оформити звіт відповідно вимогам до оформлення звітів. Оцінка зі звіту може бути один раз підвищення за умови усунення зауважень до представлених результатів і надсилання звіту у вказані терміни	до початку наступного практичного заняття	google classroom, google meet
МСО2 Карти пам'яті	Карти пам'яті розроблюються за матеріалом лекційних занятть. Розроблення карт пам'яті передбачає отримання навичок аналітичної обробки інформації. Якість карт пам'яті оцінюється двоетапно: взаємооцінювання студентами та оцінювання лектором. Остаточну оцінку визначає лектор курсу	7 тиждень кожного модульного циклу	google classroom
МСО3 Поточні контрольні роботи (проміжний модульний контроль)	Проміжний модульний контроль призначений для перевірки рівня засвоєння теоретичного матеріалу. Проводиться у форматі тестування засобами google form/ Оцінка за проміжний модульний контроль не перескладається	згідно графіку навчального процесу	google classroom
МСО5 Підсумковий контроль: диференційний залік	Диференційний залік призначень для перевірки якості отриманих знань в результаті самостійного опрацювання лекційного матеріалу курсу, основної та допоміжної літератури, МВОК. Залік проводиться у форматі тестування засобами google form. Оцінка за залік не перескладається	згідно графіку навчального процесу	google classroom, google meet

Контрольні заходи:

	Максимальна кількість балів	Мінімальна кількість балів	Можливість перескладання з метою підвищення оцінки
3 семестр	**100 балів**		
МСО1. Звіт за результатами виконання завдань практикуму з обробки великих наборів даних	**56**		
8x7	56	5	Так
МСО2. Карти пам'яті	**14**		
7x2	14	1	Так
МСО3. Поточні контрольні роботи (проміжний модульний контроль)	**10**		

		10	6	Ні
МСО5. Підсумковий контроль: диференційний залік		**20**		
		20	Не передбачено	Ні

Для отримання доступу до складання підсумкового контролю знань (диференційований залік) необхідно виконати всі практичні роботи на мінімальний рівень. За умови успішного вивчення масового відкритого онлайн курсу (отримання персоніфікованого сертифікату із зазначенням рівня успішності у відсотках) частина кредитів курсу може бути перезарахована, наприклад, таким чином: - для курсів https://www.coursera.org/learn/scala-spark-big-data або https://www.coursera.org/learn/scala2-spark-big-data може бути перезараховане виконання практикуму з обробки великих даних мовою Scala обсягом 30 балів/1 кредит

11. Ресурсне забезпечення навчальної дисципліни

11.1 Засоби навчання

ЗН1	Мультимедіа, проєкційна апаратура
ЗН2	Комп’ютери, комп’ютерні системи та мережи
ЗН3	Програмне забезпечення google apps for education, IntelLiJ EDA

11.2 Інформаційне та навчально-методичне забезпечення

Основна література	
1	Learning Spark. 2nd Edition/ Jules S. Damji, Brooke Wenig, Tathagata Das, and Denny Lee//Databricks, O’Reilly Media, Inc. - 2020. - 399 p.
Допоміжна література	
2	Frank Kane's Taming Big Data with Apache Spark and Python/Packt Publishing, 2017. - 289 p.
3	Spark for Data Science/Srinivas Duvvuri, Bikramaditya Singhal//Packt Publishing, 2016. - 339
4	Anand Rajaraman, Jure Leskovec, and Jeffrey D. Ullman Mining of Massive Datasets. 2014. - 543 p.
Інформаційні ресурси в Інтернеті	
5	Середовище навчання - https://classroom.google.com/c/NTQzMTk3NjQzMjkw?cjc=wn5sd7f
6	https://www.scala-lang.org/