І СИЛАБУС НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

1. Загальна інформація про навчальну дисципліну	
Повна назва навчальної дисципліни	Аналіз великих наборів 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, к.т.н., доцент
Рівень вищої освіти	другий рівень вищої освіти; НРК України – 7 рівень; QF-LLL – 7 рівень; FQ-EHEA – другий цикл
Семестр вивчення навчальної дисципліни	8 тижнів протягом 3-го семестру
Обсяг навчальної дисципліни	Обсяг навчальної дисципліни становить 5 кредитів ЄКТС, 150 годин, з яких 80 годин становить контактна робота з викладачем (40 годин лекцій, 40 годин практичних робіт), 70 годин становить самостійна робота
Мова(и) викладання	Дисципліна викладається українською мовою
2. Місце навчальної дисципліни в освітній програмі	
Статус дисципліни	Обов’язкова навчальна дисципліна для освітньої програми «Наука про дані та моделювання складних систем»
Передумови для вивчення дисципліни	Необхідними для вивчення дисципліни є наступні знання: • Системи баз даних • Мова структурованих запитів • Алгоритми машинного навчання.
Додаткові умови	Відсутні
Обмеження	Обмеження відсутні

3. Мета навчальної дисципліни

Предметом навчальної дисципліни є алгоритми видобування даних (data mining) з надзвичайно об'ємних баз даних. 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 технологіями Hadoop та Spark.

Тема 2. Map Reduce.
Розподілені файлові системи. Map Reduce (Згортка-Відображення). Алгоритми, побудовані 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. Нормативний спам. Хаби і авторитетні сторінки

Тема 6. Часті предметні набори.
Модель кошиків покупок. Кошики покупок і алгоритм Apriori. Обробка великих наборів даних в оперативній пам'яті. Алгоритм з обмеженою кількістю проходів. Підрахунок частних предметних наборів в потоці.

Тема 7. Кластерізація
Вступ до методів кластерізації. Ієрархічна кластерізація. Алгоритм k середніх. Алгоритм CURE. Кластерізація в неевклідових просторах. Кластреазція для потоків та паралелізм.

Тема 8. Реклама в Інтернеті.
Онлайнові алгоритми. Задача про паросумісність. Задача про ключові слова. Реалізація алгоритму Adwords.

Тема 9. Рекомендаційні системи.
Модель рекомендаційної системи. Рекомендація на основі фільтрації вмісту. Колаборативна фільтрація. Зниження розмірності.

Тема 10. Аналіз графів соціальних мереж.
Вступ до графів. Соціальні мережи як графи. Кластеризація графа соціальної мережі. Пряме знаходження спільнот. Знаходження спільнот, що перетинаються. Околиці в графах

Тема 11. Зниження розмірності
Власні значення та власні вектори. Метод головних компонент. CUR-декомпозиція. Правильний вибір рядків і стовпців.

Тема 12. 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, досягнення яких забезпечує навчальна дисципліна:	
ПРН10	Використовувати на практиці мережеві технології для експериментальної та аналітичної роботи
ПРН17	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
ПРН18	Уміти розробляти та використовувати на практиці алгоритми, пов'язані з спрощенням даних, що описують поведінку системи, класифікацією даних за певними ознаками без навчання та за попередньої наявності класів даних
ПРН19	Уміти оцінити на адекватність результат обробки великих масивів даних
7. Види навчальних занять та навчальної діяльності	
7.1 Види навчальних занять	
Видами навчальних занять при вивченні дисципліни є лекції (Л) та практичні заняття (ПЗ): **Тема 1. Вступ до великих даних:** **Л. 1** Поняття обробки, здобуття даних. Хеш-функція. Індекси Знайомство з технологіями Hadoop та Spark.	

ПЗ 1 Вправи на основі теорії до теми 1.

.

Тема 2. Map Reduce.

Л.2 Розподілені файлові системи. Map Reduce (Згортка-Відображення). Алгоритми, побудовані на MapReduce. Модель комунікативної вартості.

Л.3 Модель комунікативної вартості Теорія складності MapReduce

ПЗ 2 Вправи на основі теорії до теми 2

ПЗ 3 Вступ до мови Scala та інтерфейс командного рядка

Тема 3. Пошук схожих об'єктів.

Л. 4 Додатки пошуку найближчого сусіда. Розбиття документів на шингли. Сигнатури множин із збереженням схожості. Хешування документів. Метрики.

Л. 5 Застосування хешування з урахуванням близькості Корпоративна культура та організація бізнес-процесів

ПЗ 4 Spark - швидкий обчислювальний кластер Вступ до PySpark. Основи машинного обчислення засобами PySpark.

ПЗ 5 Вправи на основі теорії до теми 3.

Тема 4. Аналіз потоків даних

Л. 6 Потокова модель даних. Вибірка даних з потоку. Фільтрація потоків. Підрахунок різних елементів у потоці.

Л. 7. Оцінювання моментів. Моменти вищих порядків.

ПЗ 6 Spark – надбудова SQL

ПЗ 7 Spark Streaming - обробка потоків.

Тема 5. Аналіз посилань.

Л. 8 PageRank. Ефективне обчислення PageRank. Тематичний PageRank..

Л. 9. Посилальний спам. Хаби та авторитетні сторінки

ПЗ 8 Вправи на основі теорії до теми 5

ПЗ 9 Візуалізація великих даних засобами MicroStrategy. Вступ до бізнес-аналітики. Інформаційні панелі в режимі офлайн із розширеною візуалізацією

Тема 6. Часті предметні набори

Л. 10 Модель кошиків покупок. Кошики покупок і алгоритм Apriori. Обробка великих наборів даних в оперативній пам'яті.

Л. 11 Алгоритм з обмеженою кількістю проходів. Підрахунок частих предметних наборів в потоці

ПЗ 10 Вправи на основі теорії до теми 6

ПЗ 11 Візуалізація великих даних засобами MicroStrategy. Побудова простої інформаційної панелі. Аналіз показників для побудови світового сценарію викидів вуглецю

Тема 7. Кластерізація

Л. 12 Вступ до методів кластерізації. Ієрархічна кластерізація. Алгоритм k середніх. Алгоритм CURE.

Л. 13 Кластерізація в неевклідових просторах. Кластреазція для потоків та паралелізм

ПЗ 12 Вправи на основі теорії до теми 7

ПЗ 13 Візуалізація великих даних засобами MicroStrategy. Побудова комплексної інформаційної панелі. Вступ до державних курсів дистанційної освіти

Тема 8. Реклама в Інтернеті.
Л. 14 Онлайнові алгоритми. Задача про паросумісність. Задача про ключові слова. Реалізація алгоритму Adwords
ПЗ 14 Візуалізація великих даних засобами MicroStrategy. Побудова звітної інформаційної панелі. Інтеграція додатку MicroStrategy з Salesforce.com для створення звітної інформаційної панелі

Тема 9. Рекомендаційні системи.
Л. 15 Модель рекомендаційної системи. Рекомендація на основі фільтрації вмісту. Колаборативна фільтрація. Зниження розмірності.
ПЗ 15 Машинне навчання з MLLib. Створення рекомендацій фільмів

Тема 10. Аналіз графів соціальних мереж.
Л. 16 Вступ до графів. Соціальні мережи як графи. Кластеризація графа соціальної мережі.
Л 17 Пряме знаходження спільнот. Знаходження спільнот, що перетинаються. Околиці в графах
ПЗ 16 Аналіз графів з GraphX. Дослідження веб-графів і алгоритмів графів (PageRank) за допомогою GraphX.
ПЗ 17 Вправи на основі теорії до теми 10

Тема 11. Зниження розмірності.
Л. 18 Власні значення та власні вектори. Метод головних компонент.
Л. 19 CUR-декомпозиція. Правильний вибір рядків та стовпців.
ПЗ. 18 Tachyon. Розгортання надійної файлової системи в пам'яті через кластер
ПЗ. 19 Вправи на основі теорії до теми 11

Тема 12. Машинне навчання на великих даних.
Л. 20 Модель машинного навчання. Перцептрони. Метод опорних векторів. Навчання за найближчими сусідами. Порівняння методів навчання
ПЗ 20 BlinkDB. Використання SQL зі статистичною вибіркою.

7.2 Види навчальної діяльності

НД 1 Виконання базових практичних завдань.
НД 2 Виконання розширених практичних завдань.
НД 3 Розв’язання вправ за лекційним матеріалом.
НД 4 Перегляд відеолекцій он-лайн курсу.
НД 5. Розроблення інтелектуальних карт за матеріалом відеолекцій
НД 6 Розв’язання розрахункової роботи за індивідуальним завдання відповідно призначеному варіанту.
НД 7 Участь у лекціях-дискусіях

8. Методи викладання, навчання

Дисципліна передбачає навчання через:

МН1. лекції-візуалізації із використанням мультимедійних технологій, лекції з використанням студентами інтелектуальних карт (або он-лайн лекції).
МН2 репродуктивний метод, що передбачає набування практичних умінь і навичок програмування під час виконання лабораторних робіт, що сприяють використанню пізнаного за матеріалами лекцій - програмна реалізація алгоритмів, викладених у лекційних матеріалах.
МН3 частково-пошуковий метод - організації активного пошуку розв'язання висунутих викладачем (чи самостійно сформульованих) пізнавальних завдань (в тому числі їх програмна реалізація) під час виконання розрахункової роботи.
МН4 дослідницький метод, що передбачає аналіз матеріалу, постановку проблем і завдань з можливістю консультацій з викладачем, як безпосередньо, так і опосередковано через Google Classroom та використання Google Classroom за окремими освітніми компонентами (розміщення матеріалів дисципліни, літератури, тестування знань, тощо).

9. Методи та критерії оцінювання

9.1. Критерії оцінювання

1. Якщо студент під час виконування передбачених навчальним планом видів робіт до залікового тижня набрав загальний рейтинговий бал, що відповідає позитивній оцінці (60 балів і більше), цей результат заноситься в залікову екзаменаційну відомість без можливості його покращення. Підвищення оцінки на заході ПСК не передбачене. Якщо студент не набрав загальний рейтинговий бал, який відповідає позитивній оцінці (60 балів і більше), вважається, що він має заборгованість з дисципліни з процедурою її ліквідації, описаною у п. 2.

2. **Умови ліквідації заборгованостей з поточної роботи.**

   а) Протягом семестру, до залікового тижня, за рішенням викладача студенту може надаватися можливість доопрацювання завдань та контрольних робіт, що передбачені планом роботи, з метою підвищення оцінки.

   Даний пункт не розповсюджується на випадок п.3 стосовно порушень принципів академічної доброчесності.

   б) При отриманні за наслідками роботи за семестр загального рейтингового балу, що відповідає незадовільній оцінці FX (не менше 35 балів), студентові надається право на дворазове складання (викладачеві та комісії) заходу підсумкового семестрового контролю (ПСК), за правилами, що визначені у п.п. в-з;

   в) Складання заходу ПСК, відбувається після завершення екзаменаційної сесії за додатковою відомістю семестрової атестації. Студент має право на два складання заходу ПСК: викладачеві та комісії. У разі незадовільного складання заходу ПСК комісії студент отримує оцінку «незадовільно».

   г) Завдання ПСК являють собою набір тестів. Успішне складання передбачає правильні відповіді на 60 % та більше від загальної кількості питань ПСК.

   д) За умови успішного складання заходу ПСК студент отримує оцінку «задовільно, 60 балів, «Е» за шкалою ECTS, яка засвідчує виконання студентом мінімальних вимог без урахування накопичених балів та реальної кількості наданих правильних відповідей на тестові завдання ПСК.

е) Під час складання заходу ПСК оцінювання здійснюється з урахуванням рейтингових балів, отриманих за підсумком роботи за семестр, але без урахування модульних атестацій. 1 (один) рейтинговий бал прирівнюється до 1 (одного) відсотка отриманих за захід ПСК.

ж) Студенту надається право на виправлення оцінки за домашні (творчі) завдання. Отримані у такий спосіб бали будуть враховані у оцінці за ПСК у спосіб, описаний у п.п. е). Прийом виконаних або виправлених завдань припиняється не пізніше, ніж за три доби до заходу ПСК.

з) У разі незадовільного складання заходу ПСК комісії студент отримує оцінку «незадовільно» з сумою балів, яка відповідає результату, набраному за підсумком роботи за семестр з урахуванням усіх доопрацювань, але без урахування результатів відповідей на питання тестових завдань ПСК. Тобто, набрані на заході ПСК тести у разі незадовільного складання не зараховуються як підсумкові за роботу протягом семестру.

3. Дотримання принципів академічної доброчесності

У випадку порушення норм академічної доброчесності під час виконання завдання, зокрема академічного плагіату, студент отримує 0 (нуль) балів за завдання. При цьому викладач повинен надати докази факту порушення.

9.2 Методи поточного формативного оцінювання

За дисципліною передбачені такі методи поточного формативного оцінювання: тести з теорії на лекціях за допомогою Google Forms опитування (ТТ), оцінювання виконаних практичних завдань (ОВПЗ), оцінювання розроблення інтелектуальних карт (ОРІК), оцінювання вправ на основі теорії (ОВТ) та оцінювання виконання розрахункової роботи (ОВРР).

9.3 Методи підсумкового сумативного оцінювання

У відповідності до регламенту студент має можливість отримати максимальні бази у відповідності до видів завдань за таким переліком

1) Захист звітів за результатами виконання:
   - базових практичних завдань на основі онлайн- технологій (3 практичні заняття) – до 3 балів;
   - розширених практичних завдань на основі он- лайн-технологій (5 запланованих завдання) –до 10 балів;
   - практичних завдань з візуалізації даних (4 практичні заняття) – до 8 балів
2) Виконання розрахункової роботи – до 14 балів;
3) Виконання вправ на основі теорії (9 запланованих вправ) – до 18 балів
4) Розроблення інтелектуальних карт (7 карт; до 7 балів) та експрес-тести на лекційних заняттях (до 20 балів).
5) Модульна контрольна робота – 20 балів.

10. Ресурсне забезпечення навчальної дисципліни

10.1 Засоби навчання	ЗН1 - Мультимедійний проектор для проведення Л ЗН2 - Комп'ютерний клас для ПЗ 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 больших наборов данных. / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2016. – 498 с.: ил. 2. BigData Mini Course. – UC Berkeley *Додаткова література* 3. Jure Leskovec, Anand Rajaraman, Jerey D. Ullman Mining of Massive Datasets. – Cambridge University Press, 2014. – p. 513 **Інформаційні ресурси в Інтернеті:** Google Classroom