Силабус з навчальної дисципліни
“Обробка структурованих та неструктурованих даних Big Data”

Кафедра математичного моделювання

Загальна інформація про навчальну дисципліну

Назва: “Інформаційні системи та технології в системному аналізі”.
Форма навчання: денна.
Семестр IX.
Кредитів 5 (150 год.).
Лекцій – 15 год.; **лабораторних** – 30 год.; **самостійна робота** – 105 год.
Лектор: кандидат фіз.-мат. наук Горбатенко Микола Юрійович.
П.І.Б. викладача який веде лабораторні заняття: Горбатенко М.Ю.

Короткий опис навчальної дисципліни

Мета навчальної дисципліни: розглянути сучасні методи та засоби обробки структурованих та неструктурованих великих даних. Отримати навики розробляти та аналізувати математичні моделі природничих, техногенних, економічних і соціальних об`єктів та процесів. Навчитися використовувати методологію системного аналізу для прийняття рішення в складних системах різної природи.

Завдання: навчити студентів користуватися методами обробки структурованих та неструктурованих даних, використовуючи різні програмні засоби, методики та технології.

Компетенції, якими має оволодіти студент у результаті вивчення дисципліни: у результаті вивчення навчальної дисципліни студент повинен

знати: методи системного аналізу, методи математичного та інформаційного моделювання для побудови та дослідження моделей об`єктів і процесів.

Знати моделі, методи та алгоритми прийняття рішень в умовах конфлікту, нечіткої інформації, невизначеності.

Вміти: застосовувати на практиці методи системного аналізу, методи математичного та інформаційного моделювання для побудови та дослідження моделей об`єктів і процесів.

Вміти розробляти експертні системи, бази знань в умовах слабо структурованих даних різної природи.

Структура начальної дисципліни (теми занять)

Тиждень	Назва теми	к-ть год.	Теми СРС
1	**Тема 1.** Основні поняття та означення Big Data.	1	Поняття машинного навчання та використання його у Data Science. Інструменти Python для машинного навчання.
	Лабораторне заняття 1: Встановлення необхідного програмного забезпечення.	2	Процес моделювання. Створення нових показників і вибір моделі. Тренування моделі. Перевірка адекватності моделі. Прогнозування нових спостережень.

2	**Лабораторне заняття 2:** Порівняння використання реляційних й нереляційних способів збереження даних.	2	Типи машинного навчання Контрольоване та неконтрольоване навчання Частково контрольоване навчання.
3	**Тема 2**. Технології і тенденції роботи з Big Data.	2	Загальні методи обробки великих об`ємів даних. Правильний вибір алгоритма, структури даних, вибір інструментів.
	Лабораторне заняття 3: Запити за великими даними. Запити до даних з Hive. Запити до даних Excel. (Частина 1)	2	Приклади використання технологій обробки великих масивів даних та їх вплив на розвиток компаній.
	Лабораторне заняття 4: Запити за великими даними. Запити до даних з Hive. Запити до даних Excel. (Частина 2)	2	Основні концепції Map Reduce і Spark. Інструменти Hive, Sqoop.
4	**Тема 3**. Методи і техніка аналізу великих даних	2	Вивчення інструменту обробки великих даних – Apache Spar.
5	**Лабораторне заняття 5:** Вивчення інфраструктури для збереження і обробки великих об`ємів даних Hadoop. (Частина 1).	2	Алгоритми розв'язку задач машинного навчання, застосування на практиці алгоритми Spark MLlib.
	Лабораторне заняття 6: Вивчення інфраструктури для збереження і обробки великих об`ємів даних Hadoop. (Частина 2).	2	Аналіз та генерації зображень, відео, тексту, звуку та інших видів даних за допомогою таких інструментів, як Python, Keras і TensorFlow.
7	**Тема 4.** Загальні методи обробки великих даних. Вибір алгоритмів, структур даних, інструментів.	2	Розгортання кластерів HDInsight. Типи кластерів HDInsight.
	Лабораторне заняття 7: Вивчення інфраструктури для збереження і обробки великих об`ємів даних Hadoop. (Частина 3).	2	Управління кластерами HDInsight. Управління кластерами HDInsight за допомогою Power Shell.
8	**Лабораторне заняття 8:** Аналіз великих даних за допомогою Microsoft R (Частина 1)	2	Ознайомлення із онлайн курсом на платформі Prometeus «Аналіз даних та статистичне виведення на мові R» https: //courses. prometheus.org.ua/ courses/IRF/Stat101/2016_T3/about
9	**Тема 5.** Програмне забезпечення для роботи з Великими даними. Microsoft R Server та R Client. Поняття Microsoft R Server. Функції R Client.	2	Трансформація великих даних, управління великими даними.
	Лабораторне заняття 9: Аналіз великих даних за допомогою Microsoft R (Частина 2)	2	Паралельний аналіз операцій. Використання контексту обчислення RxLocalParallel та rxExcel.
10	**Лабораторне заняття 10:** Аналіз великих даних за допомогою Microsoft R (Частина 3)	2	Управління наборами даних. Категоризація даних. Імпорт даних в Azure Machine Learning. Дослідження та перетворення даних у Azure Machine Learning.

11	**Тема 6**. Математичні методи подання Великих даних.	2	Підготовка даних до використання в Azure Machine Learning. Попередня обробка даних. Обробка неповних наборів даних.
	Лабораторне заняття 11: Створення та оцінка регресійних моделей (Частина 1)	2	Використання функції інженерії та вибору.
12	**Лабораторне заняття 12:** Створення та оцінка регресійних моделей (Частина 2)	2	Робочі процес Azure Machine Learning. Оцінка моделей.
13	**Тема 7.** Графові бази даних. Метод аналізу даних в моделі «сутність-характеристика». Розроблення методу прогнозування процесів розвитку. Метод пошуку закономірностей.	2	Графові бази даних. Сфери застосування графових баз даних. Neo4j графова база даних. Мова запитів до графів Cypher.
	Лабораторне заняття 13: Створення та оцінки моделей розподілу (Частина 1).	2	Розглянути приклад прогнозування шкідливих URL- адрес. Визначення мети дослідження. Збір даних URL. Дослідження даних. Побудова моделі.
14	**Лабораторне заняття 14:** Створення та оцінки моделей розподілу (Частина 2).	2	Розглянути приклад побудови рекомендаційної системи у середині бази даних. Підбір необхідних інструментів та методів. Підготовка даних. Побудова моделі. Відображення та автоматизація.
15	**Тема 8**. Вибір типів моделей даних для представлення Великих даних. Формальний опис структури Великих даних.	2	Моделі асоціацій між сутностями та характеристиками для різних категорій NoSQL баз даних. Використання простору даних для моделювання Великих даних.
	Лабораторне заняття 15: Створення та оцінки моделей розподілу (Частина 3).	2	Використання моделей Azure Machine Learning. Розгортання та публікація моделей.

Розподіл балів, які отримують студенти

Поточне оцінювання						Кількість балів (модуль-контроль)	Сумарна к-ть балів
Змістовий модуль №1		Змістовий модуль №2		Змістовий модуль №3			
Т1-Т3	Т4-Т6	Т7	Т8	Т9	Т10	30	100
ЛР1	ЛР2	ЛР3	ЛР4	ЛР5			
10	10	20	20	10			

Індивідуальні навчально-дослідницькі завдання

Студенти, що бажають заробити додаткові бали (до 20) в рахунок ІНДЗ, можуть самостійно зареєструватися на курсі платформи Coursera "Data science"

https://www.coursera.org/professional-certificates/ibm-data-science

отримати відповідний сертифікат і показати його викладачу. Кількість балів буде виставлена пропорційно до Ваших успіхів (досягнення на курсі згідно зі статистикою Coursera).

Рекомендована література

Основна

1. Фрэнкс Б. «Революция в аналитике. Как в эпоху Big Data улучшить ваш бизнес с помощью операционной аналитики» / Билл Фрэнкс. — Москва: Альпина Паблишер, 2017. — 320 с.
2. Юрасов С. Оцифровування статистики, або Перша їжа для Bigdata [Електронний ресурс] / Стас Юрасов // Інтернет-видання Економічна правда. — Електронні дані. — [Київ, Економічна правда, 2015]. — Режим доступу: https://www.epravda.com.ua/publications/2015/08/20/554624 — Назва з екрану.
3. Юрасов С. Город разума [Электронний ресурс] / Стас Юрасов // Информационное агентство ЛІГАБізнесІнформ. — Электронные данные. — [Киев, Информационно-аналитический центр Ліга, 2017]. — Режим доступа: http://www.liga.net/projects/smart_city/ . — Название с экрана.
4. Глущенко Н. Большие данные большого города: как Big Data меняет жизнь Киева [Электронный ресурс] / Нина Глущенко // интернет-журнал AIN.UA. — Электронные данные. — [Киев: AIN.UA, 2017]. — Режим доступа: https://ain.ua/special/big-data-in-kyiv/. — Название с экрана.
5. Кулеш С. Vodafone Украина запускает проект Big Data Lab, в рамках которого откроет массив своих реальных данных IT-разработчикам [Электронный ресурс] / Сергей Кулеш // ITC.ua. — Электронные данные. — [Киев: ООО «ХОТЛАЙН», 2017]. — Режим доступа: https://itc.ua/news/vodafone-ukraina- zapuskaet-proekt-big-data-lab-v-ramkah-kotorogo-otkroet-massiv-svoih-realnyih-dannyih-it-razrabotchikam/. — Название с экрана.
6. Elie Tahari combines fashion savvy with powerful analytics [Electronic resource] / IBM Business Analytics. — Electronic data. — [NY: IBM Corporation, 2014]. — Mode of access: https://www- 01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=YTC03447USEN. — Title from the screen.
7. Золотников Я., Бондарьов О. Друга нафта. В Україні з'явиться онлайн-курс з Big data — найбільш затребуваної в світі IT-професії [Електронний ресурс] / Ярослав Золотников, Олексій Бондарьов // Новое Время: електронний журнал. — Електронні дані. — [Київ: Новое Время, 2016]. — Режим доступу: http://nv.ua/ukr/science/druga-naftu-v-ukrajini-z-javitsja-onlajn-kurs-po-big-data-najbilsh-zatrebuvanoju-v-sviti-it- profesiji-89806.html. — Назва з екрану.

Допоміжна

1. Perez C. Tecchnological revolutions and tecchno-ecoonomic paradigms [Electronic resource] / C. Perez.

// Technology Governance. — Electronic data. — [Tallinn, Tallinn University of Technology, 2009]. — Mode of access: World Wide Web: http://technologygovernance.eu/files/main/2009070708552121.pdf . — Title from the screen.
2. Cavanillas J. M. Curry E., Wahlster W. New Horizons for a Data-Driven Economy. A Roadmap for Usage and Exploitation of Big Data in Europe [Electronic resource] /José María Cavanillas, Edward Curry, Wolfgang Wahlster // Big Data Usage. — Electronic data. — [Springer, Cham, 2016]. — Mode of access: World Wide Web: https://link.springer.com/chapter/10.1007/978-3-319-21569-3_8. — Title from the screen.
3. Research Big Data [Electronic resource] / Wikibon Inc. — Electronic data. — [Wikibon, 2017]. — Mode of access: World Wide Web: https://wikibon.com/research/big-data/. — Title from the screen.
4. Черняк Л. Большие Данные — новая теория и практика [Электронний ресурс] / Л. Черняк // Открытые системы. СУБД. — Электронные данные. — [Москва: “Открытые системы”, 2011]. — № 10. — Режим доступа: https://www.osp.ru/os/2011/10/13010990/. — Название с экрана.
5. Названы причины торможения рынка больших данных [Электронний ресурс] / CNews. — Электронные данные. — [Москва: “CNews”, 2015]. — Режим доступа:

http://bigdata.cnews.ru/news/top/2015- 11-20_analitiki_otsenili_tempy_rosta_mirovogo_rynka (дата обращения 20.11.2017). — Название с экрана.

6. Аналитический обзор рынка Big Data [Электронний ресурс] / Хабрахабр. — Электронные данные. – [Москва: TechMedia, 2015]. — Режим доступа: https://habrahabr.ru/company/moex/blog/256747. — Название с экрана.

Форма контролю та оцінювання результатів навчання:
захист лабораторних робіт, модуль-контроль (усне опитування).