КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ТОРГОВЕЛЬНО-ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

СИСТЕМА УПРАВЛІННЯ ЯКІСТЮ
Система забезпечення якості освітньої діяльності та якості вищої освіти
сертифікована на відповідність ДСТУ ISO 9001:2015 / ISO 9001:2015

Кафедра цифрової економіки та системного аналізу

ЗАТВЕРДЖЕНО
Вченою радою
(прот. п. 6 від «27» 05 2021 р.)
Ректор
_____ А.А. Мазаракі

АНАЛІТИКА BIG DATA / BIG DATA ANALYTICS

ПРОГРАМА / COURSE SUMMARY

Київ 2021

Розповсюдження і тиражування без офіційного дозволу КНТЕУ заборонено

Автор: А. А. Роскладка, доктор економічних наук, професор

Програму розглянуто і затверджено на засіданні кафедри цифрової економіки та системного аналізу 06.04.2021 р., протокол № 9

Рецензент: В. В. Кулаженко, кандидат економічних наук, доцент, доцент кафедри цифрової економіки та системного аналізу

АНАЛІТИКА BIG DATA /
BIG DATA ANALYTICS

ПРОГРАМА /
COURSE SUMMARY

ВСТУП

Програма дисципліни «Аналітика Big Data» призначена для здобувачів другого (магістерського) рівня вищої освіти КНТЕУ денної форми навчання галузі знань 12 «Інформаційні технології», спеціальності 124 «Системний аналіз», освітньої програми «Інформаційні технології та бізнес-аналітика (Data Science)».

Програму підготовлено відповідно до Стандарту вищої освіти України зі спеціальності 124 «Системний аналіз» та освітньо-професійної програми КНТЕУ «Інформаційні технології та бізнес-аналітика (Data Science)» другого (магістерського) рівня вищої освіти.

Програма складається з таких розділів:

1. Мета, завдання та предмет дисципліни.
2. Передумови вивчення дисципліни як вибіркової компоненти освітньої програми.
3. Результати вивчення дисципліни.
4. Зміст дисципліни.
5. Список рекомендованих джерел.

1. МЕТА, ЗАВДАННЯ ТА ПРЕДМЕТ ДИСЦИПЛІНИ

Метою вивчення дисципліни «Аналітика Big Data» є засвоєння основних принципів роботи з великими даними у різних сферах діяльності.

Завданням вивчення дисципліни «Аналітика Big Data» є надання студентам фундаментальних теоретичних знань і набуття практичних навичок з питань методів збирання, обробки та аналізу великих даних із локальних джерел та хмарних середовищ.

Предметом вивчення дисципліни є основні положення і методи аналізу великих даних та їх комп'ютерна реалізація за допомогою аналітичних платформ та систем бізнес-аналітики, насамперед, середовища *R*.

2. ПЕРЕДУМОВИ ВИВЧЕННЯ ДИСЦИПЛІНИ ЯК ВИБІРКОВОЇ КОМПОНЕНТИ ОСВІТНЬОЇ ПРОГРАМИ

знання

- основ інформаційних технологій (операційна система *Windows*, бази даних, доступ до веб-ресурсів);
- основ вищої математики (вектори, матриці, випадкові величини та їх числові характеристики);
- основ роботи у середовищі *R*.

вміння

- вільно працювати з офісними додатками *Microsoft Word, Microsoft Excel, Microsoft PowerPoint*;
- проводити завантаження даних з локальних джерел, здійснювати первинну обробки даних в *R*, графічний аналіз даних з використанням пакетів *base* та *ggplot2*.

3. РЕЗУЛЬТАТИ ВИВЧЕННЯ ДИСЦИПЛІНИ

Дисципліна «Аналітика Big Data», як обов'язкова компонента освітньої програми «Інформаційні технології та бізнес-аналітика (Data Science)» забезпечує оволодіння студентами загальними та фаховими компетентностями і досягнення ними програмних результатів навчання за освітньо-професійною програмою:

*«*Інформаційні технології та бізнес-аналітика (Data Science)*» (ОС магістр)*

Номер в освітній програмі	Зміст компетентності	Номер теми, що розкриває зміст компетентності
Загальні компетентності за освітньою програмою		
ЗК 1	Здатність до абстрактного мислення, аналізу та синтезу	1, 2, 9, 10
ЗК 3	Здатність до пошуку, оброблення та аналізу інформації з різних джерел.	2-10
Спеціальні (фахові, предметні) компетентності за освітньою програмою		
СК 6	Здатність застосовувати теорію і методи Data Science для здійснення інтелектуального аналізу даних з метою виявлення нових властивостей та генерації нових знань про складні системи.	1-10
СК 11	*Здатність ефективно використовувати теорію і методи Data Science.*	3-8
СК 12	*Здатність до здійснення процедур дослідження, аналізу, систематизації та обробки великих даних.*	3-10
СК 13	*Здатність розробляти і впроваджувати моделі задач інтелектуального аналізу даних засобами комп'ютерного моделювання.*	3-7
Програмні результати навчання за освітньою програмою		
РН 2	Будувати та досліджувати моделі складних систем і процесів застосовуючи методи системного аналізу, математичного, комп'ютерного та інформаційного моделювання.	5, 6, 9, 10
РН 6	Застосовувати методи машинного навчання та інтелектуального аналізу даних, математичний апарат нечіткої логіки, теорії ігор та розподіленого штучного інтелекту для розв'язання складних задач системного аналізу.	10
РН 8	Здійснювати ідентифікацію та оцінювання параметрів математичних моделей об'єктів керування.	9
РН 12	*Розробляти моделі управління даними та знаннями в складних системах*	3-7, 9, 10
РН 13	*Здійснювати інтелектуальний аналіз та обробку великих даних засобами комп'ютерного моделювання.*	3-10

4. ЗМІСТ ДИСЦИПЛІНИ

Тема 1. Концепція *Big Data*.

Визначення *Big Data*. Історія розвитку *Big Data*. Масштаби великих даних. Сучасна архітектура *Big Data*. Різниця між бізнес-аналітикою та аналітикою *Big Data*. Методики аналізу великих даних. Аналітичний інструментарій для обробки великих даних. Етапи роботи з великими даними. Платформи великих даних (*Big Data Platform*). Теорія і практика великих даних у різних галузях.

Список рекомендованих джерел

*Основний:*1 [с. 1-22]; 4 [с. 7-24].
Додатковий: 9 [с. 7-17]; 12 [с. 29-59]; 18 [с. 1-7]; 20 [с. 157-197, 297-318]; 21 [с. 3-18]; 22; 25 [с. 9-20].
Інтернет-ресурси: 32; 34; 36; 37.

Тема 2. Концепція *Open Data*.

Регулювання доступу до даних. Джерела публічної інформації у формі відкритих даних. Міжнародна хартія відкритих даних. Рейтинг відкритих даних *Global Open Data Index*. Рейтинг розвитку відкритих даних *Open Data Barometer*. Закон України «Про доступ до публічної інформації». Положення Кабінету міністрів України «Про набори даних, які підлягають оприлюдненню у формі відкритих даних». Світовий портал відкритих даних *Data.world.* Єдиний державний веб-портал відкритих даних *Data.gov.ua.* Вимоги до структури наборів даних. Допустимі типи та формати даних. Успішні проекти *Open Data* в Україні. Система аналітики відкритих даних *Clarity Project.*

Список рекомендованих джерел

Основний: 1 [с. 25-36]; 4[с. 7-24, 240-311].
Додатковий: 12 [с. 75-110]; 13 [с. 140-170]; 14 [с. 695-710]; 16 [с. 2-34]; 17 [с. 45-56]; 20 [с. 1-29, 91-113]; 21 [с. 19-47].
Інтернет-ресурси: 32; 34; 36; 37.

Тема 3. Первинна обробка великих даних у середовищі R

Структури даних в *R*. Імпорт та експорт об’єктів *R.* Імпорт даних у різних форматах. Імпорт даних з використанням пакету *readr*. Виявлення пропущених значень. Виявлення джерел пропущених даних. Метод множинного відновлення пропущених значень. Категоріальні дані. Робота з факторами за допомогою пакету *forcast*. Робота з даними у бінарному форматі. Взаємодія із різними базами даних. Робота з реляційними даними та перетворення даних з використанням пакету *dplyr*. Акуратизація даних за допомогою *tidyr*. Робота з великими даними рядкового типу з використанням *stringr*. Робота з датою і часом на основі пакету *lubridate*. Створення *tibble*-фреймів. Робота з каналами великих даних з використанням пакету *magrittr*. Функції *R* для складних даних: *apply, sapply, lapply, do.call*.

Список рекомендованих джерел

Основний: 1 [с. 63-114]; 2 [с. 14-18]; 4 [с. 25-126].
Додатковий: 7 [с. 42-52]; 13 [с. 6-14, 38-70]; 14 [с. 29-72]; 15 [с. 30-80]; 18 [с. 61-78]; 24 [с. 3-13]; 25 [с. 39-80].
Інтернет-ресурси: 28; 29; 31.

Тема 4. Графічний аналіз великих даних.

Чотири графічних системи *R*. Пакет *lattice*. Пакет *ggplot2*. Інтерактивна графіка за допомогою пакетів *playwith, latticist, iplots, rggobi*. Інтерактивна візуалізація даних з використанням *plotly, googleVis, rCharts*. Коробкові графіки. Діаграми розсіювання. Корелограми. Мозаїчні діаграми. Спінограми. Скрипкові діаграми.

Побудова декількох графіків в одній системі координат. Мультиграфіки з використанням параметрів *mfrow, mfcol*. Керування кольором, типами та розмірами символів, типом та параметрами ліній графіків. Назви, підписи, легенди та інша текстова інформація в графічній системі. Паралельні графіки. Тренд у точкових даних. Візуалізація великої кількості спостережень. *Waffle*-графіки.

Вибір форми візуалізації великих даних в залежності від мети візуалізації: розподіл, взаємозв'язки, порівняння, структура даних.

Список рекомендованих джерел

Основний: 1 [с. 377-393]; 3 [с. 42-155].
Додатковий: 13 [с. 172-195, 324-348]; 14 [с. 113-232]; 15 [с. 45-71, 119-140, 373-399]; 18 [с. 121-133].
Інтернет-ресурси: 28; 29; 31.

Тема 5. Створення інтерактивних веб-додатків для аналізу великих даних у *R Shiny*.

Створення директорії і файлів веб-додатку. Запуск і зупинка веб-додатку. Керування запуском додатку. Трасування в *Shiny*.

Односторінкові макети. Функції сторінки. Сторінки з бічною панеллю. Багаторядкове введення даних. Багатосторінкові макети. Набори вкладок. Навігаційний список та навігаційна панель. Зворотній зв'язок з користувачем. Система оповіщення користувача. Індикатори ходу виконання завдань.

Додавання елементів інтерфейсу користувача. Динамічний інтерфейс користувача: ієрархічні списки, циклічні посилання, взаємопов'язані елементи введення даних. Створення інтерфейсу користувача за допомогою програмного коду.

Список рекомендованих джерел

Основний: 5 [с. 24-49, 91-226].
Інтернет-ресурси: 27-31.

Тема 6. Реактивне програмування процедур обробки великих даних у *Shiny*.

Подійно-орієнтоване програмування. Основи реактивного програмування. Зниження дублювання коду за допомогою реактивних виразів. Протистояння імперативного та декларативного програмування. Реактивні вирази та графіки. Контроль часу запуску реактивних виразів.

Складові блоки реактивного програмування. Реактивні значення. Реактивні вирази. Ізолювання коду. Функції *isolate*(), *observeEvent*(), *eventReactive*(). Інвалідація за часом.

Ефективні прийому реактивного програмування в аналізі великих даних. Функціональне програмування в *Shiny*. Інтерфейс користувача у вигляді структури даних. Серверні функції. Основи модульного програмування в *Shiny*. Кешування реактивних виразів. Оптимізація роботи веб-додатку *Shiny*.

Список рекомендованих джерел

Основний: 5 [с. 50-74, 227-368].
Інтернет-ресурси: 27-31.

Тема 7. Аналіз великих даних у *Power BI*.

Зчитування файлів у форматі *csv*. Зчитування даних з *Microsoft Excel*Імпорт даних з *SQL Server*. Завантаження даних у *Power BI* засобами API з використанням *R*. Динамічне об'єднання файлів. Фільтрація рядків на основі регулярних виразів.

Створення візуалізації в *Power BI* за допомогою *R*. Діаграми з анотаціями. Бульбашкові діаграми.Візуалізація прогнозних моделей. Лінійна діаграма із затемненням. Використання карт у *Power BI* на основі результатів аналізу даних в *R*. Діаграм квадрантів. Лінії регресії в моделях *R* у *Power BI*.

Список рекомендованих джерел

Основний: 3 [с. 157-249].
Додатковий: 12 [с. 113-136]; 17 [с. 70-153]
Інтернет-ресурси: 26.

Тема 8. Веб-скрапінг та парсінг великих даних

Категорізація веб-сторінок. Робота з HTTP в R (пакети httr, RCurl). Кирилиця і кодування URL. Аналіз посилань. Технологія PageRank. Тематичний PageRank. Спам та технологія TrustRank. Хаби й авторитетні сторінки. Аналіз графів соціальних мереж.

Елементи HTML і CSS у задачах веб-скрапінгу. Отримання та обробка документа. Пакети для веб-скрепінгу. Пакет *rvest*. Функції навігації. Керування браузером в R. Парсінг динамічних веб-сторінок за допомогою RSelenium. Парсінг html-сторінки з використанням XPath в R. PhantomJS і обробка динамічних веб-сторінок. Пакет RFacebook. Збір інформації з використанням API. Використання Twitter API.

Регулярні вирази у веб-скрапінгу та парсінгу великих даних: символи, метасимволи та квантифікатори. Створення карт на основі зібраних даних.

Список рекомендованих джерел

Основний: 1 [с. 327-356]; 2 [с. 192-204]; 4 [с. 240-369].
Додатковий: 9 [с. 178-220]; 10 [с. 33-98]; 14 [с. 435-441, 477-481].
Інтернет-ресурси: 30; 33.

Тема 9. Технології MapReduce, Hadoop і Spark в аналітиці великих даних.

Аналітика неструктурованих даних. Розподілені файлові системи. Поняття *MapReduce*. Алгоритми з використанням технології *MapReduce*. Теорія складності *MapReduce*. Екосистема *Hadoop*. Потокова обробка даних за технологією *Hadoop*. Інтеграція *R* і *Hadoop*. Спеціалізовані пакети *R* для роботи з *Hadoop*. Функції *hsTableReader*, *hsKeyValReader*, *hsLineReader*. Технологія *Spark* в аналітиці великих даних. *Spark* і *R* на багатовузловому *HDInsight*-кластері.

Список рекомендованих джерел

Основний: 1 [с. 295-323]; 2 [с. 37-141]; 4 [127-239, 370-408].
Додатковий: 9 [с. 242-259, 291-305].
Інтернет-ресурси: 30; 31; 33; 35.

Тема 10. Методи машинного навчання для надвеликих даних.

Модель машинного навчання. Перцептони. Метод опорних векторів і аналітиці надвеликих даних. Навчання методами найближчих сусідів. Ядерна регресія. Типи алгоритмів машинного навчання. Рекомендаційні алгоритми. Створення рекомендацій з *R* і *Hadoop*.

Застосування алгоритмів машинного навчання при завантаженні даних надвеликих даних в модель *Power BI*. Використання готових моделей штучного інтелекту для розширення функціоналу моделей даних. Налаштування *Cognitive Services* в *Azure*. Віртуальна машина для аналізу даних (*Data Science Virtual Machine – DSVM*). Застосування сторонніх моделей машинного навчання до моделей даних у *Power BI*. Конфігурація засобів аналізу надвеликих даних у *IBM Watson*.

Список рекомендованих джерел

Основний: 1 [с. 359-393]; 2 [с. 14-178]; 3 [с. 268-333]; 4 [409-465].
Додатковий: 7 [с. 373-413]; 14 [с. 497-672]; 20 [с. 235-277, 319-364].
Інтернет-ресурси: 30; 31; 33; 35.

5. СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

Основний

1. *Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data /EMC Education Services. Indianapolis : John Wiley & Sons, Inc, 2015. 432 p.*
2. *Prajapati V. Big Data Analytics with R and Hadoop: Packt Publishing, 2013. 238 p.*
3. *Wade R. Advanced Analytics in Power BI with R and Python: Ingesting, Transforming, Visualizing. Apress, 2020. – 440 p.*
4. *Walkowiak S. Big Data Analytics with R: Packt Publishing, 2016. 491 p.*
5. *Wickham H. Mastering Shiny: O'Reilly Media, 2021 352 p.*

Додатковий

6. Bell R. Big data in travel, consumption and marketing: big data for business. Independently published, 2021. – 142 p.
7. *Casella G., Fienberg S., Olkinand I. An Introduction to statistical learning with application in R: Springer, 2017. 440 p.*
8. Chinnici M., Pop F., Negru C. Data Science and Big Data Analytics in Smart Environments. CRC Press, 2021. – 292 p.
9. *Cuesta H., Kumar S. Practical Data Analysis: Packt Publishing, 2016. 330 p.*
10. *Danneman N., Heimann R. Social Media Mining with R: Packt Publishing, 2014. 122 p.*
11. Etaati L. Advanced Analytics with Power BI and R. - Auckland: RADACAD Systems Limited, 2017. - 179 p.
12. *Evans J. R. Business Analytics: Methods, Models, and Decisions: Pearson, 2021. 705 p.*
13. *Hafner S. An Introduction to R for Beginners. - Hafner Consulting LLC, 2019. - 360 p.*
14. *Irizarry A. Introduction to Data Science. Data Analysis and Prediction Algorithms with R: Chapman and Hall/CRC, 2020. 743 p.*
15. *Kabacoff R. R in Action. Data analysis and graphics with R. – Manning: Shelter island, 2015. 608 p*
16. *Kakulapati V. Open Data: ITexLi, 2022. 77 p.*
17. *Larson B. Data Analysis with Microsoft Power BI: Mc Graw Hill, 2020. 741 p.*
18. *Mariani M. C., Tweneboah O. K., Beccar-Varela M. P. Data Science in Theory and Practice: Techniques for Big Data Analytics and Complex Data Sets: John Wiley & Sons, Inc., 2022. 403 p.*
19. Matloff N. Probability and Statistics for Data Science: Math + R + Data. – London: Chapman & Hall, 2019. – 376 p.
20. *Mishra B.K., Kumar V., Panda S.K., Tiwari P. Handbook of Research for Big Data. Concepts and Techniques: CRC-Press, 2022. 389 p.*
21. *Moreira J. M., de Carvalho A.C.P.L.F, Horvath T., A General Introduction to Data Analytics: John Wiley & Sons, Inc., 2019. 328 p.*
22. *Rancher A. An Introduction to Big Data Concepts URL: https://www.suse.com/c/rancher_blog/an-introduction-to-big-data-concepts/*
23. Sharda R., Delen D., Turban E. Analytics, Data Science, & Artificial Intelligence: Systems for Decision Support. Pearson; 11th edition, 2019. – 832 p.
24. *Майборода Р. Є., Сугакова О. В. Аналіз даних за допомогою пакета R: навчальний посібник. – К.: ВПЦ «Київський університет», 2015. – 65 с.*
25. *Негрей М., Гнот Т. Аналітика з R. – Київ: Компринт, 2020. – 236 с.*

Інтернет-ресурси

26. Microsoft Power BI Guided Learning. URL: https://docs.microsoft.com/uk-ua/power-bi/guided-learning
27. Official portal of R Shiny. URL: https://shiny.rstudio.com/
28. RStudio. URL: https://www.rstudio.com/products/rstudio/download/#download
29. RStudio-education. URL: https://github.com/rstudio-education

30. Scraping the Web in R. URL: https://www.sscc.wisc.edu/sscc/pubs/webscraping-r/scraping-the-web.html
31. The Comprehensive R Archive Network. URL: https://cran.r-project.org
32. The Open Data World Portal. URL: https://data.world
33. Web Scraping in R with rvest. URL: https://www.dataquest.io/blog/web-scraping-in-r-rvest
34. Аналітична система *YouControl*. URL: https://youcontrol.com.ua
35. Парсінг даних у R. URL: http://rstudio-pubs-static.s3.amazonaws.com/6955_2c6760795680448f8ab9f47c08669ca0.html
36. Портал відкритих даних. URL: https://data.gov.ua
37. Система аналітики відкритих даних *Clarity Project*. URL: https://clarity-project.info

* Курсивом виділені джерела, що є у бібліотеці КНТЕУ, або наявні повнотекстові електронні версії джерел.

ЛИСТ ПОГОДЖЕННЯ
програми дисципліни «Аналітика Big Data»

Погоджено
Завідувач кафедри цифрової економіки та системного аналізу, гарант освітньої програми «Інформаційні технології та бізнес-аналітика (Data Science)»
(ОС магістр)

_____ А. А. Роскладка

«_____»_____ 2021р.