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
(пост. п. 6 від «27» 02. 2020 р.)
Ректор
_____ А. А. Мазаракі

ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ / DATA ANALYSIS TECHNOLOGIES

ПРОГРАМА / COURSE SUMMARY

Київ 2020

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

Програму розглянуто і затверджено на засіданні кафедри цифрової економіки та системного аналізу 14.02.2020 р., протокол № 13

Рецензенти: В. В. Кулаженко, кандидат економічних наук,
М. Г. Шарафутдінов, бізнес-аналітик, директор з розвитку компанії «Center Research & Development».

ТЕХНОЛОГІЇ АНАЛІЗУ ДАНИХ/
DATA ANALYSIS TECHNOLOGIES

ПРОГРАМА / COURSE SUMMARY

ВСТУП

Програма вибіркової дисципліни «Технології аналізу даних» призначена для аспірантів КНТЕУ денної форми навчання галузі знань 12 «Інформаційні технології», спеціальності 122 «Комп'ютерні науки».

Програму підготовлено відповідно до освітньо-наукової програми підготовки аспірантів КНТЕУ за спеціальністю 122 «Комп'ютерні науки».

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 ПРЕДМЕТДИСЦИПЛІНИ

Метою вивчення дисципліни «Технології аналізу даних» є надання фундаментальних теоретичних знань і набуття практичних навичок з питань формування, дослідження та всебічного аналізу інформації в різних галузях науки.

Завданням вивчення дисципліни «Технології аналізу даних» є надання аспірантам ґрунтовних знань в області аналітичних досліджень інформаційного простору, вивчення методів створення, переробки, трансформації, захисту даних.

Предметом вивчення дисципліни є основні положення та методи аналізу даних та їх комп'ютерна реалізація за допомогою аналітичних платформ.

2. ПЕРЕДУМОВИ ВИВЧЕННЯ ДИСЦИПЛІНИ ЯК ВИБІРКОВОЇ КОМПОНЕНТИ ОСВІТНЬОЇ ПРОГРАМИ

знання

- основ інформаційних технологій (операційна система *Windows*, бази даних, системи захисту інформації);
- основ теорії ймовірностей та математичної статистики (випадкові величини та їх числові характеристики, закони розподілу випадкових величин, статистичні гіпотези та методи їх перевірки);

вміння

- вільно працювати з офісними додатками *Microsoft Word, Microsoft Excel, Microsoft PowerPoint.*

3. РЕЗУЛЬТАТИ ВИВЧЕННЯ ДИСЦИПЛІНИ

Дисципліна «Технології аналізу даних» забезпечує оволодіння аспірантами загальними та фаховими компетентностями і досягнення ними програмних результатів навчання за освітньо-науковою програмою спеціальності 122 «Комп'ютерні науки»

Номер в освітній програмі	Зміст компетентності	Номер теми, що розкриває зміст компетентності
Загальнонаукові компетентності		
ЗК 2	Здатність застосовувати теоретичні та практичні знання у науковій діяльності для вирішення задач у предметній області	
ЗК 3	ЗК3. Здатність забезпечувати інноваційний характер науково-дослідної роботи та самостійно вирішувати поставлені наукові задачі.	
Спеціальні (фахові, предметні) компетентності із спеціальності (СК)		
СК1	Засвоєння основних концепцій наукових досліджень в області комп'ютерних наук.	1
СК3	Оволодіння термінологією та понятійним апаратом з досліджуваного наукового напряму.	1
СК 5	Здатність до системного мислення та аналізу при дослідженні складних проблем різної природи у галузі комп'ютерних наук, застосування методів формалізації та розв'язування системних задач, що мають суперечливі цілі, невизначеності та ризики.	
СК6	Знання механізмів застосування інтелектуального аналізу та методів обчислювального інтелекту для роботи з великими та слабо структурованими даними з метою їхньої оперативної обробки та візуалізації результатів аналізу	1-4
Програмні результати навчання		
ПРН 1	Проводедення аналітичних досліджень сучасної проблематики в області комп'ютерних наук за результатами наукової діяльності провідних зарубіжних та вітчизняних вчених, здатність формулювати мету, визначати об'єкт, предмет та завдання власного наукового дослідження	2,3
ПРН 3	Вміння здійснювати наукові дослідження у відповідності до методології наукового дослідження на основі по-етапної технології	2-4
ПРН 4	Вміти застосовувати методологію наукового пізнання, форм і методів аналізу, обробки та синтезу інформації в предметній області комп'ютерних наук.	2-4
ПРН 5	Вміння застосовувати сучасні засоби обчислювальної техніки у науковій діяльності для проведення теоретичних та експериментальних досліджень	4
ПРН 9	Застосування системного підходу та методів формалізації при дослідженні складних задач різної природи у галузі комп'ютерних наук, що характеризуються суперечливістю, невизначеністю та ризиками	2,3
ПРН 10	Вміти застосовувати механізми інтелектуального аналізу та методи обчислювального інтелекту для роботи з великими та слабо структурованими даними з метою їхньої оперативної обробки та візуалізації результатів аналізу	2-4

4. ЗМІСТ ДИСЦИПЛІНИ

Тема 1. Наука продані –*Data Science*

Глосарій *Data Science.* Історія розвитку *Artificial Intelligence* і *Business Intelligence.* Зв'язок понять «дані», «інформація» та «знання». Характеристика фахівця з аналізу даних. *Softskills* та *hardskills* аналітика даних. Приклади застосування *DataScience* у різних галузях людської діяльності.

Типи та види даних. Форми представлення даних. Вимірювання і шкали в аналізі даних. Реліяційні та багатовимірні дані. Змінні, постійні та умовно-постійні дані. Довідкові, оперативні та архівні дані. Точкові дані та дані за період. Первинні і вторинні дані. Метадані.

Формати зберігання даних. Типи наборів даних. Транзакційні дані.

Етапи розв'язування задач аналізу даних: висунення гіпотез, збір і систематизація даних, побудова моделі, яка пояснює факти, тестування моделі та інтерпретація результатів, застосування отриманої моделі.

Технологія *Knowledge Discovery in Databases*. Формування вибірки даних.

Технологія *DataMining.* Задачі *DataMining*: класифікація, регресія, кластеризація, асоціація, послідовність даних. Поняття про аналітичні системи. Актуальні бізнес-задачі аналізу даних.

Список рекомендованих джерел

Основний: 2, 4.
Додатковий: 13, 14, 19.
Інтернет-джерела: 25.

Тема 2. Консолідація та передобробка даних

Поняття консолідації. Основні критерії оптимальності консолідації даних. Джерела даних. Основні задачі консолідації даних. Схема процесу консолідації. Очищення даних. Збагачення даних.

Спеціалізовані сховища даних. *ETL*-процес. Сховища даних у системах підтримки прийняття рішень (СППР). Відмінності СППР та *OLTP*-систем. Одноплатформенні та крос-платформенні сховища даних. *OLAP*-системи.

Рівні якості даних: технічний, аналітичний та концептуальний рівень. Оцінка придатності даних до аналізу. Технології та методи оцінки якості даних. Профайлінг. Візуальна оцінка якості даних. Причини надходження в систему «брудних даних».

Передобробка даних. Очищення від шумів і згладжування рядів даних. Фільтрація даних. Відновлення пропущених значень. Редагування аномальних значень. Обробка дублікатів і протиріч. Зниження вимірності вхідних даних. Семплінг. Усунення незначущих факторів.

Трансформація даних. Основні методи трансформації даних. Перетворення часових рядів. Квантування даних. Сортування даних. Злиття даних. Об'єднання даних. Налаштування набору даних. Нормалізація даних. Трансформація впорядкованих даних. Групування та розгрупування даних. Внутрішнє та зовнішнє з'єднання даних. Нормалізація та кодування даних.

Список рекомендованих джерел

*Основний:*1-3.
Додатковий: 6, 8-12.
Інтернет-джерела: 25, 26.

Тема 3. Технології інтелектуальної обробки даних

Афінітивний аналіз. Основні поняття *Rules Mining*. Асоціативні правила. Підтримка та достовірність правил. Значущість асоціативних правил. Ліфт, левередж та покращення асоціативних правил.

Алгоритм *apriori*. Пошук предметних наборів. Генерація асоціативних правил. Практичний аспект застосування технології асоціативних правил. Секвенціальний аналіз.

Формальна постановка задачі кластеризації. Базові алгоритми кластеризації. Алгоритм кластеризації *k-means*. Критерій збіжності алгоритму. Міри відстаней у кластеризації. Міри Евкліда і Манхеттена. Алгоритм *g-means*. Кластеризація за Гюстафсоном-Кесселем. Програмні засоби кластеризації. Мережі Кохонена. Самоорганізуючі карти Кохонена.

Огляд методів класифікації. Кореляційно-регресійний аналіз. Статистичні методи аналізу. Байесівська класифікація. Простий байесівській класифікатор.

Лінійна регресія. Регресія з категоріальними вхідними змінними. Логістична регресія. Оцінки максимальної правдоподібності. Значущість вхідних змінних. Використання логістичної регресії для розв'язування задач класифікації. Тест Чоу. *ROC*-аналіз.

Візуальний аналіз даних – *Visual Mining*. Характеристики засобів візуалізації даних. Методи візуалізації.

Задача та етапи аналізу текстової інформації – *Text Mining*. Етапи *TextMining*: пошук інформації, попередня обробка документів, витяг інформації, застосування методів *TextMining*, інтерпретація результатів. Методи класифікації текстових документів. Видалення стоп-слів. Стеммінг. *N*-грами. Зведення регістра. Методи кластеризації текстових документів: ієрархічні, бінарні. Задача анотування текстів. Пошук асоціацій.

Ідея *Data Mining* у реальному часі. *Real-Time Mining*. Адаптивне добування даних. Інструменти *Data Mining* у реальному часі.

Складнощі аналізу даних з мережі Інтернет. Етапи *WebMining*. Категорії *WebMining*. Аналіз використання веб-ресурсів. Використання веб-структур та веб-контенту.

Список рекомендованих джерел

*Основний:*1-3.
Додатковий: 5-7, 15-24.
Інтернет-джерела: 25, 26.

Тема 4. Інструментальні засоби аналізу даних

Програмне забезпечення в області аналізу даних. Аналітичні платформи. Технології аналізу даних у продуктах *Microsoft Corporation*. Технологія моделювання даних у *Microsoft Power Pivot*. Інтерактивний інструмент *Microsoft Power View* для дослідження, графічного відображення та представлення даних. Надбудова *Microsoft Power Query* в задачах бізнес-аналітики.

Хмарні технології *Microsoft* для аналізу та візуалізації даних. Організація бізнес-аналітики рівня *Business Intelligence (BI)*. Платформа *Microsoft BI*. Завантаження даних *Power BI* з різних інформаційних джерел.

Аналітична платформа *Deductor Studio*. Базові навички роботи у системі *Deductor*. Візуалізатори в *Deductor Studio*. Сортування, заміна та фільтрація інформації в *Deductor Studio*. Консолідація даних в системі *Deductor*. Асоціація, кластеризація та класифікація даних в системі *Deductor*. Побудова аналітичної звітності.

Список рекомендованих джерел

*Основний:*1-3.
*Додатковий:*5-8, 11, 13, 15, 16, 18-22.
Інтернет-джерела: 25, 26.

5. СПИСОК РЕКОМЕНДОВАНИХ ДЖЕРЕЛ

Основний

1. Cuesta H., Kumar S. Practical Data Analysis.Birmingham : Packt Publishing Ltd, 2016. 316 p.
2. Data Science & Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data /EMC Education Services. Indianapolis : John Wiley & Sons, Inc, 2015. 432 p.
3. Microsoft Power BI Cookbook: Creating Business Intelligence Solutions of Analytical Data Models, Reports, and Dashboards. Birmingham : Packt Publishing Ltd, 2017. 802 p.
4. *Roskladka A., Ivanova O., Kulazhenko V. Data Scientist: a glance into the future // Зовнішня торгівля: економіка, фінанси, право*. 2019. № 3. С. 109-120

* Курсивом виділені джерела, що є у бібліотеці КНТЕУ

Додатковий

5. *Гнатієнко Г. М., Снитюк В. Є. Експертні технології прийняття рішень:монографія. Київ : Маклаут, 2008. 444 с.*
6. Матвійчук А. В. Штучний інтелект в економіці: нейронні мережі, нечітка логіка: монографія. Київ : КНЕУ, 2011. 439 с.
7. Олійник А. О., Субботін С. О., Олійник О. О. Інтелектуальний аналіз даних: навч. посібн. Запоріжжя : ЗНТУ, 2012. 278 с.
8. Панфилов А. Н., Скоба А. Н., Кузнецова А. В., Зуев В. А. Интеллектуальный анализ данных. Новочеркасск: Лик, 2016. – 76 с.
9. Ситник В. Ф.,Краснюк М. Т. Інтелектуальний аналіз даних (дейтамайнінг) : навч. посібн.Київ : КНЕУ, 2007. 376 с.

10. Субботін С. О. Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень : навч. посібн. Запоріжжя : ЗНТУ, 2008. 341 с.
11. Яковлев В. Б. Финансовый анализ данных в *Deductor Studio*. М.: ОнтоПринт, 2018. – 168 с.
12. Adamson C. Mastering Data Warehouse Aggregates: Solutions for Star Schema Performance. Wiley Publishing Inc., 2006. 318 p.
13. Albright S. C., Winston W., Zappe C. Data Analysis and Decision Making. Boston : Cengage Learning, 2016. 948 p.
14. Cao L., Yu P. S., ZhangC., Zhang H. Data Mining for Business Applications. Springer Science; Business Media, 2008. 402 p.
15. Coodley M. O. Introduction to Microsoft Power BI: bring your data to life! CreateSpace Independent Publishing Platform, 2016. 128 p.
16. Etaati L. Advance Analytics with Power BI and R. Auckland : Radacad Systems Limited, 2017. 179 p.
17. Fabrice G., Hamilton N. J.Quality Measures in Data Mining. Berlin; Heidelberg: Springer-Verlag, 2007. 361 p.
18. Ferrari A., Russo M. Introducing Microsoft Power BI. Redmond : Microsoft Press, 2016. 407 p.
19. Han J., Kamber M. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, 2006. 800 p.
20. Linoff G. S. Data Analysis Using SQL and Excel. Indianapolis: Wiley, 2015. 792 p.
21. Linoff G. S., Berry M. J. A. Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management. Indianapolis: Wiley, 2011. 888 p.
22. Milton M. Head First Data Analysis: A learner's guide to big numbers, statistics, and good decisions. Sebastopol: O'Reilly Media, 2009. 435 p.
23. RapidMiner: Data Mining Use Cases and Business Analytics Applications / Edited byMarkus Hofmann&Ralf Klinkenberg. Minneapolis : CRC Press, 2004. 518 p.
24. UptonG. Categorical data analysis by example. New Jersey: John Wiley & Sons Inc, 2017. 198 p.

Internet-ресурси

25. Microsoft Power BI Guided Learning URL: https://docs.microsoft.com/uk-ua/power-bi/guided-learning(дата звернення 18.03.2020).
26. *Deductor Studio* URL: https://basegroup.ru/deductor/description (дата звернення 18.03.2020).