ДЕРЖАВНА СЛУЖБА СТАТИСТИКИ УКРАЇНИ
НАЦІОНАЛЬНА АКАДЕМІЯ СТАТИСТИКИ ОБЛІКУ ТА АУДИТУ
Кафедра статистики, інформаційних технологій та математичних методів в економіці

МЕТОДИЧНІ РЕКОМЕНДАЦІЇ ДЛЯ САМОСТІЙНОЇ РОБОТИ

з навчальної дисципліни

Вступ до Big Data

(назва начальної дисципліни)

для студентів спеціальності 051 «Економіка»
освітньо-професійної програми
«Прикладна статистика та бізнес-аналітика»

Київ
2022 рік

Горобець О. О. Методичні рекомендації для самостійної роботи з навчальної дисципліни «Вступ до Big data» для студентів для студентів денної та заочної форми навчання, спеціальності 051 «Економіка», освітньо-професійної програми «Статистика та бізнес-аналітика». Київ: НАСОА, 2022. 13 с.

Методичні рекомендації затверджені на засіданні кафедри статистики інформаційних технологій та математичних методів в економіці. Протокол від «30» серпня 2022 року, № 1.

Схвалено Вченою радою обліково-статистичного факультету НАСОА

Протокол від «30» серпня 2022 року, № 1.

Рецензенти:

Герасименко С. С. – завідувач кафедри статистики, інформаційних технологій та математичних методів в економіці Національної академії статистики, обліку та аудиту, д.е.н., професор.

Мотузка О. М. – доцент кафедри економіки та менеджменту зовнішньоекономічної діяльності Національної академії статистики, обліку та аудиту, к.е.н., доцент.

1. ЗАГАЛЬНІ ПОЛОЖЕННЯ

Самостійна робота студентів – це форма організації навчального процесу, при якій заплановані завдання виконуються студентом під методичним керівництвом викладача, але без його безпосередньої участі, тобто самостійно.

Самостійна робота студентів займає важливе місце у системі сучасної вищої освіти, яка є основним засобом засвоєння студентом навчального матеріалу в час, вільний від обов'язкових навчальних занять. Самостійна робота студентів забезпечується всіма навчально-методичними засобами, необхідними для вивчення навчальної дисципліни чи окремої теми: підручниками, навчальними та методичними посібниками, конспектами лекцій та ін.

Основними видами самостійної роботи студента є:

- опрацювання змісту рекомендованих джерел інформації;
- опрацювання опорного конспекту лекцій;
- опрацювання питань, передбачених робочою програмою та винесених на самостійне вивчення;
- підготовка до практичних занять;
- підготовка до поточного і підсумкового контролю;
- підготовка презентацій, наукових тез, статей тощо;
- проведення досліджень із певної проблематики.

2. МЕТА ТА ЗАВДАННЯ НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Метою вивчення навчальної дисципліни є формування системи теоретичних знань у сфері big data та практичних навичок управління Big Data. Демонстрація переваг та недоліків імплементації Big Data в діяльність органів офіційної статистики, представлення наявного інструментарію, що використовується для збирання, обробки, групування та аналізу Big Data та подальшого представлення статистичної інформації.

Завданням вивчення навчальної дисципліни є вивчення основних понять у сфері Big Data, ознайомлення з принципами появи та розповсюдження Big Data, організації сховищ великих даних, із базовими алгоритмами збереження, пошуку та аналізу Big Data. Сформувати навички роботи з програмним забезпеченням Big Data та використанням статистичних методів для аналізу Big Data.

3. ПРОГРАМА НАВЧАЛЬНОЇ ДИСЦИПЛІНИ

Змістовний модуль 1.
Теоретико-методичні засади big data

Тема 1. Історія появи та поширення терміна Big Data. Предмет та завдання Big Data

Поняття, предмет та загальна концепція big data. Джерела та причини виникнення великих об'ємів даних. Переваги та недоліки big data. Завдання big data. Практичний досвід імплементації big data.

Рекомендована література:
Базова: [6, 7]
Допоміжна: [1, 2, 6, 19]

Тема 2. Глобальні групи даних. Категорії даних

Поняття про глобальні групи даних – Shallow data, Deep Data, Micro-data, Nano-data, їх структуризацію та класифікацію. Ідентифікація машинних даних, потокових даних, озер даних. Категорії структурованих, неструктурованих та напівструктурованих даних. Інструменти для роботи з big data.

Рекомендована література:
Базова: [2, 6, 7]
Допоміжна: [1]

Тема 3. Роль та значення Big Data для офіційної статистики. Забезпечення якості та конфіденційності

Big Data як альтернативне джерело статистичних даних для органів офіційної статистики. Нормативно-правове забезпечення якості та конфіденційності даних. Поняття «якість» та «конфіденційність» в контексті big data. Параметри якості даних. Вимоги до якості даних. Виміри якості даних. Поняття «життєвий цикл даних». Формування вартості даних. Типи конфіденційних та персональних даних в структурі big data.

Рекомендована література:

Базова: [6]
Допоміжна: [1, 5]
Електронні ресурси: [6]

Змістовний модуль 2.
Методи та інструменти обробки Big Data

Тема 4. Екосистема Apache Hadoop: архітектура, реплікація, читання і запис даних

Характеристика екосистеми Apache Hadoop. Класифікація алгоритмів для опрацювання big data. Найпоширеніші методи аналізу та прогнозування big data в екосистемі Apache Hadoop: метод *k*-середніх; метод опорних векторів регресійний аналіз, метод головних компонент, метод асоціативних правил, аналіз соціальних медіа (SNA). Математичний аналіз SNA.

Рекомендована література:
Базова: [1, 3, 5, 8]
Допоміжна: [4, 7, 9]
Електронні ресурси: [2]

Тема 5. Програмний каркас Hadoop MapReduce

Характеристика інтерфейсу MapReduce. Організація функціонування MapReduce. Утиліта Hadoop Streaming. Технологія опрацювання даних «Відображення» – «Згортання».

Рекомендована література:
Базова: [3]
Допоміжна: [4]
Електронні ресурси: [2]

Тема 6. Алгоритмізація даних. Обчислення даних за допомогою нейромереж

Поняття «алгоритм». Особливості алгоритмізації даних. Поняття «нейромережа». Роль та значення нейромереж в процесі обробки big data.

Рекомендована література:
Базова: [4, 9]
Допоміжна: [11]

Тема 7. Фреймворки Apache Spark та Hive для обробки та зберігання Big Data

Характеристика фреймворку Apache Spark. Особливості проєкту Hive. Роль та значення Apache Spark Hive для обробки, аналізу та зберігання Big Data.

Рекомендована література:
Базова: [1, 7]
Допоміжна: [4, 8]
Електронні ресурси: [3, 4]

Тема 8. Платформи хмарних обчислень Big Data

Платформи Amazon Web Services та Google Cloud Platform: особливості, переваги та недоліки використання. Сервіс Microsoft Azure: загальна характеристика статистичних інструментів для обробки Big Data.

Рекомендована література:
Базова: [2]
Допоміжна: [4, 8]
Електронні ресурси: [1]

Тема 9. Платформа Power BI як інструмент візуалізації Big Data

Функціонал платформи Power BI. Основні компоненти набору Power BI. Переваги та недоліки Power BI. Статистичне моделювання великих обсягів даних за допомогою інструментів Power BI.

Рекомендована література:
Базова: [7]
Допоміжна: [2, 3]

ЗАВДАННЯ ДЛЯ САМОСТІЙНОЇ РОБОТИ ЗА ТЕМАМИ

Змістовний модуль 1.
Теоретико-методичні засади Big Data

Тема 1. Історія появи та поширення терміна Big Data

Питання для самостійного вивчення:

1. Методи та інформаційні технології обробки Big Data.
2. Сфери застосування Big Data.
3. Сучасні тенденції в розвитку систем обробки великих обсягів даних.

Завдання для самостійної роботи

Підготувати ілюстративну доповідь за допомогою інструменту MicrosoftOffice – PowerPoint.

1. Досвід використання Big Data в закладах охорони здоров'я.
2. Досвід використання Big Data в телекомунікаційній сфері.
3. Досвід використання Big Data в аграрній сфері.
4. Досвід використання Big Data в роботизації.
5. Досвід використання Big Data у фінансових установах.
6. Досвід використання Big Data в умовах воєнного часу.

Рекомендована література:
Базова: [6, 7]
Допоміжна: [1, 2, 6, 19]

Тема 2. Глобальні групи даних. Категорії даних

Питання для самостійного вивчення:

1. Життєвий цикл даних. Попередня обробка даних.
2. Алгоритм ZET заповнення пробілів у таблицях даних.
3. Метадані. Життєвий цикл метаданих.

Завдання для самостійної роботи

2. Пройдіть тестування:
1. Що таке великі дані?

А) великі сукупності даних про різноманітні соціально-економічні явища і процеси, які продукуються фактично в безперервному режимі.
Б) великого обсягу таблиці, які містять сукупності даних про різноманітні

соціально-економічні явища і процеси, які продукуються фактично в безперервному режимі.
В) дані, які лаконічно представлені у наочному вигляді та уже готові до майнінгування.

2. Що таке Data Mining?
А) графічна презентація даних або інформації.
Б) процес напівавтоматизованого аналізу великих баз даних з метою пошуку корисних фактів.
В) масові системні кількісні та якісні характеристики про різноманітні соціально-економічні явища і процеси.

3. Shallow Data – це:
А) дані, які повільно змінюються та, зазвичай, стосуються об'єкта дослідження.
Б) дані, що мають надвисокий ступінь деталізації.
В) дані, які за своїми характеристиками нагадують атомарну подію.

4. Що варто розуміти, коли ми говоримо про Інтернет речей?
А) це дані з опосередкованих процесів, тобто: медична документація, банківські записи, електронна комерція.
Б) це дані згенеровані машинами, тобто: дані з датчиків, комп'ютерних систем та ін.
В) це дані згенеровані людиною, тобто: електронна пошта, відео, коментарі та ін.

5. Під час обробки потокових даних найчастіше використовують:
А) Apache Hadoop.
Б) Apache Kafka, AWS.
В) Apache Kafka

Тема 3. Роль та значення Big Data для офіційної статистики. Забезпечення якості та конфіденційності

План для самостійного вивчення

1. Принципи офіційної статистики в контексті забезпечення конфіденційності та якості даних.
2. Забезпечення конфіденційності даних в мережі Інтернет

Завдання для самостійної роботи

Підготувати презентацію на одну із запропонованих тем:

1. Роль забезпечення конфіденційності даних в офіційній статистиці

2. Забезпечення якості в офіційній статистиці
3. Недоліки Big Data у частині забезпечення якості та конфіденційності даних.

Змістовий модуль 2. Методи та інструменти обробки Big Data

Тема 4. Екосистема Apache Hadoop: архітектура, реплікація, читання і запис даних

План для самостійного вивчення

1.Архітектура та компоненти, кластери, їх вузли, файлова система HDFS.
2. Доступ до сервісів Hadoop засобами мов програмування.

Завдання для самостійної роботи

Завдання 1: Відвідайте сайт сервісу Hadoop https://hadoop.apache.org/ , ознайомтеся із структурою та архітектурою сервісу.
Завдання 2: Підготуйте презентацію щодо структури та архітектури Hadoop ґрунтуючись на отримані знання із виконаного попереднього завдання.

Тема 5. Програмний каркас Hadoop MapReduce

План для самостійного вивчення

1. Програмні оболонки, вибір інтерфейсу та мови програмування при створенні програмних продуктів.
2. MapReduce: використання бібліотек та стандартних (ринкових) програмних продуктів, потоки даних.

Завдання для самостійної роботи

З метою поглиблення знань про платформу Hadoop, відвідайте сайт сервісу Hadoop https://hadoop.apache.org/ , ознайомтеся із можливостями, які представляє користувачу MapReduce.

Тема 6. Алгоритмізація даних. Обчислення даних за допомогою нейромереж

План для самостійного вивчення

1.Особливості побудови алгоритмів.
2.Властивості штучних нейронних мереж. Біологічний прототип

3.Найпоширеніші помилки побудови алгоритмів.

Завдання для самостійної роботи

Пройдіть тестування на закріплення теми:

1.З яких процесів складається Apache Hadoop?
А)Відображення та згортання
Б)Відображення та структуризація
В)Відображення та скорочення

2.Що таке Apache Hadoop?
А)Спеціальне програмне забезпечення для обробки великих даних
Б)Фреймворк для обробки великих даних
В)Утиліта для обробки великих даних

3.За допомогою яких засобів здійснюється обробка великих даних?
А)Сталих статистичних методів
Б)Математичних та статистичних методів
В)Алгоритмів

4.Які типи алгоритмів вам відомі?
А)Ті, які одночасно завантажують увесь набір даних на пам'ять комп’ютера та ті, які у пам’яті комп’ютера зберігають проміжні копії
Б)Ті, які працюють у хмарному середовищі та ті, які які у пам’яті комп’ютера зберігають проміжні копії
В)Ті, які одночасно завантажують увесь набір даних на пам'ять комп’ютера, ті, які у пам’яті комп’ютера зберігають проміжні копії та ті, які працюють у хмарному середовищі

5.Які існують категорії алгоритмів?
А)Навчання з учителем та навчання без учителя
Б)Навчання з учителем, навчання без учителя та навчання з підкріпленням
В)Навчання з учителем та самонавчання

Тема 7. Фреймворки Apache Spark та Hive для обробки та зберігання Big Data

План для самостійного вивчення

1. Структуризація Apache Spark. Практичний досвід використання інструментарію фреймворку в практичну діяльність.
2. Особливості фреймворку Hive. Практичний досвід використання інструментарію фреймворку в практичну діяльність

Завдання для самостійної роботи

Підготувати презентацію на одну із запропонованих тем:

1. Особливості Apache Spark.
2. Практичні кейси використання Apache Spark у світі та Україні
3. Особливості фреймворку Hive.
4. Практичні кейси використання фреймворку Hive у світі та Україні.

Тема 8. Платформи хмарних обчислень Big Data
План для самостійного вивчення

1.Особливості обробки та зберігання даних на платформах Amazon Web Services та Google Cloud Platform
2. Особливості обробки та зберігання даних за допомогою сервісу Microsoft Azure.

Завдання для самостійної роботи

Підготувати презентацію на одну із запропонованих тем:

1.Унікальність платформ Amazon Web Services та Google Cloud Platform
2.Проведіть компаративний аналіз платформ Amazon Web Services та Google Cloud Platform

Тема 9. Платформа Power BI як інструмент візуалізації Big Data
План для самостійного вивчення

1. Інструментарій платформи Power BI.
2. Особливості візуалізації big data у Power BI.

Завдання для самостійної роботи

Підготувати презентацію на одну із запропонованих тем:

1. Методи візуалізації в статистичних дослідженнях
2. Унікальність платформи Power BI з точки зору статистиків

Рекомендована література

Базова

1. Bruce P., Bruce A., Gedeck P. (2020). Practical Statistics for Data Scientists: 50+ Essential Concepts Using R Python. *O'reilly Media*, 368.

2. Wengrow J. A Common-Sense Guide to Data Structures and Algorithms. 2nd Ed. *Pragmatic Bookshelf*, 456.
3. White T. (2015). Hadoop: The Definitive Guide: Storage and Analysis at Internet Scale. *O'Reilly Media*. 4th Edition, 756.
4. Бородкіна І. Л. Теорія алгоритмів: навч. посіб. Київ: Центр навчальної літератури, 2020. 184 с.
5. Мармоза А. Теорія статистики. Київ: Центр учбової літератури, 2021. 592 с.
6. Осауленко О. Г. Офіційна статистика у системі національної інформаційної безпеки: моногр. Київ: ТОВ «Август Трейд», 2017. 367 с.
7. Провост Ф. Фоусетт Т. Data Science для бізнесу. Як збирати, аналізувати і використовувати дані. Київ, Наш формат, 2019. 400 с.
8. Руденко В. Математична статистика. Київ: Центр навчальної літератури, 2019. 304 с.
9. Субботін С. О. Нейронні мережі: теорія та практика: навч. посіб. Житомир: Видавець О. О. Євенок, 2020. 184 с.

Допоміжна

1. Horobets O. (2021). Research Data as a Result of Research Activities: the Role and Significance for the Official Statistics, *Journal of the Knowledge Economy*, *Springer,* vol. 12(3), pages 1424-1436.
2. Indurkhya N. (1997). *Predictive Data Mining: A Practical Guide (The Morgan Kaufmann Series in Data Management Systems)*. 1st Edition, 228.
3. Kamiński, B., Jakubczyk, M. & Szufel, P. (2018). A framework for sensitivity analysis of decision trees. *Cent. Eur. J. Oper. Res. 26*, 135–159. https://doi.org/10.1007/s10100-017-0479-6
4. Koitzsch, K. (2017). Overview: Building Data Analytic Systems with Hadoop. In: Pro Hadoop Data Analytics. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-1910-2_1
5. Radermacher W. (2017). Official Statistics 4.0. Swiss Statistical Days 2017. doi: https://doi.org/10.13140/RG.2.2.16604.9024
6. Song D., Golin E. (1993). Fine-grain visualization algorithms in dataflow environments. *VIS '93: Proceedings of the 4th conference on Visualization '93, pp.126–133.*
7. Taleb, I., Serhani, M.A., Bouhaddioui, C. et al. (2021). Big data quality framework: a holistic approach to continuous quality management. *J Big Data*, 8, 76. https://doi.org/10.1186/s40537-021-00468-0
8. Wu C., Buyya R., Ramamohanarao K. (2016). Chapter 1. Big Data Analytics = Machine Learning + Cloud Computing. *Big Data Principles and Paradigms.* Ed. R. Buyya R. N. Calheiros A. V. Dastjerdi. Elsevier. 3-38. Retrieved from: https://www.sciencedirect.com/science/article/pii/B9780128053942000015
9. Горобець О. О. Великі дані – джерело статистичної інформації: на прикладі книговидавничої галузі. *Науковий вісник Національної академії статистики, обліку та аудиту: зб. наук. пр.*. 2019. №1-2. С. 7-13.

10. Кейт О'Нілл. Big Data: зброя математичного знищення. Як великі дані збільшують нерівність і загрожують демократії. Київ: Bookchef, 2020. 336 с.
11. Нейронні мережі : теорія та практика: навч. посіб. Житомир : Вид. О. О. Євенок, 2020. 84 с.

Електронні ресурси

1. Amazon Web Services. URL: https://aws.amazon.com
2. Apache Hadoop. URL: https://hadoop.apache.org/
3. Apache Hive. URL: https://hive.apache.org/
4. Apache Spark. URL: https://spark.apache.org/
5. Azure. URL: https://azure.microsoft.com
6. Про офіційну статистику: Закон України від 16.08.2022 р. № 2524-IX. URL: https://zakon.rada.gov.ua/laws/show/2524-20#Text