



Силабус освітнього компонента

Програма навчальної дисципліни

Аналітика даних та Big Data у видавничій справі

Шифр та назва спеціальності

G20 – Видавництво та поліграфія

Спеціалізація

-

Освітня програма

Сучасні технології у видавництві та медіаіндустрії

Рівень освіти

Другий (магістерський)

Семестр

2

Інститут

ННІ Комп'ютерних наук та інформаційних технологій

Кафедра

Системного аналізу та інформаційно-аналітичних технологій (322)

Тип дисципліни

Вибіркова

Форма навчання

Денна

Мова викладання

Українська

Викладачі, розробники

**Колбасін Вячеслав Олександрович**

viacheslav.kolbasin@khpi.edu.ua

Кандидат технічних наук, доцент, доцент кафедри системного аналізу та інформаційно-аналітичних технологій НТУ "ХПІ"

Досвід роботи – 26 років. Автор понад 40 наукових та навчально-методичних праць. Має професійні сертифікації: AWS Certified Solutions Architect – Associate, AWS Certified Machine Learning – Specialty, Oracle Certified Associate Java SE7, Oracle Certified Professional Java SE7.

[Детальніше про викладача на сайті кафедри](#)

Загальна інформація

Анотація

Дисципліна спрямована на опанування студентами платформ та фреймворків, що використовуються при обробці великих обсягів даних та їх публікації для споживача. Розглянуто фреймворки Apache Hadoop, Apache Hive, Apache Spark, основи NoSQL баз даних та архітектури, що застосовуються при обробці великих обсягів даних, типові задачі інтелектуальної обробки великих обсягів даних. В ході лабораторних робіт та курсової роботи студентами виконується аналіз великих обсягів даних.

Мета та цілі дисципліни

Мета викладання дисципліни полягає в формуванні у студентів теоретичних знань і практичних навичок застосувань технологій обробки великих обсягів даних для вирішення прикладних та наукових задач у медіаіндустрії за допомогою класичних технологій BigData, таких як Apache Hadoop, Apache Hive, Apache Spark, NoSQL баз даних, засобів машинного навчання та візуалізації.

Формат занять

Лекції, лабораторні роботи, консультації, самостійна робота. Підсумковий контроль – залік.

Компетентності

Інтегральна компетентність: здатність розв'язувати складні задачі дослідницького та/або інноваційного характеру у сфері видавництва та поліграфії.

Загальні компетентності:

ЗК 1. Здатність застосовувати знання у практичних ситуаціях.

ЗК 2. Здатність генерувати нові ідеї (креативність).

ЗК 4. Здатність до пошуку, оброблення та аналізу інформації з різних джерел.

ЗК 8. Здатність до абстрактного мислення, аналізу та синтезу.

Спеціальні компетентності:

СК 1. Здатність комплексно оцінювати вплив середовища функціонування технологічних і виробничих процесів для удосконалення параметрів продукції.

СК 2. Здатність критично осмислювати проблеми видавництва і поліграфії та на межі галузей знань, а також перспективних напрямів розвитку галузі.

СК 3. Здатність визначати головні функції і напрямки вдосконалення забезпечення виробництва розробляти заходи оперативного та перспективного управління, прогнозування і планування виробництва.

СК 7. Здатність застосовувати сучасні методи та інструменти для досліджень у сфері видавництва та поліграфії, а також забезпечення якості продукції.

СК 8. Здатність розробляти і реалізовувати наукові та прикладні проекти у сфері видавництва і поліграфії та з дотичних до неї міждисциплінарних напрямів з урахуванням технічних, економічних, соціальних, правових та екологічних аспектів.

СК 9. Здатність проектувати та розробляти архітектуру, інтерфейси (UX/UI) та контентне наповнення сучасних мультимедійних, цифрових та вебсистем у медіаіндустрії.

СК 10. Здатність інтегрувати та застосовувати методи штучного інтелекту, інтелектуального аналізу даних (Data Mining) та Big Data для оптимізації процесів підготовки контенту та прийняття рішень у видавничій справі.

Результати навчання

РН 2. Оцінювати перспективи, створювати науково-технічно обґрунтовані прогнози досліджувати й здійснювати концептуально-змістове моделювання тенденцій розвитку галузі.

РН 6. Здійснювати управління складною діяльністю у сфері видавництва та поліграфії, організовувати та вдосконалювати діяльність видавничо-поліграфічних виробництв, розробляти плани і заходи з їх реалізації, забезпечувати якість, та розраховувати техніко-економічну ефективність виробництва.

РН 7. Здійснювати комп'ютерне проектування окремих складових технологічного процесу.

РН 9. Здійснювати дослідження та/або провадити інноваційну діяльність з метою отримання нових знань та створення нових технологій та продуктів в сфері видавництва і поліграфії та в ширших мультидисциплінарних контекстах

РН 10. Будувати та досліджувати моделі технологічних процесів видавництва та поліграфії, оцінювати їх адекватність, визначати межі застосовності.

РН 11. Застосовувати сучасні експериментальні та математичні методи, інформаційні технології та спеціалізоване програмне забезпечення для досліджень і розробок у сфері видавництва та поліграфії.

РН 12. Відшуковувати необхідні дані в науковій літературі, базах даних та інших джерелах, аналізувати та оцінювати ці дані.

Обсяг дисципліни

Загальний обсяг дисципліни 120 год. (4 кредити ECTS): лекції – 32 год., лабораторні роботи – 16 год., самостійна робота – 72 год.

Передумови вивчення дисципліни (пререквізити)

Для успішного вивчення дисципліни необхідно мати знання та практичні навички з програмування та основ баз даних.

Особливості дисципліни, методи та технології навчання

Лекції проводяться з використанням мультимедійних технологій. Для виконання лабораторних робіт використовуються середовища розробки Eclipse, IntelliJ IDEA або PyCharm. Навчальні матеріали доступні студентам через OneDrive кафедри.

Програма навчальної дисципліни

Навчальні заняття

Лекції

Теми лекцій	Кількість годин
Тема 1. Вступ. Задача обробки великих обсягів даних. Технології обробки великих обсягів даних. Горизонтальне та вертикальне масштабування.	2
Тема 2. Розподілене зберігання та обробка великих обсягів даних. Файлова система HDFS. Консольні операції та API для роботи з файлами в HDFS.	2
Тема 3. Підхід до паралельної обробки даних та фреймворк Map-Reduce. Алгоритми паралельної обробки даних. Структура фреймворку Map-Reduce. Реалізація простих задач обробки даних.	2
Тема 4. Практичне використання Map-Reduce підходу. Оптимізація Map-Reduce задач. Рядкові та колонкові формати файлів. Управління сортуванням та передачею даних між вузлами.	2
Тема 5. Управління кластером. Система управління кластером YARN. Запуск задач та управління їх виконанням за допомогою YARN. Інші фреймворки управління кластером.	2
Тема 6. Фреймворк Apache Hive. Принципи побудови та архітектура Apache Hive. Виконання SQL запитів за допомогою Hive. Обмеження та особливості SQL в Hive. Направлені ациклічні графи (DAG) та фреймворк Tez.	2
Тема 7. Фреймворк Apache Spark. Принципи побудови та архітектура Apache Spark. Зберігання даних у RDD. Конвеєр обробки даних у Spark. Створення простих задач у Spark.	2
Тема 8. Робота за даними у Spark. RDD, DataFrame, DataSet. Використання Java та Python у Spark. Модель пам'яті Python застосування. Оптимізація Spark застосувань.	2
Тема 9. Spark SQL. Виконання запитів Spark SQL. Особливості використання Spark SQL. Засоби оптимізації Spark SQL запитів.	2
Тема 10. Обробка поточкових даних. Архітектура та засоби обробки поточкових даних. Система обміну повідомленнями Apache Kafka. Робота з поточковими даними в Spark Streaming.	2
Тема 11. Базы даних для роботи з великими обсягами даних. Концепція та різновиди NoSQL баз даних. Орієнтована на колонки БД Apache Cassandra. Документно-орієнтована БД MongoDB. New SQL бази даних.	2
Тема 12. Організація та моделювання даних.	2

Рівні моделювання даних. Багатовимірні моделі (multi dimensional data models). Обробка даних, що змінюються (Slowly Changing Dimensions, SCD).

Тема 13. Архітектура систем обробки даних. Типовий процес обробки. Data warehouse та datamart. Архітектура Data Lake. Лямбда-архітектура.	2
Тема 14. Інтелектуальна обробка даних. Засоби інтелектуальної обробки даних в Apache Spark. Технології машинного навчання у Spark.	2
Тема 15. Обробка текстових даних. Моделі текстових даних – модель мішка слів та векторні моделі. Тематичні моделі тексту. Аналіз текстів за допомогою тематичних моделей.	2
Тема 16. Системи візуалізації результатів. Візуалізація результатів обробки даних. Платформи Tableau та Microsoft PowerBI. Динамічні дашборди.	2
Загальна кількість годин	32

Лабораторні заняття

Теми лабораторних занять	Кількість годин	Вагові коефіцієнти a
Тема 1. Встановлення та ознайомлення з віртуальною машиною для роботи з BigData технологіями. Програмний доступ до файлів в HDFS.	2	0,1
Тема 2. Пошук слів найбільшої довжини за допомогою Map-Reduce. Дослідження впливу формату файлу на швидкість обробки даних та обсяг зайнятої дискової пам'яті.	2	0,1
Тема 3. Обчислення статистичних показників по датасету за допомогою Apache Hive.	2	0,1
Тема 4. Пошук найдовшого слова за допомогою Apache Spark.	2	0,1
Тема 5. Обчислення статистичних показників по датасету з використанням Apache Spark SQL.	2	0,1
Тема 6. Обчислення статистики по продажах мережі роздрібної торгівлі у режимі реального часу з використанням Apache Spark Streaming.	2	0,1
Тема 7. Побудова моделі даних для обліку польотів цивільної авіації та обліку вильотів пілотів. Використання архітектури Data Lake.	2	0,1
Тема 8. Побудова дашбордів для представлення статистичних даних.	2	0,1
Загальна кількість годин	16	$\sum_{i=1}^n a_i = 0,8$

Контрольні роботи

Теми контрольних робіт	Вагові коефіцієнти b
Тема 1. Основи технологій BigData.	0,1
Тема 2. Системи обробки та представлення даних.	0,1
Загалом	$\sum_{i=1}^m b_i = 0,2$

Самостійна робота

До самостійної роботи відноситься виконання розрахункового завдання, самостійне опрацювання теоретичного матеріалу та підготовка до контрольних робіт.

Тематика розрахункових завдань

Розрахункове завдання передбачає аналіз та представлення статистичних даних по індивідуально заданому зрізу датасету польотів цивільної авіації США. Здобувач отримує конкретну тему в середині семестру і має подати завдання на перевірку до заліку. Обсяг звіту: 8–12 сторінок основного тексту.

Загальна кількість годин

20

Опрацювання теоретичного матеріалу

Теми для самостійного вивчення	Кількість годин
Тема 1. Розподілене зберігання даних та масштабування Команди розподіленої файлової системи HDFS. Хмарні системи зберігання даних (AWS S3, Azure Blob Storage). Приклади практичного використання Map Reduce Framework	8
Тема 2. Apache Hive Особливості SQL в Apache Hive. Підключення сторонніх застосунків до Hive.	16
Тема 3. Apache Spark Spark Dataframe API. Особливості роботи зі Spark SQL. Виконання Spark коду в середовищі Databricks.	12
Тема 4. Робота з потоковими даними Встановлення Apache Kafka та KSQL. Основні команди KSQL. Інтеграція статичних джерел даних з KSQL.	8
Тема 5. Mongo DB Встановлення Mongo DB. Основні команди для роботи з Mongo DB. Команди агрегаційних запитів.	18
Тема 6. Датамарт та візуалізація даних Створення моделі даних для DataWarehouse інтернет магазину. Підключення сховища даних до Tableau/PowerBI. Основні засоби для створення звітів та дашбордів PowerBI.	10
Загальна кількість годин	72

Неформальна освіта

Здобувач має можливість зарахувати окремі теми або курс шляхом проходження курсів та онлайн тренінгів, професійних стажувань та професійних сертифікацій. Для зарахування потрібно надати сертифікат про закінчення тренінгу, опис його програми та звіт, що має містити результати програмних завдань курсу.

Рекомендовані курси, тренінги, стажування

1. Introduction to Big Data with Spark and Hadoop

<https://www.coursera.org/learn/introduction-to-big-data-with-spark-hadoop>

2. Power BI: Data Modeling and Data Analysis

<https://www.coursera.org/learn/power-bi-data-modeling-and-data-analysis>

3. Relational Database Support for Data Warehouses

<https://www.coursera.org/learn/dwrelational>

4. Data Software Engineering

<https://campus.epam.com/en/training/5296>

Література, навчальні матеріали та інформаційні ресурси

Основна література

1. Олещенко Л.М. Технології оброблення великих даних. Конспект лекцій. Київ: КПІ ім. Ігоря Сікорського, 2021. 227 с. URL: <https://ela.kpi.ua/bitstreams/d05e72c9-26d1-41ad-bc6f-97f88d8e0938/download>
2. Warren J., Marz N. Big Data. Manning, 2015. 328 p. URL: <https://www.manning.com/books/big-data>
3. Rioux J. Data Analysis with Python and PySpark. Manning, 2022. 456 p. <https://www.manning.com/books/data-analysis-with-python-and-pyspark>
4. Scott D., Gamov V., Klein D. Kafka in Action, 2022. 272 p. URL: <https://www.manning.com/books/kafka-in-action>
5. Wilson B. Machine Learning Engineering in Action. Manning, 2022. 576 p. URL: <https://www.manning.com/books/machine-learning-engineering-in-action>

Додаткова література

1. Chitturi P. Apache Spark for Data Science Cookbook. Packt publishing, 2016. 392 p. URL: https://www.packtpub.com/en-us/product/apache-spark-for-data-science-cookbook-9781785880100?srsId=AfmBOoLZsGdnwL_KL9fxDtZ258XrXGdiKxT-4CR1XbF78-io0Pqhf
2. Holmes A. Hadoop in Practice, 2nd Edition. Manning, 2014. 512 p. URL: <https://www.manning.com/books/hadoop-in-practice-second-edition>
3. Shapira G., Palino T., Sivaram R. Kafka: The Definitive Guide, 2nd edition. O'Reilly Media, 2021. 485p. URL: <https://www.oreilly.com/library/view/kafka-the-definitive/9781491936153/>
4. Banik R. Hands-On Recommendation Systems with Python. Packt publishing, 2018. 146 p. URL: <https://www.oreilly.com/library/view/hands-on-recommendation-systems/9781788993753/>
5. Perrin J. Spark in Action, 2nd Edition. Manning, 2020. 576 p. URL: <https://www.manning.com/books/spark-in-action-second-edition>
6. Falk K. Practical Recommender Systems, Manning. 2019. 432 p. URL: <https://www.manning.com/books/practical-recommender-systems>
7. Apache Hadoop documentation. URL: <http://hadoop.apache.org/docs/current/>.
8. Apache Hive Language Manual. URL: <https://cwiki.apache.org/confluence/display/Hive/LanguageManual>
9. Apache Spark documentation. URL: <https://spark.apache.org/docs/latest/>.
10. Kafka Documentation, URL: <https://kafka.apache.org/documentation/>

Система оцінювання

Підсумкова оцінка з освітнього компонента визначається відповідальним лектором за темами, видами занять, тощо у відповідності до силабусу і є інтегральною оцінкою результатів усіх вид навчальної діяльності здобувача вищої освіти. Підсумкова оцінка повинна відображати всі оцінки за складовими навчального процесу з урахуванням їх вагових показників k :

Поточний контроль (практичні, семінарські, лабораторні заняття), k_1	Контрольні роботи (за наявності), k_2	Індивідуальне завдання (за наявності), k_3	Підсумковий контроль (для ОК з іспитом), k_4
0,8	0,2	0	0

Розрахунок підсумкової оцінки проводиться за формулою:

$$O = P \cdot k_1 + K \cdot k_2 + I \cdot k_3 + P_k \cdot k_4$$

де: P – середньозважена середня оцінка за поточний контроль
 I – оцінка за виконання індивідуального завдання
 K – середньозважена оцінка за контрольні роботи
 P_k – оцінка за підсумковий контроль

$$P = \frac{P_1 \cdot a_1 + P_2 \cdot a_2 + \dots + P_n \cdot a_n}{\sum_{i=1}^n a_i}$$

де: a_i - ваговий коефіцієнт за кожне практичне (семінарське) або лабораторне заняття.

$$K = \frac{K_1 \cdot b_1 + K_2 \cdot b_2 + \dots + K_m \cdot b_m}{\sum_{i=1}^m b_i}$$

де: b_i - ваговий коефіцієнт за кожну контрольну роботу.

Поточні оцінки за кожну складову ($P, K, I, \dots$) виставляються за 100-бальною шкалою згідно з положенням «Про критерії та систему оцінювання знань та вмінь і про рейтинг здобувачів вищої освіти» НТУ «ХПІ».

Підсумкова оцінка виставляється відповідно до розрахованої O з округленням до найближчого цілого числа в більшу сторону.

Шкала оцінювання

Сума балів	Національна оцінка	ECTS
90–100	Відмінно	A
82–89	Добре	B
75–81	Добре	C
64–74	Задовільно	D
60–63	Задовільно	E
35–59	Незадовільно (потрібне додаткове вивчення)	FX
1–34	Незадовільно (потрібне повторне вивчення)	F

Норми академічної етики і політика курсу

Здобувач вищої освіти повинен дотримуватися «Кодексу етики академічних взаємовідносин та доброчесності НТУ «ХПІ»»: виявляти дисциплінованість, вихованість, доброзичливість, чесність, відповідальність. Конфліктні ситуації повинні відкрито обговорюватися в навчальних групах з викладачем, а при неможливості вирішення конфлікту – доводитися до відома співробітників дирекції інституту.

Нормативно-правове забезпечення впровадження принципів академічної доброчесності НТУ «ХПІ» розміщено на сайті: <http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/>

Погодження

Силабус погоджено

26.08.2026

Завідувачка кафедри
Тетяна АЛЕКСАНДРОВА

26.08.2026

Гарант ОП
Тетяна АЛЕКСАНДРОВА