Politechnika Krakowska im. Tadeusza Kościuszki

Karta przedmiotu

obowiązuje studentów rozpoczynających studia w roku akademickim 2022/2023

Wydział Informatyki i Telekomunikacji

Kierunek studiów: Matematyka Stosowana — Profil: Praktyczny

Forma sudiów: stacjonarne — Kod kierunku: MS

Stopień studiów: I

Specjalności: Analityka Danych

1 Informacje o przedmiocie

Nazwa przedmiotu	Przetwarzanie dużych zbiorów danych
Nazwa przedmiotu w języku angielskim	Big Data Processing
Kod przedmiotu	WIiT MS pIS D13 22/23
Kategoria przedmiotu	Przedmioty specjalnościowe
Liczba punktów ECTS	4.00
Semestry	6

2 Rodzaj zajęć, liczba godzin w planie studiów

Semestr	Wykład	Ćwiczenia	Laboratorium	Laboratorium komputerowe	Seminarium	Projekt
6	30	0	0	30	0	0

3 Cele przedmiotu

Cel 1 Wprowadzenie podstawowych pojęć związanych z przetwarzaniem dużych zbiorów danych typu Big Data.

Cel 2 Zapoznanie z narzędziami przetwarzania klastrowego danych i wykonywania obliczeń rozproszonych.

Cel 3 Zapoznanie z inteligentnymi algorytmami eksploracji danych wykorzystywanymi w problematyce Big Data.

4 Wymagania wstępne w zakresie wiedzy, umiejętności i innych kompetencji

1 Podstawowa wiedza z programowania w języku Python.

2 Podstawowa wiedza z metod sztucznej inteligencji i statystyki.

5 Efekty kształcenia

EK1 Wiedza Student rozumie pojęcia związane z pozyskiwaniem i przetwarzaniem dużych zbiorów danych.

EK2 Umiejętności Student wyciąga wnioski w oparciu o wyniki analiz danych.

EK3 Umiejętności Student wykorzystuje zaawansowane narzędzia informatyczne pozwalające na przetwarzanie dużych zbiorów danych typu Big Data

EK4 Kompetencje społeczne Student jest gotów do rozwiązywania problemów z zakresu analizy danych zarówno w ramach pracy indywidualnej jak i grupowej, a także poszukiwania niezbędnej w tym celu wiedzy

6 Treści programowe

Laboratorium komputerowe		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
K1	Metody przekształcania danych w języku Python.	2
K2	Budowa klastra opartego na platformie Apache Hadoop.Implementacja aplikacji w modelu MapReduce.Typy, formaty i funkcje modelu MapReduce.	8
K3	Wprowadzenie do platformy Apache Spark - instalacja i konfiguracja środowiska.Transformacje, akcje i ewaluacja w ramach rozproszonych zbiorów danych RDD.Agregacje, grupowania, złączenia i sortowania danych.Ładowanie i zapisywanie danych - pliki tekstowe, JSON, CSV, pliki sekwencyjne i obiektowe.Praca z danymi strukturalnymi w Spark SQL.Działania liczbowe na RDD.Strumieniowanie danych.	14
K4	Statystyczne systemy uczące się metody klasyfikacji, regresji, klastrowania.	6

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W1	Wprowadzenie w tematykę przetwarzania danych - czwarta rewolucja przemysłowa.Wprowadzenie podstawowych pojęć związanych z przetwarzaniem dużych zbiorów danych.Charakterystyka zbiorów danych typu Big Data - model 4V. Praktyczne aspekty przetwarzania dużych zbiorów danych.Społeczeństwo informacyjne - model DIKW (dane, informacje, wiedza, mądrość). Bazy danych NoSQL. Modele danych w bazach NoSQL.	8

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W2	Wprowadzenie do platformy Hadoop. Model MapReduce.Systemy plików i przepływ danych w Hadoopie. Projekt systemu HDFS.Menadżer klastra YARN. Budowa aplikacji w HadoopMapReduce.	6
W3	Wprowadzenie do platformy Apache Spark.Rozproszone kolekcje obiektów RDD (Resilient Distributed Dataset).Model ETL (extract, transform and load). Partycjonowanie danych.Formaty przechowywania danych: formaty i systemy plików, strukturalne źródła danych, bazy danych.Operacje na zbiorach danych w oparciu o język zapytań Spark SQL.Strumieniowanie danych.Metody klasyfikacji i regresji. Systemy uczące się w MLlib.	16

7 Narzędzia dydaktyczne

N1 Wykłady (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych)

N2 Prezentacje multimedialne

N3 Ćwiczenia laboratoryjne (w przypadku realizacji zajęć w trybie zdalnym z wykorzystaniem stosownych narzędzi teleinformatycznych)

N4 Konsultacje

N5 Platforma MS Teams

8 Obciążenie pracą studenta

Forma aktywności	Średnia liczba godzin na zrealizowanie aktywności
Godziny kontaktowe z nauczycielem akademickim, w tym:	
Godziny wynikające z planu studiów	60
Konsultacje przedmiotowe	10
Egzaminy i zaliczenia w sesji	2
Godziny bez udziału nauczyciela akademickiego wynikające z nakładu pracy studenta, w tym:	
Przygotowanie się do zajęć, w tym studiowanie zalecanej literatury	10
Opracowanie wyników	10
Przygotowanie raportu, projektu, prezentacji, dyskusji	28
Sumaryczna liczba godzin dla przedmiotu wynikająca z całego nakładu pracy studenta	**120**
Sumaryczna liczba punktów ECTS dla przedmiotu	4.00

9 Sposoby oceny

Ocena formująca

F1 Projekt indywidualny zrealizowany na podstawie laboratoriów komputerowych

Ocena podsumowująca

P1 Średnia ważona ocen formujących

P2 Egzamin pisemny

Warunki zaliczenia przedmiotu

W1 Warunkiem zaliczenia przedmiotu jest uzyskanie pozytywnej oceny z każdego efektu kształcenia, pozytywnej oceny z egzaminu oraz pozytywnych wszystkich ocen cząstkowych. Ocena końcowa to średnia arytmetyczna oceny egzaminu oraz średniej ważonej ocen formujących z laboratoriów komputerowych.

W2 Spełnienie kryteriów obecności na zajęciach laboratoryjnych

Kryteria oceny

Efekt kształcenia 1	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0

Na ocenę 3.0	Student potrafi w sposób nieformalny, lecz zrozumiały zdefiniować pojęcia związane z pozyskiwaniem i przetwarzaniem dużych zbiorów danych, takich jak: Big Data, model 4V, model DIKW, RDD, strumieniowanie danych, transformacje danych i in. W ramach egzaminu uzyskuje co najmniej 51% punktów.
Na ocenę 3.5	Student potrafi w sposób formalny i zrozumiały zdefiniować pojęcia związane z pozyskiwaniem i przetwarzaniem dużych zbiorów danych, takich jak: Big Data, model 4V, model DIKW, RDD, strumieniowanie danych, transformacje danych i in. Przedstawione definicje cechują się dopuszczalnymi błędami. W ramach egzaminu uzyskuje co najmniej 61% punktów.
Na ocenę 4.0	Student potrafi w sposób formalny i zrozumiały zdefiniować pojęcia związane z pozyskiwaniem i przetwarzaniem dużych zbiorów danych, takich jak: Big Data, model 4V, model DIKW, RDD, strumieniowanie danych, transformacje danych i in. Przedstawione definicje nie zawierają błędów. Student nie jest w stanie podeprzeć definicji realnymi przykładami obrazującymi omawiane pojęcia. W ramach egzaminu uzyskuje co najmniej 71% punktów.
Na ocenę 4.5	Student potrafi w sposób formalny, zrozumiały i bezbłędny zdefiniować pojęcia związane z pozyskiwaniem i przetwarzaniem dużych zbiorów danych, takich jak: Big Data, model 4V, model DIKW, RDD, strumieniowanie danych, transformacje danych i in. Student podpiera definicje pojedynczymi przykładami z 1-2 dziedzin obrazującymi omawiane pojęcia. W ramach egzaminu uzyskuje co najmniej 81% punktów.
Na ocenę 5.0	Student potrafi w sposób formalny, zrozumiały i bezbłędny zdefiniować pojęcia związane z pozyskiwaniem i przetwarzaniem dużych zbiorów danych, takich jak: Big Data, model 4V, model DIKW, RDD, strumieniowanie danych, transformacje danych i in. Student podpiera definicje wieloma przykładami z wielu dziedzin obrazującymi omawiane pojęcia. W ramach egzaminu uzyskuje co najmniej 91% punktów.
Efekt kształcenia 2	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0
Na ocenę 3.0	Student na podstawie zastosowania prostych metod statystycznych wysnuwa wnioski płynące z analizy prostych zbiorów danych. W ramach projektów uzyskuje co najmniej 51% punktów.
Na ocenę 3.5	Student stosuje podstawowe metody statystyczne pozwalające na dokonanie podstawowej analizy danych. Na podstawie wyników wysnuwa szersze wnioski i nieskutecznie szuka uzasadnienia otrzymanych wyników w trudniejszych zagadnieniach z zakresu analizy danych. W ramach projektów uzyskuje co najmniej 61% punktów.
Na ocenę 4.0	Student jest w stanie dokonać wielowymiarowej analizy dużych zbiorów danych i wyciągnąć płynące z niej wnioski. Student podejmuje nieudaną próbę uzasadnienia otrzymanych wyników w trudniejszych zagadnieniach z zakresu analizy danych. W ramach projektów uzyskuje co najmniej 71% punktów.
Na ocenę 4.5	Student jest w stanie dokonać wielowymiarowej analizy dużych zbiorów danych i wyciągnąć płynące z niej wnioski. Student podejmuje udaną próbę uzasadnienia otrzymanych wyników w trudniejszych zagadnieniach z zakresu analizy danych. W ramach projektów uzyskuje co najmniej 81% punktów.

Na ocenę 5.0	Student korzystając z metod sztucznej inteligencji jest w stanie dokładnie opisać relacje zachodzące w danych i na tej podstawie sformułować obszerne wnioski. W sposób celny określa przyczyny stojące za otrzymanymi wynikami w trudniejszych zagadnieniach z zakresu analizy danych. W ramach projektów uzyskuje co najmniej 91% punktów.
Efekt kształcenia 3	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0.
Na ocenę 3.0	Student z pomocą nauczyciela wykonuje prostą analizę danych stosując podstawowe narzędzia statystyczne. W ramach projektów uzyskuje co najmniej 51% punktów.
Na ocenę 3.5	Student samodzielnie wykonuje prostą analizę danych stosując podstawowe narzędzia statystyczne. W ramach projektów uzyskuje co najmniej 61% punktów.
Na ocenę 4.0	Student z pomocą nauczyciela wykorzystuje podstawowe metody inteligentne w celu wydobycia wiedzy z przetwarzanych zbiorów danych. W ramach projektów uzyskuje co najmniej 71% punktów.
Na ocenę 4.5	Student samodzielnie wykorzystuje podstawowe metody inteligentne w celu wydobycia wiedzy z przetwarzanych zbiorów danych. W ramach projektów uzyskuje co najmniej 81% punktów.
Na ocenę 5.0	Student samodzielnie lub z niewielką pomocą nauczyciela wykorzystuje zaawansowane metody inteligentne w procesie wielowymiarowej analizy dużych zbiorów danych. W ramach projektów uzyskuje co najmniej 91% punktów.
Efekt kształcenia 4	
Na ocenę 2.0	Student nie spełnia warunków określonych dla oceny 3.0
Na ocenę 3.0	Student wykonuje powierzone zadania indywidualnie - rezygnuje ze współpracy grupowej kosztem jakości rozwiązania. Prace studenta cechuje dopuszczalna niedbałość. W ramach zaliczeń uzyskuje co najmniej 51% punktów.
Na ocenę 3.5	Student wykonuje powierzone zadania indywidualnie - rezygnuje ze współpracy grupowej kosztem zakresu rozwiązania. Jego prace są wykonane w sposób staranny. W ramach zaliczeń uzyskuje co najmniej 61% punktów.
Na ocenę 4.0	Student wykonuje powierzone zadania indywidualnie, a także stara się nawiązać współpracę grupową. Współpraca grupowa nie przynosi znaczących zysków widocznych w wypracowanych rozwiązaniach. Prace studenta zawierają drobne błędy. W ramach zaliczeń uzyskuje co najmniej 71% punktów.
Na ocenę 4.5	Student wykonuje powierzone zadania indywidualnie, a także stara się nawiązać współpracę grupową. Student dostrzega zyski płynące ze współpracy grupowej, jednak jego zbyt małe zaangażowanie powoduje drobne błędy w realizowanych projektach. W ramach zaliczeń uzyskuje co najmniej 81% punktów.
Na ocenę 5.0	Student potrafi rozwiązywać problemy indywidualnie jak i grupowo; dostrzega zyski płynące ze współpracy grupowej. Prace studenta cechuje wysoka dbałość o detale. W ramach zaliczeń uzyskuje co najmniej 91% punktów.

10 Macierz realizacji przedmiotu

Efekt kształcenia	Odniesienie danego efektu do szczegółowych efektów zdefiniowanych dla programu	Cele przedmiotu	Treści programowe	Narzędzia dydaktyczne	Sposoby oceny
EK1	K_W08 K_W11	Cel 1	W1 W2 W3	N1 N2 N3 N4 N5	P1 P2
EK2	K_U02 K_U10 K_U11 K_U13 K_U19 K_U22	Cel 2 Cel 3	K1 K2 K3 K4 W2 W3	N1 N2 N3 N4 N5	F1 P1 P2
EK3	K_U16 K_U19 K_U21 K_U22	Cel 2 Cel 3	K1 K2 K3 K4 W2 W3	N3 N4	F1 P1
EK4	K_K01 K_K03 K_K04 K_K06	Cel 1 Cel 2 Cel 3	K2 K3 K4 W1 W2 W3	N3 N4 N5	F1 P1

11 Wykaz literatury

Literatura podstawowa

[1] H. Karau et. al. — *Poznajemy Sparka. Błyskawiczna analiza danych*, Warszawa, 2016, PWN

[2] N. Marz, J. Warren — *Big Data. Najlepsze praktyki budowy skalowalnych systemów obsługi danych w czasie rzeczywistym*, Gliwice, 2016, Helion

[3] T. White — *Hadoop. Komplety przewodnik. Analiza i przechowywanie danych*, Gliwice, 2016, Helion

Literatura uzupełniająca

[1] E. Matthes — *Python. Instrukcje dla programisty*, Gliwice, 2016, Helion

[2] A. Boschetti, L. Massaron — *Python. Podstawy nauki o danych*, Gliwice, 2017, Helion

[3] S. Raschka — *Python. Uczenie maszynowe*, Gliwice, 2018, Helion

12 Informacje o nauczycielach akademickich

Osoba odpowiedzialna za kartę

dr inż. Daniel Grzonka (kontakt: daniel.grzonka@pk.edu.pl)

Osoby prowadzące przedmiot

1 dr inż. Daniel Grzonka (kontakt: daniel.grzonka@pk.edu.pl)

13 Zatwierdzenie karty przedmiotu do realizacji

(miejscowość, data) (odpowiedzialny za przedmiot) (dziekan)

Przyjmuję do realizacji (data i podpisy osób prowadzących przedmiot)

.....