**Nazwa przedmiotu:**

Przetwarzanie strumieni danych i data science

**Koordynator przedmiotu:**

Rajmund Kożuszek

**Status przedmiotu:**

Fakultatywny dowolnego wyboru

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Przedmioty techniczne - zaawansowane

**Kod przedmiotu:**

PSD

**Semestr nominalny:**

3 / rok ak. 2021/2022

**Liczba punktów ECTS:**

4

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

1. liczba godzin kontaktowych – 62 godz., w tym
obecność na wykładach 30 godz.,
obecność na warsztatach 6 godz.,
spotkania projektowe 24h
obecność na egzaminie godz. 2
2. praca własna studenta – 58 godz., w tym
prace projektowe 20 godz.,
przygotowanie do prac projektowych 21 godz.
Przygotowanie do warsztatów 5 godz,
przygotowanie do egzaminu 12 godz.
Łączny nakład pracy studenta wynosi 120 godz., co odpowiada 4 pkt. ECTS.

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

2,07 pkt. ECTS, co odpowiada 62 godz. kontaktowym

**Język prowadzenia zajęć:**

polski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

1,73 pkt. ECTS, co odpowiada 20 godz. prac projektowych plus 21 godz. przygotowań do prac projektowych plus 6 godz. warsztatów i 5 godz. przygotowania do warsztatów

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład:  | 30h |
| Ćwiczenia:  | 0h |
| Laboratorium:  | 0h |
| Projekt:  | 30h |
| Lekcje komputerowe:  | 0h |

**Wymagania wstępne:**

Wskazana znajomość podstaw statystyki i sztucznej inteligencji.

**Limit liczby studentów:**

30

**Cel przedmiotu:**

Celem przedmiotu jest zapoznanie studentów z metodami przetwarzania danych o charakterze strumieniowym w czasie prawie rzeczywistym oraz metodami analizy tych danych z wykorzystaniem narzędzi data science. Analiza i przetwarzanie tego typu danych są jednym z podstawowych mechanizmów rozwiązań Big Data.

**Treści kształcenia:**

Wykład:

W1(2h): Wprowadzenie do data science (2h)
W2 (2h) Przykłady zastosowań strumieniowego przetwarzania danych np. IoT, detekcja anomalii, predictive maintenance, monitorowanie procesów, zdalny monitoring pacjentów.
W3 (2h) Przegląd zagadnień przetwarzania danych strumieniowych typu BigData.
W4 (2h) Podstawy algorytmów przetwarzania danych strumieniowych.
W5 (2h) Podstawowe platformy programistyczne przeznaczone do rozproszonego składowania i przetwarzania wielkich zbiorów danych.
W6, W7 (4h) Python jako narzędzie do przetwarzania i analizy danych strumieniowych.
W8, W9 (4h) Przykłady narzędzi do budowy systemów informacyjnych bazujących na danych strumieniowych:
• NiFi- organizacja procesu przepływu danych,
• Kafka – narzędzie do przesyłania komunikatów (ang. message) pomiędzy aplikacjami w systemach rozproszonych,
• STORM do przetwarzania strumieni danych w czasie rzeczywistym,
• SPARK do przetwarzania strumieni danych ( w trybie batch processing),
• Bazy NoSQL na przykładzie bazy Cassandra i MongoDB,
• pakiet stream systemu R.
W10, W11 (4h) Algorytmy statystyczne dla danych strumieniowych
W12, W13 (4h) Algorytmy uczenia maszynowego dla danych strumieniowych:
• algorytmy uczenia maszynowego (klasyfikacji) dla danych strumieniowych,
• algorytmy uczenia maszynowego (grupowania) dla danych strumieniowych,
• algorytmy detekcji anomalii i sytuacji nietypowych.
W14, W15 (4h) Procesy budowy systemów informacyjnych wykorzystujących dane strumieniowe
Warsztaty – zajęcia zintegrowane:
W ramach przedmiotu przewiduje się 3 spotkania warsztatowe, których celem będzie zdobycie umiejętności posługiwania się wybranymi narzędziami oraz umiejętności doboru narzędzi do postawionych zadań projektowych.
Projekt:
W ramach przedmiotu przewiduje się realizację projektu w grupach 2-3 osobowych. Projekt będzie poświęcony rozwiązaniu postawionego zadania praktycznego z wykorzystaniem narzędzi i metod przetwarzania danych strumieniowych.

**Metody oceny:**

Wykład będzie prowadzony w tradycyjnej formie: 15 spotkań po 2 godziny. Na wybranych wykładach wprowadzone zostaną techniki mające na celu sprawdzenie przyswajanej wiedzy (testy sprawdzające).
Zajęcia zintegrowane będą prowadzone metodą interaktywną, warsztatową, studenci będą mieli do wykonania pewne zadania, po krótkim wprowadzenie przez prowadzącego. Zajęcia warsztatowe będą miały na celu nabywanie umiejętności podstawowych dotyczących posługiwania się wybranymi metodami i narzędziami przetwarzania danych strumieniowych. Dodatkowo w trakcie tych zajęć zostanie przekazana wiedza z zakresu możliwości zastosowań praktycznych.
Projekt będzie prowadzony w zespołach 2-3 osobowych, dotyczyć będzie identyfikacji jednego z problemów współczesnej cywilizacji (np. problemy inteligentnych miast, zdalnego monitoringu pacjentów), jego definicji, zaprojektowania oraz implementacji wybranych elementów, korzystając z wiedzy i umiejętności uzyskanej podczas zajęć wykładowych i zintegrowanych. Zajęcia projektowe będę powiązane z zajęciami zintegrowanymi. Zakłada się, że studenci będą spotykać się regularnie ze swoim opiekunem projektu, sprawozdawać mu wyniki i napotkane problemy. Ponadto, zakładamy cztery kamienie milowe projektu (identyfikacja, projekt, implementacja i integracja), z których będą przygotowywane raporty cząstkowe.

**Egzamin:**

tak

**Literatura:**

1. Kolajo, Taiwo & Daramola, Olawande & Adebiyi, Ayodele. (2019). Big data stream analysis: a systematic literature review. Journal of Big Data. 6. 47. 10.1186/s40537-019-0210-7.
2. Arun Kejariwal, Sanjeev Kulkarni and Karthik Ramasamy, (2017). Real Time Analytics: Algorithms and Systems, CoRR, abs/1708.02621, http://arxiv.org/abs/1708.02621.
3. Gama, João, (2010). Knowledge Discovery from Data Streams. Data Mining and Knowledge Discovery. Knowledge Discovery from Data Streams. 10.1201/EBK1439826119.
4. J. Granat, J. M. Batalla, C. X. Mavromoustakis and G. Mastorakis, (2019), "Big data analytics for event detection in the IoT-multicriteria approach," in IEEE Internet of Things Journal.
5. Tom White, (2015). Hadoop: The Definitive Guide, 4th Edition, O'Reilly Media, Inc., ISBN: 9781491901687.
6. Gabriel Eisbruch, Jonathan Leibiusky, Dario Simonassi, (2012) Getting Started with Storm. Continuous Streaming Computation with Twitter's Cluster Technology O'Reilly Media.
7. Neha Narkhede, Gwen Shapira, and Todd Palino, (2017). Kafka: The Definitive Guide Real-Time Data and Stream Processing at Scale (1st. ed.). O’Reilly Media, Inc.

**Witryna www przedmiotu:**

https://usosweb.usos.pw.edu.pl/kontroler.php?\_action=katalog2/przedmioty/pokazPrzedmiot&prz\_kod=103A-INISY-MSP-PSD

**Uwagi:**

(-)

## Charakterystyki przedmiotowe