



## KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Zaawansowana eksploracja danych [S2Inf1-TPD>ZED]

### Przedmiot

Kierunek studiów  
Informatyka

Rok/Semestr  
1/2

Studia w zakresie (specjalność)  
Technologie przetwarzania danych

Profil studiów  
ogólnoakademicki

Poziom studiów  
drugiego stopnia

Język oferowanego przedmiotu  
polski

Forma studiów  
stacjonarne

Wymagalność  
obligatoryjny

### Liczba godzin

Wykład  
30

Laboratorium  
30

Inne  
0

Ćwiczenia  
0

Projekty/seminaria  
0

### Liczba punktów ECTS

5,00

### Koordynatorzy

prof. dr hab. inż. Tadeusz Morzy

dr hab. inż. Dariusz Brzeziński prof. PP  
dariusz.brzezinski@put.poznan.pl

### Wykładowcy

### Wymagania wstępne

Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę w zakresie podstawowych pojęć, technik i algorytmów eksploracji danych. Ponadto przydatna jest wiedza ze statystyki matematycznej i rachunku prawdopodobieństwa i podstawowa wiedza w zakresie teorii grafów. Student powinien posiadać umiejętność rozwiązywania podstawowych problemów z eksploracji danych oraz umiejętność pozyskiwania informacji ze wskazanych źródeł. Powinien również rozumieć konieczność poszerzania swoich kompetencji i mieć gotowość do podjęcia współpracy w ramach zespołu. Ponadto w zakresie kompetencji społecznych student musi prezentować takie postawy jak uczciwość, odpowiedzialność, wytrwałość, ciekawość poznawcza, kreatywność, kultura osobista, szacunek dla innych ludzi.

## Cel przedmiotu

1. Przekazanie studentom podstawowej wiedzy na temat zaawansowanych algorytmów eksploracji bardziej złożonych reprezentacji danych oraz realizacji różnych etapów procesu odkrywania wiedzy z danych (w tym przetwarzania wstępnego danych i oceny wyników eksploracji). 2. Rozwijanie u studentów umiejętności rozwiązywania problemów w powyższej dziedzinie (poprzez studia przypadków odnoszące się do różnorodnych reprezentacji danych i zadań uczenia maszynowego). 3. Nabycie powyższych umiejętności poprzez rozwiązywanie na ćwiczeniach laboratoryjnych praktycznych zadań klasyfikacji nadzorowanej, nienadzorowanej, predykcji z danych zależnych od czasu oraz eksploracji danych społecznościowych. 4. Przekazanie wiedzy na temat technik i algorytmów umożliwiających odkrywanie wiedzy i wzorców, ze szczególnym naciskiem na dane relacyjne i tekstowe. 5. Kształtowanie u studentów umiejętności przeprowadzania powtarzalnych eksperymentów z danymi dotyczącymi powyższych zadań przy wykorzystaniu języków programowania R i python.

## Przedmiotowe efekty uczenia się

### Wiedza:

student ma zaawansowaną wiedzę o osiągnięciach w eksploracji danych, zwłaszcza w odniesieniu do złożonych reprezentacji danych.

ma uporządkowaną, podbudowaną teoretycznie wiedzę ogólną w zakresie eksploracji danych – także wobec złożonych typów i modeli reprezentacji danych.

ma wiedzę związaną z takimi zagadnieniami jak eksploracja strumieni danych, metody wizualnej eksploracji danych, przetwarzanie języka naturalnego, rozpraszanie eksploracji danych na wiele maszyn obliczeniowych, wybranych klasyfikatorów dla strumieni danych i adaptujących się do zmiennych środowisk.

ma podbudowaną teoretycznie szczegółową wiedzę związaną z wybranymi zagadnieniami, takimi jak: metody przetwarzania wstępnego danych, budowa modeli predykcji zmiennej liczbowej (regresja, sieci neuronowe), metody wyboru i oceny klasyfikatorów oraz predykcji, ocena algorytmów skupień.

zna podstawowe metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań z zastosowaniem algorytmów eksploracji danych, zna zagadnienia przetwarzania języka naturalnego i danych relacyjnych, potrafi wdrożyć opracowywane modele do projektowanych systemów informatycznych.

### Umiejętności:

student potrafi — przy formułowaniu i rozwiązywaniu zadań inżynierskich — integrować wiedzę z różnych obszarów informatyki związanych z pozyskiwaniem danych z różnych źródeł, ich przetwarzaniem wstępnym oraz eksploracją, oceną uzyskanych wzorców i zastosowaniem znalezionych wzorców.

potrafi określić przydatność nowych algorytmów eksploracji danych, poprzez lekturę literatury naukowej i popularnonaukowej.

potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody klasyfikacji nadzorowanej, predykcji zmiennej liczbowej, grupowania danych oraz przetwarzania języka naturalnego.

potrafi pozyskiwać informacje nt. eksploracji danych z literatury, baz danych oraz innych źródeł (w języku ojczystym i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz opinie.

potrafi dokonać krytycznej analizy istniejących metod eksploracji danych oraz zaproponować ich ulepszenia.

potrafi wykonywać proste eksperymenty badawcze, wykorzystując do tego języki programowania R i python, a także komunikować wyniki eksperymentów za pomocą powtarzalnych raportów (np. knitr i jupyter notebook).

potrafi ocenić zalety i ograniczenia wybranych algorytmów eksploracji danych i ich implementacji w zależności od charakterystyki zadania.

### Kompetencje społeczne:

student rozumie istotność wykorzystywania najnowszej wiedzy z eksploracji danych i uczenia maszynowego w rozwiązywaniu problemów badawczych i praktycznych.

rozumie, że przy tworzeniu inteligentnych systemów wykorzystujących techniki eksploracji danych i podczas adaptacji tych technik do środowiska nabyta wiedza i umiejętności szczególnie szybko się zmieniają i wymagają od osoby dalszego kształcenia się z uwagi na dynamiczny rozwój dziedziny.

## Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

a) w zakresie wykładów:

- na podstawie odpowiedzi na pytania dotyczące materiału omówionego na poprzednich wykładach oraz ćwiczeń realizowanych przy tablicy

b) w zakresie laboratoriów:

- na podstawie oceny bieżącego postępu realizacji zadań – ćwiczeń oraz projektów
- kartkówek przed niektórymi zajęciami dydaktycznymi

Ocena podsumowująca:

a) w zakresie wykładów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenę wiedzy i umiejętności wykazanych na otwartym kolokwium pisemnym o charakterze problemowym (student może korzystać z dowolnych materiałów dydaktycznych). Kolokwium składa się z 5-6 zadań problemowych, za które można uzyskać 10 pkt. Łącznie można uzyskać od 50-60 pkt. Zaliczenie na ocenę 3.0 wymaga uzyskania 50% maksymalnej liczby punktów.

- omówienie wyników egzaminu

b) w zakresie laboratoriów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenę sprawozdania z realizacji dwóch projektów

- ocenę z tworzenia prostej aplikacji internetowej korzystającej z algorytmów eksploracji danych

Uzyskiwanie punktów dodatkowych za aktywność podczas zajęć, a szczególnie za:

- omówienia dodatkowych aspektów zagadnienia

- uwagi związane z udoskonaleniem materiałów dydaktycznych

- udział w międzynarodowych konkursach algorytmicznych

## Treści programowe

Charakterystyka procesu odkrywania wiedzy w bazach danych. Wstępne przetwarzanie danych.

Eksploracja sieci Web: eksploracja połączeń. Algorytmy rankingu stron. Eksploracja tekstu: modele i algorytmy. Systemy rekomendacyjne. Strumienie danych.

## Tematyka zajęć

Wykład:

Charakterystyka procesu odkrywania wiedzy w bazach danych. Główne metody przetwarzania wstępnego danych (w szczególności wykrywanie sytuacji konfliktowych podczas połączenia różnych źródeł, oczyszczenie danych z błędów, uwzględnianie niezdefiniowanych wartości atrybutów), redukcja wymiarowości danych (selekcja cech, tworzenie nowych cech, metody projekcji do przestrzeni niskowymiarowych, SVD), transformacje oraz metody dyskretyzacji. Eksploracja sieci Web: eksploracja połączeń. Algorytmy rankingu stron (PageRank, HITS). Eksploracja tekstu: modele i algorytmy. Systemy rekomendacyjne.

Laboratoria:

Przebieg ćwiczeń laboratoryjnych obejmuje naukę eksploracji danych z wykorzystaniem języków R i Python. Ponadto dwa projekty (studia przypadków), mają na celu ukierunkować studentów na aspekty praktyczne realizacji różnych etapów odkrywania wiedzy z wybranych zbiorów danych. W trakcie ćwiczeń laboratoryjnych studenci przechodzą kurs języka R i poznają pakiety R przydatne do rozwiązywania problemów regresji, klasyfikacji, analizy skupień, wizualizacji i wstępnego przetwarzania danych. Studenci wykonują również w R aplikację internetową pozwalającą stworzyć prototyp produktu wykorzystującego algorytmy eksploracji danych. Studenci zapoznają się także z bibliotekami uczenia maszynowego dla języka programowania Python, ze szczególnym naciskiem na biblioteki wspomagające przetwarzanie języka naturalnego (nltk, gensim). Wybrane laboratoria są także poświęcone technikom wizualnej oceny i eksploracji danych oraz tworzenia powtarzalnych eksperymentów za pomocą bibliotek caret, knitr (R) oraz scikit-learn, jupyter notebook (Python). Studenci poznają również sposoby przetwarzania większych zbiorów danych poprzez zrównoleglanie lub rozpraszanie obliczeń. Część wymienionych wyżej treści programowych realizowana jest w ramach pracy własnej studenta.

## Metody dydaktyczne

Wykład: prezentacja multimedialna, prezentacja ilustrowana przykładami podawanymi na tablicy, rozwiązywanie prostych zadań, demonstracja użycia wybranego oprogramowania

Ćwiczenia laboratoryjne: prezentacja multimedialna, rozwiązywanie zadań, wykonywanie

eksperymentów, dyskusja, studium przypadków, gry obliczeniowe i konkursy programistyczne, kartkówki

## Literatura

### Podstawowa

A. Rajaraman, J. Lescovec, J.D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2014

Han J., Kamber M., Data Mining: Concepts and techniques, San Francisco, Morgan Kaufmann, 2000.

B. Liu, Web Data Mining: Exporing Hyperlinks, Contents, and Usage Data, Springer, 2015

### Uzupełniająca

Tufte, Edward R. The visual display of quantitative information. Vol. 2. Cheshire, CT: Graphics press, 2001.

Wickham, Hadley. Tidy data. Journal of Statistical Software 59.10 (2014): 1-23.

Ng, Andrew. Machine Learning Yearning, 2019.

## Bilans nakładu pracy przeciętnego studenta

|                                                                                                                                                      | Godzin | ECTS |
|------------------------------------------------------------------------------------------------------------------------------------------------------|--------|------|
| Łączny nakład pracy                                                                                                                                  | 125    | 5,00 |
| Zajęcia wymagające bezpośredniego kontaktu z nauczycielem                                                                                            | 60     | 2,50 |
| Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwium/egzaminu, wykonanie projektu) | 65     | 2,50 |