

KARTA OPISU MODUŁU KSZTAŁCENIA		
Nazwa modułu/przedmiotu Zaawansowana eksploracja danych		Kod 1010512321010513981
Kierunek studiów Informatyka	Profil kształcenia (ogólnoakademicki, praktyczny) ogólnoakademicki	Rok / Semestr 1 / 2
Ścieżka obieralności/specjalność Technologie przetwarzania danych	Przedmiot oferowany w języku: polski	Kurs (obligatoryjny/obieralny) obligatoryjny
Stopień studiów: II stopień	Forma studiów (stacjonarna/niestacjonarna) stacjonarna	
Godziny Wykłady: 30 Ćwiczenia: - Laboratoria: 30 Projekty/seminaria: -		Liczba punktów 6
Status przedmiotu w programie studiów (podstawowy, kierunkowy, inny) kierunkowy		(ogólnouczelniany, z innego kierunku) z danego kierunku
Obszar(y) kształcenia i dziedzina(y) nauki i sztuki nauki techniczne nauki techniczne		Podział ECTS (liczba i %) 6 100% 6 100%
Odpowiedzialny za przedmiot / wykładowca:		
prof. dr hab. inż. Tadeusz Morzy email: Tadeusz.Morzy@put.poznan.pl tel. 61 6652906 Informatyki ul. Piotrowo 2, 60-965 Poznań		dr inż. Dariusz Brzeziński email: Dariusz.Brzezinski@cs.put.poznan.pl tel. 61 6653057 Informatyki ul. Piotrowo 2, 60-965 Poznań
Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:		
1	Wiedza:	Efekty kształcenia ze studiów I stopnia: K1st_W1-8, K1st_U2-14 weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl . 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.
2	Umiejętności:	Student powinien posiadać umiejętność rozwiązywania podstawowych problemów z eksploracji danych oraz umiejętność pozyskiwania informacji ze wskazanych źródeł.
3	Kompetencje społeczne	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 tzw. case studies 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 dot. powyższych zadań przy wykorzystaniu różnorodnych języków programowania (R, python).		
Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia		
Wiedza:		

<p>1. ma zaawansowaną wiedzę o osiągnięciach w eksploracji danych, zwłaszcza w odniesieniu do złożonych reprezentacji danych - [K2st_W1]</p> <p>2. ma uporządkowaną, podbudowaną teoretycznie wiedzę ogólną w zakresie eksploracji danych ? także wobec złożonych typów i modeli reprezentacji danych - [K2st_W2]</p> <p>3. 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 - [K2st_W3]</p> <p>4. ma wiedzę związaną z takimi zagadnieniami jak eksploracja strumieni danych, metody wizualnej eksploracji danych, przetwarzanie języka naturalnego, rozpraszania eksploracji danych na wiele maszyn obliczeniowych, wybranych klasyfikatorów dla strumieni danych i adaptujących się do zmiennych środowisk - [K2st_W5]</p> <p>5. 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 - [K2st_W6]</p>
<p>Umiejętności:</p> <p>1. 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 - [K2st_U1]</p> <p>2. potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody klasyfikacji nadzorowanej, predykcji zmiennej liczbowej, grupowania danych oraz przetwarzania języka naturalnego - [K2st_U4]</p> <p>3. 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 - [K2st_U5]</p> <p>4. potrafi określić przydatność nowych algorytmów eksploracji danych, poprzez lekturę literatury naukowej i popularnonaukowej - [K2st_U6]</p> <p>5. potrafi dokonać krytycznej analizy istniejących metod eksploracji danych oraz zaproponować ich ulepszenia - [K2st_U8]</p> <p>6. potrafi ocenić zalety i ograniczenia wybranych algorytmów eksploracji danych i ich implementacji w zależności od charakterystyki zadania - [K2st_U9]</p> <p>7. 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) - [K2st_U10]</p>
<p>Kompetencje społeczne:</p> <p>1. 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 - [K2st_K1]</p> <p>2. rozumie istotność wykorzystywania najnowszej wiedzy z eksploracji danych i uczenia maszynowego w rozwiązywaniu problemów badawczych i praktycznych - [K2st_K2]</p>

Sposoby sprawdzenia efektów kształcenia

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 (związanych z rozwiązaniem tzw. case studies) przygotowane w postaci instrukcji udostępnionych studentom, przygotowanych samodzielnych raportów z rozwiązania powyższych zadań, wykonywanie ćwiczeń przy użyciu różnych środowisk i narzędzi informatycznych na podstawie pisemnych instrukcji

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:

- kartkówki przed niektórymi zajęciami dydaktycznymi,

- ocenę sprawozdania z realizacji case study;

- 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

Program wykładu obejmuje następujące zagadnienia:

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.

Przebieg ćwiczeń laboratoryjnych obejmuje naukę eksploracji danych z wykorzystaniem języków R i Python. Ponadto dwa studia przypadków, mają na celu ukierunkować studentów na aspekty praktyczne realizacji różnych etapów w odkrywaniu 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 (AWS, Apache Spark).

Cześć wymienionych wyżej treści programowych realizowana jest w ramach pracy własnej studenta.

Metody dydaktyczne:

1. wykład: prezentacja multimedialna, prezentacja ilustrowana przykładami podawanymi na tablicy, rozwiązywanie prostych zadań, demonstracja użycia wybranego oprogramowania
2. ćwiczenia laboratoryjne: rozwiązywanie zadań, wykonywanie eksperymentów, dyskusja, studium przypadków, gry obliczeniowe i konkursy programistyczne, kartkówki

Literatura podstawowa:

1. A. Rajaraman, J. Lescovec, J.D. Ullman, Mining of Massive Datasets, Cambridge University Press, 2014
2. Han J., Kamber M., Data Mining: Concepts and techniques, San Francisco, Morgan Kaufmann, 2000.
3. B. Liu, Web Data Mining: Exporing Hyperlinks, Contents, and Usage Data, Springer, 2015

Literatura uzupełniająca:

1. Klösgen W., Żytkow J.M. (red.), Handbook of Data Mining and Knowledge Discovery. Oxford Press, 2002.
2. Torgo, L., Data Mining with R, CRC Press, 2011.
3. Spector, P., Data Manipulation with R, Springer 2008.
4. Yau, N., Visualize This: The FlowingData Guide to Design, Visualization and Statistics, Wiley, 2011.
5. Aggarwal, Ch., Social Network Data Analytics, Springer, 2011.
6. Russel, M., Mining the Social Web: Analyzing Data from Facebook, Twitter, LinkedIn, and Other Social Media Sites, O'Reilly Media, 2011.
7. Janert, P., Data Analysis with Open Source Tools, O'Reilly Media 2010.
8. Conway, D., Machine Learning for Hackers, O'Reilly Media 2012.

Bilans nakładu pracy przeciętnego studenta

Czynność	Czas (godz.)
1. Udział w zajęciach laboratoryjnych	30
2. Udział w wykładach	30
3. Udział w konsultacjach związanych z realizacją ćwiczeń laboratoryjnych (częściowo realizowane drogą elektroniczną)	2 8
4. Przygotowanie do ćwiczeń laboratoryjnych	10
5. Przygotowanie do egzaminu i obecność na kolokwium zaliczeniowym: 8 godz. + 2 godz.	5
6. Zapoznanie się ze wskazaną literaturą (50 stron)	20
7. Pisanie i testowanie programów przygotowywanych na konkursy algorytmiczne	15
8. Wypełnienie quizów i realizacja zadań samodzielnych do poszczególnych zajęć laboratoryjnych	30
9. Przygotowanie case study i powtarzalnego raportu z analizy danych	

Obciążenie pracą studenta

forma aktywności	godzin	ECTS
------------------	--------	------

Łączny nakład pracy	150	6
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	64	2
Zajęcia o charakterze praktycznym	103	4