KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu	Kod
Eksploracja dużych wolumenów danych	**1010512321010519283**

Kierunek studiów	Profil kształcenia (ogólnoakademicki, praktyczny)	Rok / Semestr
Informatyka	**ogólnoakademicki**	**1 / 2**

Ścieżka obieralności/specjalność	Przedmiot oferowany w języku:	Kurs (obligatoryjny/obieralny)
Technologie wytwarzania oprogramowania	**polski**	**obligatoryjny**

Stopień studiów:	**Forma studiów** (stacjonarna/niestacjonarna)
II stopień	**stacjonarna**

Godziny	Liczba punktów
Wykłady: **30** Ćwiczenia: **-** Laboratoria: **30** Projekty/seminaria: **-**	**5**

Status przedmiotu w programie studiów (podstawowy, kierunkowy, inny)	(ogólnouczelniany, z innego kierunku)
kierunkowy	**z danego kierunku**

Obszar(y) kształcenia i dziedzina(y) nauki i sztuki	**Podział ECTS (liczba i %)**
nauki techniczne	**5 100%**

Odpowiedzialny za przedmiot / wykładowca:

dr inż. Krzysztof Dembczyński
email: krzysztof.dembczynski@cs.put.poznan.pl
tel. 61 6652936
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań

Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

1	**Wiedza:**	Efekty kształcenia ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_W1-2, K_W4, K_W6-15, weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl Przedmiot wymaga w szczególności dobrego przygotowania w ramach programowania w języku Java, podstawowej wiedzy z zakresu systemów baz danych oraz statystyki i analizy danych.
2	**Umiejętności:**	Efekty kształcenia ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_U1-2, K_U4, K_U7-8, K_U14-20, K_U22-23, K_U26, weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl
3	**Kompetencje społeczne**	Efekty kształcenia ze studiów I stopnia zdefiniowane w Uchwale Senatu PP, a szczególnie efekty K_K1-9, weryfikowane w procesie rekrutacji na studia 2 stopnia ? efekty te prezentowane są w serwisie internetowym wydziału www.fc.put.poznan.pl 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:

Przekazanie studentom podstawowej wiedzy w zakresie eksploracji masywnych danych (bardzo dużych zbiorów danych), a dokładniej podstawowych metod organizacji, dostępu i przetwarzania masywnych danych, oraz efektywnych algorytmów eksploracji masywnych danych.

Rozwijanie u studentów umiejętności rozwiązywania problemów dotyczących zarządzania, dostępu i przetwarzania masywnych danych oraz ich eksploracji.

Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia

Wiedza:

1. Ma podbudowaną teoretycznie szczegółową wiedzę związaną z wybranymi zagadnieniami, takimi jak: efektywna organizacja, dostęp i przetwarzanie masywnych danych w technologii MapReduce, dokładne i przybliżone metody wyszukiwanie sąsiadów, eksploracja strumieni danych, metody klasyfikacji i regresji dużej skali oraz systemów rekomendacyjnych. - [K_W5]

2. Ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach w informatyce związanych z eksplozją danych i technologiami eksploracji masywnych zbiorów danych. - [K_W6]

3. Ma podstawową wiedzę o cyklu życia systemów eksploracji masywnych danych. - [K_W7]

4. Zna podstawowe metody, techniki i narzędzia stosowane do eksploracji masywnych danych. - [K_W8]

Umiejętności:

1. Potrafi pozyskiwać informacje z literatury (w języku ojczystym i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz formułować i wyczerpująco uzasadniać opinie. - [K_U1]

2. Potrafi określić kierunki dalszego uczenia się i zrealizować proces samokształcenia. - [K_U5]

3. Potrafi planować i przeprowadzać eksperymenty, w tym pomiary i symulacje komputerowe, interpretować uzyskane wyniki i wyciągać wnioski. - [K_U8]

4. Potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody analityczne oraz eksperymentalne. - [K_U9]

5. Potrafi - przy formułowaniu i rozwiązywaniu zadań inżynierskich - integrować wiedzę z różnych obszarów informatyki oraz zastosować podejście systemowe, uwzględniające także aspekty pozatechniczne. - [K_U10]

6. Potrafi formułować i testować hipotezy związane z problemami inżynierskimi i prostymi problemami badawczymi w zakresie eksploracji masywnych danych. - [K_U12]

7. Potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć (metod i narzędzi) oraz nowych produktów informatycznych w zakresie eksploracji masywnych danych. - [K_U13]

8. Potrafi - stosując m.in. koncepcyjnie nowe metody - rozwiązywać złożone zadania informatyczne, w tym zadania nietypowe oraz zadania zawierające komponent badawczy dotyczące eksploracji masywnych danych - [K_U25]

9. Potrafi odpowiednio zorganizować masywne zbiory danych i przetwarzać je za pomocą technologii MapReduce. - []

10. Potrafi zaimplementować podstawowe algorytmy eksploracji danych w środowisku Java, takie jak wyszukiwanie najbliższych sąsiadów, proste algorytmy klasyfikujące i rekomendacyjne. - []

Kompetencje społeczne:

1. Rozumie, że w informatyce, a zwłaszcza w eksploracji masywnych danych, wiedza, technologie i umiejętności bardzo szybko stają się przestarzałe. - [K_K1]

2. Zna możliwości dalszego dokształcania się w zakresie eksploracji masywnych danych. - [K_K3]

3. Zna przykłady i rozumie przyczyny wadliwie działających systemów informatycznych, które doprowadziły do poważnych strat finansowych, społecznych lub też do poważnej utraty zdrowia, a nawet życia. - [K_K4]

4. Potrafi odpowiednio określić priorytety służące realizacji określonego przez siebie lub innych zadania. - [K_K6]

Sposoby sprawdzenia efektów kształcenia

Efekty kształcenia przedstawione wyżej weryfikowane są w następujący sposób:

Ocena formująca:

a) w zakresie wykładów:

- na podstawie odpowiedzi na pytania dotyczące materiału omówionego na wykładach,

b) w zakresie laboratoriów / ćwiczeń:

- na podstawie oceny bieżącego postępu realizacji zadań.

Ocena podsumowująca:

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

- ocenę wiedzy i umiejętności wykazanych na egzaminie pisemnym o różnej charakterystyce problemów do rozwiązania: pytania egzaminacyjne dotyczą rożnego poziomu trudności, składają się na nie pytania testowe (wielokrotnego wyboru, treść do uzupełnia), proste zadania obliczeniowe (lub algorytmiczne) oraz zadania problemowe o większej złożoności; liczba pytań na egzaminie to ok. 10; wszystkie pytanie są podobnie punktowane, łącznie można otrzymać 100 punktów; zaliczenie egzaminu jest od 50 punktów; na ostateczną ocenę składa się w 60% ocena z egzaminu pisemnego, w 40% oceny z laboratorium.

- omówienie wyników egzaminu,

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

- ocenę realizacji zadań związanych z danymi zajęciami laboratoryjnymi: podczas każdych zajęć laboratoryjnych student otrzymuję listę zadań do wykonania: zadania dzielą się na niepunktowane, punktowane do realizacji na zajęciach, oraz punktowane zadania domowe; możliwe jest uzyskanie dodatkowych punktów za aktywność podczas zajęć.

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

- Przedstawienie problemu eksplozji danych we współczesnym świecie oraz rozróżnienie systemów informatycznych pod względem wykorzystywania danych na systemy operacyjne, w których dane służą do wspomagania codziennych czynności, oraz na systemy analityczne, w których stara się wydobyć wiedzę ze zgromadzonych danych. Omówienie zastosowania metod eksploracji danych oraz wskazanie pułapek związanych z przetwarzaniem dużych zbiorów danych.
- Przedstawienie historii i ewolucji systemów baz danych oraz dokładne omówienie modeli danych w rozróżnieniu na rodzaje systemów przetwarzania danych. Przypomnienie podstaw modelu relacyjnego, wprowadzenie i dokładne scharakteryzowanie modelu wielowymiarowego będącego podstawą systemów hurtowni danych, oraz modelu nierelacyjnego (NoSQL) związanego z przetwarzaniem masywnych danych w zastosowaniach internetowych.
- Wprowadzenie do MapReduce, który jest paradygmatem programowania, wywodzącym się z programowania funkcjonalnego, specjalnie stworzonym do przetwarzania masywnych danych w środowisku rozproszonych. Wykład obejmuje podstawy technologiczne związane z tym paradygmatem oraz omawia zapis podstawowych algorytmów w tym paradygmacie, takich jak zliczenia, operacje algebry relacji (projekcji, selekcji, grupowania, połączenia), oraz mnożenia macierzy.
- Wprowadzenie do problemów klasyfikacji i regresji jako podstawowych przykładów eksploracji danych. Podstawy teorii uczenia się. Podstawowe algorytmy klasyfikacji i regresji oraz elementy ich efektywnej implementacji.
- Poszukiwanie najbliższych sąsiadów, które jest podstawową operacją np. w systemach klasyfikacyjnych, systemach rekomendacyjnych oraz w zastosowaniach webowych takich jak szukanie plagiatów lub duplikatów stron www. Omówione zostaną struktury danych wykorzystywane do dokładnego wyszukiwania najbliższych sąsiadów, jak także metoda przybliżona bazująca na teorii lokalnie wrażliwych funkcji mieszających (ang. locality-sensitive hashing).
- Systemy rekomendacyjne oparte na treści oraz na filtrowaniu kolaboratywnym. Omówione zostaną algorytm dekompozycji macierzy i metoda stochastycznego spadku gradientu.

Zajęcia laboratoryjne prowadzone są w formie piętnastu dwugodzinnych ćwiczeń, odbywających się w laboratorium. Ćwiczenia realizowane są indywidualnie, z wyjątkiem niektórych zadań, które mogą być realizowane w zespołach 2-osobowych. Program laboratorium obejmuje następujące zagadnienia:

- Proste zadania z rachunku prawdopodobieństwa, które mają na celu pokazanie pułapek dotyczących analizy dużych zbiorów danych.
- Wprowadzenie do MapReduce; przedstawienie podstawowych zagadnień technicznych oraz implementacja prostych algorytmów w tym paradygmacie programowania, takich jak zliczanie, operacje algebry relacji, mnożenie macierzy.
- Algorytmy klasyfikacji i regresji: implementacja prostych algorytmów w środowisku Java.
- Implementacja algorytmu minhash oraz innych zagadnień związanych z lokalnie wrażliwymi funkcjami mieszającymi.
- Implementacja algorytmów rekomendacyjnych opartych na sąsiedztwie oraz wykorzystujących dekompozycję macierzy.

Metody dydaktyczne:

1. Wykład: prezentacja multimedialna ilustrowana przykładami podawanymi na tablicy.
2. Ćwiczenia laboratoryjne: rozwiązywanie zadań, implementacja podstawowych algorytmów w technologii MapReduce, implementacja algorytmów eksploracji danych w środowisku Java oraz w technologii MapReduce.

Literatura podstawowa:

1. Mining of Massive Datasets, A. Rajaraman, J. D. Ullman, Cambridge University Press, 2012 (http://infolab.stanford.edu/~ullman/mmds.html)
2. Systemy baz danych. Kompletny podręcznik. Wydanie II, Hector Garcia-Molina, Jeffrey D. Ullman, Jennifer Widom

Literatura uzupełniająca:

1. Hadoop in Action, Ch. Lam, Manning Publications Co., 2011.
2. Data-Intensive Text Processing with MapReduce, J.Lin, Ch. Dyer, Morgan and Claypool Publishers, 2010 (http://lintool.github.com/MapReduceAlgorithms/)
3. Elements of Statistical Learning: Second Edition, T. Hastie, R. Tibshirani, J. Friedman, Springer, 2009. (http://www-stat.stanford.edu/~tibs/ElemStatLearn/)

Bilans nakładu pracy przeciętnego studenta

Czynność	Czas (godz.)
1. Udział w zajęciach laboratoryjnych/ćwiczeniach projektowych	30
2. Zadanie domowe	15
3. Udział w konsultacjach związanych z realizacją procesu kształcenia (częściowo mogą być realizowane drogą elektroniczną)	8
4. Przygotowanie do zajęć.	15
5. Udział w wykładach	30
6. Zapoznanie się ze wskazaną literaturą i materiałami dydaktycznymi (10 stron tekstu naukowego = 1 godz.), 200 stron	20
7. Omówienie wyników egzaminu	2
8. Przygotowanie do egzaminu	20
9. Przeprowadzenie egzaminu	2

Obciążenie pracą studenta		
forma aktywności	**godzin**	**ECTS**
Łączny nakład pracy	142	5
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	72	3
Zajęcia o charakterze praktycznym	45	2