KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu	Kod
Wyszukiwanie i przetwarzanie zasobów informacyjnych	**1010514371010510091**

Kierunek studiów	Profil kształcenia (ogólnoakademicki, praktyczny)	Rok / Semestr
Informatyka	**ogólnoakademicki**	**4 / 7**
Ścieżka obieralności/specjalność	Przedmiot oferowany w języku:	Kurs (obligatoryjny/obieralny)
-	**polski**	**obieralny**

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

Godziny	Liczba punktów
Wykłady: **16** Ćwiczenia: **-** Laboratoria: **16** Projekty/seminaria: **-**	**4**

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	**4 100%**

Odpowiedzialny za przedmiot / wykładowca:

dr inż. Miłosz Kadziński
email: Milosz.Kadzinski@cs.put.poznan.pl
tel. 61 6653022
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot / wykładowca:

dr inż. Irmina Masłowska
email: Irmina.Maslowska@cs.put.poznan.pl
tel. 61 6652931
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań

Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

1	**Wiedza:**	Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z zakresu programowania obiektowego, algorytmów i struktur danych, statystyki i analizy danych, algebry liniowej oraz elementów sztucznej inteligencji.
2	**Umiejętności:**	Powinien posiadać umiejętności formułowania i rozwiązywania podstawowych problemów programowania matematycznego, stworzenia modelu obiektowego prostego systemu, programowania w co najmniej jednym języku obiektowym oraz pozyskiwania informacji ze wskazanych źródeł.
3	**Kompetencje społeczne**	Ponadto w zakresie kompetencji społecznych student musi rozumieć, że w informatyce wiedza i umiejętności bardzo szybko stają się przestarzałe, a także prezentować takie postawy jak uczciwość, odpowiedzialność, wytrwałość, ciekawość poznawcza, kreatywność, kultura osobista, szacunek dla innych ludzi.

Cel przedmiotu:

1. Przekazanie studentom wiedzy na temat podstawowych metod zbierania i indeksowania zasobów informacyjnych dla potrzeb dalszej analizy, modeli wyszukiwania informacji w odniesieniu do danych słabo-strukturalizowanych (np. tekstowych).
2. Wyjaśnienie studentom podstawowych metod przetwarzania języka naturalnego (ang. NLP - natural language processing).
3. Zapoznanie studentów z metodami rangowania zasobów internetowych pod względem adekwatności do zapytania i struktury grafu sieci, a także oceny jakości uzyskanych wyników.
4. Wyjaśnienie studentom podstawowych praw opisu struktury powiązań zasobów internetowych.
5. Zapoznanie studentów z zastosowaniami metod analizy danych i uczenia maszynowego do odkrywania wzorców w analizie zasobów informacyjnych oraz zachowania użytkowników.
6. Wyjaśnienie studentom wybranych zagrożeń funkcjonowania w sieci Internet.
7. Rozwijanie u studentów umiejętności zastosowania metod analizy danych, algebry liniowej, sztucznej inteligencji oraz uczenia maszynowego do analizy zawartości zasobów inform., struktury powiązań między tymi zasobami oraz wzorców użytkowania tych zasobów.
8. Rozwijanie u studentów umiejętności interpretacji wyników zastosowania ww. metod w kontekście analizy zawartości, struktury i użytkowania zasobów informacyjnych.

Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia

Wiedza:

1. ma szczegółową wiedzę w zakresie wybranych działów matematyki (elementy teorii macierzy, teorii prawdopodobieństwa oraz teorii grafów) - [K_W3]

2. ma uporządkowaną, podbudowaną teoretycznie wiedzę ogólną w zakresie przetwarzania i wyszukiwania informacji, algorytmów i złożoności, języków i paradygmatów programowania, elementów sztucznej inteligencji oraz narzędzi informatycznych do analizy danych - [K_W4]

3. ma podbudowaną teoretycznie szczegółową wiedzę związaną z wybranymi zagadnieniami z zakresu informatyki, takimi jak pozyskiwanie informacji (ang. information retieval), przetwarzanie języka naturalnego, analiza danych i uczenie maszynowe - [K_W5]

4. ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach w informatyce i w wybranych pokrewnych dyscyplinach naukowych w zakresie przetwarzania i wyszukiwania informacji - [K_W6]

5. ma podstawową wiedzę o cyklu życia programowych systemów informatycznych służących do przetwarzania i wyszukiwania informacji - [K_W7]

6. zna podstawowe metody, techniki i narzędzia stosowane przy rozwiązywaniu złożonych zadań inżynierskich z wybranego obszaru informatyki - [K_W8]

7. ma wiedzę i znajomość narzędzi niezbędnych do przetwarzania języka naturalnego - [-]

8. ma wiedzę niezbędną do analizy i przetwarzania zasobów informacyjnych (w tym głównie zbierania, przetwarzania oraz rangowania danych słabo-strukturalizowanych) i do dobrania właściwej metody realizacji tych zagadnień - [-]

9. ma wiedzę na temat praw opisu struktury powiązań zasobów internetowych - [-]

10. ma wiedzę na temat wybranych zagrożeń funkcjonowania w sieci Internet - [-]

Umiejętności:

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

2. posługiwać się technikami informacyjno-komunikacyjnymi wykorzystywanymi przy realizacji przedsięwzięć informatycznych - [K_U6]

3. wykorzystać do formułowania i rozwiązywania zadań informatycznych metody analityczne i eksperymentalne - [K_U8]

4. przy formułowaniu i rozwiązywaniu zadań informatycznych ? dostrzegać ich aspekty społeczne, ekonomiczne i prawne - [K_U9]

5. ocenić - przynajmniej w podstawowym zakresie - różne aspekty ryzyka związanego z przedsięwzięciem informatycznym - [K_U10]

6. ocenić złożoność obliczeniową algorytmów i problemów - [K_U13]

7. wybrać język programowania odpowiedni do danego zadania programistycznego - [K_U20]

8. zgodnie z zadaną specyfikacją - zaprojektować oraz zrealizować prosty system informatyczny, używając właściwych metod, technik i narzędzi - [K_U21]

9. ma umiejętność formułowania algorytmów i ich programowania z użyciem przynajmniej jednego z popularnych narzędzi - [K_U22]

10. zastosować wybrane metody analizy danych, algebry liniowej, sztucznej inteligencji, przetwarzania języka naturalnego oraz uczenia maszynowego do analizy zawartości zasobów informacyjnych, struktury powiązań między tymi zasobami oraz wzorców użytkowania tych zasobów - [-]

11. interpretować wyniki zastosowania ww. metod w kontekście analizy zawartości, struktury i użytkowania zasobów internetowych - [-]

Kompetencje społeczne:

1. rozumie, że w informatyce wiedza i umiejętności bardzo szybko stają się przestarzałe - [K_K1]

2. zna przykłady i rozumie przyczyny wadliwie działających systemów informatycznych, które doprowadziły do poważnych strat finansowych i społecznych - [K_K4]

3. 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 poprzednich 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 zaliczeniu pisemnym w formie testu składającego się z ok. 20 zadań otwartych: rozszerzonej odpowiedzi i/lub z krótką odpowiedzią, przy czym dla uzyskania oceny dostatecznej student musi zdobyć ponad 50% całkowitej liczby punktów,
- omówienie wyników zaliczenia,

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

- ocenę umiejętności związanych z realizacją ćwiczeń laboratoryjnych,
- ocenę sprawozdania z realizacji zadań analitycznych i symulacyjnych przygotowywanego częściowo w trakcie zajęć, a częściowo po ich zakończeniu; ocena ta obejmuje także umiejętność pracy w zespole,
- ocenę kodu źródłowego z realizacji zadań programistycznych oraz

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

Klasyfikacja zasobów internetowych i metod dostępu do informacji. Przegląd metod i zastosowań Web Mining: analiza zawartości zasobów internetowych - WCM, analiza struktury powiązań zasobów - WSM, analiza użytkowania zasobów - WUM. Charakterystyka poziomów opisu języka naturalnego i odpowiadających obszarów lingwistyki.

Etapy i metody wstępnego przetwarzania języka naturalnego na cele wyszukiwania informacji (ang. information retrieval): analiza leksykalna (wyrażenia regularne, morfemy, fleksja, wymiany głoskowe, reguły morfologiczne), identyfikacja i eliminacja słów o słabej wartości informacyjnej, lematyzacja/stemming, selekcja jednostek indeksujących, budowa struktur kategoryzujących.

N-gramy i obliczanie prawdopodobieństwa ich występowania. Poprawianie literówek (odległość Levenshteina, odległość edycyjna). Rozpoznawanie części mowy (part-of-speech (POS) tagging). Odkrywanie znaczenia słów i relacji między słowami.

Budowa reprezentacji dokumentów tekstowych, w tym szczegółowo reprezentacja w postaci wektorów TF-IDF. Miary podobieństwa dokumentów tekstowych. Klasyczne i nieklasyczne modele wyszukiwania informacji w danych tekstowych, a w szczególności: model boole'owski, modele probabilistyczne, model wektorowy VSM, rozszerzony model wektorowy GVSM, modele oparte na zbiorach rozmytych, model LSI oparty na dekompozycji macierzy term-dokument na wartości osobliwe, modele oparte na sieciach neuronowych.

Serwisy wyszukujące informacje - historia, architektura, zasady działania, metody organizacji i prezentacji wyników. Rangowanie dokumentów internetowych pod względem adekwatności do zapytania: historyczne i współczesne; idea Hubs & Authorities i algorytm HITS, algorytm PageRank i jego modyfikacje, aspekty brane pod uwagę przez współczesne wyszukiwarki podczas rangowania dokumentów wyłonionych w wynikach zapytań.

Ocena jakości wyników wyszukiwania informacji - klasyczne miary dokładności i kompletności oraz miary biorące pod uwagę odpowiedź systemu w postaci listy rankingowej. Przykładowe kolekcje testowe. Spamowanie wyników wyszukiwarek, techniki ukrywania spamu, techniki zwalczania spamu.

Indeksowanie dokumentów tekstowych, podstawowe rodzaje indeksów i ich zastosowania. Indeks odwrotny, drzewa i tablice sufiksów, złożoność czasowa i pamięciowa tworzenia i pielęgnacji poszczególnych typów indeksów. Algorytmy tworzenia indeksów odwrotnych dla dużych kolekcji tekstów - gdy rozmiar indeksu przekracza rozmiary pamięci operacyjnej. Indeksowanie rozproszone, model MapReduce, rozproszenie a replikacja.

Analiza struktury sieci Web: model Bowtie, prawo potęgowe i prawo Zipfa w opisie struktury powiązań stron/serwisów internetowych. Roboty internetowe: architektura, schemat i zasady działania, strategie crawlowania, polityka uprzejmości. Analiza użytkowania sieci Web w kontekście metodologii CRISP-DM.

Charakterystyka logów serwerów WWW i innych źródeł danych wykorzystywanych w zadaniach WUM, metody odkrywania i analizy wzorców - wykorzystanie znanych metod analizy statystycznej, data mining i uczenia maszynowego.

Automatyczna klasyfikacja i grupowanie dokumentów internetowych/ serwisów/użytkowników/wzorców zachowań użytkowników. Obserwacja zachowań użytkowników w celu personalizacji treści i usług internetowych; zastosowania w e-gospodarce, collaborative filtering i systemy rekomendacyjne.

Opinion mining - eksploracja opinii zamieszczanych w Internecie: identyfikacja opinii, klasyfikacja, sumaryzacja, wyszukiwarki opinii. Spamowanie opinii internetowych i systemów rekomendacyjnych, metody ukrywania spamu, metody identyfikacji spamu. Systemy wyszukiwania informacji w multimediach. Uwzględnianie wiedzy semantycznej w systemach wyszukujących: sieć semantyczna, metody reprezentacji i zarządzania wiedzą. Analiza nastawienia (ang. sentiment analysis)

Automatyczne udzielenie bezpośredniej odpowiedzi na pytania (ang. question answering), generowanie streszczeń dokumentu/ów.

Zajęcia laboratoryjne prowadzone są w formie piętnastu 2-godzinnych ćwiczeń, odbywających się w laboratorium. Ćwiczenia realizowane są przez studentów samodzielnie lub w 2-osobowych zespołach.

Metody dydaktyczne:

1. Wykład: prezentacja multimedialna ilustrowana przykładami podawanymi na tablicy.
2. Ćwiczenia laboratoryjne: rozwiązywanie zadań, ćwiczenia praktyczne, wykonywanie eksperymentów, dyskusja, praca w zespole, studium przypadków, demonstracja wybranych systemów przetwarzania informacji oraz pokaz multimedialny

Literatura podstawowa:

1. Eksploracja zasobów internetowych, Z.Markov, D.T.Larose, PWN, 2009

2. Speech and language processing - An Introduction to Natural Language Processing, D.Jurafsky, J.H. Martin, Computational Linguistics, and Speech Recognition?, Prentice Hall, 2008

3. Foundations of Statistical Natural Language Processing, Ch.D.Manning, H. Schütze, MIT Press, Cambridge Massachusetts, MIT Press Cambridge Mass, 1999

4. Introduction to Information Retrieval, Ch.D.Manning, P.Raghavan, H.Schütze, Cambridge University Press, 2008 (wersja poprawiona i uzupełniona w 2009 r. dostępna bezpłatnie on-line: http://nlp.stanford.edu/IR-book/)

5. Mining of Massive Datasets, Anand Rajaraman, Jeffrey David Ullman, Cambridge University Press, 2011 (wersja poprawiona i uzupełniona w 2012 r. dostępna bezpłatnie on-line: http://infolab.stanford.edu/~ullman/mmds/book.pdf)

6. Modern Information Retrieval, Ricardo Baeza-Yates, Berthier Ribeiro-Neto, Addison-Wesley, 1999

7. Data intensive text-processing with MapReduce, Jimmy Lin, Chris Dyer, University of Maryland, Morgan & Claypool Synthesis, 2010 (dostępna bezpłatnie on-line: http://beowulf.csail.mit.edu/18.337/MapReduce-book-final.pdf)

Literatura uzupełniająca:

1. Web Intelligence, Ning Zhong, Jiming Liu, Yiyu Yao (Eds.), Springer, 2003
2. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. B. Liu, Springer, 2009
3. Mining the Web: Discovering Knowledge from Hypertext Data. S. Chakrabarti, Morgan Kaufmann, 2002
4. The Text Mining Handbook. R. Feldman, J. Sanger, Cambridge University Press, 2006
5. Felietony publikowane na http://searchenginewatch.com, http://searchengineland.com/

Bilans nakładu pracy przeciętnego studenta

Czynność	Czas (godz.)
1. udział w zajęciach laboratoryjnych / ćwiczeniach :	16
2. dokończenie (w ramach pracy własnej) sprawozdań z ćwiczeń laboratoryjnych:.	16
3. udział w konsultacjach związanych z realizacją procesu kształcenia, w szczególności ćwiczeń laboratoryjnych / projektu	5
4. napisanie programu / programów, uruchomienie I weryfikacja (czas poza zajęciami laboratoryjnymi)	16
5. udział w wykładach	16
6. zapoznanie się ze wskazaną literaturą / materiałami dydaktycznymi (10 stron tekstu naukowego = 1 godz.)	10
7. przygotowanie do zaliczenia	16
8. obecność na zaliczenia	2
9. omówienie wyników zaliczenia	2

Obciążenie pracą studenta

forma aktywności	godzin	ECTS
Łączny nakład pracy	99	4
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	41	2
Zajęcia o charakterze praktycznym	48	2