KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu	Kod
Eksploracja zasobów Internetu	**1010512321010510618**

Kierunek studiów	Profil kształcenia (ogólnoakademicki, praktyczny)	Rok / Semestr
Informatyka	**ogólnoakademicki**	**1 / 2**
Ścieżka obieralności/specjalność	Przedmiot oferowany w języku:	Kurs (obligatoryjny/obieralny)
Technologie przetwarzania danych	**polski**	**obieralny**

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

Godziny	Liczba punktów
Wykłady: **30** Ćwiczenia: **-** Laboratoria: **30** 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. Irmina Masłowska
email: Irmina.Maslowska@cs.put.poznan.pl
tel. 61 6652931
Instytut Informatyki
ul. Piotrowo 2, 60-965 Poznań

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

Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

1	**Wiedza:**	Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z zakresu programowania obiektowego, aplikacji internetowych, algorytmów i struktur danych, statystyki i analizy danych, algebry liniowej oraz elementów uczenia maszynowego.
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, wstępnego przetwarzania i indeksowania zasobów internetowych dla potrzeb dalszej analizy, oraz wiedzy na temat modeli wyszukiwania informacji w odniesieniu do danych słabo-strukturalizowanych (np. tekstowych).
2. Zapoznanie studentów z metodami rangowania zasobów internetowych pod względem adekwatności do zapytania i struktury grafu sieci, a także oceny jakości uzyskanych wyników oraz z zastosowaniami metod analizy danych i uczenia maszynowego do odkrywania wzorców w analizie zasobów internetowych oraz zachowania użytkowników
3. Wyjaśnienie studentom podstawowych praw opisu struktury powiązań zasobów internetowych i wybranych zagrożeń funkcjonowania w sieci Internet.
4. Rozwijanie u studentów umiejętności zastosowania metod analizy danych, algebry liniowej, sztucznej inteligencji oraz uczenia maszynowego do analizy zawartości zasobów internetowych, struktury powiązań między tymi zasobami oraz wzorców użytkowania tych zasobów, interpretacji wyników zastosowania ww. metod w kontekście analizy zawartości, struktury i użytkowania zasobów internetowych oraz umiejętności pracy zespołowej.

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 eksploracji zasobów internetowych, 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 (informationretieval), 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 analizy oraz eksploracji zasobów internetowych - [K_W6]

5. ma podstawową wiedzę o cyklu życia programowych systemów informatycznych służących do analizy oraz eksploracji zasobów internetowych - [K_W7]

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

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

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

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

Umiejętności:

1. potrafi 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. potrafi określić kierunki dalszego uczenia się i zrealizować proces samokształcenia, - [K_U5]

3. potrafi wykorzystać do formułowania i rozwiązywania zadań inżynierskich i prostych problemów badawczych metody analityczne i eksperymentalne - [K_U9]

4. potrafi ? przy formułowaniu i rozwiązywaniu zadań inżynierskich ? integrować wiedzę z różnych obszarów informatyki (a w razie potrzeby także wiedzę z innych dyscyplin naukowych) oraz zastosować podejście systemowe, uwzględniające także aspekty pozatechniczne - [K_U10]

5. potrafi formułować i testować hipotezy związane z problemami inżynierskimi i prostymi problemami badawczymi - [K_U12]

6. potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć (metod i narzędzi) oraz nowych produktów - [K_U13]

7. potrafi ocenić przydatność metod i narzędzi służących do rozwiązania zadania inżynierskiego, polegającego na budowie lub ocenie systemu informatycznego lub jego składowych, w tym dostrzec ograniczenia tych metod i narzędzi - [K_U24]

8. potrafi wybrać język programowania odpowiedni do danego zadania programistycznego - [K_U26]

9. potrafi zastosować wybrane metod analizy danych, algebry liniowej, sztucznej inteligencji oraz uczenia maszynowego do analizy zawartości zasobów internetowych, struktury powiązań między tymi zasobami oraz wzorców użytkowania tych zasobów - [-]

10. potrafi 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. 30 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 obronę projektów przez studenta.

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

- omówienie dodatkowych aspektów zagadnienia,
- efektywność zastosowania zdobytej wiedzy podczas rozwiązywania zadanego problemu,
- wskazywanie trudności percepcyjnych studentów umożliwiające bieżące doskonalenia procesu dydaktycznego.

Treści programowe

Program wykładu obejmuje następujące zagadnienia:

Klasyfikacja zasobów internetowych i metod dostępu do informacji. Przegląd metod i zastosowań Web Mining. Charakterystyka poziomów opisu języka naturalnego i odpowiadających obszarów lingwistyki. Etapy i metody wstępnego przetwarzania tekstów na cele wyszukiwania informacji (Information Retrieval): analiza leksykalna, identyfikacja i eliminacja słów o słabej wartości informacyjnej, lematyzacja/stemming, selekcja jednostek indeksujących, budowa struktur kategoryzujących. 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, collaborativefiltering i systemy rekomendacyjne. Opinionmining - 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.

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. Program laboratorium obejmuje następujące zagadnienia:

Metody wstępnego przetwarzania dokumentów tekstowych. Praktyczne wykorzystanie modelu przestrzeni wektorowej do rangowania zasobów tekstowych pod względem adekwatności do zapytania. Miary podobieństwa zasobów tekstowych pod względem zawartości. Ocena jakości wyszukiwania. Praktyczne wykorzystanie algorytmów HITS oraz PageRank do tworzenia rankingu zasobów internetowych. Metody spamowania poprzez wpływanie na strukturę sieci. Algorytmy obliczania współczynników zaufania z rodziny TrustRank. Automatyczne rozszerzanie zapytań z wykorzystaniem metody relevance feedback, macierzy korelacji oraz słowników zewnętrznych typu WordNet. Praktyczne wykorzystanie algorytmu redukcji wymiarów przestrzeni reprezentacji zasobów tekstowych. Przetwarzanie różnych formatów plików log serwera oraz podstawy eksploracyjnej analizy danych. Wykorzystanie łańcuchów Markowa, drzewa sekwencji oraz algorytmu A-Priori do odkrywania wzorców poruszania się po stronach internetowych. Algorytmy wyboru reklamodawców. Badanie skuteczności kampanii reklamowej prowadzonej w serwisie internetowym. Algorytmy tworzenia indeksów, drzewa oraz tablicy sufiksów. Zastosowanie pakietu Lucene do indeksowania, parsowania oraz tworzenia rankingu zasobów tekstowych. Pakiet Tika w analizie oraz parsowaniu zawartości plików różnych formatów. Wykorzystanie pakietu Nutch lub Solr do gromadzenia zasobów internetowych oraz rozszerzanie ich funkcjonalności o dodatkowe moduły. Rozwój i implementacja sekwencyjnego robota internetowego. Model MapReduce. Wykorzystanie hierarchicznych oraz iteracyjno-optymalizacyjnych metod grupowania dokumentów oraz profili użytkowania i zawartości. Wykorzystanie algorytmów klasyfikacji do personalizacji zawartości serwisu internetowego, wykrywania spamu oraz sortowania wiadomości. Wykorzystanie algorytmów predykcji na podstawie zachowania i ocen wystawionych użytkowników. Podstawy działania wyszukiwarki obrazów. Podstawowe algorytmy analizy strumieni danych. Zasada działania systemów udzielających bezpośredniej odpowiedzi na zadane pytanie.

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 eksploracji zasobów internetowych oraz pokaz multimedialny

Literatura podstawowa:

1. Eksploracja zasobów internetowych, ZdravkoMarkov, Daniel T. Larose, PWN, 2009

2. Introduction to Information Retrieval, Christopher D. Manning, PrabhakarRaghavan, HinrichSchütze, Cambridge University Press, 2008 (wersjapoprawionaiuzupełniona w 2009 r. dostępnabezpłatnie on-line: http://nlp.stanford.edu/IR-book/)

3. Mining of Massive Datasets, AnandRajaraman, Jeffrey David Ullman, Cambridge University Press, 2011 (wersjapoprawionai uzupełniona w 2012 r. dostępnabezpłatnie on-line: http://infolab.stanford.edu/~ullman/mmds/book.pdf)

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

5. Automatic Text Processing: The Transformation Analysis and Retrieval of Information by Computer, Gerard Salton, Addison-Wesley, 1989

6. Data intensive text-processing with MapReduce, Jimmy Lin, Chris Dyer, University of Maryland, Morgan & Claypool Synthesis, 2010 (dostępnabezpłatnie 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 bieżąco na http://searchenginewatch.com

Bilans nakładu pracy przeciętnego studenta

Czynność	Czas (godz.)
1. udział w zajęciach laboratoryjnych / ćwiczeniach	30
2. dokończenie (w ramach pracy własnej) sprawozdań z ćwiczeń laboratoryjnych	6
3. udział w konsultacjach związanych z realizacją procesu kształcenia, w szczególności ćwiczeń laboratoryjnych / projektu (częściowo mogą być realizowane drogą elektroniczną)	4
4. napisanie programu / programów, uruchomienie i weryfikacja (czas poza zajęciami laboratoryjnymi)	16
5. udział w wykładach	30
6. przygotowanie do kolokwiów zaliczeniowych	10
7. obecność na kolokwium zaliczeniowym	2
8. omówienie wyników zaliczenia	2

Obciążenie pracą studenta

forma aktywności	godzin	ECTS
Łączny nakład pracy	100	4
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	68	3
Zajęcia o charakterze praktycznym	52	2