KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Wyszukiwanie i przetwarzanie zasobów informacyjnych

Przedmiot

Kierunek studiów
Informatyka

Rok/semestr
3 / 6

Studia w zakresie (specjalność)

Profil studiów
ogólnoakademicki

Poziom studiów
pierwszego stopnia

Język oferowanego przedmiotu
polski

Forma studiów
stacjonarne

Wymagalność
obieralny

Liczba godzin

Wykład	Laboratoria	Inne (np. online)
30	30	
Ćwiczenia	Projekty/seminaria	

Liczba punktów ECTS

4

Wykładowcy

Odpowiedzialny za przedmiot/wykładowca:
dr hab. inż. Miłosz Kadziński

Odpowiedzialny za przedmiot/wykładowca:
dr inż. Irmina Masłowska

Wymagania wstępne

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.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.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.

Przedmiotowe efekty uczenia się

Wiedza

1. ma szczegółową wiedzę w zakresie wybranych działów matematyki (elementy teorii macierzy, teorii prawdopodobieństwa oraz teorii grafów) - [K1st_W1]

2. ma uporządkowaną, podbudowaną teoretycznie wiedzę ogólną w zakresie przetwarzania i wyszukiwania informacji, algorytmów i złożoności, sztucznej inteligencji, narzędzi informatycznych do analizy danych oraz wiedzę szczegółową związaną z wybranymi zagadnieniami z zakresu informatyki, jak pozyskiwanie informacji (ang. IR), przetwarzanie języka naturalnego, analiza danych i uczenie maszynowe - [K1st_W4]

3. 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 - [K1st_W4]

4. ma wiedzę o istotnych kierunkach rozwoju i najważniejszych nowych osiągnięciach w informatyce i w wybranych pokrewnych dyscyplinach naukowych w zakresie przetwarzania i wyszukiwania informacji - [K1st_W5]

5. zna podstawowe techniki, metody i narzędzia wykorzystywane w procesie rozwiązywania zadań informatycznych, głównie o charakterze inżynierskim w zakresie przetwarzania i wyszukiwania informacji, a także ma wiedzę i znajomość narzędzi niezbędnych do przetwarzania języka naturalnego - [K1st_W7]

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

Umiejętności

1. potrafi, formułując i rozwiązując zadania informatyczne z zakresu przetwarzania i wyszukiwania informacji, zastosować odpowiednio dobrane metody, w tym metody analityczne i eksperymentalne, a także potrafi 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 - [K1st_U4]

2. potrafi interpretować wyniki zastosowania ww. metod w kontekście analizy zawartości, struktury i użytkowania zasobów internetowych, a także potrafi pozyskiwać informacje nt. przetwarzania i wyszukiwania informacji z różnych źródeł, w tym literatury i baz danych (w języku polskim i angielskim), właściwie je integrować, dokonywać ich interpretacji i krytycznej oceny - [K1st_U1]

3. potrafi odpowiednio posługiwać się technikami informacyjno-komunikacyjnymi wykorzystywanymi przy realizacji przedsięwzięć informatycznych z zakresu przetwarzania i wyszukiwania informacji - [K1st_U2]

4. potrafi - zgodnie z zadaną specyfikacją - zaprojektować oraz zrealizować system przetwarzania i wyszukiwania informacji, dobierając język programowania oraz używając właściwych metod, technik i narzędzi - [K1st_U10]

5. ma umiejętność formułowania algorytmów i ich implementacji z użyciem przynajmniej jednego z popularnych narzędzi z zakresu przetwarzania i wyszukiwania informacji - [K1st_U11]

6. potrafi planować i realizować proces własnego permanentnego uczenia się w zakresie przetwarzania i wyszukiwania informacji - [K1st_U19]

Kompetencje społeczne

1. rozumie, że w informatyce wiedza i umiejętności z zakresu przetwarzania i wyszukiwania informacji bardzo szybko stają się przestarzałe - [K1st_K1]

2. zna przykłady i rozumie przyczyny wadliwie działających systemów z zakresu przetwarzania i wyszukiwania informacji - [K1st_K2]

3. potrafi myśleć i działać w sposób przedsiębiorczy, m.in. znajdując komercyjne zastosowania dla tworzonego oprogramowania, mając na uwadze nie tylko korzyści biznesowe, ale również społeczne prowadzonej działalności - [K1st_K3]

Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Wiedza nabyta w ramach wykładu jest weryfikowana w ramach zaliczenia pisemnego w formie testu składającego się z max. 20 zadań otwartych: rozszerzonej odpowiedzi i/lub z krótką odpowiedzią, przy czym dla uzyskania oceny dostatecznej student musi zdobyć ponad 50% całkowitej liczby punktów

Umiejętności nabyte w ramach zajęć laboratoryjnych weryfikowane są na podstawie: (1) realizacji ćwiczeń laboratoryjnych, (2) 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, (3) kodu źródłowego z realizacji zadań programistycznych oraz "obrony" projektów przez studenta. Powyższa ocena może zostać podwyższona na podstawie dodatkowych punktów za aktywność podczas zajęć, w szczególności za: (1) omówienie dodatkowych aspektów zagadnienia, (2) efektywność zastosowania zdobytej wiedzy podczas rozwiązywania zadanego problemu, (3) wskazywanie trudności percepcyjnych studentów umożliwiające bieżące doskonalenie procesu dydaktycznego.

Treści programowe

Zagadnienia (wykład):

Klasyfikacja zasobów internetowych i metod dostępu do informacji. Przegląd metod i zastosowań Web Mining: analiza treści zasobów internetowych, analiza struktury powiązań zasobów, analiza użytkowania zasobów. Charakterystyka poziomów opisu języka naturalnego i odpowiadających obszarów lingwistyki. Etapy i metody wstępnego przetwarzania języka naturalnego na cele wyszukiwania informacji: 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. Poprawianie literówek (odległość Levenshteina, odległość edycyjna). Rozpoznawanie części mowy (POS-tagging). Odkrywanie znaczenia słów i relacji między słowami. Budowa reprezentacji dokumentów tekstowych, reprezentacja 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, modele oparte na zbiorach rozmytych, sieciach neuronowych, model LSI oparty na dekompozycji SVD macierzy term-dokument, inne.

Serwisy wyszukujące informacje - historia, architektura, zasady działania, metody organizacji i prezentacji wyników. Rangowanie dokumentów internetowych pod względem adekwatności do zapytania: historyczne i współczesne; algorytm HITS, algorytm PageRank i jego modyfikacje; aspekty brane pod uwagę przez współczesne wyszukiwarki podczas rangowania dokumentów. Ocena jakości wyników wyszukiwania informacji - w tym miary dokładności i kompletności. 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. Algorytmy tworzenia indeksów odwrotnych dla dużych kolekcji tekstów. Indeksowanie rozproszone, model MapReduce. Analiza struktury sieci Web: model Bowtie, prawo potęgowe i Zipfa w opisie struktury powiązań stron/serwisów internetowych. Charakterystyka logów serwerów WWW i innych źródeł danych w zadaniach WUM, metody odkrywania i analizy wzorców - wykorzystanie metod analizy statystycznej, data mining i uczenia maszynowego. Automatyczna klasyfikacja i grupowanie dokumentów/serwisów/użytkowników/wzorców zachowań użytkowników. Analiza zachowań użytkowników dla personalizacji treści i usług internetowych; zastosowania w e-gospodarce, collaborative filtering i systemy rekomendacyjne. Opinion mining&sentiment analysis - eksploracja opinii

z Internetu: identyfikacja, klasyfikacja, sumaryzacja, wyszukiwarki opinii. Spamowanie opinii internetowych i systemów rekomendacyjnych, metody ukrywania spamu, metody identyfikacji spamu.

Zajęcia laboratoryjne prowadzone są w laboratorium w formie piętnastu 2-godzinnych ćwiczeń. Studenci realizują ćwiczenia samodzielnie lub w 2-osobowych zespołach. Zagadnienia (laboratorium):

Metody wstępnego przetwarzania dokumentów tekstowych: tokenizacja, eliminacja stop words, normalizacja, stemming i lematyzacja. Praktyczne wykorzystanie modelu przestrzeni wektorowej (miara TF-IDF oraz odległość kosinusowa) do rangowania zasobów tekstowych pod względem adekwatności do zapytania. Wnioskowanie statystyczne w sprawdzaniu poprawności pisowni. Wykorzystanie pakietu OpenNLP do przetwarzania języka naturalnego, POS taggingu, parsowania i analiza nastawienia. Miary podobieństwa zasobów tekstowych pod względem zawartości. Ocena jakości wyszukiwania z wykorzystaniem miar dopasowania odpowiedzi i efektywności systemu. Praktyczne wykorzystanie algorytmów HITS oraz PageRank do tworzenia rankingu zasobów internetowych opartego na strukturze połączeń. 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 (typu Balance i AdWords) pozwalające na maksymalizację zysku twórcy wyszukiwarki. 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/Solr do gromadzenia zasobów internetowych. Rozwój i implementacja sekwencyjnego robota internetowego. Model MapReduce w indeksowaniu, rekomendacji oraz tworzeniu rankingu zasobów tekstowych. Wykorzystanie metod grupowania dokumentów oraz profili użytkowania i zawartości. Wykorzystanie algorytmów klasyfikacji do personalizacji zawartości serwisu internetowego, wykrywania spamu oraz sortowanie wiadomości. Wykorzystanie algorytmów predykcji na podstawie zachowania i ocen wystawionych użytkowników w ramach pakietu Mahout.

Metody dydaktyczne

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. 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/)

3. 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)

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

5. Data intensive text-processing with MapReduce, Jimmy Lin, Chris Dyer, University of Maryland, Morgan & Claypool Synthesis, 2010 (dostępna bezpłatnie on-line: https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf)

Uzupełniająca

1. Speech and Language Processing (3rd ed. draft), D. Jurafsky and J.H. Martin (wersja z 2018 dostępna bezpłatnie on-line: https://web.stanford.edu/~jurafsky/slp3)

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

3. Web Data Mining: Exploring Hyperlinks, Contents and Usage Data. B. Liu, Springer, 2009

4. Mining the Web: Discovering Knowledge from Hypertext Data. S. Chakrabarti, Morgan Kaufmann, 2002

5. The Text Mining Handbook. R. Feldman, J. Sanger, Cambridge University Press, 2006

6. Felietony publikowane na bieżąco na http://searchenginewatch.com, http://searchengineland.com/

Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	100	4,0
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	60	2,5
Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, napisanie programu/programów, uruchomienie i weryfikacja, dokończenie sprawozdań z ćwiczeń laboratoryjnych, przygotowanie do zaliczenia[1]	40	1,5

[1] niepotrzebne skreślić lub dopisać inne czynności