



## KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Wyszukiwanie informacji [S1S1E>WINF]

### Przedmiot

Kierunek studiów

Sztuczna inteligencja/Artificial Intelligence

Rok/Semestr

3/5

Studia w zakresie (specjalność)

–

Profil studiów

ogólnoakademicki

Poziom studiów

pierwszego stopnia

Język oferowanego przedmiotu

angielski

Forma studiów

stacjonarne

Wymagalność

obligatoryjny

### Liczba godzin

Wykład

15

Laboratorium

15

Inne

0

Ćwiczenia

0

Projekty/seminaria

0

### Liczba punktów ECTS

3,00

### Koordynatorzy

dr hab. inż. Miłosz Kadziński prof. PP  
milosz.kadzinski@put.poznan.pl

### Wykładowcy

### Wymagania wstępne

Student rozpoczynający ten przedmiot powinien posiadać wiedzę z zakresu analizy matematycznej, algebry liniowej, eksploracji danych i uczenia maszynowego. Musi też posiadać umiejętność programowania w języku Python.

### Cel przedmiotu

Celem przedmiotu jest zapoznanie studentów z zagadnieniami wyszukiwania informacji. Obejmują one systemy wyszukiwania, metody analizy użytkownika sieci, modele reprezentacji dokumentów, algorytmy eksploracji struktury sieci, metody rozszerzania zapytań, społeczne filtrowanie i rekomendację, konstrukcję i kompresję indeksów oraz przetwarzania MapReduce. Przedmiot ma przygotować studentów do udziału w kursach nt. przetwarzania języka naturalnego oraz przetwarzania równoległego.

### Przedmiotowe efekty uczenia się

Wiedza:

K1st\_W3: ma uporządkowaną wiedzę teoretyczną dotyczącą kluczowych zagadnień informatyki z zakresu wyszukiwania informacji, w tym metod eksploracji zawartości, struktury i użytkownika sieci

K1st\_W4: zna i rozumie podstawowe techniki, metody, algorytmy oraz narzędzia wykorzystywane w

procesie rozwiązywania zadań informatycznych oraz problemów związanych z wyszukiwaniem informacji (np. indeksy, modele reprezentacji stron czy miary podobieństwa dokumentów do zapytania)  
K1st\_W5: ma podstawową wiedzę o istotnych kierunkach rozwoju i najważniejszych osiągnięciach wyszukiwania informacji rozumianego jako istotna dziedzina sztucznej inteligencji czerpiąca z osiągnięć innych dyscyplin naukowych oraz dostarczająca dla nich rozwiązań o potencjale praktycznym

#### Umiejętności:

K1st\_U1: potrafi pozyskiwać informacje z właściwie dobranych źródeł o różnej charakterystyce, dokonywać ich krytycznej analizy, interpretacji i syntezy oraz wyczerpująco uzasadniać formułowane opinie

K1st\_U3: potrafi formułować i rozwiązywać złożone zadania związane z wyszukiwaniem informacji (np. pozyskiwanie informacji czy ocena adekwatności dokumentów dla zapytania), stosując odpowiednio dobrane metody

K1st\_U4: potrafi zaplanować i przeprowadzić eksperymenty, w tym pomiary i symulacje komputerowe, interpretować uzyskane wyniki oraz wyciągać z nich wnioski w kontekście zadań wyszukiwania informacji (np. dotyczących oceny jakości wyników zwracanych przez wyszukiwarki)

K1st\_U9: ma umiejętność prostej adaptacji istniejących oraz formułowania i implementacji nowych algorytmów w języku Python, w tym algorytmów typowych dla różnych zadań wyszukiwania informacji

K1st\_U10: potrafi pozyskiwać, analizować i przetwarzać dane różnego typu (w tym głównie nieustrukturalizowane), dokonywać ich syntezy do wiedzy i wniosków przydatnych do realizacji szerokiego spektrum zadań wyszukiwania informacji

K1st\_U11: potrafi wykorzystywać oraz adaptować modele zachowań inteligentnych

#### Kompetencje społeczne:

K1st\_K1: rozumie, że w informatyce ze szczególnym uwzględnieniem wyszukiwania informacji wiedza i umiejętności bardzo szybko stają się przestarzałe, dostrzegając przy tym potrzebę ciągłego dokształcania oraz podnoszenia własnych kompetencji

K1st\_K2: ma świadomość istotności wiedzy i badań naukowych związanych z informatyką i sztuczną inteligencją w rozwiązywaniu praktycznych problemów o kluczowym znaczeniu dla funkcjonowania jednostek, firm, organizacji oraz całego społeczeństwa

K1st\_K5: potrafi myśleć i działać w sposób przedsiębiorczy, m.in. znajdując komercyjne zastosowania dla tworzonych systemów sztucznej inteligencji, mając na uwadze nie tylko korzyści ekonomiczne, ale również aspekty prawne i społeczne

### Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Wykłady: zaliczenie przeprowadzone na ostatnim wykładzie. Studenci muszą rozwiązać zadania obliczeniowe dotyczące zagadnień prezentowanych na poszczególnych wykładach. Każde zadanie jest oceniane indywidualnie i za jego rozwiązanie przyznawana jest określona liczba punktów. Punkty są sumowane i następująca skala jest wykorzystywana do określenia oceny: <50% - 2.0, [50% , 60%) - 3.0, [60% , 70%) - 3.5, [70% , 80%) - 4.0, [80% , 90%) - 4.5, and [90% , 100%] - 5.0.

Laboratoria: Po każdych zajęciach, studenci rozwiązują zadania programistyczne i przedstawiają rozwiązania osobie prowadzącej laboratoria w ciągu dwóch tygodni. Każde zadanie jest oceniane na skali od 2.0 do 5.0. Ostateczna ocena jest obliczana jako średnia z ocen przyznanych za poszczególne zadania z zastrzeżeniem, że najgorsza ocena uzyskana w trakcie semestru nie jest brana pod uwagę.

### Treści programowe

Kurs wprowadza w świat wyszukiwania informacji. Jest to ściśle związane z wyszukiwaniem, rekomendacjami i eksploracją sieci. Omawiane tematy oferują przegląd klasycznych, ale różnorodnych tematów związanych z wyszukiwaniem informacji. Najważniejsze z nich obejmują przetwarzanie tekstu, model przestrzeni wektorowej, PageRank i HITS, rozszerzanie zapytań, filtrowanie społeczne, budowę i kompresję indeksów oraz Map-Reduce.

### Tematyka zajęć

Wprowadzenie do wyszukiwania informacji: wyszukiwanie Boolowskie, przetwarzanie tekstu oraz odkrywanie wzorców nawigacyjnych: zakres zainteresowania dziedziny wyszukiwania informacji, typy danych, eksploracja sieci, wyszukiwanie nieustrukturalizowane, indeksowanie - macierz wystąpień,

model wyszukiwania Boolowskiego, optymalizacja zapytań, podstawowe etapy tworzenia indeksu odwrotnego (tokenizacja, eliminacja stop words, normalizacja, stemming, lematyzacja), analiza użytkownika sieci, pliki log, formaty, identyfikacja użytkowników i sesji, uzupełnianie ścieżek, wzorce nawigacyjne, łańcuch Markowa.

Model wektorowy oraz ukryte indeksowanie semantyczne: architektura systemu wyszukiwania, reprezentacja binarna, reprezentacja bag of words, częstotliwość termów TF, odwrotna częstotliwość dokumentów IDF, model przestrzeni wektorowej, miara podobieństwa kosinusowego, macierz termy-dokumenty, analiza składowych głównych, ukryte indeksowanie semantyczne.

Ocena w systemach wyszukiwania oraz PageRank: potrzeba oceny systemów wyszukiwania, miary oceny: precyzja, czułość, dokładność, miara F, miary obcięte do k, średnia precyzja, R-precyzja, struktura sieci, algorytm PageRank, farma linków, algorytm TrustRank, Google Penguin oraz Panda.

HITS, informacja zwrotna i poprawa literówek: HITS - Hubs and Authorities, rozszerzanie i uszczegółowianie zapytań, informacja zwrotna co do istotności wyników, algorytm Rocchio, pseudo-informacja zwrotna, wykorzystania słowników, poprawianie literówek, odległość edycyjna Levenshteina, algorytm Soundex.

Systemy rekomendacyjne oraz problem Adwords: sławne systemy rekomendacyjne, rekomendacja na podstawie zawartości, społeczne filtrowanie na podstawie ocen innych użytkowników oraz obiektów, slope one predictor, problem Adwords, algorytm BALANCE.

Konstrukcja oraz kompresja indeksów: indeks odwrotny, pozycyjny indeks odwrotny, indeks K-gramów, drzewo sufiksów, naiwny algorytm konstrukcji drzewa, algorytm Ukkonena, tablica sufiksów, algorytm qsufsort, kompresja danych, prawo Heapsa, prawo Zipfa, kodowania binarne i jedynekowa, kodowanie gamma, kodowania delta.

Wprowadzenie do MapReduce: pomysły dotyczące przetwarzania dużych danych, przetwarzanie MapFold, interpretacja pojęcia MapReduce, przetwarzania par (klucz, wartość), Mapper, Reducer, Combiner, Partitioner, przykłady: zliczanie słów, średnia ocen, macierz współwystąpień, indeks odwrotny, PageRank, kiedy MapReduce jest (mniej) przydatne?

## Metody dydaktyczne

Wykład: slajdy multimedialne dotyczące wyszukiwania informacji, ilustrowane przykładami oraz zadania obliczeniowe, służące jako podsumowanie wykładu i przygotowanie do zaliczenia.

Laboratoria: rozwiązywanie zadań ilustrujących na tablicy, programowanie w języku Python, przeprowadzenie eksperymentów obliczeniowych, dyskusja wybranych metod, praca zespołowa.

## Literatura

### Podstawowa

C. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008, <http://nlp.stanford.edu/IR-book/>

A. Rajaraman, J. Ullman, Mining of Massive Datasets, Cambridge University Press, 2011  
<http://infolab.stanford.edu/~ullman/mmds/book.pdf>

R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999

J. Lin, C. Dyer, 5. Data intensive text-processing with MapReduce, J. Lin, C. Dyer, University of Maryland, Morgan & Claypool Synthesis, 2010, <https://lintool.github.io/MapReduceAlgorithms/MapReduce-book-final.pdf>

### Uzupełniająca

D. Jurafsky, J.H. Martin, Speech and Language Processing, <https://web.stanford.edu/~jurafsky/slp3>

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

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

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

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

## Bilans nakładu pracy przeciętnego studenta

|                                                                                                                                                      | Godzin | ECTS |
|------------------------------------------------------------------------------------------------------------------------------------------------------|--------|------|
| Łączny nakład pracy                                                                                                                                  | 75     | 3,00 |
| Zajęcia wymagające bezpośredniego kontaktu z nauczycielem                                                                                            | 30     | 1,50 |
| Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwium/egzaminu, wykonanie projektu) | 45     | 1,50 |