Technologie Big Data - opis przedmiotu

Informacje ogólne

Nazwa przedmiotu	Technologie Big Data
Kod przedmiotu	11.3-WE-INFD-TechBD
Wydział	Wydział Nauk Inżynieryjno-Technicznych
Kierunek	Informatyka
Profil	ogólnoakademicki
Rodzaj studiów	drugiego stopnia z tyt. magistra inżyniera
Semestr rozpoczęcia	semestr zimowy 2021/2022

Informacje o przedmiocie

Semestr	2
Liczba punktów ECTS do zdobycia	5
Typ przedmiotu	obowiązkowy
Język nauczania	polski
Sylabus opracował	• dr hab. inż. Artur Gramacki, prof. UZ

Formy zajęć

Forma zajęć	Liczba godzin w semestrze (stacjonarne)	Liczba godzin w tygodniu (stacjonarne)	Liczba godzin w semestrze (niestacjonarne)	Liczba godzin w tygodniu (niestacjonarne)	Forma zaliczenia
Wykład	30	2	18	1,2	Egzamin
Laboratorium	30	2	18	1,2	Zaliczenie na ocenę

Cel przedmiotu

Nauczenie studentów doboru odpowiednich technik analizy danych w zależności od skali rozpatrywanego problemu oraz rodzaju przeprowadzanej analizy.

Nauczenie studentów pracy z wykorzystaniem nowoczesnych platform do składowania i przetwarzania danych.

Zapoznanie studentów z wybranymi technikami analizowania dużych zbiorów danych, głównie tekstowych.

Wymagania wstępne

Bazy danych.

Znajomość podstaw statystyki.

Zakres tematyczny

Big Data: wprowadzenie do zagadnienia przetwarzania wielkich ilości danych.

Nierelacyjne bazy danych: Przypomnienie podstawowych zagadnień związanych z relacyjnymi bazami danych. Zalety i wady tych baz danych. Podstawowe problemy związane z wykorzystaniem relacyjnych baz danych do składowania i przetwarzania coraz większych ilości danych coraz bardziej rozproszonych. Skalowanie poziome oraz pionowe baz danych. Nowa koncepcja baz nie opartych o tradycyjny model relacyjny. Teoria CAP oraz BASE. Agregacyjne modele danych. Bazy danych typu klucz-wartość, kolumnowe, dokumentowe, grafowe. Replikacja baz danych. Współdzielenie zasobów w bazach danych. Metodologia Map-Reduce. Przedstawienie kilku wybranych systemów baz danych nierelacyjnych (np. MongoDB, Cassandra, Redis, Neo4J, Oracle NoSQL Database).

Wybrane systemy informatyczne: Analityka biznesowa na dużą skalę: nowoczesne rozwiązania wykorzystywane do przesyłania, składowania oraz przetwarzania dużych zbiorów danych. Architektura nowoczesnych systemów do składowania i przetwarzania Big Data na przykładzie platformy Elasticsearch. Analityka danych tekstowych w czasie rzeczywistym z wykorzystaniem platformy ElasticSearch. Podstawy przetwarzania danych z wykorzystaniem sieci splotowych (CNN, Convolutional Neural Networks). Biblioteka Keras oraz Tensorflow. Praca w środowisku chmurowym Google Colaboratory.

Text Mining: Rodzaje informacji w internecie. Wprowadzenie do tematyki Text Mining. Przeszukiwanie informacji tekstowych. Wstępne przetwarzanie dokumentów tekstowych: usuwanie zbędnych elementów z dokumentów tekstowych (stop lista, znaki interpunkcyjne, liczby itp.), sprowadzanie słów do postaci rdzenia znaczeniowego za pomocą algorytmu Portera oraz wybranych bibliotek informatycznych. Wyszukiwanie według słów kluczowych. Organizacja dokumentów w postaci macierzy term-dokument (ang. term-document matrix, TDM) oraz różne sposoby obliczania wagi poszczególnych termów (TF – term frequency, IDF – inverse document frequency). Miary podobieństwa wektorów oraz wykorzystanie ich do tworzenia rankingu wyszukanych dokumentów. Porównywanie jakości działania wyszukiwarek dokumentów tekstowych za pomocą różnych miar, np. precision-recall, krzywe ROC. Wybrane elementy algebry liniowej i zastosowanie ich do zadania aproksymacji macierzy TDM (ang. Low-rank approximation), omówienie korzyści z wykonanej aproksymacji. Różne techniki grupowania i klasyfikacji dokumentów. Ranking dokumentów oparty o strukturę połączeń: algorytm PageRank; autorytety i koncentratory. Tworzenie podsumowań dokumentów poprzez automatyczny wybór najważniejszych zdań oraz najważniejszych słów (termów). Tworzenie chmur słów (ang. wordclouds). Analiza sentymentu, jako technika badania wydźwięku dokumentów (np. pozytywny, negatywny, neutralny itp.). Omówienie wybranych narzędzi informatycznych do realizacji zadań z dziedziny Text Mining.

Metody kształcenia

Wykład konwencjonalny, ćwiczenia laboratoryjne, praca w zespole

Efekty uczenia się i metody weryfikacji osiągania efektów uczenia się

Opis efektu	Symbole efektów	Metody weryfikacji	Forma zajęć
Student potrafi samodzielnie rozwiązać postawiony problem wykorzystując do tego konwolucyjne sieci splotowe	• K_U16	• bieżąca kontrola na zajęciach • kolokwium	• Laboratorium
Student zna bazy danych określane mianem NoSQL, potrafi zdefiniować ich podstawowe cechy i rodzaje, potrafi używać najpopularniejsze systemy takich baz danych	• K_W09	• bieżąca kontrola na zajęciach • test egzaminacyjny z progami punktowymi	• Wykład • Laboratorium
Potrafi zdefiniować pojęcie Text Mining oraz podać typowe przykłady zadań z tego obszaru wiedzy	• K_W01	• test egzaminacyjny z progami punktowymi	• Wykład
Zna techniki wyszukiwania informacji tekstowych oraz tworzenia ich rankingu	• K_U16	• test egzaminacyjny z progami punktowymi	• Wykład
Potrafi wykonywać wybrane analizy statystyczne dokumentów tekstowych	• K_U06	• bieżąca kontrola na zajęciach	• Laboratorium
Student potrafi scharakteryzować cechy nowoczesnych platform Big Data.	• K_U16	• test egzaminacyjny z progami punktowymi	• Wykład
Student potrafi w praktyczny sposób wykorzystać narzędzia oferowane przez platformę Elasticsearch w celu przeprowadzenia zaawansowanej analizy danych oraz ich eksploracji w czasie rzeczywistym.	• K_U16	• bieżąca kontrola na zajęciach	• Laboratorium
Student zna możliwości konwolucyjnych sieci splotowych oraz potrafi zdefiniować zadania możliwe do rozwiązania tą techniką	• K_W08 • K_U16	• test egzaminacyjny z progami punktowymi	• Wykład

Warunki zaliczenia

Wykład - warunkiem zaliczenia jest uzyskanie pozytywnej oceny z egzaminu przeprowadzonego w formie zaproponowanej przez prowadzącego

Laboratorium - warunkiem zaliczenia jest uzyskanie ocen pozytywnych z wszystkich ćwiczeń laboratoryjnych oraz przeprowadzanych sprawdzianów

Składowe oceny końcowej = wykład: 50% + laboratorium: 50%

Literatura podstawowa

1. Larose D.T.: Metody i modele eksploracji danych, PWN, Warszawa, 2008
2. Markov Z., Larose D.T.: Eksploracja zasobów internetowych, PWN, Warszawa, 2009
3. Sadalage P. J., Fowler M.: NoSQL. Kompendium wiedzy, 2014
4. Gormley C., Tong Z.: Elasticsearch: The Definitive Guide, 2015
5. Francois Chollet: Deep Learning. Praca z językiem Python i biblioteką Keras, Helion, 2019
6. Dokumentacja systemu R

Literatura uzupełniająca

Uwagi

Zmodyfikowane przez dr hab. inż. Artur Gramacki, prof. UZ (ostatnia modyfikacja: 03-05-2021 22:28)