

Big data solutions - course description

General information	
Course name	Big data solutions
Course ID	11.3-WE-INFD-TechBD
Faculty	Faculty of Computer Science, Electrical Engineering and Automatics
Field of study	Computer Science
Education profile	academic
Level of studies	Second-cycle studies leading to MSc degree
Beginning semester	winter term 2019/2020

Course information	
Semester	2
ECTS credits to win	5
Course type	obligatory
Teaching language	polish
Author of syllabus	dr hab. inż. Artur Gramacki, prof. UZ

Classes forms					
The class form	Hours per semester (full-time)	Hours per week (full-time)	Hours per semester (part-time)	Hours per week (part-time)	Form of assignment
Lecture	30	2	18	1,2	Exam
Laboratory	30	2	18	1,2	Credit with grade

Aim of the course

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.

Prerequisites

Bazy danych.

Znajomość podstaw statystyki.

Scope

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.

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.

Teaching methods

Wykład konwencjonalny, ćwiczenia laboratoryjne, praca w zespole

Learning outcomes and methods of theirs verification

Outcome description	Outcome symbols	Methods of verification	The class form
Potrafi zdefiniować pojęcie Text Mining oraz podać typowe przykłady zadań z tego obszaru wiedzy	• K_W01	• an examination test with score scale	• Lecture
Zna techniki wyszukiwania informacji tekstowych oraz tworzenia ich rankingu	• K_U16	• an examination test with score scale	• Lecture
Potrafi wykonywać wybrane analizy statystyczne dokumentów tekstowych	• K_U06	• an ongoing monitoring during classes	• Laboratory
Student potrafi scharakteryzować cechy nowoczesnych platform Big Data.	• K_U16	• an examination test with score scale	• Lecture
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	• an ongoing monitoring during classes	• Laboratory
Student zna bazy danych określane mianem NoSQL, potrafi zdefiniować ich podstawowe cechy i rodzaje, potrafi używać najpopularniejsze systemy takich baz danych	• K_W09	• an examination test with score scale • an ongoing monitoring during classes	• Lecture • Laboratory

Assignment conditions

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%

Recommended reading

Larose D.T.: Metody i modele eksploracji danych, PWN, Warszawa, 2008

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

Gormley C., Tong Z.: Elasticsearch: The Definitive Guide, 2015

Dokumentacja systemu R

Further reading

Notes

Modified by dr hab. inż. Artur Gramacki, prof. UZ (last modification: 23-04-2019 00:14)

Generated automatically from SyllabUZ computer system