Politechnika Krakowska im. Tadeusza Kościuszki

Karta przedmiotu

obowiązuje studentów rozpoczynających studia w roku akademickim 2019/2020

Wydział Informatyki i Telekomunikacji

Kierunek studiów: Informatyka | Profil: Ogólnoakademicki

Forma sudiów: stacjonarne | Kod kierunku: I

Stopień studiów: II

Specjalności: Informatyka stosowana dla licencjatów

1 Informacje o przedmiocie

Nazwa przedmiotu	Masowe przetwarzanie danych
Nazwa przedmiotu w języku angielskim	Massive processing of data
Kod przedmiotu	WIiT I oIIS D10 19/20
Kategoria przedmiotu	Przedmioty specjalnościowe
Liczba punktów ECTS	3.00
Semestry	3

2 Rodzaj zajęć, liczba godzin w planie studiów

Semestr	Wykład	Ćwiczenia	Laboratorium	Laboratorium komputerowe	Seminarium	Projekt
3	15	0	30	0	0	0

3 Cele przedmiotu

Cel 1 Cel przedmiotu 1. Zapoznanie studentów z pojęciami, ważniejszymi algorytmami, gotowymi bibliotekami oraz metodami stosowanymi obecnie do przechowywania, przetwarzania, analizy, eksploracji i wizualizacji olbrzymich ilości danych.

Cel 2 Cel przedmiotu 2. Nabycie umiejętności zastosowania dostępnych gotowych bibliotek (tzw. pakietów) stosowanych do przechowywania, przetwarzania, analizy, eksploracji i wizualizacji olbrzymich ilości danych (zwłasz-

cza za pomocą internetowych notebooków i dashboardów udostepnionych przez RStudio dzieki bibliotekom rmarkdown i shiny), zarówno na komputerze, jak i online na istniejących darmowych kontach w Chmurze. Dodatkowo, studenci będą posiadali umiejętność poprawnej interpretacji wyników uzyskanych z algorytmów.

Cel 3 Cel przedmiotu 3. Nabycie umiejętności pracy w małych zespołach.

4 Wymagania wstępne w zakresie wiedzy, umiejętności i innych kompetencji

1 Wymaganie 1. Podstawy z Algebry liniowej

2 Wymaganie 2. Podstawy ze Statystyki matematycznej

3 Wymaganie 3. Podstawy z Rachunku prawdopodobieństwa

4 Wymaganie 4. Podstawowa znajomość języka angielskiego

5 Efekty kształcenia

EK1 Wiedza Efekt kształcenia 1. Student będzie potrafił wytłumaczyć pojęcia oraz stosowane obecnie metody służące do przechowywania, przetwarzania, analizy, eksploracji i wizualizacji olbrzymich ilości danych. Student będzie potrafił wytłumaczyć działanie oraz zinterpretować wyniki ważniejszych algorytmów stosowanych przy przetwarzaniu, analizie, eksploracji i wizualizacji dużych zbiorów danych.

EK2 Umiejętności Efekt kształcenia 2. Student będzie posiadał umiejętność zastosowania do przechowywania, przetwarzania, analizy, eksploracji i wizualizacji olbrzymich ilości danych ważniejszych algorytmów zawartych w wybranych bibliotekach (pakietach) języka R, środowiska zintegrowanego RStudio oraz w ekosystemie Apache SPARK (czyli w tzw. uniwersalnym silniku Big Data), zarówno na komputerze, jak i online na istniejących darmowych kontach w Chmurze.

EK3 Umiejętności Efekt kształcenia 3. Umiejętności: Student będzie potrafił poprzez RStudio łączyć się internetowo zarówno ze SPARKiem oraz z darmowymi serwerami RStudio, wspomagającymi przetwarzanie dużych zbiorów danych w Chmurze. Student będzie również posiadał umiejętność samodzielnego programowania w języku R w środowisku RStudio.

EK4 Kompetencje społeczne Efekt kształcenia 4. Student będzie umiał razem współpracować w małych zespołach oraz zauważy potrzebę samokształcenia i potrzebę ciągłego uaktualniania swej wiedzy.

6 Treści programowe

Laboratorium		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
L1	Treści programowe 1. Instalacja RStudio oraz R oraz ekosystemu Apache SPARK oraz podstawowych pakietów RStudio oraz R, takich jak sparklyr. Proste eksperymenty w środowisku RStudio z wykorzystaniem języka R i próby łączenia się z ekosystemem SPARK.	2
L2	Treści programowe 2. Instalacja bogatej rodziny tidyverse złożonej z kilkunastu nowoczesnych pakietów RStudio oraz eksperymentowanie z tymi pakietami wzorując się na przykładach i tutoriałach dostępnych online na stronach: http://r4ds.had.co.nz/ ; https://github.com/tidyverse ; http://tidyverse.org ; https://github.com/hadley/r4ds ; https://github.com/hadley/dplyr ; https://github.com/tidyverse/ggplot2	4

Laboratorium		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
L3	Treści programowe 3. Instalacja pakietu sparklyr, dplyr i magrittr, data.table, DBI i knitr oraz eksperymentowanie z bazodanowymi funkcjami tych pakietów: select(), filter(), arrange(), rename(), mutate(), group_by(), oraz z operatorem pipeline, tj. %>%.	2
L4	Treści programowe 4. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na tworzenie internetowych notebooków. Instalowane pakiety - to: magrittr, Rmarkdown i kilka innych mniejszych.	2
L5	Treści programowe 5. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na tworzenie internetowych dashboardów. Instalowany pakiet, to pakiet shiny. Eksperymenty będą bazowały na licznych przykładach zawartych w online tutoriałach dostępnych na stronach: http://www.rstudio.com/shiny/ ; http://rstudio.github.io/shiny/tutorial/ ; http://www.rstudio.com/shiny/lessons/Intro/	4
L6	Treści programowe 6. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na dokonywanie wzajemnych porównań (podobieństwa i odróżnialności) zbiorów tekstowych zawierających olbrzymie ilości danych. Instalowane pakiety - to: tm, lda, topicmodels, RTextTools, wordcloud, RTextTools, tau i kilka innych dodatkowych.	2
L7	Treści programowe 7. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na tworzenie grafów i sieci społecznościowych, bazując na zależnościach istniejących w dużych zbiorach danych. Instalowane pakiety - to: graphframe, igraph, rgl, snowfall, network, tmap, sna, SocialNetworks, tmap, spnet i kilka innych dodatkowych.	2
L8	Treści programowe 8. Instalacja i eksperymentowanie z pakietami języka R - pozwalającymi na tworzenie rekomenderów: recommenderlab, arules, readr, plyr.	2
L9	Treści programowe 9. Założenie darmowego konta w Chmurze (oferowanego przez twórców Apache Spark -Databricks) lub przez środowisko programistyczne RStudio. Przegląd tutoriałów z przykładowymi eksperymentami w chmurze i ich naśladowanie. Importowanie Notebooków w Chmurze (z adresu podanego w Literaturze) i ich wykonywanie.	2
L10	Treści programowe 10. Instalacja i eksperymentowanie z pakietem ISLR. Eksperymenty będą bazowały na licznych przykładach zawartych w licznych online lablatoryjnych przykładach i tutoriałach dostępnych na stronie, gdzie są wszystkie kody źródłowe z książki [5]: http://www-bcf.usc.edu/gareth/ISL/code.html); Książka dostępna jest pod adresem: http://www-bcf.usc.edu/gareth/ISL Dodatkowo dużo przykładów autora książki jest na jego stronie: http://web.stanford.edu/hastie/StatLearnSparsity/	2
L11	Treści programowe 11. Eksperymenty w RStudio z wykorzystaniem platformy H2O i ważniejszych pakietów związanych z tą platformą, tj. pakietem h2o, rsparkling, data.table, dplyr, itp.	2

Laboratorium		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
L12	Treści programowe 12. Eksploracja danych strumieniowych (mining data streams); eksperymentowanie z pakietem stream w języku R, do przetwarzania strumieniowego dużych zbiorów danych.	2
L13	Treści programowe 13. Eksperymentowanie z dodatkowymi, wybranymi pakietami: RAdwords (Web Advertising), manipulate, dpmr, itp.Skorzystanie z alternatywnej do pakietu "sparklyr" (z RStudio) możliwości połączenia R'a ze Sparkiem - tj. pakietu SparkR, jako "frontend'u" prowadzącego z R'a do SPARK'a. Eksperymentowanie z operacjami na ramkach danych (tj. DataFrames), stosowanie operacji: select, filter, aggregate, korzystanie z operacji na kolumnach, funkcji collect() i subset()).	2

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W1	Treści programowe 1. Charakterystyka pojęcia "Big data" i sumaryczne omówienie zagadnień związanych z tym pojęciem. Wprowadzenie do paradygmatu MapReduce. Przedstawienie i opisanie architektury i funkcji, jakie spełnia ekosystem Apache SPARK i jego podsystemy. Wprowadzenie do środowiska RStudio.Opisanie pakietu sparklyr (autorstwa twórców środowiska RStudio), pozwalającego na połączenie się z programu w języku R ze SPARK'iem oraz pozwalającego na wykonywanie dowolnego pakietu R'a w środowisku skalowalnym i rozproszonym. Omówienie pojęcia ramki danych (tj. DataFrame) oraz ważniejszych operacji na dużych zbiorach danych. Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem pakietu sparklyr.	2
W2	Treści programowe 2. Opisanie środowiska RStudio, jego ważniejszych pakietów, ich funkcji i operacji, oraz zalet i metody pracy w tym środowisku. Omówienie metod tworzenia notebook'ów - korzystając z pakietów RStudio: magrittr, markdown, R markdown oraz dplyr.	2
W3	Treści programowe 3. Omówienie metod tworzenia dashboard'ów - korzystając z pakietów RStudio: magrittr, markdown, R markdown oraz shiny.	1
W4	Treści programowe 4. Omówienie ważniejszych operacji bazodanowych udostępnionych przez funkcje ważniejszych pakietów bazo-danowych środowiska RStudio, np. operacji select, filter, aggregate, operacji na kolumnach, funkcji collect() oraz subset().	1
W5	Treści programowe 5. Opisanie ważniejszych pakietów R'a służących do eksploracji danych tekstowych (text mining): tm, lda, topicmodels, RTextTools, tau, wordcloud. Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem tych pakietów.	1

Wykład		
Lp	Tematyka zajęć Opis szczegółowy bloków tematycznych	Liczba godzin
W6	Treści programowe 6. Omówienie zasady działania systemów rekomendacji. Opisanie ważniejszych pakietów R'a służących jako rekomendery. Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem pakietów: recommenderlab, arules, readr, plyr.	2
W7	Treści programowe 7. Opisanie ważniejszych pakietów R'a służących do eksploracji sieci społecznościowych (social networks). Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem pakietów: sna, network, igraph, SocialNetworks, tmap, spnet.	1
W8	Treści programowe 8. Opisanie ważniejszych pakietów R'a służących do eksploracji danych strumieniowych (mining data streams). Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem pakietu stream, do przetwarzania strumieniowego dużych zbiorów danych.	1
W9	Treści programowe 9. Opisanie ważniejszego pakietu R'a służącego do tworzenia webowych ogłoszeń (Web Advertising) - pakiet RAdwords. Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem tego pakietu, oraz omówienie pakietów: manipulate oraz dpmr.	1
W10	Treści programowe 10. Opisanie platformy H2O i ważniejszych pakietów związanych z tą platformą, tj. pakietem h2o, rsparkling, data.table, dplyr, itp. Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem tych pakietów.	1
W11	Treści programowe 11. Omówienie zasad i reguł potrzebnych do eksperymentowania w Chmurze na dużych zbiorach danych.Omówienie metody tworzenia i importowania notebook'ów w Chmurze.	1
W12	Treści programowe 12. Omówienie pakietu SparkR - jako alternatywnej opcji dawanej przez pakiet "sparklyr" (z RStudio) i umożliwiającej połączenie R'a ze Sparkiem i pozwalającej na wykonywanie obliczeń w języku R w Chmurze w ekosystemie SPARK.Przedstawienie i opisanie przykładów skryptów napisanych z wykorzystaniem pakietu SparkR.	1

7 Narzędzia dydaktyczne

N1 Narzędzie 1. Wykłady

N2 Narzędzie 2. Ćwiczenia laboratoryjne

N3 Narzędzie 3. Prezentacje multimedialne

N4 Narzędzie 4. Konsultacje

N5 Narzędzie 5. Dyskusja

N6 Narzędzie 6. Praca w 2-3 osobowych grupkach

8 Obciążenie pracą studenta

Forma aktywności	Średnia liczba godzin na zrealizowanie aktywności
Godziny kontaktowe z nauczycielem akademickim, w tym:	
Godziny wynikające z planu studiów	45
Konsultacje przedmiotowe	5
Egzaminy i zaliczenia w sesji	0
Godziny bez udziału nauczyciela akademickiego wynikające z nakładu pracy studenta, w tym:	
Przygotowanie się do zajęć, w tym studiowanie zalecanej literatury	30
Opracowanie wyników	10
Przygotowanie raportu, projektu, prezentacji, dyskusji	5
Sumaryczna liczba godzin dla przedmiotu wynikająca z całego nakładu pracy studenta	**95**
Sumaryczna liczba punktów ECTS dla przedmiotu	3.00

9 Sposoby oceny

Ocena formująca

F1 Ocena 1. Ćwiczenia praktyczne

F2 Ocena 2. Odpowiedzi ustne

F3 Ocena 3. Sprawozdania z umiejętności wykorzystania wybranych bibliotek języka R wywoływanych w środowisku RStudio i Apache SPARK

Ocena podsumowująca

P1 Ocena 1. Średnia ważona ocen formujących

Warunki zaliczenia przedmiotu

W1 Ocena 1. Warunkiem otrzymania zaliczenia z przedmiotu Masowe przetwarzanie danych" jest uzyskanie wystarczającej liczby punktów za aktywnie wykonywane ćwiczenia praktyczne w klasie, za odpowiedzi ustne podczas zajęć laboratoryjnych oraz za oddanie wszystkich zleconych do napisania sprawozdań wykonanych w środowisku programistycznym RStudio i Apache SPARK, z wykorzystaniem wspomnianych bibliotek. Wszystko to będzie punktowane.

Ocena aktywności bez udziału nauczyciela

B1 Ocena 1. Oddanie wszystkich zleconych do napisania sprawozdań wykonanych w środowisku programistycznym RStudio i Apache SPARK. Wszystkie sprawozdania będą punktowane.

B2 Ocena 2. Ocena za odpowiedzi ustne podczas zajęć.

B3 Ocena 3. Ocena za aktywność podczas wykonywania ćwiczeń praktycznych w klasie.

Kryteria oceny

Efekt kształcenia 1	
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%
Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%
Efekt kształcenia 2	
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%.
Efekt kształcenia 3	
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%.
Efekt kształcenia 4	
Na ocenę 3.0	Opanowanie zagadnienia w stopniu powyżej 50%.
Na ocenę 3.5	Opanowanie zagadnienia w stopniu powyżej 60%.
Na ocenę 4.0	Opanowanie zagadnienia w stopniu powyżej 70%.
Na ocenę 4.5	Opanowanie zagadnienia w stopniu powyżej 80%.
Na ocenę 5.0	Opanowanie zagadnienia w stopniu powyżej 90%.

10 Macierz realizacji przedmiotu

Efekt kształcenia	Odniesienie danego efektu do szczegółowych efektów zdefiniowanych dla programu	Cele przedmiotu	Treści programowe	Narzędzia dydaktyczne	Sposoby oceny
EK1	I2_W01 I2_W02 I2_W03 I2_W05 I2_W06	Cel 1	W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12	N1 N2 N5 N6	F1 F2 F3 P1
EK2	I2_W01 I2_W02 I2_W03 I2_W05 I2_W06 I2_U01b I2_U03b I2_U04b I2_U05 I2_U06 I2_U07 I2_U08 I2_U11 I2_U12	Cel 2	L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 L11 L12 L13 W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12	N1 N2 N3 N4 N5 N6	F1 F2 F3 P1
EK3	I2_W01 I2_W02 I2_W03 I2_W06 I2_U01b I2_U02b I2_U03b I2_U04b I2_U05 I2_U06 I2_U07 I2_U08 I2_U12	Cel 2	L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 L11 L12 L13 W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12	N1 N2 N3 N4 N5 N6	F1 F2 F3
EK4	I2_K01 I2_K02 I2_K03 I2_K04	Cel 3	L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 L11 L12 L13	N2 N3 N4 N5 N6	F1 F2 F3

11 Wykaz literatury

Literatura podstawowa

[1] Garrett Grolemund, Hadley Wickham, R for Data Science, 2016, OReilly, książka dostępna w Internecie na str.: http://r4ds.had.co.nz/ ; dodatkowe tutoriały i kody są dostępne na stronach: https://github.com/tidyverse http://tidyverse.org https://github.com/hadley/r4ds https://github.com/hadley/dplyr https://github.com/tidyverse/ggp

[2] Y. Zhao, "R and Data Mining: Examples and Case Studies", 2014, książka dostępna w WWW z licznymi materiałami, np. książką: http://www.rdatamining.com/docs/introduction-to-data-mining-with-r oraz przykłady w R: http://www.rdatamining.com/examples oraz wiele inn. wartościowych plików: http://www.rdatamining.com)

[3] Liczne przykłady gotowych internetowych Notebooków wraz z tutoriałami i możliwość tworzenia nowych (swoich) online w środowisku programistycznym RStudio: https://rnotebook.io/, https://rstudio.cloud/ , https://spark.rstudio.c oraz pod adresem: https://rpubs.com (lub w ekosystemie SPARK w Chmurze z wykorzystaniem języka R: https://docs.databricks.com/spark/latest/training/index.html)

[4] Online dokumentacja ekosystemu Apache SPARK w Chmurze (stale przez autorów uaktualniana): http://spark.apache.org/do (oraz online tutoriał pakietu R'a SparkR, pozwalający na bezpośrednio z R'a łączenie się z Apache SPARK (bez udziału RStudio): http://spark.apache.org/docs/latest/sparkr.html)

[4]] R.D. Peng, "R programming for data science", Publ. by Leanpub, 2015, książka dostępna w WWW pod adr.: http://www.cs.upc.edu/robert/teaching/estadistica/rprogramming.pdf

[5] Online tutoriały tworzenia Notebooków w ekosystemie Apache SPARK w Chmurze (stale przez autorów uaktualniane): https://docs.databricks.com/spark/latest/training/index.html

Literatura uzupełniająca

[1] G. James, D. Witten, Trevor Hastie, Robert Tibshirani, An Introduction to Statistical Learning with Applications in R, Springer Series in Statistics, 2016, Stanford, CA, książka dostępna w Internecie na stronie: http://www-bcf.usc.edu/gareth/ISL (a wszystkie kody źródłowe z książki są dostępne na stronie: http://www-bcf.usc.edu/gareth/ISL/code.html); dużo przykładów autora książki jest na jego stronie: http://web.stanford.edu/hastie/Sta

[2] Materiały w wersji elektronicznej dostarczone studentom na pierwszych laboratoriach.

12 Informacje o nauczycielach akademickich

Osoba odpowiedzialna za kartę

dr Barbara Borowik (kontakt: `bborowik@pk.edu.pl`)

Osoby prowadzące przedmiot

1 Dr Barbara Borowik (kontakt: `bborowik@pk.edu.pl`)

13 Zatwierdzenie karty przedmiotu do realizacji

(miejscowość, data) (odpowiedzialny za przedmiot) (dziekan)

Przyjmuję do realizacji (data i podpisy osób prowadzących przedmiot)

. . . . . .