

POLITECHNIKA KRAKOWSKA IM. TADEUSZA KOŚCIUSZKI

KARTA PRZEDMIOTU

obowiązuje studentów rozpoczynających studia w roku akademickim 2017/2018

Wydział Fizyki, Matematyki i Informatyki

Kierunek studiów: Informatyka

Profil: Ogólnoakademicki

Forma studiów: stacjonarne

Kod kierunku: I

Stopień studiów: II

Specjalności: Informatyka stosowana dla inżynierów

1 INFORMACJE O PRZEDMIOCIE

NAZWA PRZEDMIOTU	Masowe przetwarzanie danych
NAZWA PRZEDMIOTU W JĘZYKU ANGIELSKIM	Massive processing of data
KOD PRZEDMIOTU	WFMiI I oIIS D9 17/18
KATEGORIA PRZEDMIOTU	Przedmioty specjalnościowe
LICZBA PUNKTÓW ECTS	4.00
SEMESTRY	3

2 RODZAJ ZAJĘĆ, LICZBA GODZIN W PLANIE STUDIÓW

SEMESTR	WYKŁAD	ĆWICZENIA	LABORATORIUM	LABORATORIUM KOMPUTERO- WE	SEMINARIUM	PROJEKT
3	15	0	15	0	0	0

3 CELE PRZEDMIOTU

Cel 1 Zapoznanie studentów z pojęciami, ważniejszymi algorytmami, gotowymi bibliotekami oraz metodami stosowanymi obecnie w eksploracji dużych zbiorów danych.

Cel 2 Nabycie umiejętności zastosowania dostępnych gotowych bibliotek (tzw. pakietów) eksplorujących masowe zbiory danych zarówno na komputerze, jak i na istniejących darmowych kontach w Chmurze.

Cel 3 Nabycie umiejętności pracy w małych zespołach.

4 WYMAGANIA WSTĘPNE W ZAKRESIE WIEDZY, UMIEJĘTNOŚCI I INNYCH KOMPETENCJI

- 1 Podstawy z Algebry liniowej.
- 2 Podstawy ze Statystyki matematycznej.
- 3 Podstawy z Rachunku prawdopodobieństwa.
- 4 Podstawy z języka R (podstawy te łatwo można nadrobić samemu na początku semestru).

5 EFEKTY KSZTAŁCENIA

EK1 Wiedza Student będzie rozumiał i będzie potrafił wytłumaczyć pojęcia, zasadę działania ważniejszych algorytmów oraz metody stosowane obecnie przy eksploracji dużych zbiorów danych.

EK2 Wiedza Student będzie znał metodę implementacji ważniejszych algorytmów stosowanych przy eksploracji dużych zbiorów danych na komputerze oraz w Chmurze oraz będzie wiedział, jakie są podstawowe zasady przy eksperymentowaniu w Chmurze.

EK3 Umiejętności Student będzie potrafił praktycznie zastosować posiadane informacje i nabędzie umiejętność zastosowania dostępnych gotowych bibliotek lub będzie potrafił samemu zaimplementować algorytmy eksplorujące masowe zbiory danych, zarówno na komputerze, jak i na istniejących darmowych kontach w Chmurze.

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

6 TREŚCI PROGRAMOWE

WYKŁAD		
LP	TEMATYKA ZAJĘĆ OPIS SZCZEGÓŁOWY BŁOKÓW TEMATYCZNYCH	LICZBA GODZIN
W1	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.	1
W2	Opisanie pakietu sparklyr (autorstwa twórców środowiska RStudio), pozwalającego na połączenie się z programem 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.	1
W3	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.	1
W4	Omówienie metod tworzenia dashboard'ów - korzystając z pakietów RStudio: magrittr, markdown, R markdown oraz shiny.	1

WYKŁAD		
LP	TEMATYKA ZAJĘĆ OPIS SZCZEGÓŁOWY BLOKÓW TEMATYCZNYCH	LICZBA GODZIN
W5	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
W6	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.	2
W7	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.	2
W8	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
W9	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
W10	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
W11	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.	1
W12	Omówienie zasad i reguł potrzebnych do eksperymentowania w Chmurze na dużych zbiorach danych. Omówienie metody tworzenia i importowania notebook'ów w Chmurze.	1
W13	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

LABORATORIUM		
LP	TEMATYKA ZAJĘĆ OPIS SZCZEGÓŁOWY BLOKÓW TEMATYCZNYCH	LICZBA GODZIN
L1	Nauczenie się korzystania z RStudio podczas pracy z językiem R. Poznanie ważniejszych nowszych pakietów w języku R utworzonych przez RStudio.	1

LABORATORIUM		
LP	TEMATYKA ZAJĘĆ OPIS SZCZEGÓŁOWY BLOKÓW TEMATYCZNYCH	LICZBA GODZIN
L2	Korzystanie z bogatego asortymentu funkcji dostępnych w pakiecie sparklyr, pozwalającego na połączenie platformy programistycznej R'a ze SPARK'iem.	1
L3	Tworzenie notebook'ów - korzystając z pakietów RStudio: magrittr, markdown, R markdown oraz dplyr. Eksperymenty z ważniejszymi pakietami RStudio i R działającymi na dużych zbiorach danych: z pakietem data.table, knitr, rsparkling.	1
L4	Tworzenie dashboard'ów - korzystając z pakietów RStudio: magrittr, markdown, R markdown oraz shiny.	1
L5	Eksperymenty bazodanowe z ważniejszymi pakietami RStudio i R działającymi na dużych zbiorach danych: z pakietem magrittr, data.table, sparklyr, DBI, knitr, dplyr.	1
L6	Eksperymenty w RStudio z wykorzystaniem platformy H2O i ważniejszych pakietów związanych z tą platformą, tj. pakietem h2o, rsparkling, data.table, dplyr, itp.	1
L7	Eksploracja danych tekstowych (text mining); eksperymentowanie z pakietami w języku R: tm, lda, topicmodels, RTextTools, tau, wordcloud, itp.	2
L8	Eksploracja danych tekstowych (text mining); sieci społecznościowe; eksperymentowanie z pakietami: sna, network, igraph, SocialNetworks, tmap, spnet.	2
L9	Eksploracja danych strumieniowych (mining data streams); eksperymentowanie z pakietem stream w języku R, do przetwarzania strumieniowego dużych zbiorów danych.	1
L10	Eksperymentowanie z dodatkowymi, wybranymi pakietami: RAdwords (Web Advertising), manipulate, dpmr, itp.	1
L11	Eksperymentowanie z ważniejszymi pakietami R'a służącymi jako rekomendery: recommenderlab, arules, readr, plyr.	1
L12	Założenie darmowego konta w Chmurze (oferowanego przez twórców Apache Spark - Databricks). Przegląd tutorialów z przykładowymi eksperymentami w chmurze i ich naśladowanie. Importowanie Notebooków w Chmurze (z adresu podanego w Literaturze) i ich wykonywanie.	1
L13	Korzystanie 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()).	1

7 NARZĘDZIA DYDAKTYCZNE

N1 Wykłady

N2 Ćwiczenia laboratoryjne

N3 Prezentacje multimedialne

N4 Konsultacje

N5 Dyskusja

N6 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	30
Konsultacje przedmiotowe	8
Egzaminy i zaliczenia w sesji	2
Godziny bez udziału nauczyciela akademickiego wynikające z nakładu pracy studenta, w tym:	
Przygotowanie się do zajęć, w tym studiowanie zalecanej literatury	20
Opracowanie wyników	15
Przygotowanie raportu, projektu, prezentacji, dyskusji	15
implementacja wybranych algorytmów	30
SUMARYCZNA LICZBA GODZIN DLA PRZEDMIOTU WYNIKAJĄCA Z CAŁEGO NAKŁADU PRACY STUDENTA	120
SUMARYCZNA LICZBA PUNKTÓW ECTS DLA PRZEDMIOTU	4.00

9 SPOSOBY OCENY

OCENA FORMUJĄCA

F1 Ocena 1. Sprawozdania z ćwiczeń laboratoryjnych

F2 Ocena 2. Ćwiczenia praktyczne

F3 Ocena 3. Odpowiedzi ustne

F4 Ocena 4. Kolokwium

OCENA PODSUMOWUJĄCA

P1 Ocena 1. Średnia ważona ocen formujących

WARUNKI ZALICZENIA PRZEDMIOTU

W1 Zaliczenie uzyskają ci studenci, którzy uzyskają wystarczającą liczbę punktów za: a) przysłane sprawozdania z ćwiczeń laboratoryjnych; b) aktywność podczas wykonywania ćwiczeń praktycznych w klasie; c) odpowiedzi ustne podczas zajęć; oraz za d) poprawne napisanie wcześniej zapowiedzianego kolokwium.

OCENA AKTYWNOŚCI BEZ UDZIAŁU NAUCZYCIELA

B1 Ocena przysłanych sprawozdań z zadanych wcześniej prac do wykonania w domu.

KRYTERIA OCENY

EFEKT KSZTAŁCENIA 1	
NA OCENĘ 3.0	Opanowanie zagadnień związanych z Efektem kształcenia 1 i poruszanych na wykładach w stopniu powyżej 50%.
NA OCENĘ 4.0	Opanowanie zagadnień związanych z Efektem kształcenia 1 i poruszanych na wykładach w stopniu powyżej 70%.
NA OCENĘ 5.0	Opanowanie zagadnień związanych z Efektem kształcenia 1 i poruszanych na wykładach w stopniu powyżej 90%.
EFEKT KSZTAŁCENIA 2	
NA OCENĘ 3.0	Opanowanie zagadnień związanych z Efektem kształcenia 2 i poruszanych na wykładach w stopniu powyżej 50%.
NA OCENĘ 4.0	Opanowanie zagadnień związanych z Efektem kształcenia 2 i poruszanych na wykładach w stopniu powyżej 70%.
NA OCENĘ 5.0	Opanowanie zagadnień związanych z Efektem kształcenia 2 i poruszanych na wykładach w stopniu powyżej 90%.
EFEKT KSZTAŁCENIA 3	
NA OCENĘ 3.0	Posiadanie umiejętności opisanych w Efekte kształcenia nr 3 w stopniu powyżej 50%.
NA OCENĘ 4.0	Posiadanie umiejętności opisanych w Efekte kształcenia nr 3 w stopniu powyżej 70%.
NA OCENĘ 5.0	Posiadanie umiejętności opisanych w Efekte kształcenia nr 3 w stopniu powyżej 90%.
EFEKT KSZTAŁCENIA 4	
NA OCENĘ 3.0	Posiadanie kompetencji opisanych w Efekte kształcenia nr 4 w stopniu powyżej 50%.
NA OCENĘ 4.0	Posiadanie kompetencji opisanych w Efekte kształcenia nr 4 w stopniu powyżej 70%.
NA OCENĘ 5.0	Posiadanie kompetencji opisanych w Efekte kształcenia nr 4 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_W05 I2_W06	Cel 1	W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13	N1 N2 N3 N4 N5 N6	F3 F4 P1
EK2	I2_W01 I2_W02 I2_W03 I2_W05 I2_W06	Cel 1	W1 W2 W3 W4 W5 W6 W7 W8 W9 W10 W11 W12 W13	N1 N2 N3 N4 N5 N6	F3 F4 P1
EK3	I2_U01 I2_U02 I2_U03 I2_U04 I2_U05 I2_U06 I2_U07 I2_U10 I2_U11	Cel 2	L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 L11 L12 L13	N1 N2 N3 N4 N5 N6	F1 F2 F3 F4
EK4	I2_U01 I2_U02 I2_U03 I2_U04 I2_U05 I2_U06 I2_U07 I2_U10 I2_U11 I2_K01 I2_K02 I2_K03 I2_K04	Cel 3	L1 L2 L3 L4 L5 L6 L7 L8 L9 L10 L11 L12 L13	N1 N2 N3 N4 N5 N6	F1 F2 F3 F4 P1

11 WYKAZ LITERATURY

LITERATURA PODSTAWOWA

- [1] Y. Zhao — *"R and Data Mining: Examples and Case Studies"*, książka dostępna w WWW z licznymi materiałami, 2014, <http://www.rdatamining.com/docs/introduction-to-data-mining-with-r>
- [2] R.D. Peng — *"R programming for data science"*, <http://www.cs.upc.edu/robert/teaching/estadistica/rprogramming.pdf>, 2015, Publ. by Leanpub, książka dostępna w WWW
- [3] **Liczne przykłady gotowych Notebooków w ekosystemie SPARK w Chmurze** — *z wykorzystaniem języka R i środowiska RStudio*, <http://spark.rstudio.com/examples.html>, 2017, oraz <http://spark.rstudio.com/>
- [4] **Online tutoriały tworzenia Notebooków w ekosystemie Apache SPARK w Chmurze** — *(stale przez autorów uaktualniane)*, <https://docs.databricks.com/spark/latest/training/index.html>, 2017, oraz <http://spark.apache.org/docs/latest/>

- [5] Materiały w wersji elektronicznej — *dostarczone studentom na pierwszych laboratoriach.*, , 0,
- [6] Liczne przykłady gotowych Notebooków — *w ekosystemie SPARK w Chmurze z wykorzystaniem języka R i środowiska RStudio*; <http://spark.rstudio.com/examples.html>, 2017, oraz <http://spark.rstudio.com/>

12 INFORMACJE O NAUCZYCIELACH AKADEMICKICH

OSOBA ODPOWIEDZIALNA ZA KARTĘ

dr Barbara Borowik (kontakt: cnborowi@cyf-kr.edu.pl)

13 ZATWIERDZENIE KARTY PRZEDMIOTU DO REALIZACJI

(miejscowość, data)

(odpowiedzialny za przedmiot)

(dziekan)