**Nazwa przedmiotu:**

Przetwarzanie danych w językach R i Python

**Koordynator przedmiotu:**

Dr inż. Marek Gągolewski

**Status przedmiotu:**

Obowiązkowy

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Wspólne

**Kod przedmiotu:**

1120-INPAD-MSP-0111

**Semestr nominalny:**

1 / rok ak. 2017/2018

**Liczba punktów ECTS:**

5

**Liczba godzin pracy studenta związanych z osiągnięciem efektów uczenia się:**

.

**Liczba punktów ECTS na zajęciach wymagających bezpośredniego udziału nauczycieli akademickich:**

.

**Język prowadzenia zajęć:**

polski

**Liczba punktów ECTS, którą student uzyskuje w ramach zajęć o charakterze praktycznym:**

.

**Formy zajęć i ich wymiar w semestrze:**

|  |  |
| --- | --- |
| Wykład: | 0h |
| Ćwiczenia: | 0h |
| Laboratorium: | 30h |
| Projekt: | 0h |
| Lekcje komputerowe: | 0h |

**Wymagania wstępne:**

Znajomość zagadnień programowania strukturalnego i obiektowego, algorytmów i struktur danych oraz programowania równoległego i rozproszonego

**Limit liczby studentów:**

.

**Cel przedmiotu:**

Kurs omawia najbardziej przydatne techniki programowania w językach R i Python 3 w pracy analityka danych. W ramach zajęć student uzyskuje praktyczną umiejętność zbierania i przygotowywania danych do analizy, oraz stosowania i implementacji wybranych algorytmów maszynowego uczenia się. Uczestnicy mają możliwość dogłębnego poznania technik programowania w tych językach, a także najbardziej popularnych i użytecznych bibliotek. W trakcie zajęć szczególną uwagę poświęca się sposobom przetwarzania dużych zbiorów danych (filtrowanie danych, agregacja danych w podgrupach, złączenia) przy użyciu Apache Spark.

**Treści kształcenia:**

Wykład:
Podstawowe atomowe typy danych w języku R.
Działania na wektorach w R.
Implementacja wybranych algorytmów maszynowego uczenia się przy użyciu tzw. wektoryzacji.
Typ podstawowy lista. Funkcje. Atrybuty obiektów. Podstawy programowania obiektowego w stylu S3. Typy złożone w R: macierz, czynnik, ramka danych.
Działania na ramkach danych (filtrowanie danych, złączenia, agregacja danych w podgrupach).
Instrukcja sterująca i pętle. Testy jednostkowe, profilowanie wydajności kodu.
Przetwarzanie napisów przy użyciu wyrażeń regularnych, obiektów typu data i czas oraz plików.
Środowiska. Leniwa ewaluacja. Niestandardowa ewaluacja. Środowiskowy model obliczeń. Dynamiczny a leksykalny zasięg zmiennych. Programowanie obiektowe w stylu S4, RefClasses i R6.
Podstawy programowania w języku Python 3. Typy skalarne i sekwencyjne, iteratory.
Słowniki, zbiory. Funkcje, instrukcje sterujące.
Obliczenia na wektorach, macierzach i innych tablicach (NumPy).
Przygotowywanie danych do analizy, praca z ramkami danych (Pandas).
Przetwarzanie napisów i plików, serializacja obiektów, dostęp do baz danych SQL. Ekstrakcja danych ze stron WWW, parsowanie plików XML.
Cython i Rcpp – tworzenie modułów/pakietów rozszerzających przy użyciu C++.
Programowanie rozwiązań wsadowych przy użyciu Apache Spark – wprowadzenie.
PySpark – wprowadzenie do interfejsu Apache Spark dla języka Python. Podstawowe operacje: scalanie zbiorów danych na podstawie wartości klucza, agregacja potencjalnie dużych zbiorów danych.
Laboratorium:
Laboratorium obejmuje praktyczne zastosowanie wiedzy zdobytej na wykładzie oraz rozwój umiejętności jej użycia w problemach analizy danych i algorytmach maszynowego uczenia się.

**Metody oceny:**

Na zaliczenie składają się oceny zdobyte za rozwiązanie 7 prac domowych (3 po 20 p., 4 po 10 p.). Na zaliczenie konieczne jest uzyskanie łącznie ponad 50 p. na 100 p. możliwych. Ostateczna ocena z przedmiotu wynika z sumy zdobytych punktów: [0, 50] – 2,0; (50, 60] – 3,0; (60, 70] – 3,5; (70, 80] – 4,0; (80, 90] – 4,5; (90, 100] – 5,0.

**Egzamin:**

nie

**Literatura:**

1. M. Gągolewski, M. Bartoszuk, A. Cena, Przetwarzanie i analiza danych w języku Python, PWN, Warszawa, 2016 (w druku)
2. M. Gągolewski, Programowanie w języku R. Analiza danych, obliczenia, symulacje, Wydawnictwo Naukowe PWN, 2016 (wyd. II)
3. W. McKinney, Python for Data Analysis. Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012
4. W. Richert, L.P. Coelho, Building Machine Learning Systems with Python, Packt Publishing, 2013
5. E. Bressert, SciPy and NumPy, O'Reilly Media, 2012
6. J.M. Chambers., Programming with Data, Springer, 1998
7. J.M. Chambers, Software for Data Analysis. Programming with R, Springer, 2008
8. N.S. Matloff, The Art of R Programming: A Tour of Statistical Software Design, No Starch Press, 2011
9. W.N. Venables, B.D. Ripley, S Programming, Springer, 2000
10. H. Wickham, Advanced R, Chapman&Hall/CRC
11. M. Hamstra, Learning Spark, O'Reilly, 2013
12. M. Frampton, Mastering Apache Spark, Packt, 2015
13. R 3.3, RStudio, gcc, g++, Python 3.5 (Anaconda), Cython, Jupyter
14. http://www.gagolewski.com/teaching

**Witryna www przedmiotu:**

e.mini.pw.edu.pl

**Uwagi:**

.

## Efekty przedmiotowe

### Profil ogólnoakademicki - wiedza

**Efekt W2\_01:**

Zna kluczowe języki programowania wykorzystywane w analizie danych – R i Python

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_W13

**Powiązane efekty obszarowe:**

**Efekt W2\_02:**

Zna metody filtrowania, czyszczenia, podsumowywania i łączenia zbiorów danych

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_W05, PD\_W09, PD\_W15

**Powiązane efekty obszarowe:** , ,

### Profil ogólnoakademicki - umiejętności

**Efekt U2\_01:**

Potrafi projektować wydajne metody przetwarzania i analizy danych

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_U09

**Powiązane efekty obszarowe:**

**Efekt U2\_02:**

Potrafi wykorzystać możliwości równoległego przetwarzania danych do rozwiązywania złożonych problemów

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_U11

**Powiązane efekty obszarowe:**

**Efekt U2\_03:**

Potrafi zaprojektować system wykorzystujący metody klasyfikacji i regresji

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_U12

**Powiązane efekty obszarowe:**

**Efekt U2\_04 :**

Dostrzega ograniczenia i słabe strony istniejących narzędzi informatycznych

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_U17

**Powiązane efekty obszarowe:**

### Profil ogólnoakademicki - kompetencje społeczne

**Efekt K2\_01:**

Posiada zdolność do kontynuacji kształcenia oraz świadomość potrzeby samokształcenia w ramach procesu kształcenia ustawicznego

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_K01

**Powiązane efekty obszarowe:**

**Efekt K2\_02:**

Potrafi posługiwać się językiem angielskim w stopniu umożliwiającym bezproblemową komunikację w zakresie zagadnień zawodowych oraz studiowanie dokumentacji technicznej

Weryfikacja:

ocena prac domowych

**Powiązane efekty kierunkowe:** PD\_K08

**Powiązane efekty obszarowe:**