**Nazwa przedmiotu:**

Przetwarzanie danych ustrukturyzowanych

**Koordynator przedmiotu:**

dr hab. inż. Marek Gągolewski, prof. PW

**Status przedmiotu:**

Obowiązkowy

**Poziom kształcenia:**

Studia I stopnia

**Program:**

Inżynieria i Analiza Danych

**Grupa przedmiotów:**

Wspólne

**Kod przedmiotu:**

1120-DS000-ISP-0124

**Semestr nominalny:**

2 / rok ak. 2017/2018

**Liczba punktów ECTS:**

4

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

1. godziny kontaktowe – 65 h; w tym
 a) obecność na wykładach – 30 h
 b) obecność na laboratoriach – 30 h
 c) konsultacje – 5 h
2. praca własna studenta – 50 h; w tym
 a) przygotowanie do laboratoriów – 20 h
 b) zapoznanie się z literaturą – 5 h
 c) przygotowanie prac domowych – 25 h
Razem 115 h, co odpowiada 4 pkt. ECTS

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

1. obecność na wykładach – 30 h
2. obecność na laboratoriach – 30 h
3. konsultacje – 5 h
Razem 65 h, co odpowiada 2 pkt. ECTS

**Język prowadzenia zajęć:**

polski

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

obecność na laboratoriach – 30 h
przygotowanie do laboratoriów – 20 h
przygotowanie prac domowych – 25 h
Razem 75 h, co odpowiada 3 pkt. ECTS

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

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

**Wymagania wstępne:**

Podstawy programowania i przetwarzania danych

**Limit liczby studentów:**

.

**Cel przedmiotu:**

Celem przedmiotu jest przegląd najczęściej stosowanych w praktyce metod przetwarzania danych i przygotowywania ich do analizy. Szczególny nacisk położony jest na omówienie i ćwiczenie technik filtrowania, przekształcania i agregacji zmiennych lub całych zbiorów danych, także w podgrupach. Uczestnicy kursu poznają najbardziej podstawowe metody czyszczenia danych, ich wizualizacji i podsumowywania, a także imputacji braków danych. Ponadto nabywają umiejętność przetwarzania danych w postaci tekstowej, m.in. z różnego rodzaju API i innych zasobów internetowych. Do osiągnięcia powyższych celów używany jest język R wraz z bogatym zbiorem pakietów dla tego środowiska. Szczególnie jednak eksponowane są techniki i funkcje obecne w innych środowiskach, m.in. Python 3.

**Treści kształcenia:**

1. Podstawowe atomowe typy danych w R: Wektory i NULL
2. Zwektoryzowane operacje na wektorach atomowych. Przekształcanie i filtrowanie zmiennych. Agregacja i graficzna prezentacja zmiennych
3. Przetwarzanie danych tekstowych. Wyrażenia regularne
4. Listy. Funkcje
5. Instrukcja sterująca i pętle
6. Atrybuty obiektów. Programowanie obiektowe w stylu S3
7. Typy złożone: obiekty reprezentujące czas, czynniki, szeregi czasowe, macierze i ramki danych oraz podstawowe operacje na nich
8. Niestandardowa ewaluacja. Formuły
9. Filtrowanie, przekształcanie i czyszczenie ramek danych. Imputacja braków danych
10. Agregacja i inne operacje na danych w podgrupach. Scalanie ramek danych
11. Operacje na plikach i katalogach. Pobieranie danych z API. Wydobywanie informacji ze stron WWW

**Metody oceny:**

Na zaliczenie składają się oceny zdobyte za rozwiązania 5-7 prac domowych.

**Egzamin:**

nie

**Literatura:**

1. Gągolewski M., Programowanie w języku R. Analiza danych, obliczenia, symulacje, Wydawnictwo Naukowe PWN, wydanie II, 2016
2. Wickham H., Grolemund G., R for Data Science, O'Reilly, 2016
3. Chambers J.M., Programming with Data, Springer, 1998
4. Chambers J.M., Software for Data Analysis. Programming with R, Springer, 2008
5. Matloff N.S., The Art of R Programming: A Tour of Statistical Software Design, No Starch Press, 2011
6. Venables W.N., Ripley B.D., S Programming, Springer, 2000
7. Wickham H., Advanced R, Chapmah & Hall/CRC, 2014
8. Gagolewski M., Bartoszuk M., Cena A., Przetwarzanie i analiza danych w języku Python, Wydawnictwo Naukowe PWN, Warszawa, 2016
9. McKinney W., Python for Data Analysis. Data Wrangling with Pandas, NumPy, and IPython, O'Reilly Media, 2012
10. Richert W., Coelho L.P., Building Machine Learning Systems with Python, Packt Publishing, 2013

**Witryna www przedmiotu:**

http://www.gagolewski.com/teaching/

**Uwagi:**

.

## Charakterystyki przedmiotowe

### Profil ogólnoakademicki - wiedza

**Charakterystyka W01:**

Zna podstawy programowania w języku R i jego najważniejsze typy danych (w tym wektory atomowe, listy, funkcje, czynniki, szeregi czasowe, macierze oraz ramki danych)

Weryfikacja:

ocena prac domowych

**Powiązane charakterystyki kierunkowe:** DS\_W14

**Powiązane charakterystyki obszarowe:** I.P6S\_WG

**Charakterystyka W02:**

Zna podstawowe metody i algorytmy wykorzystywane w przetwarzaniu danych i przygotowywaniu ich do analizy.

Weryfikacja:

ocena prac domowych

**Powiązane charakterystyki kierunkowe:** DS\_W08, DS\_W09

**Powiązane charakterystyki obszarowe:** I.P6S\_WG

### Profil ogólnoakademicki - umiejętności

**Charakterystyka U01:**

Umie przeprowadzić wstępną analizę danych, m.in. podsumować wartości zmiennych oraz przedstawić je w postaci graficznej

Weryfikacja:

ocena prac domowych

**Powiązane charakterystyki kierunkowe:** DS\_U03, DS\_U04, DS\_U15

**Powiązane charakterystyki obszarowe:** I.P6S\_UW, I.P6S\_UK

**Charakterystyka U02:**

Umie zaimplementować proste metody przetwarzania i analizy danych oraz ocenić ich złożoność pamięciową i czasową

Weryfikacja:

ocena prac domowych

**Powiązane charakterystyki kierunkowe:** DS\_U11, DS\_U13, DS\_U14

**Powiązane charakterystyki obszarowe:** I.P6S\_UW

**Charakterystyka U03:**

Potrafi pozyskiwać dane ze źródeł tekstowych i zasobów i Internecie, wyczyścić je i przygotować do analizy

Weryfikacja:

ocena prac domowych

**Powiązane charakterystyki kierunkowe:** DS\_U22

**Powiązane charakterystyki obszarowe:** I.P6S\_UW

### Profil ogólnoakademicki - kompetencje społeczne

**Charakterystyka K01:**

Zna potrzebę uczenia się przez całe życie i podnoszenia kompetencji zawodowych

Weryfikacja:

ocena prac domowych

**Powiązane charakterystyki kierunkowe:** DS\_K01

**Powiązane charakterystyki obszarowe:** I.P6S\_KK