

FACULTY OF MANAGEMENT

SUBJECT CARD

Name of subject in Polish Pozyskiwanie i analiza danych stron www

Name of subject in English Web scraping and data analysis

Main field of study (if applicable): Business Engineering

Specialization (if applicable): Business Intelligence

Profile: academic

Level and form of studies: 2nd level, full-time

Kind of subject: obligatory

Subject code W08IZZ-SM8031G

Group of courses YES

	Lecture	Classes	Laboratory	Project	Seminar
Number of hours of organized classes in University (ZZU)	30		15		
Number of hours of total student workload (CNPS)	50		25		
Form of crediting			crediting with grade		
For group of courses mark final course with (X)			X		
Number of ECTS points			4		
including number of ECTS points for practical classes (P)			2		
including number of ECTS points corresponding to classes that require direct participation of lecturers and other academics (BU)			1,96		

PREREQUISITES RELATING TO KNOWLEDGE, SKILLS AND OTHER COMPETENCES

1. Basic knowledge and ability to use R program
2. Basic knowledge of HTML and CSS

SUBJECT OBJECTIVES

C1: Technical knowledge necessary to quickly obtain a large amount of data, automate this process.

C1: Mastering the ability to process such data into useful information supporting management processes.

C3: Mastering the ability to use the R program throughout the process: from data acquisition to analysis

SUBJECT EDUCATIONAL EFFECTS

Relating to knowledge:

PEU_W01: Basic knowledge to obtain and analyze data from websites.

Relating to skills:

PEU_U01: Ability to design and implement a procedure for obtaining data from websites, and then apply statistical methods to analyze such data.

PROGRAM CONTENT

Lectures		Number of hours
Lec1	Course assessment criteria. The Internet as a source of data supporting decision-making processes	1
Lec1-2	Review and expansion of R language topics	3
Lec3	<i>Tidyverse</i> ecosystem packages	2
Lec4	Functional programming	2
Lec5	Methods of text data processing (strings)	2
Lec6	Pattern searching, regular expressions	2
Lec7	Models and techniques for data extraction	2
Lec8-9	Parsing web pages	3
Lec9-10	Creating web crawlers. Case study	3
Lec11-12	Parsing dynamic web pages	4
Lec13-14	Data extraction through API	4
Lec15	Written exam	2
	Total hours	30

Laboratory		Number of hours
Lab1	Course assessment criteria. Rules and safety procedure for laboratory. R as a web scraping environment	1
Lab2	Selected data operations, functional programming, visualization	2
Lab3	String processing, regular expressions	2
Lab4	Task discussion: string processing using a selected web page example	1
Lab4-7	Creating web crawlers for a chosen decision-making problem. Report preparation	7
Lab8	Discussion and report review	2
	Total hours	15

TEACHING TOOLS USED

N1. Presentation

N2. Solving problems, case study

N3. Statistical program R, scripts in R

EVALUATION OF SUBJECT LEARNING OUTCOMES ACHIEVEMENT

Evaluation (F – forming (during semester), P – concluding (at semester end))	Learning outcomes number	Way of evaluating learning outcomes achievement
F1	PEU_W01	Written test
F2	PEU_U01	Assignment
F3	PEU_U01	Report
$P = 0.3 \times F1 + 0.7 \times (0.3 \times F2 + 0.7 \times F3)$		

PRIMARY AND SECONDARY LITERATURE

PRIMARY LITERATURE:

- [1] Kapłon R. *Lecture notes* [available on ePortal/Teams]
- [2] Mitchell R. *Web Scraping with Python*, 2nd Edition, O'Reilly Media, 2018.
- [3] Wickham H., Çetinkaya-Rundel M., Grolemund G., *R for Data Science*, 2nd Edition, O'Reilly Media, 2023.

SECONDARY LITERATURE:

- [4] Aydin O. *R Web Scraping Quick Start Guide*, Packt Publishing, 2018.
- [5] Fitzgerald M. *Introducing Regular Expressions*, O'Reilly Media, 2012.

SUBJECT SUPERVISOR (NAME AND SURNAME, E-MAIL ADDRESS)

Dr inż. Robert Kapłon; robert.kaplon@pwr.edu.pl