

SYLABUS
 DOTYCZY CYKLU KSZTAŁCENIA 2020-2024
 Rok akademicki 2022/2023

1. PODSTAWOWE INFORMACJE O PRZEDMIOCIE

Nazwa przedmiotu	<i>eksploracja danych internetowych</i>
Kod przedmiotu	
Nazwa jednostki prowadzącej kierunek	<i>Kolegium Nauk Przyrodniczych</i>
Nazwa jednostki realizującej przedmiot	<i>Kolegium Nauk Przyrodniczych</i>
Kierunek studiów	<i>informatyka</i>
Poziom studiów	<i>studia I stopnia</i>
Profil	<i>ogólnoakademicki</i>
Forma studiów	<i>stacjonarne</i>
Rok i semestr/y studiów	<i>rok III, semestr 6</i>
Rodzaj przedmiotu	<i>przedmiot specjalnościowy</i>
Język wykładowy	<i>język polski</i>
Koordynator	<i>dr inż. Piotr Lasek</i>
Imię i nazwisko osoby prowadzącej / osób prowadzących	<i>dr inż. Piotr Lasek, dr inż. Wiesław Paja</i>

1.1. Formy zajęć dydaktycznych, wymiar godzin i punktów ECTS

Semestr (nr)	Wykł.	Ćw.	Konw.	Lab.	Sem.	ZP	Prakt.	Inne (jakie?)	Liczba pkt. ECTS
6	15			30					3

1.2. Sposób realizacji zajęć

zajęcia w formie tradycyjnej

1.3 Forma zaliczenia przedmiotu (z toku)

egzamin

2. WYMAGANIA WSTĘPNE

Obowiązuje wiedza i umiejętności wymagana na przedmiotach: algorytmy i struktury danych, sztuczna inteligencja, programowanie obiektowe

3. CELE, EFEKTY UCZENIA SIĘ, TREŚCI PROGRAMOWE I STOSOWANE METODY DYDAKTYCZNE

3.1 Cele przedmiotu

C1	Zapoznanie z podstawowymi pojęciami statystycznymi, NLP i ML dotyczącymi reprezentacji danych tekstowych
C2	Nabycie praktycznej znajomości dotyczącej najważniejszych pakietów i narzędzi do przetwarzania danych tekstowych (Python, Jupyter Notebook, Pandas, Scikit)
C3	Zapoznanie z podstawowymi metodami analizy danych tekstowych (czyszczenie tekstu, wyrażenia regularne, tokenizacja, stop-words)
C4	Wprowadzenie do wektoryzacji (count-vectorizer, TF-IDF, N-grams)
C5	Inżynieria cech (tworzenie transformacje), k-krotna walidacja skrośna, miary jakości klasyfikacji
C6	Metody kombinowane (bagging, boosting), grid search, wybór optymalnego modelu

3.2 Efekty uczenia się dla przedmiotu

EK (efekt uczenia się)	Treść efektu uczenia się zdefiniowanego dla przedmiotu	Odniesienie do efektów kierunkowych
EK_o1	Student posiada wiedzę specjalistyczną, z zakresu podstawowych technik, metod oraz narzędzi wykorzystywanych w procesie przetwarzania danych tekstowych, w tym metod weryfikacji i wyboru optymalnego modelu przetwarzania danych tekstowych	K_Wo4, K_Wo7
EK_o2	Student posiada umiejętność samodzielnej implementacji potoku przetwarzania danych tekstowych	K_U2o

3.3 Treści programowe

A. Problematyka wykładu

Zapoznanie z podstawowymi pojęciami statystycznymi, NLP i ML dotyczącymi reprezentacji danych tekstowych
Nabycie praktycznej znajomości dotyczącej najważniejszych pakietów i narzędzi do przetwarzania danych tekstowych (Python, Jupyter Notebook, Pandas, Scikit)
Zapoznanie z podstawowymi metodami analizy danych tekstowych (czyszczenie tekstu, wyrażenia regularne, tokenizacja, stop-words)
Wprowadzenie do wektoryzacji (count-vectorizer, TF-IDF, N-grams)
Inżynieria cech (tworzenie transformacje), k-krotna walidacja skrośna, miary jakości klasyfikacji
Metody kombinowane (bagging, boosting), grid search, wybór optymalnego modelu

B. Problematyka ćwiczeń laboratoryjnych

Instalacja środowiska do analizy danych tekstowych, zapoznanie się z podstawowymi obiektami (np. DataFrame)
Usuwanie znaków interpunkcyjnych, tokenizacja, usuwanie stop-words
Wykorzystanie możliwości wyrażeń regularnych na przykładzie wskazanych funkcji
Wektoryzacja wybranego zbioru danych przy użyciu Count Vectorizer, N-Gram, TF-IDF

Tworzenie nowych cech. Wybór optymalnej transformacji Boxa-Coxa.

Porównywanie modeli klasyfikacyjnych (TF-IDF, Count, N-gram)

3.4 Metody dydaktyczne

Wykład: wykład z prezentacją multimedialną

Laboratorium komputerowe: wykonywanie zadań programistycznych

4. METODY I KRYTERIA OCENY

4.1 Sposoby weryfikacji efektów uczenia się

Symbol efektu	Metody oceny efektów uczenia się (np.: kolokwium, egzamin ustny, egzamin pisemny, projekt, sprawozdanie, obserwacja w trakcie zajęć)	Forma zajęć dydaktycznych (w, ćw, ...)
Ek_01	Kolokwium zaliczeniowe, egzamin	Wykład
Ek_02	Sprawozdania z zajęć laboratoryjnych	Laboratorium

4.2 Warunki zaliczenia przedmiotu (kryteria oceniania)

Laboratorium

Warunkiem pozytywnego zaliczenia laboratorium jest przedstawienie przynajmniej 80% sprawozdań. Ocenę wystawia prowadzący laboratorium uwzględniając kompletność zadań przedstawionych w przygotowanych przez studenta sprawozdaniach.

Wykład

Warunkiem zaliczenia wykładu jest uczestnictwo w kolokwium zaliczeniowym i uzyskanie przynajmniej połowy punktów.

Egzamin

Warunkiem uzyskania oceny pozytywnej jest uczestnictwo w egzaminie i uzyskanie przynajmniej połowy punktów. Oceny wystawia się według tabeli:

% punktów	Ocena
< 50	2
51 – 60	3
61 – 70	3,5
71 – 80	4,0
81 – 90	4,5
91 – 100	5,0

**5. CAŁKOWITY NAKŁAD PRACY STUDENTA POTRZEBNY DO OSIĄGNIĘCIA ZAŁOŻONYCH EFEKTÓW
W GODZINACH ORAZ PUNKTACH ECTS**

Forma aktywności	Średnia liczba godzin na zrealizowanie aktywności
Godziny kontaktowe wynikające z harmonogramu studiów	45
Inne z udziałem nauczyciela akademickiego (udział w konsultacjach, egzaminie)	10
Godziny niekontaktowe – praca własna studenta (przygotowanie do zajęć, egzaminu, napisanie referatu itp.)	35
SUMA GODZIN	90
SUMARYCZNA LICZBA PUNKTÓW ECTS	3

6. PRAKTYKI ZAWODOWE W RAMACH PRZEDMIOTU

wymiar godzinowy	-
zasady i formy odbywania praktyk	-

7. LITERATURA

<p>Literatura podstawowa:</p> <ul style="list-style-type: none"> • Larose D. T., Metody i modele eksploracji danych, Wydawnictwo Naukowe PWN, Warszawa 2012. • Morzy T., Eksploracja danych. Metody i algorytmy, Wydawnictwo Naukowe PWN, 2013.
<p>Literatura uzupełniająca:</p> <ul style="list-style-type: none"> • Data Mining: Concepts and Techniques, J. Han, M. Kamber, Morgan Kaufman, 2000 • Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, I. H. Witten, E. Frank, Morgan Kaufman, 2000 • Principles of Data Mining, J. Han, H. Mannila, P. Smyth, MIT Press, 2001 • Systemy uczące się, P. Cichosz, WNT, 2000 • Odkrywanie asocjacji: Algorytmy i struktury danych, T. Morzy, OWN, 2004.

Akceptacja Kierownika Jednostki lub osoby upoważnionej