

Karta przedmiotu

Nazwa i kod przedmiotu	Text mining, PG_00157366						
Kierunek studiów	Informatyka i ekonometria (O)						
Data rozpoczęcia studiów	październik 2024 r.		Rok akademicki realizacji przedmiotu		2025/2026		
Poziom kształcenia	II stopnia		Grupa zajęć		Grupa zajęć obowiązkowych z zakresu kierunku studiów Grupa zajęć powiązanych z prowadzonymi badaniami naukowymi w dziedzinie nauki związanej z kierunkiem - profil ogólnoakademicki		
Forma studiów	stacjonarne		Sposób realizacji		na uczelni		
Rok studiów	2		Język wykładowy		polski		
Semestr studiów	4		Liczba punktów ECTS		1.0		
Profil kształcenia	ogólnoakademicki		Forma zaliczenia		zaliczenie		
Jednostka prowadząca	Rektor -> Wydział Zarządzania -> Katedra Statystyki						
Imię i nazwisko wykładowcy (wykładowców)	Odpowiedzialny za przedmiot		dr hab. Krzysztof Najman				
	Prowadzący zajęcia z przedmiotu						
Formy zajęć	Forma zajęć	Wykład	Ćwiczenia	Laboratorium	Projekt	Seminarium	RAZEM
	Liczba godzin zajęć	0.0	15.0	0.0	0.0	0.0	15
	W tym liczba godzin zajęć na odległość: 0.0						
Aktywność studenta i liczba godzin pracy	Aktywność studenta	Udział w zajęciach dydaktycznych, objętych planem studiów		Udział w konsultacjach		Praca własna studenta	RAZEM
	Liczba godzin pracy studenta	15		5.0		10.0	30
Cel przedmiotu	Zapoznanie studentów z teoretycznymi i praktycznymi aspektami eksploracji danych tekstowych. Zapoznanie studentów z procesem analizy text mining, przeszukiwaniem dokumentów niestrukturalnych i praktycznym zastosowaniem text mining.						

Efekty uczenia się przedmiotu	Efekt kierunkowy	Efekt z przedmiotu	Sposób weryfikacji i oceny efektu
	[liEMU2_W04] Ma pogłębioną wiedzę o zaawansowanych metodach matematycznych, statystycznych, ekonometrycznych oraz informatycznych umożliwiających pozyskiwanie, przetwarzanie i analizę danych odzwierciedlających funkcjonowanie i wzrost gospodarki narodowej i jej składowych oraz zjawisk i procesów zachodzących w ich otoczeniu.	Ma pogłębioną wiedzę o zaawansowanych metodach analizy danych z zastosowaniem wybranych metod Text Mining.	[SW1] wypowiedź ustna/rozmowa/ dyskusja [SW2] prezentacja/projekt/referat/ raport
	[liEMU2_U02] Potrafi sprawnie pozyskiwać szczegółowe informacje o procesach i zjawiskach ekonomicznych w drodze obserwacji bezpośredniej, planowanego eksperymentu lub kwerendy baz danych oraz gromadzić je i przetwarzać za pomocą nowoczesnych narzędzi informatycznych.	Student potrafi zbierać dane statystyczne ze źródeł pierwotnych i wtórnych na potrzeby realizacji badań społecznych z zastosowaniem modeli Text Mining.	[SU1] wypowiedź ustna/rozmowa/ dyskusja [SU2] prezentacja/projekt/referat/ raport
	[liEMU2_K02] Potrafi swobodnie komunikować się z otoczeniem na specjalistyczne tematy z zakresu informatyki i ekonometrii w miejscu pracy i poza nim, przekazywać swoją wiedzę oraz dzielić się swoimi umiejętnościami za pomocą różnych środków przekazu. Kulturalnie uczestniczy w dyskusji, nie boi się zadawać pytań i umie wyrażać konstruktywną krytykę.	Potrafi dyskutować, komunikować się i dzielić wiedzą w zakresie teorii i praktyki stosowania modeli Text Mining.	[SK1] wypowiedź ustna/rozmowa/ dyskusja [SK2] prezentacja/projekt/referat/ raport
	[liEMU2_K04] Potrafi pracować w zespole, współtworzyć go, efektywnie nim zarządzać i sprawować nad nim nadzór; sprawnie dostosowuje swoje zachowania i sposób postępowania do roli w nim odgrywanej; jest gotowy do brania odpowiedzialności za zespół i ponoszenia konsekwencji; rozumie konieczność systematyczności i konsekwencji w działaniu; jest otwarty na innych członków zespołu oraz krytyczny wobec siebie.	Student potrafi odpowiedzialnie pracować w zespole, pełnić w nim różne role, krytycznie oceniać pracę swoją i innych członków zespołu.	[SK2] prezentacja/projekt/referat/ raport
	[liEMU2_K01] Rozumie potrzebę ustawicznego uzupełniania i pogłębiania nabytej wiedzy; inspiruje i organizuje proces uczenia się innych osób.	Rozumie potrzebę stałego uzupełniania wiedzy w zakresie Text Mining.	[SK1] wypowiedź ustna/rozmowa/ dyskusja [SK2] prezentacja/projekt/referat/ raport
	[liEMU2_K03] Potrafi swobodnie komunikować się z otoczeniem w miejscu pracy i poza nim, przekazywać swoją wiedzę oraz dzielić się swoimi umiejętnościami za pomocą różnych środków przekazu.	Potrafi komunikować się i przekazywać wiedzę i umiejętności w zakresie Text Mining.	[SK1] wypowiedź ustna/rozmowa/ dyskusja [SK2] prezentacja/projekt/referat/ raport
	[liEMU2_W05] Ma pogłębioną wiedzę o źródłach danych społeczno-ekonomicznych, ich bazach oraz sposobie ich tworzenia.	Student zna źródła i metody wyszukiwania danych wtórnych wykorzystywanych w badaniach społecznych z zastosowaniem metod Text Mining.	[SW1] wypowiedź ustna/rozmowa/ dyskusja [SW2] prezentacja/projekt/referat/ raport

Treści przedmiotu	Definicja, istota, cel, zakres i korzyści ze stosowania text miningu (eksploracja danych tekstowych). Powiązania text miningu z dziedzinami pokrewnymi, interdyscyplinarny charakter text miningu, dyskusja nad związkami z data mining, uczeniem maszynowym, przetwarzaniem języka naturalnego (NLP Natural Language Processing), tłumaczeniem maszynowym (MT Machine Translation), systemami wyszukiwania informacji (IR Information Retrieval), ekstrakcją informacji (IE Information Extraction), statystyką i informatyką. Obszary i przykłady zastosowań tekst miningu: pozyskiwanie informacji z dokumentów, identyfikacja wiadomości zawierających określone treści, klasyfikacja wzorcowa, klasyfikacje bezwzorcowe, identyfikacja powiązań, wizualizacja, generowanie odpowiedzi na pytania. Proces analizy text mining. Określenie celu, zakresu i kosztów analizy, określenie źródeł danych nieustrukturyzowanych, przekształcanie dokumentów źródłowych do postaci pliku tekstowego (preprocessing): transformacja dokumentów do postaci tekstowej, usunięcie znaków formatujących, ujednolicenie sposobu kodowania znaków, tokenizacja identyfikacja podstawowych fraz tekstowych, podział tekstu na zdania, tokeny, słowa. Proces leksykalny oznaczanie części mowy. Ekstrakcja rdzeni wyrazów (stemming) automatyczne odnajdowanie rdzeni lub pni wyrazów, przypisywanie słowom ich bazowej formy. Parsing wyłączenie wyrazów o niskiej wartości informacyjnej. Tagowanie (tagging) wybór opisu morfoskładniowego, który jest właściwy w konkretnym kontekście użycia danej formy. Lematyzacja analiza morfologiczna ograniczana do znalezienia podstawowej formy wyrazu (identyfikacja leksemu). Stop lista (stop list) usuwanie słów (tagów) nieistotnych i gromadzenie nieistotnych semantycznie części mowy na osobnej liście. Przycinanie (pruning) usuwanie niepotrzebnych słów: podejście oparte na usuwaniu wyrazów występujących najczęściej (most frequent) i najrzadziej (least frequent). Sposoby przechowywania w systemie komputerowym informacji występujących w dokumencie: oparte na częstości występowania poszczególnych wyrazów w dokumencie, dążenie do nadania informacjom pozyskanym z dokumentów postaci ustrukturyzowanej. Format danych wejściowych: .txt, .html, rzadziej .doc czy .pdf. Macierz częstości występowania wyrazów dokumencie, filtrowanie tekstu i ekstrakcja faktów (macierz dokumentów-wyrazów, term-document matrix). Reprezentacja dokumentów tekstowych reprezentacja wektorowa (vector space model VSM) oraz reprezentacja grafowa: standardowa, prosta, n-odległości, prostej n-odległości, bezwzględnej częstości, względnej częstości. Przeszukiwanie dokumentów niestrukturalnych: wyszukiwanie pełnotekstowe FTS (Full Text Searching), wykorzystujące specjalny rodzaj indeksów tzw. indeksy pełnotekstowe. Macierz term-by-document frequency. Ważenie słów w macierzy: Frequency Weight metoda określania częstości występowania określonych zwrotów dokumencie: metoda binarna, logarytmiczna, bez modyfikacji (none) i Term Weight metoda określania liczby wystąpień danego wyrażenia w zbiorze dokumentów (kolekcji), metody: Entropy, IDF (Inverse Document Frequency), GF-IDF (Global Frequency - Inverse Document Frequency), Normal, Chi-Squared, Mutual Information, Information Gain.		
Wymagania wstępne i dodatkowe	Znajomość podstaw algebry liniowej, statystyki, statystycznej analizy wielowymiarowej.		
Sposoby i kryteria oceniania osiągniętych efektów uczenia się	Sposób oceniania (składowe)	Próg zaliczeniowy	Składowa oceny końcowej
	projekt zaliczeniowy	51.0%	100.0%
Zalecana lista lektur	Podstawowa lista lektur	Feldman R., Sanger J., 2006, Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, Cambridge University Press New York Berry M.W., Kogan J., 2010, Text Mining, Applications and Theory, A John Wiley and Sons, Ltd. Maimon O., Rokach L., 2005, The Data Mining and Knowledge Discovery Handbook, Springer Solka J. L., 2008, Text Data Mining: Theory and Methods, Statistic Survey	
	Uzupełniająca lista lektur	P.Lula, Statystyczne modelowanie zawartości dokumentów tekstowych, WUEK, 2018.	
	Adresy eZasobów		
Przykładowe zagadnienia/ przykładowe pytania/ realizowane zadania			
Praktyki zawodowe w ramach przedmiotu	Nie dotyczy		

Dokument wygenerowany elektronicznie. Nie wymaga pieczęci ani podpisu.