



KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu

Eksploracja danych tekstowych (Text Mining) [S1DSwB1>EDT]

Przedmiot

Kierunek studiów

Data Science w biznesie

Rok/Semestr

3/5

Studia w zakresie (specjalność)

–

Profil studiów

ogólnoakademicki

Poziom studiów

pierwszego stopnia

Język oferowanego przedmiotu

polski

Forma studiów

stacjonarne

Wymagalność

obieralny

Liczba godzin

Wykład

15

Laboratorium

0

Inne

0

Ćwiczenia

15

Projekty/seminaria

0

Liczba punktów ECTS

3,00

Koordynatorzy

dr hab. Grzegorz Pawłowski

grzegorz.pawlowski@put.poznan.pl

Wykładowcy

Wymagania wstępne

Studenci powinni posiadać podstawową wiedzę z zakresu statystyki, algebry liniowej i rachunku prawdopodobieństwa, a także umiejętność pracy z danymi. Wskazana jest znajomość podstaw programowania (preferowany Python) oraz podstawowe umiejętności analizy danych. Przydatne będzie także ogólne zrozumienie baz danych i SQL oraz podstaw uczenia maszynowego.

Cel przedmiotu

Celem przedmiotu jest wyposażenie uczestników w teoretyczne podstawy i praktyczne umiejętności w zakresie eksploracji danych tekstowych, umożliwiające efektywną analizę, interpretację i wykorzystanie informacji zawartych w dużych zbiorach tekstu przy zastosowaniu narzędzi NLP i technik uczenia maszynowego.

Przedmiotowe efekty uczenia się

Wiedza:

1. Charakteryzuje podstawowe metody eksploracji danych tekstowych, w tym tokenizację, normalizację, wektoryzację oraz analizę semantyczną [DSB1_W01].
2. Identyfikuje kluczowe techniki przetwarzania języka naturalnego (NLP), takie jak analiza sentymentu,

modelowanie tematów oraz ekstrakcja informacji [DSB1_W02].

3. Wyjaśnia zastosowanie metod eksploracji tekstu w kontekście uczenia maszynowego, big data oraz systemów rekomendacyjnych [DSB1_W03].

Umiejętności:

1. Analizuje dane tekstowe, stosując odpowiednie metody NLP i eksploracji tekstu, takie jak TF-IDF, Word Embeddings oraz klasyfikacja i klasteryzacja dokumentów [DSB1_U02].
2. Projektuje i implementuje modele przetwarzania tekstu przy użyciu narzędzi open-source i środowisk big data [DSB1_U03].
3. Wykorzystuje techniki eksploracji danych tekstowych do analizy opinii, klasyfikacji dokumentów oraz budowania systemów wyszukiwania informacji [DSB1_U08].

Kompetencje społeczne:

1. Uwzględnia etyczne i prawne aspekty przetwarzania danych tekstowych, w tym kwestie prywatności i rzetelności analizy [DSB1_K01].
2. Podejmuje odpowiedzialność za jakość i interpretację wyników eksploracji danych tekstowych w kontekście zastosowań biznesowych i naukowych [DSB1_K04].

Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:

Wykład:

Dwa kolokwia, za które studenci otrzymują oceny formułujące w postaci punktów - po 50 punktów za każde kolokwium. Ocena końcowa stanowi sumę punktów z dwóch ocen formułujących. Pierwsze kolokwium odbywa się w połowie kursu, a drugie na jego zakończenie. Próg zaliczeniowy to 50 punktów łącznie z obu kolokwiów.

Laboratoria:

Studenci uzyskują dwie oceny formułujące. Pierwsza z nich dotyczy przygotowania opracowania dotyczącego rozwiązania wybranego problemu biznesowego dotyczącego eksploracji danych tekstowych wskazanego przez prowadzącego. Za opracowanie studenci mogą uzyskać 50 punktów. Druga z nich dotyczy kolokwium, za które studenci mogą uzyskać również 50 punktów. Próg zaliczeniowy to 50 punktów łącznie ze wszystkich aktywności.

Treści programowe

Program przedmiotu obejmuje kompleksowe zagadnienia z zakresu przetwarzania języka naturalnego, w tym tokenizacji, normalizacji i reprezentacji tekstu za pomocą metod takich jak TF-IDF i Word Embeddings. Kurs skupia się na technikach analizy tekstu, klasyfikacji, modelowaniu tematów, analizie sentymentu oraz wykorzystaniu sieci neuronowych i głębokiego uczenia w NLP. Uczestnicy poznają metody ekstrakcji informacji, analizy semantycznej i wizualizacji wyników, a także praktyczne zastosowania narzędzi open-source w środowiskach big data, przy jednoczesnym uwzględnieniu etycznych i prawnych aspektów przetwarzania danych tekstowych.

Tematyka zajęć

Wprowadzenie do eksploracji danych tekstowych
Podstawy przetwarzania języka naturalnego (NLP)
Techniki tokenizacji i segmentacji tekstu
Normalizacja tekstu: usuwanie stop-słów, stemming, lematyzacja
Reprezentacja tekstu: wektoryzacja, TF-IDF, Word Embeddings
Analiza częstości słów i n-gramy
Klasyfikacja tekstu - metody nadzorowane i nienadzorowane
Analiza sentymentu i emocji w tekście
Tematyczne modelowanie i wykrywanie tematów
Ekstrakcja informacji: rozpoznawanie encji nazwanych
Modele językowe i generowanie tekstu
Sieci neuronowe i głębokie uczenie w NLP
Analiza semantyczna i relacje między słowami
Wykrywanie duplikatów i klastrowanie dokumentów
Wykorzystanie eksploracji tekstowej w analizie opinii

Techniki wyszukiwania informacji i systemy rekomendacyjne
Przetwarzanie języka w środowiskach big data
Narzędzia i biblioteki open-source w eksploracji danych tekstowych
Wizualizacja wyników analizy tekstu
Etyczne i prawne aspekty przetwarzania danych tekstowych

Metody dydaktyczne

Wykłady: wykład problemowy, prezentacja case studies
Laboratoria: analiza rzeczywistych danych, zadania praktyczne w grupach, studium przypadku

Literatura

Podstawowa:

Spinczyk, D., Dzieciatko, M. (2016). Text Mining: metody, narzędzia, zastosowania. PWN
Weiss, D. (2006). Descriptive clustering as a method for exploring text collections. Unpublished doctoral dissertation, Poznan University of Technology, Poznan, Poland.
Russell, A., Klassen, M. (2019). Data Mining. Eksploracja danych w sieciach społecznościowych. Wydanie III, Helion
Rajnish, R., Srivastava, M. (2023). Web Data Mining z użyciem języka Python. Odkrywaj i wyodrębniaj informacje ze stron internetowych za pomocą języka Python. APN Promise

Uzupełniająca:

Zastosowania statystyki i data mining w badaniach naukowych, Red. Jakubowski, J., Wątroba, J. (2017), StatSoft Polska. Kraków

Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	75	3,00
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	30	1,00
Praca własna studenta (studia literaturowe, przygotowanie do zajęć laboratoryjnych/ćwiczeń, przygotowanie do kolokwium/egzaminu, wykonanie projektu)	45	2,00