**Nazwa przedmiotu:**

Przetwarzanie i analiza danych tekstowych

**Koordynator przedmiotu:**

dr inż. Anna Wróblewska

**Status przedmiotu:**

Obowiązkowy

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Inżynieria i Analiza Danych

**Grupa przedmiotów:**

Wspólne

**Kod przedmiotu:**

.

**Semestr nominalny:**

3 / rok ak. 2019/2020

**Liczba punktów ECTS:**

6

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

1. godziny kontaktowe – 85 h; w tym
a) obecność na wykładach – 30 h
b) obecność na zajęciach projektowych – 45 h
c) obecność na egzaminie – 5 h
d) konsultacje – 5 h
2. praca własna studenta – 70 h; w tym
a) przygotowanie do zajęć projektowych – 35 h
b) przygotowanie raportu i prezentacji projektu – 10 h
c) zapoznanie się z literaturą – 10 h
d) przygotowanie do egzaminu – 15 h
Razem 155 h, co odpowiada 6 pkt. ECTS

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

a) obecność na wykładach – 30 h
b) obecność na zajęciach projektowych – 45 h
c) obecność na egzaminie – 5 h
d) konsultacje – 5 h
Razem 85 h, co odpowiada 3 pkt. ECTS

**Język prowadzenia zajęć:**

polski

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

a) przygotowanie do zajęć projektowych – 35 h
b) obecność na zajęciach projektowych – 45 h
Razem 80 h, co odpowiada 3 pkt. ECTS

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

|  |  |
| --- | --- |
| Wykład:  | 15h |
| Ćwiczenia:  | 15h |
| Laboratorium:  | 0h |
| Projekt:  | 45h |
| Lekcje komputerowe:  | 0h |

**Wymagania wstępne:**

Wstęp do uczenia maszynowego, Programowanie w R i Python

**Limit liczby studentów:**

.

**Cel przedmiotu:**

Celem przedmiotu jest przedstawienie podstawowej teorii, modeli i metod przetwarzania, analizy i odkrywania wiedzy w różnych zbiorach danych tekstowych. W trakcie nauki będzie można dowiedzieć się o podstawowych i zaawansowanych technikach systemów informacyjnych oraz rozwiązaniach przetwarzania tekstu metodami uczenia głębokiego.
W ramach przedmiotu, studenci powinni zdobyć praktyczną umiejętność wykorzystania wybranych metod do analizy danych tekstowych.

**Treści kształcenia:**

trakcie nauki będzie można dowiedzieć się o podstawowych i zaawansowanych technikach systemów informacyjnych: ekstrakcja informacji, indeksowanie tekstu, mierzenie jakości systemów wyszukiwania i eksploracji, architektury współczesnych systemów wyszukiwania wiedzy i zarządzania informacją. Podane będą także współczesne rozwiązania przetwarzania tekstu metodami uczenia głębokiego, m.in. najnowsze architektury tj. sieci rekurencyjne LSTM, GRU, modele reprezentacji tekstu - word embeddings.
Wykład i projekty będą zawierały wybrane tematy z poniższej listy:
1. Wstęp do wyszukiwania informacji: teoria informacji, metody NLP/text mining, statystyka, lingwistyka, zagadnienia w procesie przetwarzania tekstu: lematyzacja, stemming
2. Słowa i zdania: wyrażenia regularne, tokenizacja, odległość edytorska, lingwistyka: poziomy opisu języka, morfologia: słowotwórstwo, fleksja, stemmery, części mowy, algorytm soundex , błędy ortograficzne
3. Źródła danych: korpusy tekstu
4. Modele reprezentacji tekstu, m.in. modele wektorowe (word embeddings) - word2vec, gloVe, FastText
5. Architektury głębokich modeli neuronowych - rekurencyjne, rekursywne, konwolucyjne i dynamiczne sieci do zastosowań rozpoznawania tekstu: modelowania języka, analizy opinii, parsowania tekstu, klasyfikacji zdań
6. Statystyczne metody przetwarzania języka naturalnego, modelowanie języka, n-gramy, kolokacje, ujednoznacznianie (word sense disambiguation)
7. Analiza gramatyczna (HMM, POS tagging, parsowanie)
8. Ekstrakcja informacji, NER (named-entity recognition), ekstrakcja relacji, semantyka informacji (ontologie, budowa ontologii z tekstu)
9. Wyszukiwanie informacji: indeks odwrócony, miary podobieństwa, ranking wyników, analiza linków (PageRank, HITS), architektury komercyjnych systemów, mierzenie jakości zwracanych wyników, wizualizacja wyników wyszukiwania, architektury komercyjnych systemów wyszukiwania informacji/baz wiedzy
10. Zastosowania:
a. Kategoryzacja i grupowanie dokumentów (grupowanie hierarchiczne, LDA – latent dirichlet allocation)
b. Analiza zabarwienia emocjonalnego tekstu (sentiment analysis)
c. Odpowiadanie na zapytania (question answering)
d. agenci dialogowi (chatbots)
e. Streszczanie dokumentów
f. Tłumaczenia automatyczne
g. Rekomendacje oparte na treści
Projekt:
Opracowanie metod i aplikacji z zakresu tematyki przedmiotu, aplikacje będą miały na celu przetwarzanie tekstu i danych pochodzących ze stron internetowych lub korpusów tekstu.

**Metody oceny:**

Zaliczenie przedmiotu oparte jest o wyniki realizacji zadań realizowanych w ramach projektu (w sumie maksymalnie 60 pkt) oraz wyniki egzaminu (maksymalnie 40 pkt). W przypadku zadań realizowanych w laboratorium oceniana jest zarówno jakość rozwiązań jak i terminowość ich realizacji. Szczegółowe zasady przyznawania punktów przedstawiane są studentom podczas pierwszych zajęć projektowych (dot. zaliczenia projektu) oraz wykładowych (dot. egzaminu).
Ocena końcowa zależy od łącznej liczby punktów uzyskanych z projektów oraz egzaminu i jest wyznaczana zgodnie z poniższymi regułami: 0-50 pkt – 2.0, 51-60 pkt – 3.0, 61-70 pkt – 3.5, 71-80 pkt – 4.0, 81-90 pkt – 4.5, 91-100 pkt – 5.0. Warunkiem koniecznym uzyskania oceny pozytywnej jest zdobycie co najmniej 30 pkt z laboratorium oraz 20 pkt z egzaminu.

**Egzamin:**

tak

**Literatura:**

1. D. Jurafsky, J. Martin, Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition, Prentice Hall Series in Artificial Intelligence, Pearson/Prentice Hall 2009
2. H. Schutze, C. Manning, P. Raghavan, Introduction to information retrieval, 2008 http://nlp.stanford.edu/IR-book/pdf/irbookprint.pdf
3. J. Hirschberg, C. Manning, Advances in natural language processing”, Science (New York, N.Y.), 17 July 2015, Vol.349(6245), pp.261-6
4. M. Kłopotek, Inteligentne wyszukiwarki internetowe, Warszawa: Akademicka Oficyna Wydawnicza EXIT, 2001
5. Python NLTK (Natural Language Toolkit) http://nltk.sourceforge.net
6. Open NLP http://opennlp.sourceforge.net/
7. Deep Learning (Adaptive Computation and Machine Learning series) – November 18, 2016, Ian Goodfellow, Yoshua Bengio , Aaron Courville

**Witryna www przedmiotu:**

.

**Uwagi:**

Program 4 semestralny - 3 semestr
Program 3 semestralny - 2 semestr

## Charakterystyki przedmiotowe

### Profil ogólnoakademicki - wiedza

**Charakterystyka W01:**

Zna teoretyczne podstawy metod odkrywania wiedzy w zbiorach danych tekstowych

Weryfikacja:

ocena z egzaminu, ocena z projektu i prezentacji projektu, ocena z egzaminu

**Powiązane charakterystyki kierunkowe:** DS2\_W04

**Powiązane charakterystyki obszarowe:** I.P7S\_WG

**Charakterystyka W02:**

Zna podstawowe metody reprezentacji danych tekstowych niestrukturalnych

Weryfikacja:

ocena z egzaminu

**Powiązane charakterystyki kierunkowe:** DS2\_W04

**Powiązane charakterystyki obszarowe:** I.P7S\_WG

### Profil ogólnoakademicki - umiejętności

**Charakterystyka U01:**

Potrafi zaprojektować algorytmy rozwiązujące określony problem posiadający praktyczne znaczenie z obszaru eksploracji i wizualizacji danych tekstowych

Weryfikacja:

ocena z projektu i prezentacji projektu, ocena z egzaminu

**Powiązane charakterystyki kierunkowe:** DS2\_U04, DS2\_U05, DS2\_U06, DS2\_U07, DS2\_U21

**Powiązane charakterystyki obszarowe:** I.P7S\_UW, I.P7S\_UK

**Charakterystyka U02:**

Potrafi wybrać właściwe narzędzia programistyczne do zaprojektowania algorytmu dotyczącego danych tekstowych

Weryfikacja:

ocena z projektu i prezentacji projektu, ocena z egzaminu

**Powiązane charakterystyki kierunkowe:** DS2\_U05, DS2\_U06, DS2\_U01, DS2\_U02

**Powiązane charakterystyki obszarowe:** I.P7S\_UW

**Charakterystyka U03:**

Posiada umiejętność korzystania ze źródeł literaturowych oraz zasobów internetowych dotyczących rozwiązywanego zadania

Weryfikacja:

ocena z egzaminu, ocena z projektu i prezentacji projektu, ocena z egzaminu

**Powiązane charakterystyki kierunkowe:** DS2\_U07

**Powiązane charakterystyki obszarowe:** I.P7S\_UW, I.P7S\_UK

### Profil ogólnoakademicki - kompetencje społeczne

**Charakterystyka K01:**

Potrafi pracować indywidualnie i w zespole, oraz kierować niedużym zespołem

Weryfikacja:

ocena z projektu

**Powiązane charakterystyki kierunkowe:** DS2\_K03, DS2\_K04

**Powiązane charakterystyki obszarowe:** I.P7S\_KR

**Charakterystyka K02:**

Posiada zdolność do kontynuacji kształcenia oraz świadomość potrzeby samokształcenia w ramach procesu kształcenia ustawicznego

Weryfikacja:

ocena z projektu i prezentacji projektu, ocena z egzaminu

**Powiązane charakterystyki kierunkowe:** DS2\_K01

**Powiązane charakterystyki obszarowe:** I.P7S\_KK

**Charakterystyka K03:**

Ma świadomość odpowiedzialności za wspólnie realizowane zadania w ramach pracy zespołowej

Weryfikacja:

ocena z projektu

**Powiązane charakterystyki kierunkowe:** DS2\_K04

**Powiązane charakterystyki obszarowe:** I.P7S\_KR