**Nazwa przedmiotu:**

Wprowadzenie do eksploracji danych tekstowych w sieci WWW

**Koordynator przedmiotu:**

Piotr Andruszkiewicz

**Status przedmiotu:**

Fakultatywny ograniczonego wyboru

**Poziom kształcenia:**

Studia II stopnia

**Program:**

Informatyka

**Grupa przedmiotów:**

Przedmioty techniczne - zaawansowane

**Kod przedmiotu:**

WEDT

**Semestr nominalny:**

3 / rok ak. 2018/2019

**Liczba punktów ECTS:**

4

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

30 godzin wykładu, 24 godziny przygotowania do kolokwiów,
15 godzin zajęć projektowych, 30 godzin realizacji zadania projektowego
w sumie 99 godzin

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

30 godzin wykładu,
15 godzin zajęć projektowych
co daje ok. 2 ECTS

**Język prowadzenia zajęć:**

polski

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

15 godzin zajęć projektowych,
30 godzin realizacji zadania projektowego
co daje ok. 2 ECTS

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

|  |  |
| --- | --- |
| Wykład: | 30h |
| Ćwiczenia: | 0h |
| Laboratorium: | 0h |
| Projekt: | 15h |
| Lekcje komputerowe: | 0h |

**Wymagania wstępne:**

Znajomość podstawowych języków programowania (Java / C++)
Znajomość koncepcji SZBD

**Limit liczby studentów:**

48

**Cel przedmiotu:**

Treść wykładu obejmuje zagadnienia związane z szeroko rozumianą eksploracją danych w nieustrukturalizowanych źródłach tekstowych oraz zasobach sieci Internet. W ramach wykładu omówione zostaną podstawowe pojęcia i metody wykorzystywane przy przetwarzaniu dokumentów tekstowych, ze szczególnym zwróceniem uwagi na metody statystyczne oraz ich praktyczne zastosowanie w eksploracji danych. Przedstawione zostaną m.in. metody takie jak analiza statystyczna tekstu (analiza współwystepowania słów itp.) reprezentacje dokumentów tekstowych (przestrzenie wektorowe dokumentów, n-gramy, zbiory odwrócone itp.) oraz analiza gramatyki (oznaczanie części mowy, budowa modeli języka, stemming itp.). Wykorzystanie przedstawionych metod zostanie zaprezentowane w praktycznych zastosowaniach w grupowaniu i klasyfikacji dokumentów, w tłumaczeniu automatycznym, w automatycznej generacji streszczeń i wyszukiwaniu słów kluczowych. W drugiej części wykładu omówione zostanie wykorzystanie technik text mining w środowisku sieci Internet. Przedstawione zostaną także algorytmy służące analizie grafu hiperpołączeń tej sieci.

**Treści kształcenia:**

1. Tematy wstępne
Omówione zostaną podstawowe wiadomości ze statystyki matematycznej i lingwistyki.
2. Metody eksploracji dokumentów tekstowych
Praca z korpusami tekstowymi: Omówione zostaną najpopularniejsze korpusy tekstu w języku angielskim (TREC, Brown Corpus, kolekcja Reutersa, Penn TreeBank, Wordnet). Przedstawione zostaną popularne sposoby kodowania metainformacji dotyczących dokumentów zgromadzonych w korpusach oraz typowe zastosowania repozytoriów (w tłumaczeniu automatycznym, jako sieci semantyczne itp.).
Przetwarzanie informacji tekstowej: Omówione zostanie zagadnienie wstępnego przetwarzania dokumentów tekstowych i wyodrębniania pojedynczych jednostek takich jak zdania i słowa. Omówione zostaną wyrażenia regularne jako narzędzie niskopoziomowej analizy treści tekstowej. Następnie przedstawione zostaną podstawowe wiadomości dotyczące morfologii, w tym przede wszystkim metody sprowadzania słów do postaci kanonicznej. Zostaną omówione metody analizy dokumentów tekstowych bazujące na miarach statystycznych (porównywanie rozkładu częstości wystąpień słów, wyszukiwanie kolokacji, automatyczne wyszukiwanie synonimów i budowanie tezaurusów).
Modele dokumentów tekstowych: Dokonany zostanie przegląd metod reprezentacji dokumentów tekstowych. Przedstawiony zostanie model wektorowy oparty na częstościach występowania unigramów i n-gramów oraz jego implementacje (m.in. zbiór odwrócony). Omówione zostaną metody wzbogacania modeli o informację nietekstową (formatowanie tekstu) oraz ich przetwarzania (stosowanie list słów nieistotnych - ang. stoplist, prawo Zipfa, metoda TF-IDF).
Analiza gramatyczna: Przedstawione zostaną metody analizy gramatycznej zdań, w tym automatycznego oznaczania części mowy (ang. part-of-speech tagging). Omówione zostaną modele Markowa, w tym HMM oraz gramatyki bezkontekstowe.
3. Zastosowania
Omówione zostaną zagadnienia związane z problemem wyszukiwania dokumentów w repozytoriach pełnotekstowych. Zostaną przedstawione sposoby analizy i wzbogacania zapytań, przedstawiania wyników wyszukiwania, omówione zostaną także metody oceny jakości działania narzędzi wyszukiwawczych (m.in. miary precision, recall). Przedstawione zostaną algorytmy grupowania i automatycznej kategoryzacji dokumentów tekstowych. Omówiona zostanie problematyka automatycznego tłumaczenia dokumentów i automatcznej identyfikacji języka. Szczególny nacisk zostanie położony na metody statystyczne oparte na wykorzystaniu korpusów tekstów dwujęzycznych (ang. statistical text alignment). Przedstawione zostaną metody wyszukiwania istotnych słów kluczowych w dokumentach i generowania streszczeń.
Omówione zostaną także algorytmy analizy grafu połączeń hipertekstowych (HITS, Page Rank, HillTop). Przedstawione zostaną podstawy działania wyszukiwarek sieciowych (Google), oraz tworzenia własnych robotów sieciowych. Poruszone zostaną także problemy manipulowania pozycjonowaniem stron internetowych w rankingach wyszukiwarek poprzez wykorzystanie specyficznych cech algorytmów stosowanych przez wyszukiwarki. Omówiona zostanie koncepcja Web Farming, jako stałego i ustrukturalizowanego procesu pozyskiwania informacji z sieci WWW.

**Metody oceny:**

Dwa kolokwia oraz projekt

**Egzamin:**

nie

**Literatura:**

1. Chris Manning, Hinrich Schütze, „Foundations of Statistical Natural Language Processing”, MIT Press, 1999
2. Dan Jurafsky, James H. Martin „Speech and Language Processing”, Prentice-Hall, 2000
3. Mieczysław Kłopotek, „Inteligentne wyszukiwarki internetowe”, Exit, 2001
oraz szereg dostępnych na stronach internetowych publikacji z dziedziny text mining oraz web mining, o których wykładowca będzie informował na bieżąco w trakcie zajęć.

**Witryna www przedmiotu:**

https://usosweb.usos.pw.edu.pl/kontroler.php?\_action=katalog2/przedmioty/pokazPrzedmiot&prz\_kod=103A-INxxx-MSP-WEDT

**Uwagi:**

## Charakterystyki przedmiotowe

### Profil ogólnoakademicki - wiedza

**Charakterystyka WEDT\_W01:**

Student zna podstawowe algorytmy i struktury danych służace do eksploracji danych pełnotekstowych

Weryfikacja:

Kolokwia

**Powiązane charakterystyki kierunkowe:** K\_W12, K\_W06, K\_W08, K\_W11

**Powiązane charakterystyki obszarowe:** I.P7S\_WK, III.P7S\_WK.o, I.P7S\_WG, III.P7S\_WG.o

### Profil ogólnoakademicki - umiejętności

**Charakterystyka WEDT\_U01:**

Student posiada umiejętność samodzielnego analizowania i wykorzystywania artykułów naukowych z dziedziny NLP

Weryfikacja:

Kolokwia oraz tworzenie własnego projektu

**Powiązane charakterystyki kierunkowe:** K\_U01, K\_U03, K\_U06, K\_U07, K\_U13

**Powiązane charakterystyki obszarowe:** I.P7S\_UK, I.P7S\_UW, III.P7S\_UW.2.o, III.P7S\_UW.3.o

### Profil ogólnoakademicki - kompetencje społeczne

**Charakterystyka WEDT\_K01:**

Student rozumie znaczenie narzędzi automatycznego wyszukiwania i przetwarzania informacji pełnotekstowej dla rozwoju społeczno-gospodareczego

Weryfikacja:

Kolokwia

**Powiązane charakterystyki kierunkowe:** K\_K02

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