KARTA OPISU PRZEDMIOTU - SYLABUS

Nazwa przedmiotu
Inżynieria lingwistyczna

Przedmiot

Kierunek studiów
Informatyka
Studia w zakresie (specjalność)
Inteligentne Technologie Informatyczne
Poziom studiów
drugiego stopnia
Forma studiów
stacjonarne

Rok/semestr
II/3
Profil studiów
ogólnoakademicki
Język oferowanego przedmiotu
polski
Wymagalność
obligatoryjny

Liczba godzin

Wykład
30
Ćwiczenia
30
Laboratoria
Projekty/seminaria
Inne (np. online)

Liczba punktów ECTS

4

Wykładowcy

Odpowiedzialny za przedmiot/wykładowca:
Mateusz Lango

email: mateusz.lango@cs.put.poznan.pl

tel. 61 665 21 24

Wydział Informatyki i Telekomunikacji

Piotrowo 2, 60-965 Poznań

Odpowiedzialny za przedmiot/wykładowca:

Wymagania wstępne

Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z rachunku prawdopodobieństwa i statystyki (rozkład normalny, dwumianowy, Bernoulliego, estymacja maksymalnej wiarygodności, estymatory nieobciążone, zgodne, efektywne), a także pogłębioną wiedzę z uczenia maszynowego (klasyfikatory złożone, algorytmy k-NN, Naive Bayes, SVM), a w szczególności

uczenia głębokiego (architektury wielowarstwowe, sieci rekurencyjne i splotowe, wsteczna propagacja błędu). Dodatkowo zakłada się podstawową wiedzę z zakresu przetwarzania tekstu, ekwiwalentną do przedmiotu "Przetwarzanie i wyszukiwanie informacji" lub "Przetwarzanie języka naturalnego" (wyrażenia regularne, stemming, lematyzacja, stopwords, model bag-of-words, miary podobieństwa tekstu).

Student powinien posiadać umiejętność rozwiązywania podstawowych problemów ze statystyki oraz rachunku prawdopodobieństwa, programowania w języku Python wraz z odpowiednią biblioteką do uczenia głębokiego oraz umiejętność pozyskiwania informacji ze wskazanych źródeł.

W zakresie kompetencji społecznych student musi rozumieć, że w informatyce wiedza i umiejętności bardzo szybko stają się przestarzałe, a także prezentować takie postawy jak uczciwość, odpowiedzialność, wytrwałość, ciekawość poznawcza, kreatywność, kultura osobista, szacunek dla innych ludzi.

Cel przedmiotu

Celem przedmiotu jest zapoznanie studentów z metodologią, zasobami i narzędziami stosowanymi w inżynierii lingwistycznej. Zajęcia skupiają się na omówieniu klasycznych metod statystycznych oraz technik opartych na nowych osiągnięciach głębokiego uczenia maszynowego do problemów takich jak przekład automatyczny, analiza wydźwięku, klasyfikacja tekstów, konstrukcja systemów dialogowych, rozpoznawanie jednostek nazewniczych, analiza składniowa czy modelowanie tematyczne. Ponadto dodatkowym celem przedmiotu jest kształtowanie umiejętności analizowania modeli statystycznych i uczenia maszynowego pod różnymi względami (złożoność obliczeniowa, rodzaj danych uczących i rozmiar próbki, założenia/ograniczenia modelu, metody wnioskowania) oraz ich wykorzystania do rozwiązywania nietrywialnych problemów dot. zasobów tekstowych.

Przedmiotowe efekty uczenia się

Wiedza

1. ma zaawansowaną i pogłębioną wiedzę w zakresie konstrukcji systemów informatycznych przetwarzających język naturalny metodami statystycznymi - [K2st_W3]

2. ma pogłębioną wiedzę o architekturach głębokich sieci neuronowych stosowanych w inżynierii lingwistycznej (w szczególności architektury rekurencyjne i rekursywne) - [K2st_W3]

3. ma zaawansowaną i pogłębioną wiedzę związaną z wybranymi zagadnieniami, takimi jak: modelowanie języka, analiza składniowa, semantyka dystrybucyjna, wykrywanie jednostek nazewniczych, tłumaczenie maszynowe, systemy konwersacyjne - [K2st_W3]

4. ma wiedzę o trendach rozwojowych i najistotniejszych nowych osiągnięciach inżynierii lingwistycznej (w tym nowoczesnych architektur głębokiego uczenia maszynowego) - [K2st_W4]

5. zna zaawansowane metody, techniki i narzędzia stosowane przy budowie systemów dialogowych, translatorów, analizatorów składniowych oraz systemów odpowiadających na pytania - [K2st_W6]

6. zna zaawansowane metody stosowane przy prowadzeniu prac badawczych w zakresie inżynierii lingwistycznej - [K2st_W6]

Umiejętności
1. potrafi pozyskiwać informacje nt. technik inżynierii lingwistycznej z literatury oraz innych źródeł (w języku polskim i angielskim), integrować je, dokonywać ich interpretacji i krytycznej oceny, wyciągać wnioski oraz formułować i wyczerpująco uzasadniać opinie - [K2st_U1]

2. potrafi pozyskiwać odpowiednie zbiory danych do poszczególnych zadań inżynierii lingwistycznej (np. z bazy CLARIN) - [K2st_U1]

3. potrafi planować i przeprowadzać eksperymenty obliczeniowe na danych tekstowych, interpretować uzyskane wyniki i wyciągać wnioski - [K2st_U3]

4. potrafi - przy formułowaniu i rozwiązywaniu zadań inżynierskich - integrować wiedzę z różnych obszarów systemów uczących się, inżynierii oprogramowania oraz lingwistyki. - [K2st_U5]

5. potrafi ocenić przydatność i możliwość wykorzystania nowych osiągnięć uczenia maszynowego do rozwiązywania problemów inżynierii lingwistycznej - [K2st_U6]

6. potrafi określić kierunki dalszego uczenia się i zrealizować proces samokształcenia - w szczególności w zakresie poznawania nowych technik "state-of-the-art" inżynierii lingwistycznej - [K2st_U16]

Kompetencje społeczne
1. rozumie, że w inżynierii lingwistycznej wiedza i umiejętności bardzo szybko stają się przestarzałe - [K2st_K1]

2. rozumie znaczenie wykorzystywania najnowszej wiedzy z zakresu inżynierii lingwistycznej i uczenia maszynowego w rozwiązywaniu problemów badawczych i praktycznych - [K2st_K2]

Metody weryfikacji efektów uczenia się i kryteria oceny

Efekty uczenia się przedstawione wyżej weryfikowane są w następujący sposób:
Ocena formująca:

a) w zakresie wykładów:

- na podstawie odpowiedzi na pytania dotyczące materiału omówionego na poprzednich wykładach

b) w zakresie ćwiczeń:

- na podstawie oceny bieżącego postępu realizacji zadań oraz rozwiązywania zadań przy tablicy

Ocena podsumowująca:

a) w zakresie wykładów weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenę wiedzy i umiejętności wykazanych na testach pisemnych zawierających proste zadania problemowe, pytania otwarte oraz pytania w formie testu wielokrotnego wyboru - test może liczyć od około 5 do kilkunastu takich pytań w zależności od ich formy,

- omówienie wyników testu,

b) w zakresie ćwiczeń weryfikowanie założonych efektów kształcenia realizowane jest przez:

- ocenianie ciągłe, na każdych zajęciach (odpowiedzi ustne przy tablicy) premiowanie przyrostu umiejętności posługiwania się poznanymi zasadami i metodami oraz narzędziami programowymi,

- ocenę i obronę sprawozdań z realizacji zestawów zadań obejmujących zadania obliczeniowe jak i implementacyjne (wymagające wykonania eksperymentów oraz analizy i interpretacji uzyskanych wyników),

- ocenę przygotowanej przez studenta prezentacji omawiającej wybrane zagadnienia z inżynierii lingwistycznej.

Uzyskiwanie punktów dodatkowych za aktywność podczas zajęć, a szczególnie za:

- omówienia dodatkowych aspektów zagadnienia np. poprzez krótkie prezentacje artykułów naukowych,

- uwagi związane z udoskonaleniem materiałów dydaktycznych,

- wskazywanie trudności percepcyjnych studentów umożliwiające bieżące doskonalenia procesu dydaktycznego.

Zarówno w zakresie wykładów jak i ćwiczeń stosuje się następującą skalę ocen: powyżej 51% punktów - dostateczny, 61% - dostateczny plus, 71% - dobry, 81% - dobry plus, 91% - bardzo dobry

Treści programowe

1. Język naturalny jako system: próba zdefiniowania języka, poziom formalny i semantyczny języka (signe, signifiant, signifie), podwójna artykulacja systemu językowego, wariatywność języka w ujęciu synchronicznym, relatywizm językowy, teorie uniwersalistyczne. Wybrane zagadnienia z semantyki: denotacja, referencja, konotacja. Relacje semantyczne i ich użycie w konstrukcji leksykonów komputerowych: antonimia, homonimia, synonimia, polisemia, homonimia, hiponimia, hiperonimia. Słowosieć.

2. Statystyczne modelowanie języka: modele Markova, model 3-gramowy, estymacja największej wiarygodności, ewaluacja modeli języka, interpolacja liniowa modelu 3-gramowego, metoda kubełkowania, metody rozmywania estymat, model back-off Katza oraz ogólny zarys modelu Knesser-Ney'a . Znaczenie wyrazów a ich własności dystrybucyjne. Zaawansowane modele języka: model n-gramów klas, grupowanie semantyczne Brown'a, zależności semantyczne w dendogramie grupowania, neuronowe modelowanie języka (neuronowy model 3-gramowy), problem skalowania modeli neuronowych do dużych słowników (próbkowanie ważone, softmax hierarchiczny). Reprezentacje

rozproszone słów: metody iteracyjne (word2vec), metody globalne (HAL, GloVE), metody dla języków bogatych morfologicznie (FastText, ELMO). Analogie semantyczne i syntaktyczne, problem słów spoza słownika, problem polisemii.

3. Rozpoznawanie jednostek nazewniczych (NER) i rozpoznawanie części mowy (PoS): definicja problemu i sposoby kodowania. Modele generatywne: 3-gramowe ukryte modele Markova (Trigram HMMs), estymacja parametrów modelu, algorytm Viterbiego. Modele dyskryminacyjne: warunkowe modele losowe (CRF), modele Markova o maksymalnej entropii (MEMM), algorytm backward-forward. Neuronowe rozpoznawanie encji nazwanych: rekurencyjne sieci neuronowe (architektura Elmana i Jordana) wykorzystujące reprezentacje rozproszone, przegląd neuronów GRU i LSTM, wykorzystanie warstw CRF, modele dwukierunkowe.

4. Analiza składniowa: drzewo wyprowadzania, drzewo zależnościowe, gramatyki bezkontekstowe, problem wieloznaczności, probabilistyczne gramatyki bezkontekstowe (definicja, estymacja, algorytm CKY, forma normalna Chomskiego), wprowadzenie do zleksykalizowanych probabilistycznych gramatyk bezkontekstowych. Rekursywne sieci neuronowe: RecNN, algorytm wstecznej propagacji błędu przez strukturę, stosowanie błędu rankingowego (definicja, analiza wad i zalet), metody rekurencyjne.

5. Tłumaczenie maszynowe: źródła trudności związane z automatyzacją przekładu, piramida Vauquois, model IBM 1, model IBM 2, estymacja parametrów z korpusu zawierającego przypisania wyrażeń do ich tłumaczenia, estymacja parametrów z korpusu równoległego, algorytm maksymalizacji oczekiwań, wstęp do tłumaczenia frazowego, ewaluacja systemów tłumaczenia maszynowego (ocena ekspercka i automatyczna - BLEU). Neuronowe metody tłumaczenia maszynowego: podejścia typu enkoder/decoder, podejścia z atencją, reprezentacje rozproszone niezależne od języka, współdzielenie enkodera, technika backtranslation. Transfer lingwistyczny.

6. Klasyfikacja tekstu. Reprezentacja worka słów (bag-of-words) z reprezentacji wektorowej, klasyfikacja z ekstremalną liczbą cech (haszowanie cech n-gramowych, metoda tokenów spersonalizowanych). Sieci splotowe do klasyfikacji tekstu: warstwa splotu 1D (na znakach i słowach), warstwa pooling-over-time, idea wielu kanałów w kontekście reprezentacji rozproszonej. Studia przypadku: identyfikacja języka, przypisywanie autorstwa.

7. Analiza wydźwięku: klasyczne podejścia nienadzorowane, model sentymentu Ossgood'a, tworzenie cech dla algorytmów uczenia maszynowego, problem negacji, leksykony sentymentu, reprezentacje rozproszone słów i ich sentymentu, analiza wydźwięku krótkich wypowiedzi użytkowników na przykładzie sieci Twitter.

8. Transfer wiedzy w inżynierii lingwistycznej: metody mapowania zagnieżdżeń słów w sposób nadzorowany i nienadzorowany, wykorzystanie wiedzy z modeli języka do klasyfikacji tekstu i innych zadań, architektura transformer – BERT, Universal Sentence Encoder, GPT-3 i podobne.

9. Przegląd wybranych zagadnień inżynierii lingwistycznej (wybór wg zainteresowań studentów): metody text-to-speech, techniki rozpoznawania mowy (ASR), budowanie grafów wiedzy z tekstów, question answering, information retrieval (modele DSSM), systemy dialogowe, modelowanie tematyczne.

Powyższy spis zagadnień obejmuje zarówno wykłady jak i ćwiczenia - obie formy zajęć są integralną częścią przedmiotu tj. zagadnienia omawiane w czasie ćwiczeń/wykładów nie są ponownie przerabiane na innej formie zajęć. Rozkład materiału pomiędzy wykłady i ćwiczenia odbywa się dynamicznie, w zależności od tempa pracy grupy.

Metody dydaktyczne

1. Wykład: prezentacja multimedialna, ilustrowana przykładami podawanymi na tablicy

2. Ćwiczenia: prezentacja multimedialna, ilustrowana przykładami podawanymi na tablicy, ćwiczenia praktyczne (m.in. obliczeniowe na tablicy), dyskusja zagadnień i rozwiązań

Literatura

Podstawowa

1. Jurafsky D., Martin J.H.: Speech and Language Processing, III edycja, Pearson/Prentice Hall, 2018 (dostęp online: https://web.stanford.edu/~jurafsky/slp3/)

2. Li Deng, Yang Liu: Deep Learning in Natural Language Processing. Springer, 2018 (dostęp poprzez eZasoby biblioteki PP)

Uzupełniająca

1. Goodfellow I., Yoshua B., Courville A.: Deep Learning. Systemy uczące się., PWN, 2018

2. Lango M., Brzeziński D., Stefanowski J.: PUT at SemEval-2016 Task 4: The ABC of Twitter Sentiment Analysis, Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), 2016

3. Mykowiecka, A: Inżynieria lingwistyczna : komputerowe przetwarzanie tekstów w języku naturalnym, Wydawnictwo PJWSTK, 2007

Bilans nakładu pracy przeciętnego studenta

	Godzin	ECTS
Łączny nakład pracy	120	4
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	60	2
Praca własna studenta (studia literaturowe, przygotowanie do ćwiczeń, przygotowanie do testów, wykonanie zadań domowych i prezentacji)[1]	60	2

1 niepotrzebne skreślić lub dopisać inne czynności