KARTA OPISU MODUŁU KSZTAŁCENIA

Nazwa modułu/przedmiotu	Kod
Inżynieria lingwistyczna	**1010512331010510206**

Kierunek studiów	Profil kształcenia (ogólnoakademicki, praktyczny)	Rok / Semestr
Informatyka	**ogólnoakademicki**	**2 / 3**

Ścieżka obieralności/specjalność	Przedmiot oferowany w języku:	Kurs (obligatoryjny/obieralny)
Inteligentne technologie informatyczne	**polski**	**obligatoryjny**

Stopień studiów:	**Forma studiów** (stacjonarna/niestacjonarna)
II stopień	**stacjonarna**

Godziny	Liczba punktów
Wykłady: **30** Ćwiczenia: **30** Laboratoria: **-** Projekty/seminaria: **-**	**4**

Status przedmiotu w programie studiów (podstawowy, kierunkowy, inny)	(ogólnouczelniany, z innego kierunku)
kierunkowy	**z danego kierunku**

Obszar(y) kształcenia i dziedzina(y) nauki i sztuki	**Podział ECTS (liczba i %)**
nauki techniczne	**4 100%**
nauki techniczne	**4 100%**

Odpowiedzialny za przedmiot / wykładowca:

Mateusz Lango
email: mateusz.lango@cs.put.poznan.pl
tel. 61 665 21 24
Wydział Informatyki
Piotrowo 2, 60-965 Poznań

Wymagania wstępne w zakresie wiedzy, umiejętności, kompetencji społecznych:

1	**Wiedza:**	Student rozpoczynający ten przedmiot powinien posiadać podstawową wiedzę z rachunku prawdopodobieństwa i statystyki, a także pogłębioną wiedzę z uczenia maszynowego (klasyfikatory złożone, algorytmy k-NN, Naive Bayes, SVM), a w szczególności uczenia głębokiego (architektury wielowarstwowe, sieci rekurencyjne, 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).
2	**Umiejętności:**	Student powinien posiadać umiejętność rozwiązywania podstawowych problemów ze statystyki oraz rachunku prawdopodobieństwa, programowania w co najmniej jednym języku obiektowym wraz z odpowiednią biblioteką do uczenia głębokiego oraz pozyskiwania informacji ze wskazanych źródeł.
3	**Kompetencje społeczne**	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 encji nazwanych, 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) oraz ich wykorzystania do rozwiązywania nietrywialnych problemów dot. zasobów tekstowych.

Efekty kształcenia i odniesienie do kierunkowych efektów kształcenia

Wiedza:

1. ma zaawansowaną wiedzę szczegółową 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]

Sposoby sprawdzenia efektów kształcenia

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" przez studenta sprawozdań z realizacji zestawów zadań obejmujących zadania obliczeniowe jak i implementacyjne (wymagające wykonania eksperymentów oraz analizy i interpretacji uzyskanych wyników),

- ocenę sprawozdania z realizacji zadań.

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, znaczenie leksykalne a użycie

leksemu. 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 model 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, modele rekurencyjne RNNLM, model Colloberta&Wetsona, problem skalowania do dużych słowników, hierarchiczny model logarytmiczno-biliniowy). Reprezentacje rozproszone słów: metody iteracyjne (word2vec), metody globalne (GloVE, Hellinger-PCA), metody dla języków bogatych morfologicznie (FastText, ELMO). Analogie semantyczne i syntaktyczne, problem słów spoza słownika, problem polisemii.

3. Rozpoznawanie encji nazwanych (NER) i rozpoznawanie części mowy (PoS): definicja problemu, modele generatywne, 3-gramowe ukryte modele Markova (Trigram HMMs), estymacja parametrów modelu, algorytm Viterbiego. Warunkowe modele losowe (CRF) i ich neuronowe rozszerzenie. Neuronowe rozpoznawanie encji nazwanych: rekurencyjne sieci neuronowe (architektura Elmana i Jordana) wykorzystujące reprezentacje rozproszone, przegląd neuronów GRU i LSTM.

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, strategie bezpośrednie, transferowe i wykorzystujące interlingua, 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 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. Metody semantyki kompozycyjnej do tworzenia reprezentacji zdań (model RNTN). 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, dedykowane metody selekcji cech dla analizy sentymentu, metody anotacji zasobów lingwistycznych (MaxDiff), analiza wydźwięku krótkich wypowiedzi użytkowników na przykładzie sieci Twitter.

8. Modelowanie tematyczne (topic modelling): algorytm LDA (założenia, estymacja), neuronowe modelowanie tematów na przykładzie Relationship Modelling Network, zastosowania modelowania tematycznego do aspektowej analizy sentymentu oraz wykrywania zmian rodzaju relacji pomiędzy bohaterami w tekstach literackich.

9. Systemy dialogowe: systemy typu chatbot a systemy zadaniowe, architektura systemu dialogowego. Moduł rozumienia języka: pojęcie ramki semantycznej, modele neuronowe do: wykrywania domeny, rozpoznawania zamiaru, wypełniania slotów (m.in. pointer networks), modele przywidujące całą ramkę semantyczną, śledzenie stanu konwersacji (modele rekurencyjne, modele deleksykalizowane, model Neural Belief Tracker). Moduł zarządzania dialogiem: trzy rodzaje modeli (regułowe, nadzorowane i ze wzmocnieniem), symulowanie użytkownika. Moduł generacji języka (podejścia rekurencyjne, metody multiplikacyjne, metody oparte na schemacie enkoder/dekoder, metody z atencją, zbieranie danych: symulowanie regułowe, metoda czarnoksiężnika z Oz). Systemy dialogowe jednomodułowe.

10. Question Answering: odpowiadanie na pytania oparte na wyszukiwarce (automatyczna formulacja zapytania, wykrycie typu odpowiedzi, wybór odpowiedzi), odpowiadanie oparte na bazie wiedzy (modele regułowe i nadzorowane, przykład IBM Watson), miary ewaluacji (MRR, WEBQUESTIONS)

11. Przegląd wybranych zagadnień inżynierii lingwistycznej: metody text-to-speech, techniki rozpoznawania mowy (ASR), budowanie grafów wiedzy z tekstów.

Literatura podstawowa:

1. Jurafsky D., Martin J.H.: Speech and Language Processing, III edycja, online draft, 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)

Literatura uzupełniająca:

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

2. Aurélien Géron: Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow, Helion, 2018

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

4. 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

5. Charu C. Aggarwal, Machine Learning for Text, Springer, 2018 (dostęp poprzez eZasoby Biblioteki PP)

Bilans nakładu pracy przeciętnego studenta

Czynność	Czas (godz.)

1. udział w ćwiczeniach	30
2. przygotowanie do ćwiczeń	8
3. rozwiązywanie (w ramach pracy własnej) zbiorów zadań z ćwiczeń	20
4. udział w konsultacjach związanych z realizacją procesu kształcenia (częściowo mogą być realizowane drogą elektroniczną)	2
5. udział w wykładach	30
6. zapoznanie się ze wskazaną literaturą / materiałami dydaktycznymi	5
7. przygotowanie do testów zaliczeniowych	10

Obciążenie pracą studenta

forma aktywności	godzin	ECTS
Łączny nakład pracy	105	4
Zajęcia wymagające bezpośredniego kontaktu z nauczycielem	62	2
Zajęcia o charakterze praktycznym	50	2