

OBDELAVA NARAVNEGA JEZIKA

UČNI NAČRT PREDMETA/COURSE SYLLABUS

Predmet:	Obdelava naravnega jezika
Course title:	Natural language processing
Članica nosilka/UL Member:	UL FF

Študijski programi in stopnja	Študijska smer	Letnik	Semestri	Izbirnost
Digitalno jezikoslovje, druga stopnja, magistrski	Ni členitve (študijski program)	1. letnik	2. semester	obvezni

Univerzitetna koda predmeta/University course code: 0642876

Predavanja /Lectures	Seminar /Seminar	Vaje /Tutorials	Klinične vaje /Clinical tutorials	Druge oblike študija /Other forms of study	Samostojno delo /Individual student work	ECTS
45	10	20	0	0	75	5

Nosilec predmeta/Lecturer: Marko Robnik Šikonja

Vrsta predmeta/Course type: Obvezni/Compulsory

Jeziki/Languages:

Predavanja/Lectures:	Angleščina, Slovenščina
Vaje/Tutorial:	Angleščina, Slovenščina

Pogoji za vključitev v delo oz. za opravljanje študijskih obveznosti:

Prerequisites:

/

/

Vsebina:

Vsebina predmeta temelji na izboru sodobnih statističnih tehnik obdelave naravnega jezika podkrepjenih s praktično rabo. V predavanjih predstavimo glavne pristope in pojasnimo delovanje posameznih metod in njihovo teoretično ozadje. V okviru laboratorijskih vaj znanje povežemo s praktično rabo in ga utrdimo z uporabo odprtokodnih sistemov za obdelavo naravnega jezika. Študenti rešujejo naloge, ki temeljijo na realnih raziskovalnih in

Content (Syllabus outline):

The syllabus is based on a selection of modern statistical natural learning techniques and their practical use. The lectures introduce the main tasks and techniques, explain their operation and theoretical background. During practical sessions and seminars the gained knowledge is applied to language practical task using open source tools. Student investigate and solve assignments,

praktičnih problemih, pretežno v slovenskem in angleškem jeziku. Uvod: motivacija, razumevanje jezika, Turingov test, tradicionalni in statističen pristop. Jezikovni viri: korpusi, slovarji, tezavri, omrežja in semantične baze, pregled orodij. Lingvistika: fonologija in morfologija, sintaktična analiza, formalne gramatike. Uporaba avtomatov in gramatik: avtomati in algoritmi za iskanje nizov, prepoznavanje sintakse, gramatično razčlenjevanje. Oblikoslovno označevanje besedil: vrste oznak, lematizacija, ngrami, skriti markovski model, označevanje s pravili. Računska in leksikalna semantika: predstavitev pomena, metode s pravili, leksikalna semantika. Razvrščanje besedil in mere podobnosti: kosinusna razdalja, jezikovna omrežja in grafi, WordNet, vektorska predstavitev, uteževanje vektorjev, semantična korelacija. Tekstovno rudarjenje: prilagojene klasifikacijske metode, metoda podpornih vektorjev na dokumentih, izbira atributov. Globoka omrežja in besedila: predstavitev besedil za uporabo v globokih nevronske mrežah, avtoenkoderji, rekurzivne nevronske mreže. Povzemanje: predstavitev besedil, matrična faktorizacija, ekstrakcijske metode, povpraševane metode. Strojno prevajanje: jezikovni model, prevajalni model, poravnava jezikov, parametri modelov, izzivi v prevajanju. Dopolnjevanje besedil z drugimi viri informacij: heterogena omrežja, predstavitev word2vec, heterogeni ansambli klasifikatorjev, analiza povezav. Metodologija in evalvacija pri obdelavi naravnega jezika	based on real-world research and commercial problems from English and Slovene languages. Introduction to natural language processing: motivation, language understanding, Turing test, traditional and statistical approaches. Language resources: corpora, dictionaries, thesauruses, networks and semantic data bases, overview of tools. Linguistics: phonology and morphology, syntactical analysis, formal grammars. Using automata and grammars: automata and algorithms for searching strings, syntax parsing, dependency parsing. Part-of-speech tagging: types of tags, lemmatization, ngrams, Hidden Markov model, rule-based tagging. Computational and lexical semantics: semantic representations, rule-to-rule approaches, semantic role labelling. Clustering words and text similarity measures: cosine distance, language networks and graphs, WordNet, vector representation, vectorweighting, semantic correlation. Text mining: adaptation of classification methods to the specifics of text, support vector machines for language, feature selection. Deep networks for text: document representations for deep neural networks, autoencoders, recurrent neural networks. Text summarization: text representations, matrix factorization, multi-document summarization, extractive methods, query based methods. Machine translation: language model, translation model, alignment model, challenges in machine translation. Augmenting text with other data sources: heterogeneous networks, word2vec representation, heterogeneous ensembles of classifiers, link analysis. Methodology and evaluation in NLP.
--	--

Temeljna literatura in viri/Readings:

Jurafsky, David and Martin, James H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, 2nd and 3rd draft. Upper Saddle River, NJ: Prentice-Hall, 2009 and 2017.

Aggarwal, Charu C., and Zhai, ChengXiang. Mining text data. Springer Science & Business Media, 2012.

Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python. O'Reilly Media, Inc., 2009.

Cilji in kompetence:

Študenti se bodo naučili teorije in rabe osnovnih algoritmov in pristopov na področju obdelave naravnega jezika.

Objectives and competences:

Upon completion of the course, students shall be able to explain and apply fundamental algorithms and techniques in the area of natural language processing.

Študenti bodo: razumeli pristope k analizi sintakse in semantike na področju obdelave naravnega jezika; razumeli pristope k povzemanju dokumentov; razumeli delovanje statističnih pristopov k strojnemu prevajanju, razumeli uporabo metod strojnega učenja v obdelavi naravnega jezika: skritega Markovskega modela, verjetnostnih kontekstno neodvisnih gramatik in algoritma EM, znali uporabiti orodja za obdelavo naravnega jezika.	In particular, students will: understand approaches to syntax and semantics in NLP, understand approaches to summarization understand statistical approaches to machine translation, understand machine learning techniques used in NLP, including hidden Markov models, probabilistic context-free grammars, and the EM algorithm as applied within NLP, know how to apply standard natural language processing tools.
---	--

Predvideni študijski rezultati:

Ob zaključku predmeta bodo študenti: razumeli pristope k analizi sintakse in semantike na področju obdelave naravnega jezika; znali ovrednotiti pristope k povzemanju dokumentov; razlikovali med različnimi statističnimi pristopi k strojnemu prevajanju, uporabljali in prilagajali metode strojnega učenja za obdelavo naravnega jezika, uporabljali in kritično vrednotili orodja za obdelavo naravnega jezika, poznali obstoječe in znali zasnovati nove jezikovne vire, uporabljali vektorske vložitve besedil in jih prilagajali novi okoliščinam	Intended learning outcomes: Upon completion of the course, students will: understand approaches to syntax and semantics in NLP, evaluate approaches to summarization differentiate between statistical approaches to machine translation, use and adapt machine learning techniques for NLP, apply and critically evaluate natural language processing tools, know the existing language resources and be able to design new ones, use vector embeddings of text and adapt them to new contexts
---	---

Metode poučevanja in učenja:

Predavanja, laboratorijske vaje, delo v majhnih skupinah, javne predstavitve projektov	Learning and teaching methods: Lectures, lab work, work in small groups, public presentations of projects.
---	---

Načini ocenjevanja:

Delež/Weight

Assessment:

Sprotno preverjanje (domače naloge, projektno delo, javne predstavitve)	50,00 %	Continuing (homework, project work, public presentations)
Končno preverjanje (pisni in ustni izpit)	50,00 %	Final (written and oral exam)

Reference nosilca/Lecturer's references:

ROBNIK ŠIKONJA, Marko, KONONENKO, Igor. Theoretical and empirical analysis of ReliefF and RReliefF. Mach. learning, 2003, vol. 53, pp. 23-69.

ROBNIK ŠIKONJA, Marko. Data generators for learning systems based on RBF networks. IEEE transactions on neural networks and learning systems, May 2016, vol. 27, no. 5, pp. 926-938.

ROBNIK ŠIKONJA, Marko, VANHOOF, Koen. Evaluation of ordinal attributes at value level. Data mining and knowledge discovery, 2007, vol. 14, no. 2, pp. 225-243.

ROBNIK ŠIKONJA, Marko, KONONENKO, Igor. Explaining classifications for individual instances. IEEE trans. knowl. data eng. May 2008, vol. 20, no. 5, pp. 589-600.

KRANJC, Janez, ORAČ, Roman, PODPEČAN, Vid, LAVRAČ, Nada, ROBNIK ŠIKONJA, Marko. ClowdFlows: online workflows for distributed big data mining. FGCS, 2017, vol. 68, pp. 38-58

Celotna bibliografija je dostopna na SICRISu: / Complete bibliography is available in SICRIS:
<http://sicris.izum.si/search/rsr.aspx?lang=slv&id=8741>.