

| | | | | | | |
|------|---------------------------|-------------------------|---|---|---|-----------|
| I051 | Izborni 2. i 3. godina | Računalno jezikoslovlje | P | S | V | ECTS 6 |
| | | | 2 | 0 | 2 | |

Cilj predmeta. Stjecanje znanja o prirodnom jeziku i njegovim računalnim obradbama, preko algoritama i programa za morfologiju, sintaksu i semantiku, s posebnim naglaskom na izvlačenju informacije iz dokumenata, te stvaranju povezanih *LLOD* podataka (*linguistic linked open data*). Upoznati *Python* module za obradu teksta, te ovladati tehnikama prepoznavanja uzoraka. Primijeniti matematička i statistička znanja u modeliranju i računalnoj obradbi hrvatskoga jezika.

Potrebna predznanja. Uvod u računalnu znanost. Strukture podataka i algoritmi I.

Sadržaj predmeta.

1. Riječ, oblici i vrste riječi, fraze, rečenice. Služba riječi u rečenici. Morfologija/oblikoslovlje, sintaksa/skladnja i semantika/smisao.
2. Uvod u računalno jezikoslovlje. Gramatike umjetnih i prirodnih jezika. Jezik kao podskup slobodnog monoida (Kleene closure).
3. Kategorijska ili C-gramatika; Generativna ili PS gramatika (phrase structure grammar).
4. Gramatičke izvedenice; LA (left-associative) gramatika.
5. Morfosintaktičko označivanje (lemma tagging, syntactic SPO tagging, PoS - part of speech tagging, category annotation). Algoritmi u oblikoslovlju, XML i vizualno sintaktičko označivanje.
6. Semantičke označiteljske klase (role & sense tagging). Semantička stabla. Parseri
7. Regularni izrazi. Izvlačenje informacije iz rečenice. Nadgledano i nenadgledano strojno učenje.
8. Rudarenje teksta. Klasifikacija i grupiranje (clustering) dokumenata, vjerojatnostni modeli.
9. Frekvencijski (tf-idf) pristup, latentna semantička analiza (LSA).
10. Mrežne ontologije i povezani podaci, SparQL upiti, triplestore spremišta (s-p-o: subjekt-objekt-predikat) trojaca informacije.

ISHODI UČENJA

| R.b. | ISHODI UČENJA |
|------|---|
| 1. | Demonstrirati znanje i razumijevanje koje osigurava temelj za originalni razvoj i primjenu ideja matematike i računalstva u jezikoslovlju |
| 2. | Primijeniti svoje znanje, razumijevanje i sposobnosti rješavanja problema u širem kontekstu vezanom uz jezikoslovlje. |
| 3. | Integrirati nova znanja za uspješno rješavanje programskih problema iz dizajniranja i modeliranja lingvističkih podataka. |
| 4. | Jasno i nedvosmisleno obrazložiti svoje zaključke stručnjacima i laicima, zasnovanima na znanju i argumentima. |
| 5. | Primijeniti stečene vještine učenja na cjeloživotno obrazovanje iz ovog područja. |

**POVEZIVANJE ISHODA UČENJA, ORGANIZACIJE NASTAVNOG PROCESA I
PROCJENA ISHODA UČENJA**

| ORGANIZACIJA A NASTAVNOG PROCESA | ECTS | ISHOD UČENJA ** | AKTIVNOST STUDENATA* | METODA PROCJENE | BODOVI | |
|---|------|-----------------------|--|--|--------|-----|
| | | | | | min | max |
| Pohađanje predavanja | 1 | 1-5 | Prisutnost na nastavi, rasprava, timski rad i samostalan rad na zadacima | Potpisne liste, praćenje aktivnosti na nastavi | 0 | 10 |
| Zadace | 1 | 1-4 | Samostalno rješavanje problemskih zadataka | Provjera točnih rješenja (ocjenjivanje) | 18 | 30 |
| Provjera znanja (kolokvij) | 2 | 1-4 | Priprema za pismenu provjeru znanja | Provjera točnih odgovora (ocjenjivanje) | 16 | 30 |
| Završni ispit | 2 | 1-4 | Izrada projekta | Usmeni ispit | 16 | 30 |
| UKUPNO | 6 | | | | 50 | 100 |

Izvođenje nastave i vrednovanje znanja. Predavanja će bit ilustrirana programima pisanima u *Python + NLTK (natural language toolkit)* alatu, uz pomoć više razvijenih *Python* modula za hrvatski jezik (korpus, oblikoslovlje, semantika). U praktični dio nastave uvest će se upoznavanje *web2py* <http://www.web2py.com> MVC okvira (*model-view-controller framework*), kako bi svi studentski uradci automatski bili uključeni u mrežno (*web*) okruženje. Predavanja i vježbe su obavezne. Ispit se sastoji od uspješno izrađenog projektnog uradka, a polaže se nakon odslušanih predavanja i obavljenih vježbi te uspješno položenih domaćih zadaća i kolokvija. U projektnim radovima studenti će koristiti dio programskih modula s poznatih izvora računalnog jezikoslovlja: <http://www.nltk.org/>, <http://www.clips.ua.ac.be/> i <http://scikit-learn.org/stable/>.

Može li se predmet izvoditi na engleskom jeziku: Da

Osnovna literatura:

1. S. Bird, E. Klein, E. Loper: "Natural Language Processing with Python", <http://www.nltk.org/book/>, O'Reilly Media, 2009.

Dopunska literatura:

1. C. Chiarcos, S. Nordhoff, S. Hellmann: "Linked Data in Linguistics", Springer-Verlag, 2012.
2. M. Essert, K. Štrkalj Despot: "Python za jezikoslovce (*u pripremi*)", IHJJ, 2016.
3. R. Hausser: "Foundations of Computational Linguistics: Human-Computer Communication in Natural Language", 3Ed, Springer, 2014.

4. R. Grishman: "Computational linguistics: an introduction. (Studies in natural language processing)" , Cambridge University Press, third edition, 1994.
5. D. Jurafsky, J. H. Martin: "Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition", Pearson Education, 2009.
6. R. Delmonte: "Computational linguistic text processing: lexicon, grammar, parsing and anaphora resolution", Published by Nova Science Publishers, Inc., New York. 2008.
7. S. P. Abney: "Semisupervised learning in computational linguistics", Taylor & Francis Group/ Chapman & Hall/CRC, 2008.
8. C. J. Fillmore: "Semantics of Natural Language", Springer Netherlands, 1973.
9. M. W. Berry: "Survey of text mining: clustering, classification, and retrieval", Springer-Verlag New York, Inc., 2004.
10. C. C. Aggarwal, C. Zhai: "Mining Text Data", Springer Science+Business Media, 2012.
11. P. Harrington: "Machine Learning in Action", Manning Publications Co., 2012.