D – Charakteristika studijního předmětu

Název studijního předmětu	**Úvod do počítačové lingvistiky**				
Typ předmětu	povinně volitelný			**doporučený ročník / semestr**	2/L
Rozsah studijního předmětu	2 + 2	**hod. za týden**	4	**kreditů**	5
Jiný způsob vyjádření rozsahu	—				
Způsob zakončení	zkouška			**Forma výuky**	př + cv

Další požadavky na studenta

Podmínkou zápočtu je aktivní účast na cvičeních. Zkouška je ústní.

Vyučující

doc. Ing. Josef Chaloupka, Ph.D.

Stručná anotace předmětu

Předmět je úvodem do problematiky počítačové lingvistiky respektive obecněji do problematiky efektivního zpracování textů na počítači. Kromě skriptovacích programovacích jazyků (např. Perl) se studenti seznámí se základními metodami formálního popisu přirozené jazyka, technikami jeho statistického zpracování, značkování a syntaktické analýzy, a také praktickými aspekty souvisejícími s používáním regulárních výrazů pro vyhledávání a zpracování textů.

Témata přednášek:

1. Základní pojmy, praktické aplikace a motivační úlohy. Úvod do Perlu.
2. Další důležité vlastnosti jazyka Perl (vstup a výstup, cykly, pole, asociativní pole, podprogramy).
3. Reference a tvorba datových struktur pomocí referencí, moduly a jejich využití.
4. Regulární výrazy, jazyky a konečné automaty. Využití regulární výrazů v jazyku Perl a dalších programech - hledání, nahrazování atd.
5. Připomenutí vybraných pojmů z teorie pravděpodobnosti a informace. Podmíněná a sdružená pravděpodobnost, entropie a její vlastnosti, sdružená a podmíněná entropie, Kullback-Leiblerova vzdálenost a vzájemná informace.
6. Úloha zašuměného kanálu a její řešení pomocí Bayesova teorému. Jazykové modely, jejich výhody a využití v různých rozpoznávacích systémech.
7. Metody vyhlazování jazykových modelů.
8. Kolokace a metody jejich vyhledávání. Slovní třídy a jejich aplikace v jazykových modelech. Princip automatické tvorby slovních tříd.
9. Lingvistika - úrovně popisu a zpracování přirozeného jazyka. Syntax a bezkontextové gramatiky.
10. Textové korpusy a principy metod pro značkování textu (tagging).

Informace ke kombinované nebo distanční formě		
Rozsah konzultací (soustředění)		**hodin za týden**
Rozsah a obsahové zaměření individuálních prací studentů a způsob kontroly		

Studijní literatura a studijní pomůcky

Doporučená literatura:

Manning C. D. and H. Schütze. *Foundations of Statistical Natural Language Processing.*. The MIT Press. ISBN 0-262-13360-1

Online kurzy a tutoriály:

https://www.coursera.org/course/nlp
https://www.coursera.org/course/nlangp