Universitatea „Alexandru Ioan Cuza” din Iași
Facultatea .....Informatică.....
Departamentul .....Informatică.....
Domeniul de studii.....Informatică.....

FISA DISCIPLINEI

DENUMIREA DISCIPLINEI	**Procesarea statistică a limbajului natural**				COD: MLC2102
CICLUL DE STUDII (L-licență/M-master/D-doctorat) ŞI ANUL DE STUDIU (1,2,3,4)	**M**	Semestrul		STATUTUL DISCIPLINEI (OB-obligatorie/OP-opţională/F-facultativă)	**OB**

NUMĂRUL ORELOR PE SAPTĂMÂNĂ				TOTAL ORE SEMESTRU	TOTAL ORE ACTIVITATE INDIVIDUALA*	NUMĂR DE CREDITE	TIPUL DE EVALUARE (**P**-pe parcurs, **C**-colocviu, **E**-examen, **M**-mixt)	LIMBA DE PREDARE
C	S	L	Pr.					
2	2			56	124	8	M	Mixt: Română/Engleză

TITULARUL ACTIVITĂȚILOR DE CURS	GRADUL DIDACTIC ŞI ŞTIINȚIFIC, PRENUMELE, NUMELE	DEPARTAMENTUL
	LECTOR DR. ANCA VITCU	Informatică

TITULARUL ACTIVITĂȚILOR DE SEMINAR/L.P.	GRADUL DIDACTIC ŞI ŞTIINȚIFIC, PRENUMELE, NUMELE	DEPARTAMENTUL
	LECTOR DR. ANCA VITCU	Informatică

DISCIPLINE ABSOLVITE ANTERIOR	Probabilități și statistică (Anul I)

OBIECTIVE*	Utilizarea teoriei probabilităților și statisticii în NLP: Prelucrare: modele probabilistice; Învățare; Evaluare
COMPETENȚE SPECIFICE ACUMULATE	
COMPETENȚE PROFESIONALE**	Informatică - C3
COMPETENȚE TRANSVERSALE	Informatică - CT3
CONTINUTUL CURSULUI	Cursul este alcătuit din două module obligatorii plus un modul opțional: **Modulul 1**. Statistică – elemente de teoria probabilităților și elemente de statistică clasică (descriptivă și inferențială – teste de semnificație, analiza exploratorie a datelor - analiza cluster, analiza componentelor principale, MDS; clasificare și predicție; procese stohastice (procese Markov, HMM)) **Modulul 2**. Aplicații ale metodelor statistice în NLP – prelucrare, învățare, evaluare: **Probability Language Modeling -** String probabilities; Types of models: N-gram models, N-class models, probabilistic context-free grammars (PCFG); Parameter estimation (training): maximum likelihood estimation (MLE), sparse data, smoothing techniques (additive smoothing, held-out estimation, Good-Turing reestimation, back-off smoothing, linear interpolation); Managing the size of the model; Evaluation methods: probability, entropy and related measures **Part-of-Speech Tagging -** The n-class model (lexical probabilities, contextual probabilities, HMM); Parameter estimation: tagged training data (MLE), untagged training data (EM), sparse data and smoothing; Algorithms for HMMs : string probability (Forward, Backward, Forward-Backward), optimal state sequence (Viterbi), unsupervised parameter estimation: Baum-Welch **Syntactic Parsing -** Probabilistic context-free grammars (PCFG): probabilities for rules, strings and parse trees, parameter estimation, parsing algorithms; Variations of stochastic grammars (Markov grammars, lexicalized grammars); Alternative models: decision tree parsing, stochastic automata, data- orientated parsing **Word Sense Disambiguation -** Word sense disambiguation based on Bayesian classifiers (supervised learning, bilingual corpora, thesaurus classes, unsupervised learning); Combining statistics with a priori constraints: (one sense per collocation, one sense per discourse); Word sense discrimination (clustering) **Machine Translation -** Translation as a noisy channel problem: translation of source text into target text, language model, translation model; Alignment: sentence alignment, word alignment; Translation modeling: translation, fertility, distortion. **Tree-Based Models -** Synchronous Grammars; Learning Synchronous Grammars; Decoding by Parsing **Evaluation** (Manual Evaluation, Automatic Evaluation, Hypothesis Testing, Task- Oriented Evaluation); Empirical evaluation of accuracy: Independent test data, Supervised or unsupervised, Gold standard evaluation; Descriptive statistics/measures of accuracy: Accuracy rate (percent correct), Recall, Precision, Logprob; Statistical inference: Confidence intervals, Hypothesis testing (significance) **Modulul 3 (opțional)**. Metode bayesiene utilizate în NLP

BIBLIOGRAFIE (SELECTIVĂ)	1. Berry Michael, Kogan Jacob (2010) - Text mining : applications and theory, Wiley 2. Feldman Ronen, Sanger James (2007) – The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data, (second ed.), Cambridge University Press 3. Manning D. Christopher, Raghavan Prabhakar, Schütze Hinrich (2009) – An Introduction to Information Retrieval. Cambridge University Press 4. Manning D. Christopher , Schütze Hinrich (1999) - Foundations of Statistical Natural Language Processing, MIT Press. http://nlp.stanford.edu/fsnlp/ 5. Witten H. Ian, Frank Eibe (2005) - Data Mining:Practical Machine Learning Tools and Techniques. Second edition. Elsevier. http://www.cs.waikato.ac.nz/~ml/weka/book.html
CONȚINUTUL LUCRĂRILOR DE SEMINAR/LABORATOR	Seminariile sunt adaptate temelor abordate în lucrările de dizertație din cadrul masterului de lingvistică.
BIBLIOGRAFIE (SELECTIVĂ)	1. Kao Anne, Poteet R. Stephen (Eds) (2007) - Natural Language Processing and Text Mining, Springer 2. Bird Steven, Klein Ewan, Loper Edward (2009) – Natural Language Processing with Python, O'Reilly Media 3. Hall N. Joseph, Schwartz L. Randal (1998) - Effective Perl Programming Writing Better Programs with Perl, Addison-Wesley 4. Perkins Jacob (2010) - Python Text Processing with NLTK 2.0 Cookbook, Packt Publishing ***Text mining with R (tutorials) ***Journal of Statistical Software, Vol 25, Issue 5, 2008
REPERE METODOLOGICE***	Modulul teoretic este construit pornind de la exemple și studii de caz.

EVALUARE	Metodele	Construirea și implementarea modelului statistic asociat temei studiate în cadru lucrării de dizertație.
	Forme	Activitate de seminar: alegerea modelului, parcurgerea etapelor de implementare. Prezentarea finală a proiectului
	ponderea formelor de evaluare în formula notei finale	60% activitatea de seminar, 40% prezentarea finală
	standardele minime de performanță****	Înțelegerea statisticii ca instrument de analiză în NLP și abilitatea de a construi modele proprii adaptate unui anumit context.

* obiectivele sunt formulate în funcție de grila competențelor profesionale pentru programul de studii
** la nivel de descriptor
*** strategia didactică, materiale, resurse
**** raportate la competențele formulate la Obiective sau la Standardele minime de performanta din grila 1L/1M după caz

Data completării Semnătura titularului de curs Semnătura titularului de seminar/l.p.

Data avizării în departament Semmnătura directorului de departament