MD-2068, CHIŞINĂU, STR. STUDENŢILOR, 9/7, TEL: 022 50-99-63, www.utm.md

DATA MINING

1. Date despre unitatea de curs

Facultatea	Calculatoare, Informatică și Microelectronică				
Catedra/Departamentul	Ingineria Software și Automatică				
Ciclul de studii	Studii superioare de master, ciclul II				
Programul de studii	Tehnologia informației pentru afaceri				
Anul de studii	**Semestrul**	**Tip de evaluare**	**Categoria formativă**	**Categoria de opţionalitate**	**Credite ECTS**
II (învăţământ cu frecvenţă);	3	E-examen	S – unitate de curs de specialitate	A - unitate de curs opțională	5

2. Timpul total estimat

Total ore în planul de învăţământ	**Din care**				
	Ore auditoriale		**Lucrul individual**		
	Curs	**practice**	**Proiect de an**	**Studiul materialului teoretic**	**Pregătire aplicaţii**
150	20	20	-	110	-

3. Precondiţii de acces la unitatea de curs

Conform planului de învăţământ	Analiza si exploratorie a datelor, managmentul datelor, fundamente ale tehnologiei informatiei
Conform competenţelor	Posedarea cunoștințelor medii de utilizare a instrumentelor Mirosoft Office. Posedarea abilitatilor de folosire a calculatorului. Competențe de înțelegere și aplicare a formulelor matematice.

4. Condiţii de desfăşurare a procesului educaţional pentru

Curs	Pentru prezentarea materialului teoretic în sala de curs este nevoie de proiector, PC/laptop şi acces la internet. Nu vor fi tolerate întârzierile studenților, precum și convorbirile telefonice în timpul cursului
Laborator/Seminar	Studenţii vor perfecta rapoarte conform condiţiilor impuse de indicaţiile metodice. Termenul de predare a lucrării de laborator - o săptămână după finalizarea acesteia. Pentru predarea cu întârziere a lucrării aceasta se depunctează cu 1pct./săptămână de întârziere.

5. Competenţe specifice acumulate

Competenţe profesionale	C1 Operarea cu concepte și metode ale domeniului Tehnologiei Informației C2 Aspecte organizaționale și informaționale ale sistemelor C3 Modelarea sistemelor informaționale complexe și implementarea lor prin sisteme informatice C5 Managementul produselor și al serviciilor TIC în concordanța cu cerințele pieței
Competenţe transversale	CT1. Aplicarea principiilor, normelor și valorilor eticii profesionale CT2. Identificarea, descrierea și derularea activităților organizate într-o echipă cu dezvoltarea capacităților de comunicare și colaborare, dar și cu asumarea diferitelor roluri (de execuție și conducere) CT3. Demonstrarea spiritului de inițiativă și acțiune pentru actualizarea propriilor cunoștințe profesionale, economice și de cultura organizațională

6. Obiectivele unităţii de curs

Obiectivul general	Utilizarea metodelor de căutare şi extragere a informaţiei în domeniul dobîndirii datelor (data mining)permite reducerea timpului de rezolvare a problemelor informaţionale, mareşte precizia dobîndirii informaţiei nestructurate şi oferă soluţii practic admisibile acolo unde soluţii exacte sînt imposibil de abordat. Metode şi tehnici de text mining trebuie să de vină unele din instrumentele curente cercetări
Obiectivele specifice	La terminatea cursului „Sisteme de cautare a informaţiei" studentul trebuie să cunoască: * Tehnologiile construirii şabloanelor(modelelor „paterns") pentru formarea diferitor structuri de text. * Metode statistice şi deterministice pentru sumarizarea textului. * Aplicarea metodelor de căutare a informaţiei (Information Retrieval) pentru rezolvarea problemelor „Text Mining" * Tehnologii multivariante (word corelation, word covariance şi component analysis) pentru procesearea textului. * Aplicarea metodelor de clasterizare a textului pentru dobindirea datelor nestructurate. * Utilizarea limbajului PERL pentru rezolvarea problemelor „Text Mining"

7. Conţinutul unităţii de curs

Tematica cursului	**Numărul de ore învăţământ cu frecvenţă**
T.1. Text paterns. Introduction. Regular Expression. Finding words in a text. Decomposing Poe's „The Tell- Tale Heart" into words.	1
T.2.Text Paterns. Simple concordance First Attempt at Extracting Sentences. Regular expresion Odds and Ends	1
T.3.Quantitative Text Summaries . Introduction Scalars, Interpolaction and conect in Perl. Word Lengths in Poe's „The Tell- Tale Heart". Array and Functions.	1
T.4. Quantitative Text Summaries. Using Hash. Zipf' law for „A Chrismas Carol". Word Anagrams. Finding Words in setof letters. Complex Data Structures	1
T.5. Probability and text Sampling. Probability. Conditional probability. Mean and Variance of random, Variable. The Bag-of-Words Model for Poe's „The Blaak Cat	1
T.6. Applying Information. Retrieval to text mining. Introduction. Counting letters and Words.Text counts and vectors. The Term-Document Matrix Applied to Poe.	1
T.7. Applying Information. Retrieval to text mining. Matrix Multiplication. Function of Counts. Document Similarity. Invers Document Frequency. Poe Story Angles Revisired	1
T.8. Concordance Lines and Corpus Linguistics. Text Sampling. Statistical survey sampling. Corpus as Baseline. Sorting concordance lines.	1
T.9. Concordance Lines and Corpus Linguistics. Aplication: Word Usage Differentces between London and Sheley. Word morphology of adverbs. Collocations and Concordance Lines. More ways to sort concordance lines.	2
T.10. Multivariance Techniques with text. Basic Statistics.Z-scores apllied to poe. Correlation and Cosines. Correlation and Covariances. 2 by 2 Correlation and Covariance	2
T.11. Multivariate Techniquies with text.Principal Components Analysis(PCA) Finding the principal Components. PCA Applied to the N Poe short stories	2
T.12. Text Clustering. Two variable Example of K-Means. K-Means with language R. Poe Cluster using Eight Prononns.	2
T.13. Text Clustering. Clustering Poe using principal components. Hierarchical clustering of Poe's short Stories. Decision Trees and Overfinding.	2
T.14. A Simple of aditional Topics. Perl Module: module for number words; the stop-moduler the senteces segmentation module; an object oriented Module for tagging. Miscellaneous modules. Conclusions	2
Total ore curs:	**20**

Tematica lucrărilor de laborator/seminarelor	Numărul de ore învăţământ cu frecvenţă
L1. Aplicarea instrucţiunilor de bază a limbajului Perl pentru procesarea propoziţiilor (array, legths, function: split, join, substr, pos ş.a.	4
L2. Construirea mini-maşinii de cautare a informaţiei în baza textelor electronice date	4
L3. Testarea şi obţnerea rezultatelor funcţionării programelor în limbajul Perl pentru concordance sofisticate în baza modelelor specifice pentru procesarea structurilor lingvistice date.	4
L4. Elaborarea algoritmului clasterizării textelor în baza metodei K-Means	4
L5. Oformarea şi prezentarea finală a rapoartelor pentru lucrările de laborator 1-6	4
Total ore practice:	**20**

8. Referinţe bibliografice

Principale	1. Ronen Feldman. The Text Mining Handbook. Cambridge. 2007. 2.Roger Bilisoly . Practical Text Mining With Perl. Canada 2008. 3. Chris Manning, Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA. 2. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008 4. Stuart Russel, Peter Norving , Artificial Intelligence a Modern Approach. New Delhi.2005
Suplimentare	1. Roger Bilisoly. Practical text mining with Perl. Willey. 2008. 295p. 2. Cristopher Dmeaning Prabhacar, Ragharan. Hinrich Schutze. An introduction to Information Retrieval. Cambridge University Press. 2009. 3. Randal L. Schwartz & Tom Christianseu. Learning Perl. O'REILLY. 1997. 267p.

9. Evaluare

Periodică		Curentă	Studiu individual	Proiect/teză	Examen
EP 1	EP 2				
10%	10%	10%	30%	-	40%
Standard minim de performanţă Prezența și activitățile la prelegeri și lucrări practice. Obținerea notei minime de „5” la fiecare dintre lucrări și examen					

10. Criterii de evaluare

Activitate	Componente evaluare	Metodă de evaluare, Criterii de evaluare	Pondere în nota finală a activităţii	Ponderea în evaluarea disciplinei
Evaluare periodică I	Conținut teoretic, teme 1-7	Test pe MOODLE	100%	**10%**
Evaluare periodică II	Conținut teoretic, teme 8-14	Test pe MOODLE	100%	**10%**
Evaluare curentă	Lucrări practice	Discuții în cadrul lecțiilor practice	50%	**10%**
		Dosar completat cu Rapoarte pentru fiecare Studiu de caz în discuție	50%	
Studiul individual	Teme individuale	Prezentare/discurs public	100%	**30%**
Evaluarea finală	Conținut teoretic și practic	Test pe MOODLE. Notare conform baremului	100%	**40%**