FIȘA DISCIPLINEI

1. Date despre program

1.1. Instituția de învățământ superior	Universitatea „Sapientia" din municipiul Cluj-Napoca
1.2. Facultatea/ DSPP	Facultatea de Științe Tehnice și Umaniste din Târgu Mureș
1.3. Domeniul de studii	Calculatoare și tehnologia informației
1.4. Ciclul de studii	Licență
1.5. Programul de studiu	Calculatoare
1.6. Calificarea	Inginer

2. Date despre disciplină

2.0. Departamentul				Matematică-Informatică			
2.1 Denumirea disciplinei				Regăsirea informației			
2.2 Titularul disciplinei / activităților de curs				Conf.dr. Iclănzan Dávid			
2.3 Titularul (ii) activităților de	seminar			-			
	laborator			Conf.dr. Iclănzan Dávid			
	proiect			-			
2.4 Anul de studiu	3	2.5 Semestrul	6	2.6. Tipul de evaluare	EX	2.7 Regimul disciplinei	DO

3. Timpul total estimat (ore pe semestru al activităților didactice)

3.1 Număr de ore pe săptămână	4	Din care: 3.2 curs	2	3.3 seminar /laborator/ proiect	0 /2/ 0
3.4 Total ore din planul de învățământ	56	Din care: 3.5 curs	28	3.6 seminar / laborator / proiect	0 /28 /0

Distribuția fondului de timp:	ore
a) Studiul după manual, suport de curs, bibliografie și notițe	10
b) Documentare suplimentară în bibliotecă, pe platformele electronice de specialitate și pe teren	2
c) Pregătire seminarii/laboratoare, teme, referate, portofolii și eseuri	5
d) Tutoriat	-
e) Examinări	2
f) Alte activități:	-

3.7 Total ore studiu individual	19
3.8 Total ore pe semestru	75
3.9 Numărul de credite	3

4. Precondiții (acolo unde este cazul)

4.1 de curriculum	
4.2 de competențe	

5. Condiții (acolo unde este cazul)

5.1 De desfășurare a cursului	• Sală de curs cu tablă și videoproiector
5.2 De desfășurare a seminarului/laboratorului/proiectului	• Laborator echipat cu calculatoare, cu mediul de programare în Python instalat.

6. Competențele specifice acumulate

Competențe profesionale	• C6. Proiectarea sistemelor inteligente • C6.1 Descrierea componentelor sistemelor inteligente • C6.2 Utilizarea de instrumente specifice domeniului pentru explicarea funcționării sistemelor inteligente • C6.3 Aplicarea principiilor si metodelor de baza pentru specificarea de soluții la probleme tipice utilizând sisteme inteligente • C6.4 Alegerea criteriilor si metodelor de evaluare a calității, performantelor si limitelor sistemelor inteligente • C6.5 Dezvoltarea și implementarea de proiecte profesionale pentru sisteme inteligente
Competențe transversale	

7. Obiectivele disciplinei (reieșind din grila competențelor acumulate)

7.1. Obiectivul general al disciplinei	Obținerea de cunoștințe și abilitați pentru a proiectarea, implementarea și testarea unor sisteme de regăsire de informații in cadrul colecțiilor de date structurate și nestructurate.
7.2. Obiectivele specifice	Cunoașterea conceptelor specifice domeniului stocării informației și regăsirii informației. Abilitatea de a identifica și aplica tehnici specifice domeniului regăsirii informației pentru diferite probleme din viată reală. Dezvoltarea capacității de evaluare a sistemelor de regăsire a informațiilor.

8. Conținuturi

8.1. Curs	Metode de predare	Observații
Introducere în problematica extragerii de cunoștințe din date. Factorii care au favorizat dezvoltarea domeniului. Comunități și domenii implicate în data mining. Tipuri de date explorate în data mining. Procesul KDD (Knowledge Discovery in Databases), prelucrarea datelor de tip text, prelucrarea limbajului natural.	expunere	2 ore
Etape în procesul de extragere a cunoștințelor din date: colectarea datelor, curățarea datelor, extragerea proprietăților, extragerea șabloanelor și descoperirea, vizualizarea datelor, evaluarea rezultatelor, prezentarea concluziilor. Șabloane și șabloane interesante. Teorema lui Bonferroni. P hacking.	expunere	2 ore
Explorarea și cunoașterea datelor. Tipuri de atribute. Descrierea statistică. Tehnici și metode de vizualizare a datelor. Măsurarea similitudinii și disimilarității datelor.	expunere	2 ore
Preprocesarea datelor. Curățarea datelor. Transformarea și integrarea datelor. Reducerea datelor. Discretizare și generarea ierarhiilor de concepte.	expunere	2 ore
Descoperirea regulilor de asociere frecvente (analiza asocierilor Algoritmi pentru extragerea regulilor de asociere unidimensionale, booleene, din baze de date de tranzacții - Apriori, FP-Growth. Algoritmi pentru	expunere	2 ore

extragerea regulilor de asociere multinivel, multidimensionale, cu constrângeri Analiza corelațiilor. Metode de evaluare a regulilor de asociere.		
Clasificare și predicție. Definirea problemei. Clasificare folosind arbori de decizie. Clasificare Bayes. Alte metode de clasificare. Predicție - regresie lineară.	expunere	2 ore
Evaluarea și selecția clasificatorilor. Erori de tipul I și de tipul II. Matricea de confuzie. Metrici de evaluare a clasificatorilor. Metoda holdout. Subeșantionare aleatorie. Validarea încrucișată. Metode de Bootstrap. Compararea clasificatorilor bazată pe analiza cost-beneficii și curbe ROC.	expunere	2 ore
Tehnici de îmbunătățire a acurateței clasificării. Metode ansamblu, ensemble learning. Bagging, Boosting, AdaBoost, Random forests. Îmbunătățirea acurateței clasificării datelor cu reprezentarea claselor dezechilibrată.	expunere	2 ore
Clusterizare (analiza clusterilor). Definirea problemei. Tipuri de date în analiza clusterilor. Clasificarea metodelor de clusterizare. Clase de metode de clusterizare: de partiționare, ierarhice, bazate pe densitate, bazate pe griduri, bazate pe modele. Determinarea numărului de clustere. Evaluarea calității clusterilor, metode extrinseci și intrinseci - coeficientul de siluetă.	expunere	2 ore
Detectarea valorilor aberante (outlier). Definirea problemei. Tipuri de metode de detectare a valorilor aberante. Abordări statistice. Metode bazate: pe măsuri de proximitate, clasificare, clustering. Extragerea valorilor aberante contextuale și colective. Detectarea valorilor aberante în date multi-dimensionale.	expunere	2 ore
Modelul de regăsire binară, procesarea interogărilor booleene. Dicționare și regăsire tolerantă. Modelul probabilistic de regăsire a informațiilor. PRP (Probability Ranking Principle). Modelul binar de independență.	expunere	2 ore
Caracteristici statistice ale textelor și a cuvintelor. Structuri de date pentru IR. Indexarea colecțiilor de documente. Extragerea criteriilor de căutare. Funcții și metode de căutare a colecțiilor indexate. Măsuri de similaritate si ierarhizare	expunere	2 ore
Clasificarea documentelor. Probleme și provocări în clasificarea documentelor text. Selectarea atributelor. Clasificarea documentelor cu metoda Naive Bayes, clasificarea în spațiul vectorial, mașini cu suport vectorial. Metode de învățare automată în scenarii ad-hoc de extragerea informațiilor.	expunere	2 ore
IR și WWW. Webmining. Roboți Web.	expunere	2 ore

Bibliografie obligatorie

1. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Adatbányászat: alapvetés, 2012, Panem Könyvkiadó, Budapest.

2. Jiawei Han, Micheline Kamber, Adatbányászat: koncepciók és technikák, 2004, Panem Könyvkiadó, Budapest.

3. Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

Bibliografie opțională

4. Pieter Adriaans, Dolf Zantinge, Adatbányászat, 963-545-367-1, 2002, Panem Könyvkiadó, Budapest.

5. Max Bramer, Principles of data mining (2nd ed.), 978-1-4471-4883-8, 2013, Springer.

6. Rada Mihalcea, Dragomir Radev, Graph-based natural language processing and information retrieval, 978-0-521-89613-9, 2012, Cambridge University Press.

7. W. Bruce Croft, John Lafferty, Language modeling for information retrieval, 2003, Kluwer Academic Publishers.

8.2. Laborator	Metode de predare	Observații
Instrumente de calcul științific pentru Python. Ecosistemul SciPy (SciPy.org). Anaconda Python Distribution. Resurse ale cursului, acces premium Datacamp, cursuri și exerciții on-line în ritm propriu. Utilizarea Jupyter Notebook și Google Colab.	Prezentare resurse, rezolvare de probleme.	2 ore
Citirea, agregarea, curățarea, transformarea și exportul datelor. Pandas. Analiza exploratorie a datelor. Numpy. Vizualizarea datelor cu Matplotlib.	Programare, rezolvare de probleme.	4 ore
Descoperirea regulilor de asociere frecvente. Algoritmii Apriori și FP-Growth.	Programare, rezolvare de probleme.	2 ore
scikit-learn - metode de clasificare.	Programare, rezolvare de probleme.	4 ore
scikit-learn - metode de clustering.	Programare, rezolvare de probleme.	4 ore
Procesarea textelor, indexarea colecțiilor de documente text.	Programare, rezolvare de probleme.	2 ore
Procesarea textelor, implementarea funcțiilor de căutare.	Programare, rezolvare de probleme.	2 ore
PageRank, analiza relevanței.	Programare, rezolvare de probleme.	2 ore
Webscrapping, clasificarea textelor.	Programare, rezolvare de probleme.	2 ore
Clusterizarea documentelor text	Programare, rezolvare de probleme.	2 ore
Verificare finală	Programare, rezolvare de probleme.	2 ore

Bibliografie

1. Mark Lutz, Programming Python, 978-0-596-15810-1, 2011, O'Reilly.

2. https://nlp.stanford.edu/IR-book/information-retrieval-book.html

3. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, Adatbányászat: alapvetés, 2012, Panem Könyvkiadó, Budapest.

4. Max Bramer, Principles of data mining (2nd ed.), 978-1-4471-4883-8, 2013, Springer.

5. W. Bruce Croft, John Lafferty, Language modeling for information retrieval, 2003, Kluwer Academic Publishers.

9. Coroborarea conținuturilor disciplinei cu așteptările reprezentanților comunității epistemice, asociațiilor profesionale și angajatori reprezentativi din domeniul aferent programului

- Conținutul disciplinei este în concordanță cu cerințele planului de învățământ, cu conținutul disciplinei predate la alte Universități din țară (Universitatea Tehnică „Gheorghe Asachi" din Iași, Universitatea Politehnica din București, Facultatea de Automatică și Calculatoare, master anul II) și de peste hotare (Stanford University, University of Stuttgart).

10. Evaluare

A. Condiții de îndeplinit pentru prezentarea la evaluare:

Efectuarea tuturor lucrărilor de laborator și promovarea testului de evaluare la laborator

B. Criterii, metode și ponderi în evaluare:

Tip activitate		10.1 Criterii de evaluare	10.2 Metode de evaluare	10.3 Pondere din nota finală
10.4 Curs		Cunoașterea principilor de bază și aplicarea conceptelor studiate la curs, rezolvarea de probleme.	examen scris cu subiect de teorie și rezolvare de probleme	40%
10.5	Seminar	-	-	-
	Laborator	Înțelegerea problemei, formularea formală, implementarea algoritmilor studiați.	Verificarea activității de laborator pe parcurs / și la sfârșitul semestrului	10% / 50%
	Proiect	-	-	-

10.6 Standard minim de performanță
1. Definirea problematicii a regăsirii informațiilor. 2. Cunoașterea caracteristicilor statistice ale textelor și a cuvintelor. 3. Cunoașterea tehnicilor de extragere a criteriilor de căutare. 4. Abilitatea de a exemplifica măsuri de similaritate. 5. Cunoașterea algoritmilor de descoperire a regulilor de asociere frecvente, de clasificare și de clustering. 6. Abilitatea de a implementa un robot web simplu 7. Abilitatea de a implementa o aplicație care dezvolta un index invers calitativ

Semnătura titularului disciplinei: Semnătura titularului /rilor de aplicații:

Conf.dr. Iclănzan Dávid Conf.dr. Iclănzan Dávid

Data: Semnătura directorului de departament: Semnătura coordonatorului programului de studii:

16. 09. 2019 Conf. dr. ing. Kátai Zoltán Ș.l.dr.ing. Szabó László Zsolt