

M078	MR-izborni - 2.god	Matematičke metode za klasifikaciju teksta	P+V+S 1+0+1	ECTS 3
------	--------------------	---	----------------	-----------

Cilj predmeta. Dati pregled osnovnih pojmova i standardnih matematičkih metoda koje se koriste u klasifikaciji teksta te time uvesti studente u ovu modernu problematiku. Osigurati da studenti na osnovi novije literature mogu samostalno analizirati slične probleme kao i implementirati odgovarajuće algoritme.

Potrebna predznanja. Preddiplomski studij matematike. Strojno učenje

Sadržaj predmeta.

1. Uvod i motivacija.
2. Metoda naivnog Bayesovog klasifikatora: opis metode, analiza nedostataka, ilustrativni primjeri.
3. Metoda vektorskog prostora: vektorski prikazi dokumenta, funkcije sličnosti, metoda najbližih k-susjeda, metoda Rocchio klasifikacije.
4. Primjene mehanizama potpornih vektora (Support vector machine SVM) na klasifikaciju teksta.
5. Klasifikacija teksta i grupiranje podataka. Mjere za kvalitetu grupiranja. Primjena algoritma k-sredina. Kvazimetričke funkcije. Primjene hijerarhijskih metoda
6. Primjene matricnih dekompozicija na klasifikaciju teksta: singularna dekompozicija, najbolja aproksimacija s matricom nižeg ranga.

Očekivani ishodi učenja.

Očekuje se da nakon položenog kolegija studenti:

- prepoznaju probleme na koje se mogu primijeniti matematičke metode za klasifikaciju teksta;
- razumiju metode i mogućnosti primjena i nedostatke matematičkih metoda za klasifikaciju teksta;
- matematički dokazuju utemeljenost postupaka i formula kojima se služe u zaključivanju;
- koriste matematičku literaturu različitih izvora i primjenjuju barem jedan programski alat u svrhu ilustracije različitih primjera;
- prezentiraju rezultate metoda za klasifikaciju teksta laicima i stručnjacima.

Izvođenje nastave i vrednovanje znanja.

Predavanja i seminari su obavezni. Ispit se sastoji od pismenog i usmenog dijela, a polaže se nakon odslušanih predavanja. Prihvatljivi rezultati postignuti na kolokvijima, koje studenti pišu tijekom semestra, zamjenjuju pismeni dio ispita. Studenti mogu utjecati na ocjenu tako da tijekom semestra pišu domaće zadaće ili izrade seminarski rad.

Može li se predmet izvoditi na engleskom jeziku: Da

Osnovna literatura:

1. C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

Dopunska literatura:

1. J. Kogan, Introduction to Clustering Large and High-Dimensional Data, Cambridge University Press, 2007.
2. H. Zhang, Statistical Clustering Analysis: An Introduction, in S. Butenko, W. A. Chaovalitwongse, and P. M. Pardalos (eds.), Clustering Challenges in Biological Networks, World Scientific, 2009, 101-126

3. M. W. Berry, J.Kogan, Text Mining. Applications and Theory, Wiley, 2010.
4. M. W. Berry, M. Castellanos, Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition, Springer, New York, 2007.
5. I. Dhillon, S. Mallela, R. Kumar, A divisive information-theoretic feature clustering algorithm for text classification, Journal of Machine Learning Research, 3(4):1265–1287, 2003.
6. B. Schölkopf, A. J. Smola, Learning with Kernels. Support Vector Machines, Regularization, Optimization, and Beyond, MIT Press, 2002.
7. G. Gan, C.Ma, J.Wu, Data clustering : theory, algorithms, and applications, SIAM, Philadelphia, 2007.
8. B. S. Everitt, S. Landau, M. Leese, Cluster analysis, Wiley, London, 2001.