

## Tárgymutató

- χ-négyzet statisztika 57, 160
- 10-pontosság 72
- 11 pontos átlagos pontosság 71, 137
- AdaBoost 134
- adaptív szűrés 108, 110
- adat
  - gyengén strukturált 20, 21
  - strukturálatlan 20, 21
  - strukturált 20
  - tanító 109
  - teszt 109
  - validációs 110
- adattárház-migráció 75
- adathordozó 26
- adatszövegség 233
- adattisztítás 75
- Aliweb 180
- álláskereső portál 83
- általánosító képesség 135
- Altavista 213
- alultövezés 43
- alultövezési index 43
- anaforafeloldás 86, 99
- annotálás 107
- anonimizáció 90
- aratórobot 184, 187
- Archie 180
- Ask 199, 214, 220
- AskJeeves 199
- átlagos kapcsolódás 155
- átlagos pontosságok átlaga (MAP) 72
- ATN 222
- B-fa 191
- bagging 134
- Baum–Welsh-eljárás 97
- Bayes-módszer
  - naiv 119, 139
- belső szorzat 112
- bitmap index 197
- biword index 194
- bizottság
  - osztályozóké 133
  - tagok 133
- block merge 196
- boosting 134
- boosting eljárások
  - AdaBoost 134
- BOTW kereső 211
- Callimachus 63
- célkereső 230
- centroid 113, 149
  - sűrűsége 151
- centroid kapcsolódás 155, 156
- címkézés
  - csoportosítás 159
  - differenciális 160
- CLEF 50
- Clementine 238
  - csomópont 238
  - feldolgozási folyamat 238
- Completeplanet 215
- CoNLL-adattárház 92
- CONTAINS 262, 266
- Contains 255
- CONTAINSTABLE 263
- crawler 184
- CTX\_DDL csomag 254

- C5.0 243
- csoportosítás 126, 145, 244  
 alkalmazásai 147  
 címkézése 159  
 definíció 146  
 hierarchikus 146  
 alulról-felfelé 153  
 egyesítő 153  
 felosztó 153  
 felülről-lefelé 153  
 jellegzetességek 145  
*k*-átlag 149, 164  
 kettészelő 152, 164  
 magjai 150  
*k*-medoid 151  
 lágy 146  
 particionáló 146, 148  
 szigorú 146  
 típusai 146
- csoportosító motor 198
- Datastore 252  
 Dawson-szótövező 47  
 demóciós konstans 118  
 dendogramm 153  
 Dewey tizedes osztályozás 63  
 Dexter 177  
 Dictionary 252  
 dimenziócsökkentés 55  
 globális 56  
 lokális 56
- dinamikus weboldal 228
- dokumentum  
 ábrázolása 32  
 elérési helye 27  
 formátuma 28  
 hordozó médiuma 26  
 jellemzői 26  
 karakterkódolása 29  
 mérete 27  
 metaadatai 28  
 mondatokra bontása 38  
 normálása 34  
 prototípusa 112  
 reprezentációja
- bináris 33  
 csoportosításnál 148  
 statisztikai jellemzői 27  
 stílusa 28  
 tokenizálása 39
- dokumentum-prototípus 112
- dokumentumgyakoriség 35  
 inverze 35
- dokumentumgyakoriség alapú szűrés 56
- dokumentumgyűjtemény 26  
 reprezentálása 32
- dokumentumok  
 átlaga 149  
 centroidja 149
- dokumentumok csoportosítása 145
- dokumentumrendszerezés 107
- dokumentumszűrés 108  
 adaptív 108, 110
- dokumentumtábla 252
- dokumentumterkép 37
- döntési fa 258  
 metszése 124  
 szövegosztályozó 122
- döntési szabály  
 szövegosztályozó 122
- dzsókerkarakter 51
- e-mail feldolgozás 84
- egyedi szavak száma 32
- egyensúlyi pont 136
- egyszerű kapcsolódás 155  
 láncolási effektus 157
- együttes hasonlóság 153
- elfogultság 134
- eliminálhatóság 179
- előfeldolgozás 25
- előfordulás 33, 204  
 kiemelt 204
- EM-algoritmus 152
- EntireWeb 214
- entrópiaszűrés 36
- érthetőség 179
- eseménykeret 86, 87
- ETO 63
- Expectation Maximization *lásd* EM-algoritmus 152

- F-mérték 93, 162  
szintenkénti 141  
fájlkeresők 180  
fedés *lásd* felidézés 69  
félautomatikus osztályozás 106, 142  
feldolgozási folyam 238  
felidézés 69, 93, 136, 147, 175, 178  
formula 69  
fuzzy 72  
szintenkénti 141  
félreelemzés 43  
felszíni háló 202, 229  
felszíni jellemzők 94  
feltételes függetlenségi feltevés  
szavak előfordulására 120  
szavak pozícióira 122  
feltételes valószínűségi mező 98  
felügyelet nélküli tanulás 145  
felügyelt tanulás 91, 104  
Filter 252  
finomítható keresés 208  
fokozatos tanulás 115, 140  
fókuszált robot 187  
fontossági forrás 205  
formátum 28  
PDF 30  
forward-backward algoritmus 98  
forward index 204  
főkomponens-analízis 60  
FrameNet 88  
FREETEXT 263  
FREETEXTTABLE 263  
frekvencia 34  
frekvenciainformációk 95  
Frobenius-norma 133  
funkció szó *lásd* stopszó 35  
futásidő 73  
fuzzy illeszkedés 74, 242  
független adatforrás 233
- gépi tanulás 103  
Gigablast 214  
Gini-index 123  
Google 185, 202, 203, 211
- gyakoróság 33, 34
- relatív 34  
gyengítő változó 129  
gyűjteménytámogatottság 35, 95
- HAM *lásd* Hipertext Absztrakt Gép 176  
Hamming-távolság 75  
módosított 75  
harvester 184  
hasonlósági mérték 148  
hatékonyság  
előfeldolgozás 25  
hatékonyság mérése  
csoportosításánál 161  
szövegosztályozás  
egyszerű 136  
hierarchikus 141  
tulajdonnév-felismerés 93
- hiba 71  
hibavezérelt tanulás 115, 116  
hierarchikus csoportosítás 146  
egyesítő 153  
felosztó 153  
inverzió 155, 158  
hierarchikus szövegosztályozás 139  
hiperlink 108, 176  
hipertext 176  
Hipertext Absztrakt Gép 176  
HITEC 139  
hivatkozás alapú indexelés 200  
HMM *lásd* rejtett Markov-modell 97  
horgony 160, 200  
Hotbot 214  
hozzárendelési elv 189  
statikus 190
- HunLex 52  
HunMorph 54  
HunSpell 54  
HunStem 54  
HunToken 38  
HunTools 54
- IBM DB2 Text Extender 265  
idf 35  
időbeliség 89  
illeszkedés  
MySQL 264  
indexelés

- hivatkozás alapú 200
- kifejezés alapú 194, 199
- metaadat alapú 200
- szó alapú 199
- tartalom alapú 200
- Indexing Engine 252, 253
- indexszekvenciális szervezés 191
- információ-visszakeresés 63, 217
- információkinyerés 81
  - nyelvközi 82
- információ lokalizálása 81
- információnyereség 56, 123, 160
- inkrementális tanulás *lásd* fokozatos tanulás 115
- invertált indexstruktúra 192
- inverzió 155, 158
- jegy 88
- jellemzők csoportosítása 58
- jellemzőkinyerés 56, 58
  - csoportosítás alapján 58
  - LSI 59
- jellemzőkiválasztás 55, 160
  - $\chi$ -négyzet statisztika 57
  - dokumentumgyakorosság alapján 56
  - információnyereség 56
  - kölcsönös információ 57
- $k$ -átlag módszer 149, 164
  - kettészélő 152, 164
- $k$ -medoid módszer 151
- $k$ -NN *lásd* legközelebbi szomszédok osztályozó 124
- kanonikus alak 41
- karakterkódolás 29
  - unicode 29
  - UTF-8 29
- karakter  $n$ -gramm 39, 109
- kategória 104
- kategóriaösvény 141
- kategóriaprofil 112
- kategóriarendszer 104
- kategóriavektor 112
- kategorizálás *lásd* osztályozás 102
- keresés
  - finomítható 208
  - kifejezés alapú 207
  - kiterjesztett 208
  - kiterjesztett, statisztikai alapú 208
  - klaszter alapú 208
  - metaadat szerint szűkített 209
  - összetett feltétellel 208
  - szekció szerint szűkített 209
  - szemantikus háló alapú 210
  - szó alapú 207
  - szótő alapú 209
  - taxonómia alapú 207
  - témaorientált 208
  - természetes nyelvi 209
- keresőkifejezés
  - természetes nyelvi 209
- keresőmotor
  - struktúrája 183
- keresőmotorok 180
- keresőmotorok láthatósági szintje 229
- kereszthivatkozás 98
- kereszthivatkozás-feloldás 98
- keresztvalidáció
  - $k$ -szoros 109
- keret 87
- kernel 130
  - kétrétegű perceptron 131
  - polinomiális 131
  - RBf 131
- kifejezés alapú
  - indexelés 194, 199
  - keresés 207
- kifejezéssablon 224
- kifejezőerő 178
- kiterjesztett keresés 208
  - statisztikai alapú 208
- kiválasztási elv 186
- kivonatolás 166
  - csoportosítás alapú módszerek 171
  - definíció 167
  - hatékonyságának mérése 175
  - jellemzők 168
  - klasszikus módszer 169
  - MEAD-módszer 173
  - MMR-módszer 172
  - mondatkiválasztással 168
  - tf-idf alapú módszer 171

- klasszifikáció *lásd* osztályozás 102  
 klaszter alapú keresés 208  
 klaszterezés *lásd* csoportosítás 145  
 klaszterhipotézis 146  
 korpusz 26  
 koszinusztávolság 112  
 kovarianciamátrix 61  
     elemzése 62  
 kölcsönös információ 57, 160  
 kölcsönös információtartalom 208  
 köteget tanulás 115, 140  
 követelmény  
     megjelenítés 183  
     naprakészség 182  
     rangsorolás 183  
     széleskörűség 182  
 KR-kódolás 54  
 Kronecker-szimbólum 75  
 kvadratikus optimalizálás 129  
 KWIC 64
- label bias probléma 97  
 LADDER 222  
 Lagrange-multiplikátor 129  
 láncolás 157  
 Laplace-simítás 120  
 látens szemantikus indexelés (LSI) 59, 174  
 láthatatlan háló 229  
 legközelebbi szomszédok osztályozó 124  
 lekérdezőnyelv 219  
 lemma 41  
 lemmatizálás 41  
 Levenshtein-távolság 76  
 Lexer 252  
 lexikon *lásd* szótár 32  
 lineárisan szeparálható 116  
 lineáris legkisebb négyzetek módszere 132  
 lineáris osztályozó 111  
 linkindex 201  
 Lovins-szótövező 45  
 LSI *lásd* látens szemantikus indexelés 59,  
     174  
 LUNAR 222  
 lusta tanuló 124  
 Lycos 214
- magfüggvény *lásd* kernel 130  
 makro-átlagolás 136  
 Mamma 214  
 Manhattan-távolság 78  
 manuális osztályozás  
     Oracle Text 257  
 MAP *lásd* átlagos pontosságok átlaga 72  
 Markov-modell  
     maximum entrópia 97  
     rejtett 97  
 Masque/SQL 225  
 maximum entrópia Markov-modell 97  
 medoid 151  
 medoid kapcsolódás 155  
 megbízhatóság *lásd* pontosság 69  
 megjósolhatóság 179  
 mélyháló 187, 201, 215, 229  
 MEMEX 176  
 mérték  
     belső 161  
     külső 161, 175  
 metaadat-generálás 107  
 metaadat alapú indexelés 200  
 metaadatok adatbázisa 185  
 Metacrawler 214  
 metakereső 214, 230  
     szerver oldali 230, 231  
     ügyfél oldali 230, 231  
 metrika 74  
     háromszög-egyenlőtlenség 75  
 metszés  
     döntési fáé 124  
 Microsoft Search Service 259  
 Microsoft SQLSERVER *lásd* SQLSERVER  
     258  
 mikro-átlagolás 136  
 minta 74  
 mintaillesztés 74  
     hibatűró 74, 242  
 modell 25  
 modellalkotás 25  
 mohó algoritmus 139  
     gyengített 140  
 mondatokra bontás 38  
 morphdb.hu 52, 53  
 MRR *lásd* reciprok rangok átlaga 72

- MSN 213
- mySQL Fulltext Search 264
- n*-gramm 39, 109
- n*-gramm index 195
- naiv Bayes-feltevés 120
- naiv Bayes-módszer 119, 124
  - binomiális 122
  - hierarchikus osztályozás 139
  - multinomiális 122
  - működési vázlata 120
- Needleman–Wunch-távolság 77
- neurális hálózat 115
- névelem 91
- névelem-összerendelés 99
- névelemosztály 91
- névelemrendszer
  - hierarchikus 92
- NLI 198
- NLIDB alapú mélyhálókereső 235
- normalizált tf-idf 36
- nyelők 205
- nyelvfelismerés *lásd* nyelvmeghatározás 109
- nyelvmeghatározás 28, 109
- nyelvtechnológia 22
- nyílt osztály 93
- optikai karakterfelismerés 27
- Oracle Media Server 201
- Oracle Text 250
  - Contains 255
  - Datastore 252
  - Dictionary 252
  - dokumentumtábla 252
  - Filter 252
  - Indexing Engine 252, 253
  - indextípusok 254
    - Context 254
    - CTXRULE 256
  - komponensei 252
  - Lexer 252
  - manuális osztályozás 257
  - osztályozás 258
  - particionálási módszerek 257
  - Sectioner 252
  - szabályleíró tábla 252
  - szövegkereső funkciói 251
- „oszd meg és uralkodj” stratégia 123
- osztály 104
- osztályozás 102, 244
  - alesetei
    - eredmény szerint 106
    - fókusz szerint 105
    - kategóriák száma szerint 104
  - bináris 105
  - definíció 104
  - dokumentumvezérelt 105, 127
  - egycímkés 104
  - egyszerű 105
  - félautomatikus 106, 142
  - hierarchikus 105, 139, 141
  - kategóriavezérelt 105, 127
  - kiválasztó 106
  - Oracle Text 258
  - rangsoroló 106
  - szabadalmi hivatalokban 107
  - támogató 106, 142
  - többszintű 104
  - többszintű 105
- osztályozó 104
  - bizottság 133
  - döntési fa alapú 122, 258
  - döntési szabály alapú 122
  - HITEC 139
  - k*-NN 124
  - legközelebbi szomszédok 124
  - lineáris 111
  - minta alapú 124
  - naiv Bayes-módszer 119, 139
  - nemlineáris 127
  - neurális hálózat alapú 115
  - Rocchio- 113
  - SVM- 127, 258
  - szavazásos 133
- óvatos szótövező 50
- öregedési algoritmus 188
- összegzéskészítés 166
  - általános 168
  - független 168
  - indikatív 168

- informatív 168
- kérdésvezérelt 168
- témaspecifikus 168
- PageRank 204
- Paice–Husk-szótövező 47
- párhuzamos feldolgozási elv 186, 189
- passzus 172
- PCA *lásd* főkomponens-analízis 60
- perceptron 115
- permuterm index 195
- perplexitás 163
- perzisztens 157
- Petri-hálók 178
- pillanatkép 184
- pontosság 69, 93, 136, 175, 178
  - formula 69
  - fuzzy 72
  - szintenkénti 141
- Porter-szótövező 45
- pozícióindex 194
- Precise 226
- promóciós konstans 118
- P10 (10-pontosság) 72
- Rand-index *lásd* szabatoság 162
- rangsorolási jellemzők 183
- RCV1-korpusz 143
- reciprok rang (RR) 72
- reciprok rangok átlaga (MRR) 72
- redundanciaszűrés 83
- reflexív metrika 75
- reguláris kifejezés 79
- regularizációs faktor 130
- rejtett Markov-modell 97
- reprezentáció 25
  - bináris 33
  - előfordulás alapú 34
  - gyakoriság alapú 34
  - logaritmikus súlyozással 34
- Reuters-gyűjtemény 137
- Rocchio-osztályozó 113
- Sectioner 252
- selejt 71
- Sellers-algoritmus 77
- shrinkage 139
- Skafe 211
- skalárszorzat 112
- skiplista 193
- SMART 64
- Smith–Waterman-távolság 77
- Snawball 45
- SPSS Clementine 238
- SQLSERVER 258
  - CONTAINS 262
  - CONTAINSTABLE 263
  - Filter 260, 261
  - Filter Daemon Manager 260
  - FREETEXT 263
  - FREETEXTTABLE 263
  - Full-text Query Engine Processor 260
  - Gatherer 260
  - indexelés 260
  - Indexer 260
  - Key map 260
  - noise (stop) words 260
  - Query 260
  - Stemmer 260
  - stopszósűrés 262
  - szövegkezelése 258
  - Thesaurus 260
  - Word Breaker 260
- START 224
- statikus weboldal 228
- Statistica 243
  - Text Miner 244
- stemmer 41
- stop szó 35, 40, 148, 244
- stopszósűrés 40
  - mySQL 264
  - SQLSERVER 262
- strukturálatlan adat 20, 21
- strukturált adat 20
- strukturált előrejelzés 96
- strukturált információ 81
- súlybeállítás
  - additív 116
  - multiplikatív 117
- súlyozás
  - bináris 33, 245
  - előfordulás alapú 34, 245

- gyakoriság alapú 34
- logaritmikus 34, 245
- normalizált logaritmikus 35
- TF- 34
- tf-idf 36, 171, 245
  - normalizált 36
- SVD *lásd* szinguláris értékfelbontás 59, 133, 174
- SVM 127, 258
  - kernel 130
  - nemlineáris 130
  - nemszeparábilis 129
  - szeparábilis 128
  - tartalék 127
- Sybase Verity Full Text Search Engine 266
- szabályleíró tábla 252
- szabatosság 71, 136, 162
- számítógépes nyelvészet 22
- szavak egyértelműsítése 108
- szavazásos osztályozás 133
  - többségi döntéssel 133
- Szeged Korpusz 49
- szekvencia alapú modell 96
- szemantika 93
- szemantikai elemzés 222
- szeparálhatóság
  - lineáris 116
- szereplő 86
- szereplők közti reláció 86
- szerkesztési elv 171
- szimbolikus tanuló 122
- szimmetrikus metrika 75
- szinguláris értékfelbontás (SVD) 59, 133, 174
- szó
  - lemmája 41
  - szótöve 41
- szó-dokumentum mátrix 32
- szó alapú indexelés 199
- szó alapú keresés 207
- szóelőfordulás 204
- szótár 32
- szótő 41
- szótő alapú keresés 209
- szótövezés 41, 148, 244
- szótövező
  - Dawson- 47
  - erős 44
  - gyenge 44
  - Lovins- 45
  - óvatos 50
  - Paice–Husk- 47
  - Porter- 45
  - Tordai-féle 49
- szózsákmodell 33, 122, 200
- szövegbányászat
  - általános modellje 22
  - definíció 22
  - feladata 20
- szövegfeldolgozás
  - biológiai 92
  - orvosi 92
- szövegosztályozás 102
- sztring hasonlósági metrika 74
  - Hamming-távolság 75
    - módosított 75
  - Levenshtein-távolság 76
  - Manhattan-távolság 78
  - Needleman–Wunch-távolság 77
  - Smith–Waterman-távolság 77
- szupportvektor 127
- szupportvektor-gépek *lásd* SVM 127
- támogatottság 33
- tanítóadat
  - negatív 56, 109
  - pozitív 56, 109
- tanítóhalmaz 109
- tanítókönyvet 103, 109
- tanulás
  - felügyelet nélküli 145
  - felügyelt 91, 104
  - fokozatos 115, 140
  - hibavezérelt 115, 116
  - köteget 115, 140
- tanulási ráta 116
- tartalék 127
- tartalom alapú indexelés 200
- tartalomjegyzés 178
- taxonómia 102, 139
- taxonómia alapú keresés 207



- teljes kapcsolódás 155, 156
- teljesség *lásd* felidézés 69
- témaorientált keresés 208
- természetes nyelvek megértése 219
- természetes nyelvű adatbázis-interfész 218
- tesztadat 109
- teszthalmaz 109
- tesztkorpuszok
  - csoportosításhoz 163
- Text Extender 265
- CONTAINS 266
- tf-idf 36, 171, 245
  - normalizált 36
- TF-súlyozás 34
- tiltott szó *lásd* stopszó 35
- típus 39
- token 39
- tokenizálás 39, 74, 253
- Tordai-féle szótövező 49
- többségi döntés 133
- töltelékszó *lásd* stopszó 35
- tövezés *lásd* szótövezés 41
- Tradewave Galaxy 181
- Trellis 178
- tulajdonnév-felismerés 90
- tulajdonnévkorpusz 92
- tulajdonnévszótár 95
- túltanulás 55, 110, 124
- túltövezés 43
- túltövezési index 43
  
- udvariassági elv 186, 189
- ugrás 194
- ugró pointer 193
- újralátogatás 182
  
- újralátogatási arány 187
- újralátogatási elv 186, 187
  - arányos 188
  - uniform 188
- újralátogatási gyakoriság 188
- unicode 29
- UTF-8 29
  
- válaszkereső rendszerek 217
- validációs halmaz 110
- variancia 134
- vektortér
  - dimenziója 32
- vektortérmodell 32, 156, 190
  - dimenziócsökkentés 55
- véletlen szörföző modell 205
- visszaulató névszói csoport 99
- Viterbi-algoritmus 97
- Vivísimo 147, 198
  
- Wandex 181
- Webcrawler 198
- WebFountain 210
- webkorpusz 95
- webrobot 184
- WINNOWER 117, 143
  - kiegyensúlyozott 118
  - pozitív 117
- WiseNut 214
- Word Wide Web Worm 200
- Wow 211
  
- XML 179
  
- Yahoo! 213