

Tartalomjegyzék

Jelölésjegyzék	9
Előszó	14
1. Bevezetés	20
1.1. A szövegbányászat feladata	20
1.2. A szövegbányászat alkalmazási területei	23
2. Előfeldolgozás, modellalkotás, reprezentáció	25
2.1. Az előfeldolgozásnál vizsgált dokumentumjellemzők	26
2.1.1. Alapvető jellemzők	26
2.1.2. A dokumentum formátuma és karakterkódolása	28
2.2. Dokumentum reprezentálása vektortérmodellben	30
2.2.1. Dokumentumreprezentációs modellek	30
2.2.2. A vektortérmodell	32
2.2.3. Súlyozási sémák	33
2.2.4. A szöveg felbontása és a szótár felépítése	37
2.2.5. Lemmatizálás és szótövezés	41
2.2.6. Morphdb.hu alapú magyar nyelvi erőforrások	52
2.3. A vektortérmodell dimenziójának csökkentése	55
2.3.1. Jellemzőkiválasztó módszerek	56
2.3.2. Jellemzőkinyerő módszerek	58
3. Az információ-visszakeresés alapjai	63
3.1. Az információ-visszakeresés modellje	63
3.2. Az információvisszakereső-rendszerek értékelési módszerei	67
3.2.1. Az egyes komponensek szerepe	67
3.2.2. A relevancia mérése	69
3.2.3. Egyéb hatékonysági mértékek	73
3.3. Mintaillesztés	74
3.3.1. Hibatűrő mintaillesztés sztringmetrikákkal	74
3.3.2. Mintaillesztés reguláris kifejezésekkel	79
4. Információkinyerés	81
4.1. Bevezető	81
4.1.1. Példák alkalmazott IE-re	82

4.1.2.	Az információkinyerés és -visszakeresés összehasonlítása	84
4.2.	Az információkinyerés tipikus részfeladatai	85
4.3.	Szabály alapú és statisztikai megközelítések az IE-ben	87
4.4.	IE során felmerülő nyelvészeti problémák	89
4.5.	Tulajdonnév-felismerés	90
4.5.1.	Névelem	91
4.5.2.	A tulajdonnév-felismerés problémaköre	92
4.5.3.	A tulajdonnév-felismerésben hasznosítható jellemzők	94
4.5.4.	Szekvencia és token alapú modellek	96
4.5.5.	Ingyenes tulajdonnév-felismerő rendszerek	98
4.6.	Kereszthivatkozások feloldása	98
5.	Osztályozás	102
5.1.	Az osztályozás definíciója és alosztályozásai	104
5.1.1.	Az osztályozás fajtái kategóriák száma szerint	104
5.1.2.	Dokumentum- és kategóriavezérelt osztályozás	105
5.1.3.	Az eredmény típusa: kiválasztó és rangsoroló osztályozás	106
5.2.	Az osztályozás alkalmazásai	107
5.3.	A tanítókörnyezet és dokumentummodell	109
5.3.1.	A dokumentumgyűjtemény particionálása	109
5.3.2.	Dokumentummodell	110
5.4.	Osztályozó algoritmusok	111
5.4.1.	Rocchio-osztályozó	112
5.4.2.	Neurális hálózat alapú módszerek	115
5.4.3.	Valószínűség alapú osztályozás: a naiv Bayes-módszer	119
5.4.4.	Döntési fa alapú szövegosztályozók	122
5.4.5.	Legközelebbi szomszédokon alapuló osztályozó (k -NN)	124
5.4.6.	Szupportvektor-gépek (SVM)	127
5.4.7.	Regressziós modellek	132
5.4.8.	Osztályozók kombinációja	133
5.5.	Osztályozók elemzése	134
5.5.1.	Elfogultság és variancia közötti kompromisszum	134
5.5.2.	Hatékonyságmérés	136
5.5.3.	Osztályozók összehasonlítása	137
5.6.	Hierarchikus osztályozás	139
5.6.1.	A taxonómia felhasználása	139
5.6.2.	HITEC osztályozó	139
5.6.3.	Hatékonyságmérés	141
5.6.4.	Hierarchikus osztályozók összehasonlítása	142
6.	Csoportosítás	145
6.1.	A csoportosító módszerek típusai	146
6.2.	A csoportosítás alkalmazásai	147
6.3.	Reprezentáció	148

<i>Tartalomjegyzék</i>	7
6.4. Particionáló módszerek	148
6.4.1. A k -átlag módszer	149
6.4.2. További particionáló módszerek	152
6.5. Hierarchikus csoportosítók	153
6.5.1. Egyesítő és felosztó módszerek, illusztráció	153
6.5.2. Egyesítő módszerek	154
6.6. Csoportok címkézése	159
6.7. A csoportosító módszerek elemzése	161
6.7.1. A hatékonyság mérése	161
6.7.2. Dokumentumgyűjtemények	163
6.7.3. Csoportosító algoritmusok összehasonlítása	164
7. Kivonatolás	166
7.1. Az összegzéskészítő eljárások típusai	166
7.2. A kivonatolásnál használt jellemzők	168
7.3. Kivonatoló módszerek	169
7.3.1. A klasszikus módszer	169
7.3.2. A tf-idf alapú módszer	171
7.3.3. Csoportosítás alapú módszerek	171
7.3.4. Gráfelméleti megközelítések	173
7.3.5. Az LSI használata a kivonatolásban	174
7.4. A kivonatolás hatékonyságának mérése	175
8. Tartalomkeresés webdokumentumokban	176
8.1. Történeti áttekintés	176
8.1.1. Hipertext-dokumentumok kialakulása	176
8.1.2. A keresőmotorok kialakulása	180
8.2. Követelmények a keresőmotorokkal szemben	182
8.3. A keresőmotorok struktúrája	183
8.3.1. Webrobot – webes begyűjtő	185
8.4. A dokumentumok indexelése	190
8.4.1. Adatstruktúrák	190
8.4.2. Az indexelés gyakorlati kérdései	197
8.4.3. Alkalmazott indexelési technikák	199
8.5. A Google áttekintése	202
8.5.1. A Google indexelési mechanizmusa	203
8.5.2. PageRank-módszer	204
8.6. A keresési technikák áttekintése	207
8.7. A piaci keresőrendszerek működésének áttekintése	210
8.7.1. Taxonómia alapú keresők	211
8.7.2. Általános keresők	211
8.7.3. Metakeresők	214
8.7.4. Mélyhálókeresők	215
8.7.5. Keresőmotorok funkcióinak összefoglalása	215

9.	Válaszkereső rendszerek	217
9.1.	Természetes nyelvű adatbázis-interfészek	218
9.1.1.	Egy rövid történeti áttekintés	221
9.2.	Keresés a mélyhálóban	228
9.2.1.	Keresés metakeresővel	230
9.2.2.	Kooperációs megoldások	233
9.2.3.	A mélyháló és a válaszkereső rendszerek	234
10.	Szövegbányász-szoftverek bemutatása	237
10.1.	SPSS Clementine	238
10.1.1.	Kezelői felület, működés	238
10.1.2.	Szöveges állományok kezelése	240
10.1.3.	A korpusz szavainak feltérképezése	240
10.1.4.	Szavak szűrése, a szó-dokumentum mátrix létrehozása	242
10.1.5.	Analízis	243
10.2.	Statistica Text Miner	243
10.2.1.	A Text Miner modul áttekintése	244
10.2.2.	A Text Miner modul kezelőfelülete	245
10.3.	Oracle Text	250
10.3.1.	Tipikus alkalmazások	250
10.3.2.	A funkciók áttekintése	251
10.3.3.	Feldolgozási lépések	252
10.3.4.	Az Oracle Text CONTEXT indexelési eljárása	254
10.3.5.	További indextípusok	256
10.3.6.	Megjelenítési lehetőségek	256
10.3.7.	A dokumentumok particionálása	257
10.4.	Microsoft SqlServer szövegkezelő modulja	258
10.4.1.	Áttekintés	258
10.4.2.	Feldolgozási lépések	260
10.4.3.	Indexelés	260
10.4.4.	Kezelőfelület	262
10.5.	Egyéb adatbáziskezelő-rendszerek szövegbányászati elemei	264
10.5.1.	mySQL Fulltext Search	264
10.5.2.	DB2 Text Extender	265
10.5.3.	Sybase Verity Full Text Search Engine	266
	Irodalomjegyzék	269
	Tárgymutató	286