Szövegbányászat

Szövegbányászat

Tikk Domonkos (szerkesztő)
tikk@tmit.bme.hu

TYPOTEX Kiadó
Budapest, 2007

A könyv az Oktatási és Kulturális Minisztérium támogatásával, a Felsőoktatási Tankönyv- és Szakkönyv-támogatási Pályázat keretében jelent meg.

Farkas Richárd: 4.5–6. szakaszok
Kardkovács Zsolt Tivadar: 9. fejezet
Kovács László: 2.3.2.3 alpont, 8. és 10. fejezet (kivéve 10.1. szakaszt)
Répási Tibor: 8. fejezet
Szarvas György: 4.1–4. szakaszok
Szaszkó Sándor: 10.1. szakasz
Tikk Domonkos: 1., 2. (kivéve 2.3.2.3.), 3.2.2–3., 5–7. fejezetek
Vázsonyi Miklós: 3. fejezet (kivéve 3.2.2–3.)

ISBN 978 963 9664 45 6

Kedves Olvasó!

Önre gondoltunk, amikor a könyv előkészítésén munkálkodtunk. Kapcsolatunkat szorosabbra fűzhetjük, ha belép a Typoklubba, ahonnan értesülhet új kiadványainkról, akcióinkról, programjainkról, és amelyet a www.typotex.hu címen érhet el. Honlapunkon megtalálhatja az egyes könyvekhez tartozó hibajegyzéket is, mert sajnos hibák olykor előfordulnak.

Kiadja a Typotex Elektronikus Kiadó Kft., az 1795-ben alapított
Könyvkiadók és Könyvterjesztők Egyesületének tagja
http://www.typotex.hu
Felelős kiadó: Votisky Zsuzsa
Felelős szerkesző: Gerner József
A borítót tervezte: Tóth Norbert
Terjedelem: 20,56 (B/5) ív
Nyomtatta és kötötte: Séd Nyomda Kft., Szekszárd
Felelős vezető: Katona Szilvia

Tartalomjegyzék

Jelölésjegyzék

Az alábbi táblázat tartalmazza a könyvben használt fontosabb jelöléseket. Amennyiben ettől eltérünk, azt külön jelezzük.

$\mathbb{R}$	a valós számok teste
$\mathbb{N}$	természetes számok halmaza
$\mathbf{A} \in \mathbb{R}^{N \times M}$, a_{ij}	$N \times M$ méretű valós mátrix, ill. i-edik sorának j-edik eleme
$\mathbf{v} = \langle v_1, \ldots, v_n \rangle \in \mathbb{R}^n$	n elemű (valós) vektor
$\mathbf{v} \in \mathbb{R}^{1 \times n}$, $\mathbf{w} \in \mathbb{R}^{n \times 1}$	sorvektor, illetve oszlopvektor (ha hangsúlyozni akarjuk a vektor alakját)
$\langle \mathbf{u}, \mathbf{v} \rangle$	$\mathbf{u}$ és $\mathbf{v}$ vektorok skalárszorzata
$\lvert A \rvert$	A halmaz elemszáma
c; $c_j \in C$	kategória; a kategóriarendszer egy eleme
$\mathbf{c}$	a c kategória kategóriaprofilját megadó vektor
$C = \{c_1, \ldots, c_{\lvert C \rvert}\}$	kategóriák halmaza
cf_k	a t_k szó gyűjteménytámogatottsága
d; $d_i \in D$	dokumentum; a dokumentumgyűjtemény egy eleme
$\mathbf{d}$	a d dokumentum vektorreprezentációja
$D = \{d_1, \ldots, d_N\}$	dokumentumgyűjtemény (korpusz) és elemei
$d(\cdot, \cdot)$	távolságfüggvény
df_k	a t_k szó dokumentumgyakorisága a korpuszban
l_d, l_t	tanító-, ill. tesztdokumentumok átlagos vektormérete (ritka vektorként)
L_d, L_t	tanító-, ill. tesztdokumentumok átlagos hossza (szavak száma)
M, N	egyedi szavak, ill. dokumentumok száma
n_k	a korpusz t_k szót tartalmazó dokumentumainak száma
n_{ki}	a t_k szó előfordulásainak száma d_i dokumentumban
Neg_j	a c_j kategóriába nem tartozó tanítóadatok
Pos_j	a c_j kategóriába tartozó tanítóadatok
$s(\cdot, \cdot)$	hasonlóságfüggvény
$S_k(\mathbf{d}_i)$	$\mathbf{d}_i$-hez legközelebbi k szomszéd halmaza
t, t_k	szó (terminus); a vektortér k-adik dimenziójához

Fontosabb szakkifejezések rövidítésekkel magyarul és angolul

magyar	angol	rövidítés
adaptív szűrés	adaptive filtering	
alulról-felfelé	bottom-up	
alultövezési index	under-stemming index	UI
anaforafeloldás	anaphora resolution	AR
aratórobot	harvester	
átlagos kapcsolódás	group-average link	
balelemző	top-down parser	
csomópont	node	
dokumentumszűrés	text filtering	
dokumentumvezérelt osztályozás	document-pivoted categorization	DPC
döntési fa alapú osztályozó	decision tree classifier	DT-classifier
döntési szabály alapú osztályozó	decision rule classifier	DR-classifier
dzsókerkarakter	wildcard	
egycímkés osztályozás	single-label classification	
egyszerű kapcsolódás	single-link	
eltolás	bias	
erőforrás-leíró keretrendszer	resource description framework	RDF
feldolgozási folyam	stream	
feltételes valószínűségi mező	conditional random fields	CRF
felügyelet nélküli tanulás	unsupervised learning	
felügyelt tanulás	supervised learning	
felülről-lefelé	top-down	
fokozatos tanulás	incremental learning	
fontossági forrás	source of rank	
főkomponens-analízis	principal components analysis	PCA
frázissablon	phrasal template	
gyűjteménytámogatottság	collection frequency	CF
hibavezérelt tanulás	mistake driven learning	
hierarchikus (szöveg)osztályozás	hierarchical text categorization	HTC
információkinyerés	information extraction	IE
információnyereség	information gain	IG
információ-visszakeresés	information retrieval	IR
jellemzőkinyerés	term extraction	
jellemzőkiválasztás	term selection	
jobbelemző	bottom-up parser	
k-átlag	*k*-means	
kategóriavezérelt osztályozás	category-pivoted categorization	CPC
kereszthivatkozás	co-reference	
kereszthivatkozás-feloldás	co-reference resolution	

magyar	angol	rövidítés
keresztvalidáció, *k*-szoros	cross-validation, *k*-fold	
keret	frame	
kéretlen levelek szűrése	spam filtering	
kifejezéssablon	phrasal template	
kiterjesztett vagy bővített átmenet-háló	augmented transition network	ATN
kiválasztási elv	selection policy	
kötegelt tanulás	batch learning	
látens szemantikus indexelés	latent semantic indexing	LSI
legközelebbi szomszéd osztályozó (*k*-NN osztályozó)	nearest neighbor classifier	*k*-NN
lineáris legkisebb négyzetek módszere	linear least-squares fit	LLSF
lusta tanuló	lazy learner	
maximum entrópia Markov-modell	maximum entropy Markov modell	MEMM
meredekségi faktor	slope factor	
metszés (döntési fáé)	pruning	
minta alapú osztályozó	example-based classifier	
mintaillesztés	pattern matching	
névelem-felismerés	named entity recognition	NER
nyelő	rank sink	
nyelvközi információkinyerés	cross-language information extraction	CLIE
osztályozó bizottság	classifier committee, ensemble classifier	
öregedési algoritmus	aging algorithm	
összegzéskészítő eljárás	text summarization method	
párhuzamos feldolgozási elv	parallelization policy	
pillanatkép	snapshot	
radiális bázisfüggvény	radial basis function	RBF
rangsoroló eljárás	ranking algorithm	
rejtett Markov-modell	hidden Markov model	HMM
relevancia-visszacsatolás	relevance feedback	
szekvencia alapú modell	structured prediction	SP
szinguláris értékfelbontás	singular value decomposition	SVD
szó-dokumentum mátrix	term-document matrix	TD matrix
szógyakoriság alapú súlyozás (TF-súlyozás)	term frequency	TF
szótövező	stemmer	
szózsákmodell	bag of words model	

magyar	angol	rövidítés
szövegosztályozás	text categorization/classification	TC
szupportvektor gép	support vektor machine	SVM
támogató osztályozás	categorization assistance	
tanítóhalmaz	training set	
teljes kapcsolódás	complete-link	
természetes nyelvek megértése	natural language understanding	NLU
természetes nyelvű adatbázis-interfész	natural language interfaces to databases	NLIDB
természetes nyelvű mélyhálókereső-interfész	natural language interface to deep web searcher	NLIDW
terminusfrekvencia és inverz dokumentumfrekvencia	term frequency & inverse document frequency	tf-idf
teszthalmaz	test set	
tisztaság	purity	
többcímkés osztályozás	multi-label classification	
többértelmű szavak egyértelműsítése	word sense disambiguation	
többségi döntés	majority voting	
többszintes osztályozás	multi-level classification	
válaszkereső rendszerek	question answering systems	QAS
udvariassági elv	politeness policy	
ugró pointer	skip pointer	
újralátogatási elv	re-visit policy	
újraparametrizálás	re-parametrization	
űrlap/nyomtatvány	form	

Egyéb alkalmazott angol rövidítések, és az esetlegesen kapcsolódó honlapcímek

rövidítés	jelentés	URL
ACE	Automatic Content Extraction	`www.itl.nist.gov/iad/894.01/tests/ace/`
ANSI	American National Standards Institute	`www.ansi.org`
CART	Classification and Regression Trees	`www.salfordsystems.com/cart.php`
CoNLL	Conference on Computational Natural Language Learning	`ifarm.nl/signll/conll/`
ETO	Egyetemes Tizedes Osztályozás	
HITEC	HIerarchical TExt Categorizer	`categorizer.tmit.bme.hu`
ID3	Interactive Dichotomizer 3	
IPC	Internatial Patent Classification (Nemzetközi Szabadalmi Osztályozás)	`www.wipo.int/classifications/ipc/en/`
ISO	International Organization for Standardization	`www.iso.org`
KWIC	Key Word in Context	
MUC	Message Understanding Conferences	`www.itl.nist.gov/iaui/894.02/related_projects/muc/`
OPAC	Open Public Access Catalog	
SMART	Salton's Magical Automatic Retriever of Text	
SQL	Structured Query Language	`www.ncb.ernet.in/education/modules/dbms/sql99index.html`
TREC	Text REtrieval Conference	`trec.nist.gov`
WIPO	World Intellectual Property Organization (Nemzetközi Szellemi Tulajdonok Szervezet)	`www.wipo.int`

Előszó

A szövegbányászat a számítástudomány szöveges elektronikus dokumentumok feldolgozásával és elemzésével foglalkozó szakterülete. Az internet korának egyik jelentős trendje az elektronikus adatok rohamosan növekvő mennyisége, melyek nagy része szöveges. Ez a jelenség a mindennapjainkban is jelentkezik az üzleti- és magánszféra, valamint a tudományos, gazdasági és mérnöki élet számos területén: az írásos kommunikáció, az adminisztráció, a dokumentálás folyamatainak jelentős részében elektronikus szövegeket gyártunk. A nagy mennyiségű szöveges adathalmazok hatékony kezelésében kínál segítséget a szövegbányászat. Módszereivel nemcsak az adatok közti eligazodás és keresés válik lehetővé, hanem támogatást is nyújt a dokumentumokban lévő rejtett összefüggések feltárására és kinyerésére.

Könyvünk az első olyan magyar nyelven megjelenő kötet, amely a szövegbányászat feladataira és módszerekre fókuszál. A szövegbányászat alkalmazásorientált szakterület, ezért fontosnak tartjuk, hogy az eljárások elméleti alapjainak széleskörű és alapos ismertetése mellett gyakorlati feladatok megoldásában is segítséget nyújtsunk az Olvasónak. Ez a törekvésünk megmutatkozik egyrészt abban, hogy az anyag tárgyalása során az algoritmusok gyakorlati megvalósításaival kapcsolatos tényezőknek külön figyelmet szentelünk, másrészt pedig hogy külön fejezetben tárgyaljuk néhány jelentősebb, szövegbányászati módszereket tartalmazó szoftvercsomag vonatkozó részét.

A könyvet egyaránt haszonnal forgathatják tehát a szövegbányászati megoldások bevezetését és alkalmazását tervező szakemberek, döntéshozók, informatikusok, valamint az informatikában jártas, a téma algoritmikus és elméleti alapjai iránt érdeklődő Olvasók is. A kötet tankönyvként és oktatási segédletként is szolgál. Anyaga részben a BME Villamosmérnöki és Informatikai Karán a könyv szerkesztője által tartott azonos című választható tárgy tematikájára és oktatási tapasztalataira, valamint a szerzők szövegbányászattal kapcsolatos kutatási és üzleti munkáira épül.

A kötet tartalma

A bevezető fejezet meghatározza a szövegbányászat feladatát, pozícionálja a szakterületet a kapcsolódó témakörökhöz képest, valamint bemutat néhány tipikus alkalmazási példát.

A 2. fejezet a szövegbányászatban alkalmazott alapvető előfeldolgozási módszereket tárgyalja. Megismertetjük az Olvasót a dokumentumok reprezentálására szolgáló numerikus modellekkel, amelyek közül részletesen foglalkozunk a vektortérmodellel. A dokumentumok vektorreprezentációinak létrehozásánál kitérünk a nyelvspecifikus feldolgozás kérdéseire (pl. szótövezés), külön pontban tárgyalva a magyar vonatkozású eredményeket és eszközöket. Jelentős terjedelemben mutatjuk be a vektortérmodell dimenziójának csökkentésére vonatkozó jellemzőkiválasztó és -kinyerő módszereket.

A 3. fejezetben röviden tárgyaljuk az információ-visszakeresésnek a szövegbányászattal szoros kapcsolatban lévő területeit, különös tekintettel az eredmények relevanciájának, ill. a rendszerek hatékonyságának mérésére. Szintén ez a rész foglalkozik a mintaillesztés alapvető technikáival.

A 4. fejezet elsőként néhány tipikus alkalmazási példán keresztül bemutatja az információkinyerés célját és jelentőségét, valamint összeveti tulajdonságait az információ-visszakeresésével. Ezután röviden elemezzük a legfontosabb részfeladatait: a névelem-felismerést, a kereszthivatkozások, szereplők és köztük lévő kapcsolatok azonosítását, illetve az eseménykeretek illesztését. A továbbiakban a szabály alapú és statisztikai megközelítések tulajdonságait, valamint a nyelvspecifikus problémákat vizsgáljuk. A fejezetet a névelem-felismerés, illetve azon belül a tulajdonnév-felismerés problematikájának tárgyalása zárja.

A tematikus osztályozás a dokumentumok rendszerezésének leggyakrabban alkalmazott módszere. Az 5. fejezet elsőként az osztályozási feladat különböző aleseteit veszi számba, majd néhány jellemző példán keresztül bemutatja az alkalmazási területek sokszínűségét. Ezután a felügyelt tanulási paradigma alapjait tárgyalja a fejezet, amit az osztályozó algoritmusok részletes ismertetése, majd elemzése követ. Külön szakaszban foglalkozunk a hierarchikus osztályozás kérdéseivel.

A dokumentumok tematikus rendszerezésének alternatívája a csoportosítás, ennek módszereit a 6. fejezet veszi górcső alá. A fejezet szerkezete hasonló az előzőhöz. Először a csoportosítási problémák és eljárások fajtáit, valamint az alkalmazási példákat tárgyaljuk, amit a felügyelet nélküli tanulási modell ismertetése követ. A particionáló és hierarchikus csoportosítási eljárásokat külön szakaszok-

ban tárgyaljuk, majd kitérünk a csoportok címkézésének kérdésére. Végül összehasonlító elemzés keretében vizsgáljuk az egyes módszerek hatékonyságát.

A 7. fejezet a dokumentumok tartalmi összegzésével, ezen belül főleg a kivonatolással — azaz a szöveg legrelevánsabb mondatainak meghatározásával — foglalkozik. Először megvizsgáljuk, hogy milyen jellemzők alapján tudjuk meghatározni a mondatnak a dokumentum tartalmára vonatkozó relevanciáját, majd néhány fontosabb módszert ismertetünk. A fejezetet a módszerek összehasonlítása zárja.

A 4–7. fejezetekben olyan módszereket ismertetünk, amelyek a szövegekben lévő nemtriviális vagy rejtett információk kinyerésére nyújtanak megoldásokat; ezeket a feladatokat tekintjük a szövegbányászat legalapvetőbb területeinek. A 8–9. fejezetek a dokumentumkeresés feladatával foglalkoznak, amely témakör szorosan kapcsolódik az információ-visszakeresés területéhez. Ennek ellenére úgy gondoltuk, hogy a szöveges dokumentumok kezelésének teljes körű tárgyalása mindenképpen megkívánja, hogy számottevő terjedelemben tárgyaljuk ezt a témát is.

A 8. fejezet az internetes keresőmotorokkal foglalkozik. A történeti áttekintés után a keresőmotorokkal szemben támasztott követelményeket mutatjuk be. Ezt követi a keresőmotorok felépítésének és a dokumentumok indexelését végző technikáknak az áttekintése. Külön fejezetben tárgyaljuk a piacvezető Google keresési technológiájának alapjait és a PageRank módszert, végül összevetjük a piacon található keresőmotorok hatékonyságát és funkcióit.

A 9. fejezet az információkeresésnek egy magasabb szintű módjával, a válaszkereső rendszerekkel foglalkozik. Előbb a természetes nyelvű adatbázis-interfészek megközelítését ismertetjük, majd pedig az internetes adatbázisok tartalmában, az ún. mélyhálóban való keresés problematikájával foglalkozunk.

A könyv zárófejezete néhány szövegbányászati szoftvercsomagot ismertet. Az első két szakaszban statisztikai és adatbányászati elemzőszoftverek szövegbányászati kiegészítéseit elemezzük: az SPSS Clementine szoftver Text Mining for Clementine modulját és a StatSoft Statistica Text Mining komponensét. A következő szakaszokban az adatbázis-kezelő szoftverek szövegbányász funkcióit tekintjük át. Nagyobb terjedelemben foglalkozunk az Oracle Text komponenssel és a MicroSoft SqlServer szövegkezelő moduljával, majd röviden ismertetjük a mySQL, a DB2 és a Sybase adatbázis-kezelők szöveges dokumentumok kezelésére vonatkozó támogatását. A kötetet gazdag irodalomjegyzék és részletes tárgymutató zárja.

Útmutató a könyv olvasásához

A kötet a szövegbányászat területének elméleti és gyakorlati oldalát egyaránt igyekszik bemutatni. Az elméleti részek tárgyalásánál feltételezzük, hogy az Olvasó legalább alapszintű ismeretekkel rendelkezik a lineáris algebra, a valószínűségszámítás, az adatbázis-kezelés, és a bonyolultság-, valamint az információelmélet területein.

A könyv felépítése lehetővé teszi, hogy bizonyos fejezetek önmagukban is érthetőek legyenek azok számára, akik csak néhány témakör iránt érdeklődnek, vagy már rendelkeznek előismeretekkel. Mindenképpen javasoljuk a 2. fejezet áttanulmányozását, hiszen az ebben tárgyalt részekre a későbbiekben gyakran támaszkodunk.[1] Szintén sokszor használjuk a 3.2.2. pontban tárgyalt mértéket. A többi fejezet egymástól függetlenül is érthető, ezekben hivatkozással jelezzük, ha más fejezetben tárgyalt ismeretekre építünk.

Mint az összes informatikai szakterületnek, a szövegbányászatnak is főleg angol nyelvű a szakirodalma. Könyvünkben ezért a fontosabb fogalmaknál az angol megfelelőt is megadjuk, hogy az Olvasót ezzel is segítsük a téma részletesebb tanulmányozásában. A kiemelt terminológiák magyar és angol megfelelői összegyűjtve is megtalálhatóak a jelölésjegyzékben a 10. oldalon. Bizonyos esetekben nem feltétlenül ragaszkodtunk a terminológia magyarításához, különösen ha a magyar kifejezés használata nem terjedt el, vagy nem egyértelmű.[2]

A különböző jellegű kifejezések kiemelését egymástól eltérő szedéssel jelöljük. *Kurzív* betűtípussal szedjük a fontosabb, tárgymutatóban is szereplő fogalmak előfordulásait, valamint olykor ezt használjuk nyomatékosításra is. *Dőlt* betűvel emeljük ki a példák szövegét, illetve a példákban használt szöveges konstansokat. Betűtalp nélküli (sanserif) betűvel szedjük a programkódrészleteket és utasításokat. KISKAPITÁLIS fonttal emeljük ki a kettőnél több karaktert tartalmazó nagybetűs rövidítéseket. Végül az internetes címeket `írógépes` betűtípussal jelöljük, ahol a `http` protokollt alapértelmezésnek tekintettük, és csak az ettől eltérőeket írtuk ki. A szintaktialag helytelen példaszövegeket *-gal jelöljük.

A könyv terjedelmi korlátai miatt számos érdekes és hasznos anyagrész, illetve példa kiszorult a nyomtatott anyagból. Úgy gondoltuk azonban, hogy a téma iránt érdeklődő Olvasók nagy része rendelkezik internet-hozzáféréssel, ezért a könyvhöz készítettünk egy webes mellékletet is, ahol az említett anyagrészeken kívül

[1] Ez alól talán csak a 2.3. kivétel, amelynek anyagára főleg az 5–6. fejezetekben építünk.

[2] Például *karakterfüzér* vagy *-lánc* helyett a *sztring* kifejezést használjuk, a *funkció-/töltelék-/tiltott szó* kifejezések helyett salamoni döntéssel a *stopszó*t alkalmazzuk.

még számos hasznos forrást és linket találhat az érdeklődő. Az alábbiakban ismertetjük a részleteket.

A könyv honlapjáról

A könyv honlapja a

`szovegbanyaszat.typotex.hu`

oldalon található. A honlap az alábbi — a könyvhöz szorosan kapcsolódó — menüpontokat tartalmazza:

- a könyvhöz kapcsolódó példák, anyagrészek és kiegészítések fejezetenként rendezve; a könyv nyomdába adásáig az alábbi anyagok készültek el, illetve vannak előkészületben:

  2. fejezet Mondatokra bontó algoritmus működése (Tikk Domonkos); Porter-, Paice–Husk- és Tordai-féle szótövező részletes leírása példákkal (Tikk Domonkos); MATLAB példa a PCA-algoritmusra (Kovács László)

  4. fejezet Rejtett Markov-modellek és a Viterbi-algoritmus; Maximum entrópia Markov-modell; Feltételes valószínűségi mezők (előkészületben, Farkas Richárd)

  5. fejezet Karakter *n*-gramm alapú nyelvfelismerés (Tikk Domonkos)

  5. fejezet EM-algoritmus részletes leírása (előkészületben, Tikk Domonkos)

  7. fejezet Esettanulmány: böngészés támogatása kivonatolással kézi számítógépeken (Tikk Domonkos)

  10. fejezet Statistica mintapélda dokumentumok osztályozására; Az Oracle Text által nyújtott további keresési lehetőségek és mintapélda; Három példa az SQLSERVER keresési lehetőségeinek illusztrálására (Kovács László)

  Egyéb Tipogenetika; Spektrális szövegbányászat (előkészületben; Vázsonyi Miklós)

  Az elkészült anyagokra a kötet megfelelő pontján utalunk.

- a könyv előszava és tartalomjegyzéke;
- a könyv internetes linkekkel ellátott irodalomjegyzéke, amelynek segítségével a könyvbeli hivatkozások publikusan hozzáférhető része közvetlenül elérhető;
- a könyvben hivatkozott programcsomagok, algoritmusok, dokumentumgyűjtemények, szabványok stb. linkgyűjteménye;
- rövid ismertető a szerzőkről;

- hibajegyzék;
- a könyvről megjelent kritikák, recenziók, visszajelzések.

A honlap céljának tekinti a szövegbányászat népszerűsítését, valamint hogy megjelenési és publikációs fórumot nyisson a szövegbányászat iránt érdeklődőknek, illetve a területen dolgozó hazai szakembereknek, kutatóknak.

A kötet szerzői

A könyv 1–2. (kivéve a 2.3.2.3. alpontot), 5–7. fejezeteit, valamint a 3.2.2–3. pontokat Tikk Domonkos (BME, Távközlési és Médiainformatikai Tanszék; TMIT) írta. A 3. fejezet fennmaradó része Vázsonyi Miklós (BME, Kognitív Tudományi Tanszék) munkája. A 4.1–4. szakaszokat Szarvas György (Szegedi Tudományegyetem, Informatikai Tancsékcsoport; SZTE IT), a 4.5–6. szakaszokat Farkas Richárd (SZTE IT) jegyzi. A 8. és a 10. fejezet (kivéve 10.1. szakaszt), valamint a 2.3.2.3. alpont szerzője Kovács László (Miskolci Egyetem, Általános Informatikai Tanszék; ME ÁIT), a 8. fejezet társszerzője Répási Tibor (ME ÁIT). A 9. fejezet Kardkovács Zsolt Tivadar (BME TMIT) munkája, a 10.1. szakaszt pedig Szaszkó Sándor (BME TMIT) írta.

Köszönetnyilvánítás

A szerzők szeretnének köszönetet mondani mindazoknak, akik segítették a könyv létrejöttét. Külön köszönet jár azoknak, akik részt vettek a könyv kéziratának javításában, és értékes megjegyzéseikkel segítették munkánkat: Bodon Ferenc, Gál Viktor, Halácsy Péter, Körmendy György, Lopata Antal, Pilászy István, Szidarovszky Ferenc P., Takács Gábor. Szintén köszönjük Kiss Ferenc, Pléh Csaba és Infopark Alapítvány szakmai támogatását.

A Clementine és a Text Mining for Clementine adat- és szövegbányászati programcsomagokat az SPSS Hungary bocsátotta rendelkezésünkre, a Statistica szoftvert és Text Mining kiegészítését a StatSoft Hungary Kft-től kaptuk.

Köszönettel tartozunk az *Oktatási és Kulturális Minisztériumnak* a *Felsőoktatási Tankönyv- és Szakkönyvtámogatási Pályázat* keretében nyújtott segítségéért, valamint a TypoTEX Kiadó minden érintett munkatársának a könyv megjelenésében való segítségéért.

Minden igyekezetünk ellenére maradhattak hibák a könyvben. Kérjük, hogy amennyiben hibára bukkan, tájékoztasson bennünket a

`szovegbanyaszat@typotex.hu`

e-mail címen.

1. fejezet

Bevezetés

1.1. A szövegbányászat feladata

Az írástudó emberi civilizációk kialakulása óta a tudást szöveges dokumentumok formájában tárolják. Az ősi egyiptomiak is szöveges dokumentumokat hagytak az utókorra, azonban hieroglifikus írásuk megfejtése korántsem bizonyult könnyű feladatnak. A szöveg megértését végül az segítette elő, hogy a feliratok több nyelven szerepeltek ugyanazon a kövön, amelyek közül az egyik a görög volt, a másik kettő pedig egyiptomi. A görög nyelv kulcsként szolgált a hieroglifák megfejtéséhez. Az ősi egyiptomi hieroglifák megfejtéséből két tanulságot vonhatunk le: (1) a szöveges dokumentumok az emberiség egyik ősi emlékezeti mechanizmusa, ezért fontos biztonságosan és ugyanakkor visszanyerhető módon tárolni az adatokat; (2) a dokumentumokhoz való hozzáférés a tudás feltárásához nem elegendő, ez speciális gyakorlatot és erőforrást igényel.

Napjainkban, amikor a dokumentálási és adminisztrációs folyamatok túlnyomó része elektronikusan valósul meg — és ezáltal rendkívül nagy mennyiségű elektronikus dokumentum keletkezik — megfigyelhető az a trend, hogy az adminisztratív munkát végzők munkaidejük egyre növekvő hányadát fordítják (elektronikus) dokumentumok kezelésére. Míg ez 1997-ben csupán 20%-ot tett ki, addigra 2003-ra már a 30–40%-ot is elérte a Gartner Group becslése szerint. A Meryll Lynch elemzése szerint az üzleti információk 85%-a *strukturálatlan*, illetve *gyengén strukturált adat*, pl. e-mailek, emlékeztetők, üzleti és kutatási beszámolók, prezentációk, hírek, reklámanyagok, weboldalak, ügyfélszolgálati tevékenység jegyzetei stb. formájában áll rendelkezésre [26].

Adatbányászati módszerekkel *strukturált*, gyakran adatbázisokban tárolt adatokból nyerhetők ki összefüggések. Ezek a módszerek többnyire kihasználják, hogy a tárolásra szolgáló adatstruktúra információt ad az adat szemantikájára vonatkozóan is. Például egy személyek adatait tároló adattáblában a *születési év* mezőben szereplő évszámból sokkal könnyebben kinyerhető az életkor, mint amikor ugyanez az évszám a megfelelő kontextusban egy önéletrajzban, szabad szöveg-

ben fordul elő.[1] Ez utóbbi adattípust *strukturálatlan*nak nevezzük. Ezen azt értjük, hogy az adat szemantikájára nem utal a tároló adatstruktúra. *Gyengén strukturált adat*nak tekinthető pl. az XML, ahol bizonyos szemantikus vagy szerkezeti információk rendelkezésre állhatnak. Az adatbányászati módszerek közvetlenül nem alkalmazhatóak a jellemzően strukturálatlan, általános típusú, szöveges adatokra, amelyek tehát más megoldásokat tesznek szükségessé. Az ezzel foglalkozó szakterületet szövegbányászatnak nevezzük. Az 1.1. táblázatban összehasonlítjuk a szöveg- és adatbányászat alapvető ismérveit.

1.1. táblázat. Az adat- és szövegbányászat összehasonlítása ([100] alapján)

	Adatbányászat	Szövegbányászat
az elemzés tárgya	numerikus és kategorikus	szabad formátumú szöveges dokumentum
az adatok jellege	strukturált	strukturálatlan, gyengén strukturált
az adatok tárolási helye	(relációs) adatbázis	tetszőleges dokumentumgyűjtemény
feladat	összefüggések feltárása, jövőbeni szituációk előrejelzése	szövegelemzés, információkinyerés, osztályozás, csoportosítás, összegzéskészítés, vizualizálás, kereséstámogatás stb.
módszerek	neurális hálózatok, döntési fák, statisztikai modellek, klaszteranalízis, idősorok elemzése stb.	dokumentumindexelés, felügyelt és felügyelet nélküli gépi tanulók, számítógépes nyelvészeti eszközök, ontológiák
a világpiac jelenlegi mérete	100 000 elemző közepes és nagyvállalatoknál	100 000 000 vállalati munkatárs és egyéni felhasználó
széleskörű piaci megjelenés	1994-től	2000-től

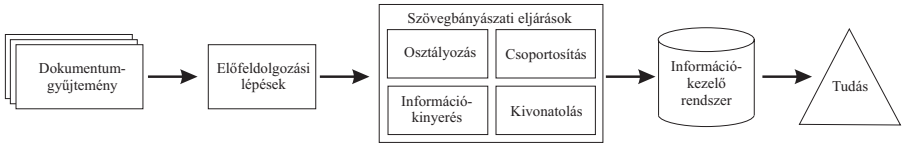
A *szövegbányászat*ot szöveges adatokon végzett feldolgozási és elemzési tevékenységként definiáljuk, melynek célja a dokumentumokban rejtetten meglévő

[1] Ez persze nem meglepő, hiszen az adat tárolási formája a megcélzott felhasználástól függ. A szöveges önéletrajz olvasásra, az adatbázis gépi feldolgozásra készül. A két tárolási mód közötti átjárás jelentős transzformációs költséggel jár.

új információk feltárása, azonosítása és elemzése. Ez a meghatározás analóg az adatbányászat definíciójával.

A szövegbányászat alapvető problémája nyilvánvaló: a természetes nyelvek emberek közötti — elsősorban szóbeli, majd később írásbeli — kommunikáció céljára alakultak ki és fejlődtek, nem a számítógépes feldolgozás szempontjai szerint. Az emberek könnyedén felismerik és alkalmazzák a nyelvi mintákat, és általában nem okoznak gondot nekik olyan, a számítógépek számára nehezen megoldható feladatok, mint pl. a különböző helyesírási variációk kezelése, a kontextus felismerése vagy a stilisztikai jelleg azonosítása. Nyelvi tudásunk lehetővé teszi a strukturálatlan szövegek megértését, ugyanakkor nincs meg bennünk a számítógépeknek az a képessége, hogy a szöveget nagy mennyiségben, vagy nagy sebességgel dolgozzuk fel. A szövegbányászat általános célja tehát az emberi nyelvi tudás ötvözése a számítógép nagy feldolgozási kapacitásával [65].

A szövegbányászat interdiszciplináris alkalmazásorientált szakterület. A szövegbányászati feladatok megoldása során egyaránt szükség van a matematikai, az informatikai, azon belül főként a gépi tanulással kapcsolatos eszköztárak alkalmazására, valamint emellett a természetes nyelvek feldolgozásával foglalkozó területek, a *számítógépes nyelvészet*, a *nyelvtechnológia* eredményeire. Fontos látni, hogy a szövegbányászat és az utóbbi szakterületek céljai különbözőek: míg a nyelvtechnológia alapvetően a nyelvészeti feladatok — pl. morfológiai, szintaktikai, ill. szemantikai elemzés — automatizálását tekinti feladatának, addig a szövegbányászat szövegekkel kapcsolatos informatikai problémák algoritmikus megoldásait keresi, amihez gyakran felhasznál nyelvtechnológiai eszközöket is.



1.1. ábra. A szövegbányászat általános modellje

A szövegbányászat általános modellje az 1.1. ábrán látható. A dokumentumokon először előfeldolgozási lépéseket hajtunk végre, ennek eredményeként előáll a dokumentumhalmaz adott feladatnak megfelelő reprezentációja. A reprezentáció legtöbbször numerikus, esetleg strukturált (pl. XML) szöveges formátumú. Az előfeldolgozás során gyakran alkalmazunk nyelvtechnológiai eszközöket is. Ezután hajtjuk végre a szövegbányászati eljárásokat. Az eredményeket célszerű a hatékony hozzáférés érdekében információkezelő rendszerben tárolni.

1.2. A szövegbányászat alkalmazási területei

Az üzleti élet szereplői és az átlagos felhasználók egyaránt gyakran találkoznak olyan problémákkal, amelyekre a szövegbányászat nyújthat megoldást. Az alábbiakban ízelítőt nyújtunk azon területekből, amelyek az utóbbi években a legjellemzőbb alkalmazói voltak a szövegbányászati megoldásoknak, illetve ahol a közeljövőben várható a szövegbányászat eszköztárának elterjedése. Ezen kívül a könyv több fejezetében ismertetünk az adott problémakörhöz kötődő alkalmazási példákat és lehetőségeket, pl. az információkinyerésnél (82. oldal), a szövegosztályozásnál (107. oldal), a szövegek csoportosításánál (147. oldal), a kivonatolásnál (166. oldal), ill. a szövegbányászati szoftverek ismertetésénél (250. oldal).

Ügyfélszolgálati tevékenység A nagy forgalmat lebonyolító ügyfélszolgálatoknál hatalmas mennyiségű ügyféllel történő beszélgetés zajlik naponta. Ezek jellemző tartalma, fontosabb témái, az ügyfélkör igényeinek változása a szolgáltatónak fontos információt jelent, amellyel hatékonyan reagálhat a piac változásának kihívására.

Biztonság, bűnüldözés A terrorveszély elhárítása érdekében szövegbányászati módszereket is bevetnek a biztonsági szervek. A lefoglalt ill. megszerzett szöveges digitális adatok nagy mennyisége miatt információkinyerő és szövegelemző eljárásokat alkalmaznak az adatok átvizsgálásánál, amelynek segítségével hatékonyan tudnak neveket, helyszíneket, kapcsolatokat, egyéb összefüggéseket azonosítani. A technológiát a bűnüldözés egyéb területein — pl. gazdasági csalások — is hatékonyan alkalmazzák.

Üzleti intelligencia és információszerzés Az üzleti életben rendkívül fontos az információhoz jutás sebessége: a releváns adatoknak minél gyorsabban kell a megfelelő formában a döntéshozók, ill. az elemzők rendelkezésére állnia. Jelentős piaci előnyre tehet szert az, aki idejében tud reagálni egy gazdasági információra vagy kiszivárogtatott hírre — ugyanez fennáll a tőzsdei ügyleteknél is. Mivel az információtermelés sebessége gazdasági és üzleti dokumentumok esetén is folyamatosan növekszik, ezért ezt emberi kapacitással nem, vagy csak igen költségigényesen lehet követni. Helyette automatikus szövegfeldolgozást is tartalmazó hírfigyelő és -elemző rendszereket alkalmaznak. A technológia ugyancsak felhasználható az olyan üzleti hírekről való automatikus értesítésre, amelyek konkurens cégekről, ill. termékekről tartalmaznak információt.

Gyógyszerkutatás A farmakológia és a kapcsolódó orvosi tudományok az egyik tipikus alkalmazási területe a szövegbányászatnak, mivel itt rendkívül nagy

mennyiségű szöveges dokumentum keletkezik a különböző publikációkban, beszámolókban, feljegyzésekben, jelentésekben stb. A szakterület dokumentumainak egy része nagy méretű adatbázisokban található (pl. Medline). Az ezekben való hatékony keresésre szövegbányászati eszközöket alkalmaznak, amelyek képesek pl. bizonyos betegségfajták, tünetek, gyógymódok stb. együttes vagy kombinált előfordulását is kinyerni az adatokból.

Államigazgatás, e-kormányzat Ezen a területen szintén nagy mennyiségű dokumentumot kell hatékonyan kezelni. A szövegbányászati technológiát alkalmazni lehet pl. az írásbeli beadványok megfelelő ügyintézőhöz való irányítására, az egyszerűbb kérdések automatikus megválaszolására. Ennek segítségével a fenntartási költség és az ügyfelek várakozási ideje egyaránt csökkenthető.

Internetes keresés A kulcsszó alapú keresés korlátaival a legtöbb felhasználó szembesült már. Ha többértelmű keresőkifejezést használunk — a tipikus példák: *jaguár* (állat, autómárka), *saturn* (bolygó, elektronikai cég, autótípus), *tus* (zuhany, írószer, vívás, zene)[2] —, akkor a kívánt információ megszerzéséhez a keresés finomítására van szükség. Ennek kiküszöbölésére egyes keresőszolgáltatások lehetővé teszik a keresés kontextusának megadását, amelyet szövegbányászati eljárásokkal valósítanak meg. Alternatív megoldást jelent a keresés megkönnyítésére a tartalmak tematizált tárolása, itt a keresők szövegosztályozási módszereket használnak a téma szerinti besorolásra.

A kereséstámogatás másik tipikus példáját az motiválja, hogy a találatok gyakran nagyméretű, akár több száz oldalas dokumentumok, amelyek több témát is tárgyalnak, és nem feltétlenül relevánsak a kereső számára. Ahhoz, hogy a felhasználó megtalálja a neki fontos információt, el kell mélyednie a szövegben, ami rendkívül időigényes. Erre a problémára a szövegbányászat az összegzéskészítő módszereket kínálja megoldásként, amelyek automatikusan összefoglalják a dokumentum tartalmát, így segítve a gyorsabb keresést, böngészést.

[2] Érdekes, hogy a nemzetközi keresők erre a keresőszóra a nyomtatóval kapcsolatos cikkeket is találnak a *tűs* szó ékezetnélküli reprezentációja miatt. Ez a példa is jól mutatja, hogy a hatékony szövegbányászati alkalmazások — bizonyos fokig — nyelvfüggők.