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.