<http://www.predictiveanalyticstoday.com/top-11-free-software-for-text-analysistext-mining-text-analytics/>

<http://stackoverflow.com/questions/15857384/accuracy-annie-vs-stanford-nlp-vs-opennlp-with-uima>

<http://www.jstatsoft.org/v25/i05/paper>

<http://stats.stackexchange.com/questions/7211/what-are-the-text-mining-packages-for-r-and-are-there-other-open-source-text-min>

<http://stackoverflow.com/questions/8695552/algorithm-or-c-sharp-library-for-identifying-keywords-in-a-set-of-messages>

<http://www.codeproject.com/Articles/9461/DotLucene-Full-Text-Search-for-Your-Intranet-or-We>

<http://128.206.119.157/nlp/papers/NLPTools.pdf>

<http://stackoverflow.com/questions/2303098/java-open-source-text-mining-frameworks>

<http://www.searchenginecaffe.com/2007/03/java-open-source-text-mining-and.html>

<http://butleranalytics.com/5-free-text-mining-tools/>

Szövegbányászat

Bevezető gondolatok a szövegbányászatról

Az emberek, már a kezdeti, ősi civilizációkban is, a szóbeli mellett, jellemzően szövegesen tárolták, és adták át egymásnak az ismereteiket. Napjainkban a rögzített tudásanyagainknak egyszerű szöveges dokumentumokban található. Ezt a feltevést támasztják alá, többek között a Merill Lynch elemzései is, melyek becslése szerint az üzleti információk körülbelül 85%-a található strukturálatlan, illetve gyengén strukturált szövegekben. Az általunk kezelt szövegek növekvő arányban digitálisan tárolt dokumentumok.

Így talán nem meglepő, hogy a szövegek, és főképp a bennük lévő információk, kezelésének hatékonyabbá tétele, napjaink egyre fontosabbá váló informatikai tevékenysége. Az egyik tudományág mely ezzel foglalkozik, a szövegbányászat.

A szövegbányászatot definiálhatjuk úgy, mint szöveges adatokon végzett feldolgozási és elemzési tevékenység, mely célja a dokumentumban rejtett információk feltárása, azonosítása, és elemzése. A szövegbányászat interdiszciplináris szakterület, mely olyan informatikai eszközök mellett, mint a gépi tanulás és a hatékony algoritmusok, a matematika és a nyelvészet eszközeit is felhasználja.

A szövegbányászat egy általános modellje:

kép?

Dokumentum gyűjtemény ->

Szövegbányászati alapfogalmak

Ontológia

Natural Language Process

A szövegbányászat és az adatbányászat összehasonlítása

Szövegbányászati előfeldolgozás

Számos szövegbányászati feladat megoldható, már létező, adatbányászati eszközökkel, és algoritmusokkal. Ehhez a szöveges adatokat úgy kell transzformálni, olyan alakra kell hozni, hogy ezek a bejáratott eljárások, lehetőleg hatékonyan, képesek legyenek működni rajta. Ezt a transzformációs lépést szokták előfeldolgozásnak is nevezni. A folyamat végeredménye a dokumentumot reprezentáló modell. Az előfeldolgozás egységesítési, formalizációs és normalizációs feladatokat is tartalmaz.

A dokumentum alapvető jellemzői

dokumentumot hordozó médium:

Ha digitális, gondoskodni kell az elérhetőségről, ha pedig analóg, akkor előbb digitalizálni kell. ha képi, akkor a képen látható szövegeket, ki kell nyerni, lehetőleg automatikus karakterfelismeréses vagy akár manuális begépeléses módszerrel.

dokumentum elérési helye:

A hagyományos világban lehet egy könyvtár egy polca, míg a digitális világban jellemzően lehet egy URL (Uniform Resource Locator) mely egy egységes erőforrás-azonosító és az interneten keresztüli adatelérésre szolgál, vagy lehet egy háttértárolón egy fájl elérési útvonala. Szövegbányászatnál mindenképpen fontos, hogy rendelkezzünk minimum olvasási jogokkal az adott dokumentumon.

Mérete

A nyers digitális állományok legelterjedtebb általános méretmegadási módja az adattárolón elfoglalt terület alapján, bájt alapon történik. A szövegfájlok mérete megadható még benne szereplő karakterek, szavak vagy az oldalaik számával. A szöveges korpuszok méretét a benne szereplő dokumentumok összesített mérete adja.

Az előfeldolgozási fázis után a méretet jellemzően szavaik, ritkábban karaktereik számával adhatjuk meg. A modellbeli reprezentációjuk méretét pedig olyan a felírásukra használt elemek számával jellemzik, mint a tokenek vagy az egyedi szavak száma, esetleg a modell indextömbjének mérete.

statisztikai jellemzői

A későbbi feldolgozás során szükségünk lehet egyes statisztikai adatokra is a dokumentumunkról. A legfontosabb ilyen jellemző lehet a szavak eloszlása, de esetenként fontos lehet még a karaktereinek eloszlása, a szóhosszoknak az eloszlása vagy átlaga. A magán és mássalhangzók száma és eloszlása, vagy a mélyebb nyelvi elemzéssel kideríthető zöngés és zöngétlen magán- illetve mássalhangzók száma, melyekből például következtetni lehet az adott dokumentum nyelvére.

Metaadatai

A dokumentum adathordozón való tárolásakor keletkeznek metaadatok, mint a keletkezési ideje, az utolsó módosítás ideje, a fájlnév, illetve a hozzáférést módosító attribútumok.

Szövegbányászati szempontból egyéb jellemzők is fontosa, ilyen például a dokumentum tartalmára vagy a keletkezési körülményeire vonatkozó metaadatok, mint a nyelv, a keletkezési idő/hely, a dokumentum témája vagy más szempont szerinti besorolása, a stílusa, a nyelvezete vagy zsánere. Ezek osztályozáskor vagy csoportosításkor lehetnek segítségünkre. Ezek közül az adatok közül nem áll rendelkezésünkre minden, de néhányat, ha hiányzik, szövegbányászati eszközökkel meg tudunk határozni. A nyelv vagy a téma is lehet ilyen.

Formátuma

A dokumentum formátumát többnyire a készítésére használt eszköz típusa bekorlátozza. egyes eszközök egyszerű szövegként (plain text, txt) mentenek, mígy másik eszközök formátumai különböző formázási és szerkezetre vonatkozóadatok, de akár a dokumentum régebbi változatai is lehetnek. Az ilyen tárolási formátumok, a teljesség igénye nélkül, például a doc/docx, a pdf, rtf, de akár a HTML és az XML is ide sorolható.

A dokumentum készítői által kiemelt szövegrészek lehetnek külön formázásúak, mint például a félkövéren szedett, a dőlt, vagy az aláhúzott szövegrészek, és lehetnek szerkezetileg meghatározó helyen, mint a különböző szintű címekben, vagy például a webes elérésű dokumentumok az elérési útvonalában is szerepelhetnek. A dokumentumban kiemelt kifejezések a dokumentum szempontjából relevánsabbak lehetnek, ezért érdemes lehet, ezeket az iformációkat a modell reprezentációba is átvinni.

Karakterkódolása

Az általunk használt adatok különböző karakterkódolásokban fordulhatnak elő, hiszen a számítógép, a szöveges adatokat is kettes számrendszerű számokként tárolja, ezért szükséges annak a meghatározása, hogy egy-egy szám milyen karaktert kódol.

Az egyik első széles körben használt kódtábla az ASCII (*American Standard Code for Information Interchange*), amit az Amerikai Egyesült államokban alkották meg, először hét biten ábrázolt, ami 128 karakter ábrázolását teszi lehetővé, melyek némelyike vezérlő karakter, mint az új sor, vagy a tabulátor, illetve vannak nem látható vezérlő karakterek is melyek a korabeli nyomtatókat vezérelték. Az ASCII 8. bitjét eleinte paritásbitnek szánták, de később általánossá vált, hogy a karakterkészlet kibővtésére használták, az így nyert további 128 lehetséges állapottal az [ISO 8859-1](http://hu.wikipedia.org/wiki/ISO_8859-n) Latin-1-es kódolás a nyugat európai, míg az [ISO 8859-](http://hu.wikipedia.org/wiki/ISO_8859-n)2 inkább a kelet európai nyelvek speciális ékezetes karaktereit írja le. Napjainkban egyre inkább az Unicode veszi át a vezető karakterkódolási szabvány szerepét, mely arra hivatott, hogy legyen képes az ismert nyelvek összes karakterét kódolni. Legelterjedtebb megvalósítása az UTF-8, mely változó hosszon kódolja a karaktereket, ezért tömör kódolást biztosít, mindeközben kompatibilis a 7 bites ASCII-vel. A karakterkódolások különbözősége mellett olyan dolgok is nehezíti a feladatunkat, mint hogy a a sor végének jelölése egyes platformokon eltérhet.

Az ismert kódolások részben, vagy teljesen inkompatibilisek, ezért a konvertálásnál kiemelt kérdés a karakterkódolás, mivel ennek a helytelen használata, a beolvasott szöveg torzulásához vezethet, vagy az akár teljesen használhatatlanná is válhat.

A karakterkódolási információt esetenként maga a fájl tartalmazza, ha nem ismerjük, akkor az előfeldolgozás helyessége nem garantálható.

Modellalkotás

Az ismertebb dokumentum reprezentációs modellek bemutatása

A legelterjedtebb modellek jellemzően valamilyen numerikus objektumok. Ez számos előnnyel jár, mint a kisebb tárolási méret. Ugyanis ha a szavakat karakterenkénet letároljuk, és jellemzően egy-két bájtos egy karakter, akkor nagyobb helyet vennénk igénybe, mint a numerikus tárolásnál, ahol szavanként egy darab 2 vagy akár 4 bájtos számmal számolunk. Mindemellett, a számok használatának van egy másik jelentős előnye, méghozzá az, hogy matematikai műveleteket, és transzformációkat hajthatunk végre az ilyen modelleken. Ráadásul a modellben a matematikai eszközökkel való munka elősegíti a dokumentumok hatékony kezelését.

Szövegbányászat főbb feladattípusai

Kereséstámogatás, információ-visszakeresés

!!! Információkinyerés !!!

Osztályozás

Csoportosítás

Összegzéskészítés, kivonatolás

Szövegelemzés

Webes tartalomkeresés

Válaszkereső rendszerek

A legismertebb ingyenes szövegbányászati csomagok és rövid bemutatásuk.

StanfordNLP

OpenNLP

GATE

<https://gate.ac.uk/>

UIMA

R nyelv

LingPipe

MagyarLánc

Hol lehet fontos a projectemben