Logana Tárgyszó-Kigyűjtés

(Szórabontás+SzokatlanSzavakKigyűjtése+Redukálás+Rendezés)

Nagy István (analog@logana.com)

BEVEZETŐ

A kereső kulcsszavak (vagy legalábbis az azokhoz "eléggé" hasonló szavak) előfordulásainak, vagy egy szöveg szignifikáns szavainak (a tárgyszavainak) automatikus kigyűjtése a könyvtártudomány egyik legizgalmasabb területe. E feladat még ismeretlen nyelvű szöveg esetén is megoldható a Logana Team által kifejlesztett hasonlósági szövegfeldolgozás módszerével!

A szöveghasonlítás módszere egyrészt az összehasonlítandó szövegek azonos karaktereinek menynyisége, és azok sorrendje által meghatározott függvény használatán alapul, másrészt azon, hogy a hasonlóságot nem tekintjük korlátlannak. Bevezetjük a határhasonlóság fogalmát, és csak azokat a szövegeket tekintjük hasonlóknak, melyeknek hasonlósága e határhasonlóság értéknél nagyobb. A gyakorlatban a hasonlóság értékét százalékban adjuk meg, ahol a 100% a teljes azonosságot jelöli.

Az alábbiakban bemutatjuk e módszer felhasználásával készült automatikus tárgyszó-kigyűjtést végző eszközt. A szöveghasonlóság általános bemutatása, és egyéb demó-alkalmazások megtalálhatók a **www.logana.com** címen.

ALKALMAZÁSI MEGJEGYZÉSEK

a.) Az alkalmazás könyvtárában található $ANAL0G kezdetű szövegfájlok a különböző nyelvek általánosan használt szavait tartalmazzák. Ezeket nevezzük Referencia szógyűjteményeknek.

b.) A különböző típusú feldolgozandó mintaszövegek esetén alkalmazott jelölések:
T : Text típusú szöveg (ANSI kódolású).
D, E, Es, Fr, H, I: Német, Angol, Spanyol, Francia, Magyar, Olasz nyelvű szöveg
T, L: Technikai, Irodalmi szöveg
(tehát például TET: egy Text típusú Angol nyelvű Technikai szöveget jelöl)

d.) A könyvtár elemei:

!!!!_LoganaTárgyszóKigyűjtés_HasználatiLeírás_151205.doc
Jelen leírás

!!_0-Szórabont+Redukál+Rendez_THL1.bat
!!_1-DisszocKigyűjt+Redukál+Rendez_THL1.bat
!!_2-DisszocKigyűjt+Redukál+Rendez_THT1.bat
!!_3-DisszocKigyűjt+Redukál+Rendez_TET1.bat
Az alábbiakban részletesen bemutatott demó-programok

BACK, BackFritter, BackReduc, BackSort
Az egyes részfeladatokat megvalósító eszközök munkakönyvtárai

TET1-SimilarityTheory.txt
Angol nyelvű technikai mintaszöveg

THL1-JózsefAttilaÖsszes.txt
Magyar nyelvű irodalmi mintaszöveg

THT1-Kvantumszámítógép.txt
Magyar nyelvű technikai mintaszöveg

e.) Megjegyezzük, hogy a Szórabontott szövegek soronként egy szót tartalmaznak, a Folyamatos szövegek a hagyományos szövegek.

MEGJEGYZÉSEK A DEMÓ-PROGRAMOK HASZNÁLATÁHOZ

a.) A demó-programok egyszerű futtatása után célszerű saját szövegeken kipróbálni a használatukat. Ügyeljünk arra, hogy a vizsgált szövegek valóban ANSI kódolású text-típusú szövegek legyenek. Ez azt jelenti, hogy ha például egy word típusú szöveget szeretnénk vizsgálni, akkor azt konvertálnunk kell ANSI kódolású text-típusú szöveggé. (Ennek egyik módja, hogy a word dokumentum megjelenítése után Ctrl-A – Ctrl-C módon kijelöljük a teljes szöveget, majd egy egyszerű szövegállományba Ctrl-V módon bemásoljuk.)

b.) Különböző jellegű szövegek esetén célszerű a Határhasonlósági értékeket módosítani a számunkra megfelelő eredmény létrehozásának érdekében. Az alábbi demó-programokban vizsgált irodalmi és technikai szövegekhez hasonló jellegű szövegek esetén, az ott alkalmazott Határhasonlósági értékek általában elfogadható eredményt adnak, de azoktól lényegesen különböző szövegek esetén lehet, hogy érdemes más értékekkel is próbálkozni. Ugyanez érvényes az egyéb nyelvű szövegeket vizsgálata esetén.

c.) Az automatikus tárgyszó-kigyűjtés nagyobb szövegfájlok esetén akár több percet is igényelhet.

d.) Megjegyezzük, hogy a fenti demó-programok a hasonlósági szövegvizsgálati módszer szemléltetésére készültek. Speciális feladatok megoldása érdekében célszerű felvenni velünk a kapcsolatot **(analog@logana.com)**.

A Demó-programok bemutatása

1.) !!_0-Szórabont+Redukál+Rendez_THL1.bat

Ez egy THL típusú szöveget (THL1-JózsefAttilaÖsszes.txt) dolgoz fel. A feldolgozás nem használ semmilyen szöveghasonlítási technikát. Egyszerűen csak szavakra bontja a szöveget (egy sor - egy szó), kiszűri az azonos (teljesen egyező!) szavakat (ez a redukció), végül pedig az ábécé szerint rendezi a szógyűjteményt.

Megjegyezzük, hogy e demó-program futtatásának eredményeként a fenti mindhárom részművelet részeredmény szógyűjteménye megmarad, tehát megtekinthető.

2.) !!_1-DisszocKigyűjt+Redukál+Rendez_THL1.bat

Ez is egy THL típusú szöveget (THL1-JózsefAttilaÖsszes.txt) dolgoz fel. A feldolgozás itt már szöveghasonlítási technikát alkalmaz.

2.1.) Először disszociatív módon (szórabontva) kigyűjti a nem általánosan használt szavakat a magyar referencia szógyűjtemény ($ANAL0G-HU41.DIC) alapján 46%-os határhasonlóság szerint. Ez azt jelenti, hogy a vizsgált szöveg szavai közül csak azokat tartja meg, melyek a referencia szógyűjtemény minden szavához képes 46%-nál kevésbé (!) hasonlók.

2.2.) A második lépés a kigyűjtött szavak disszociatív redukciója 60%-os határhasonlóság szerint. Ez azt jelenti, hogy a szógyűjteményből csak azokat a szavakat tartja meg, melyek hasonlósága 60%-nál kisebb. Ezt a vizsgálatot a szógyűjtemény első szavától kezdve végzi el.

2.3.) Végül az ábécé szerint rendezi a szógyűjteményt.

Megjegyezzük, hogy e demó-program is megtartja mindhárom részművelethez tartozó részeredmény szógyűjteményét.

3.) !!_2-DisszocKigyűjt+Redukál+Rendez_THT1.bat

Ez egy THT típusú szöveget (THT1-Kvantumszámítógép.txt) dolgoz fel, szintén szöveghasonlítási technika alkalmazásával.

Működése teljesen megegyezik az előző demó-programmal, különbség a vizsgált szöveg kijelölésén túl csak az egyes műveleteknél megadott határhasonlóság-értékekben van (80%, 60%).

4.) !!_3-DisszocKigyűjt+Redukál+Rendez_TET1.bat

Ez egy TET típusú (tehát angol nyelvű) szöveget (TET1-SimilarityTheory.txt) dolgoz fel, szintén szöveghasonlítási technika alkalmazásával.

Működése szintén megegyezik az előző két demó-programmal, különbség a vizsgált szöveg kijelölésén túl az angol referencia szógyűjtemény ($ANAL0G-E31.DIC) megadásában, és az egyes műveleteknél megadott határhasonlóság-értékekben van (67%, 70%).

Az egyes részfunkciók működése, használata

Az automatikus tárgyszó-kigyűjtés egy összetett feladat, mely az asszociatív, illetve a disszociatív szöveghasonlítás technikájára épül. Ennek részfeladatait az alábbiakban mutatjuk be. Ez elsősorban azok számára lehet érdekes, akiket informatikai felkészültségük révén e részfeladatok megvalósítása, és azok működése, valamint használata is érdekel.

1.) Egyszerű Szórabontás

Használata:

Call $$01-Fritter.bat AAA.txt BBB.txt

ahol

AAA.txt: Feldolgozandó szöveg (Text típusú Folyamatos szöveg)
BBB.txt: Eredmény szöveg (Text típusú Szórabontott szöveg)

Működés:

A folyamatos Feldolgozandó szöveget szavakra bontja olymódon, hogy minden sorba csak egyetlen szó kerüljön. Ezt nevezzük szórabontásnak.

2.) Disszociatív Szórabontás

Használata:

Call $$02-Dissoc.bat AAA.txt BBB.txt CCC.txt DDD

ahol

AAA.txt: Feldolgozandó szöveg (Text típusú Folyamatos szöveg)
BBB.txt: Referencia szógyűjtemény (Text típusú Szórabontott szöveg)
CCC.txt: Eredmény szógyűjtemény (Text típusú Szórabontott szöveg)
DDD: Határhasonlóság (0 és 100 közötti egész szám)

Működés:

A Feldolgozandó szövegből csak azok a szavak kerülnek az Eredmény szógyűjteménybe, melyek a Referencia szógyűjtemény minden szavához Határhasonlóságnál kevésbé hasonlóak. Az eljárás egyúttal a folyamatosnak feltételezett szöveg szórabontását is elvégzi.

3.) Disszociatív Redukció

Használata:

BackReduc\ANATXRED.exe AAA.txt BBB.txt CCC

ahol

AAA.txt: Feldolgozandó szöveg (Text típusú Szórabontott szöveg)
BBB.txt: Eredmény szógyűjtemény (Text típusú Szórabontott szöveg)
CCC: Határhasonlóság (0 és 100 közötti egész szám)

Működés:

A Feldolgozandó szövegből csak azok kerülnek az Eredmény szógyűjteménybe, melyek ez utóbbi szavainak mindegyikéhez a Határhasonlóságnál kevésbé hasonlóak. A vizsgálat a Feldolgozandó szöveg első szavától kezdődik.

4.) Szövegrendezés

Használata:

BackSort\ANATXSORT.exe AAA.txt BBB.txt

ahol

AAA.txt: Rendezendő szógyűjtemény (Text típusú Szórabontott szöveg, Max 150000 szó)

BBB.txt: Rendezett szógyűjtemény (Text típusú Szórabontott szöveg)

Működés:

A Rendezendő szógyűjtemény szavait ábécé (ANSI kódtábla) szerint rendezi.