KOPI – A fordítási plágiumok keresője

Pataki Máté, Kovács László
MTA SZTAKI, 1111 Budapest, Lágymányosi utca 11
{mate.pataki, laszlo.kovacs}@sztaki.hu

A 2011-es Networkshop Konferencián beszámoltunk az MTA SZTAKI Elosztott Rendszerek Osztályán folyó, a fordítások felismerésére irányuló kutatásunkról. Év végére a kutatásból online szolgáltatás lett, amelyet bárki kipróbálhat, használhat a KOPI Plágiumkereső Portálon (http://kopi.sztaki.hu/). Az alábbiakban ezt az új szolgáltatást mutatjuk be részletesen.

Mint azt számos külföldi és hazai eset is bizonyítja, a későn felismert plágiumok komoly kárt okozhatnak nem csak egyes emberek karrierjében, hanem egy teljes oktatási intézmény hírnevében is. Ezért éreztük fontosnak, hogy az egynyelvű plágiumkeresőnket kiegészítsük – a világon elsőként – fordítások felismerésére is képes algoritmussal. Az így kibővített rendszer már különböző nyelven írt dokumentumokat is össze tud hasonlítani – például angol, magyar és német nyelvű dokumentumokat a magyar vagy angol Wikipédiával – és pontosan kijelezni, ha egyezést vagy fordítást talál a feltöltött dolgozat és a másik nyelvű szöveg között. További nyelvekkel és tartalmakkal folyamatosan bővítjük a rendszert: a francia és német nyelvű Wikipédia már feldolgozásra került és év végéig a SZTAKI Szótárban lévő összes nyelvet is igyekszünk beépíteni a Plágiumkeresőbe, ezáltal is segítve az új szolgáltatás minél szélesebb körű használatát.

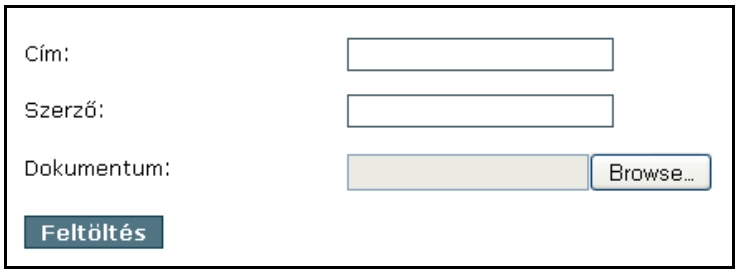
1. Történet

A KOPI Online Plágiumkereső elkészítése 2003-ban kezdődött és a rendszer 2004 óta áll a hazai internetező közönség rendelkezésére. 2006-ban a felhasználók száma meghaladta az ezret, 2009-ben pedig már több mint 10 000 felhasználója volt a KOPI-nak: a plágiumkereső szolgáltatás számos tanár, egyetemi kar mindennapos eszköze lett. Ma már 20 000-hez közelít a felhasználók száma, több mint ötmillió Wikipédia cikkben és körülbelül harmincezer dokumentumban, diplomában, szakdolgozatban keres a rendszer.

A fordítási plágiumok felismerésére irányuló kutatások 2010-ben kezdődtek, és egy évvel később – a világon elsőként – már beépítettük ezt a tudást az online rendszerbe. A http://kopi.sztaki.hu/ címen bárki elérheti, kipróbálhatja az új fordítási plágiumkereső szolgáltatást.

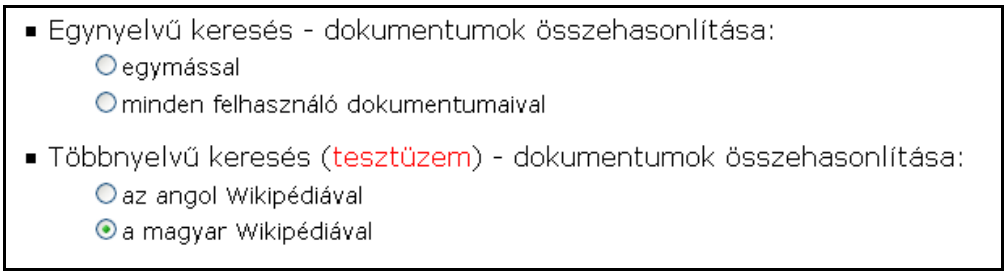
2. Használat

A plágiumkeresés öt fő lépésből áll, melyet a felhasználói felület is tükröz. A felhasználónak először **fel kell töltenie azt a dokumentumot**, amelyet össze szeretne hasonlítani más forrásokkal. A rendszer jelenleg html, doc, docx, rtf, txt és pdf formátumú dokumentumokat kezel. Érdemes kitölteni a dokumentum címét és szerzőjét, hogy pontosan lehessen látni a keresés eredményénél, hogy ugyanaz a dokumentum szerepel kétszer a rendszerben, vagy tényleges egyezésről van szó.

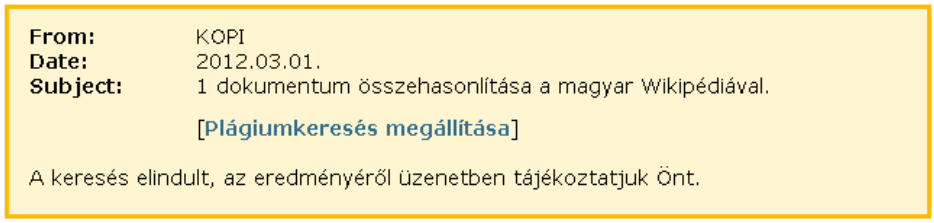


A feltöltés után **ki kell választani** egy **dokumentumot**, amelyet az adatbázissal, vagy több dokumentumot, amelyeket egymással szeretnénk összehasonlítani.

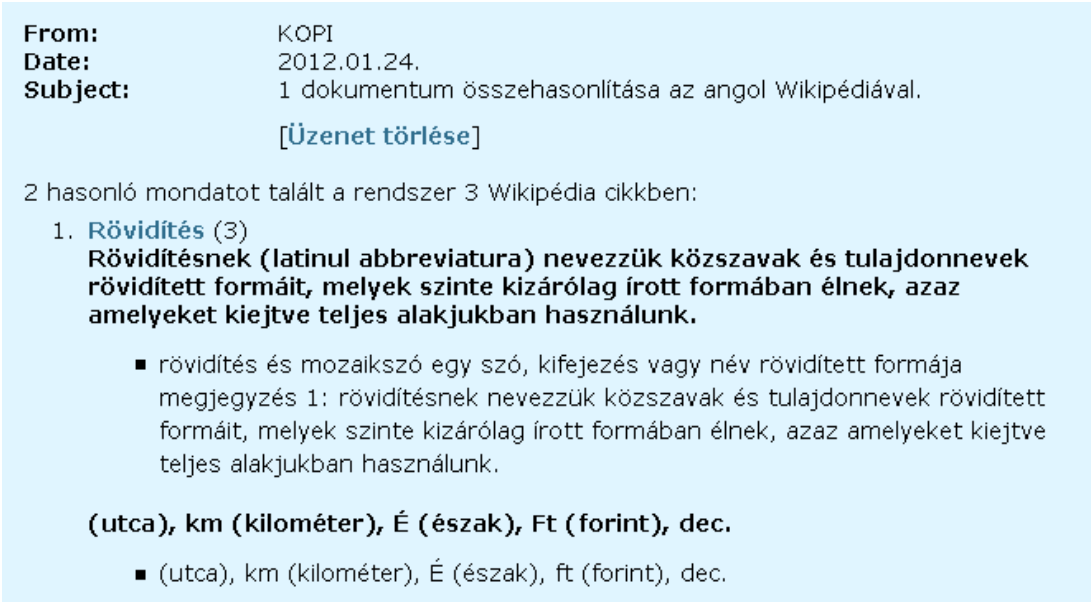
A következő oldalon attól függően jelennek meg a **választható keresési lehetőségek**, hogy hány dokumentumot választottunk ki. Egy dokumentum esetén azt összehasonlíthatjuk a KOPI adatbázisával (minden felhasználó dokumentumával), ez jelenleg körülbelül 30 000 dokumentumot jelent, ez a szám azonban folyamatosan növekszik. Ugyancsak lehetőségünk van a dokumentumot összehasonlítani az angol vagy magyar Wikipédiával. Amennyiben több dokumentumot választottunk ki, akkor azokat egymással is összehasonlíthatjuk - ez a funkció alkalmas egy dolgozatban található szakirodalmak mennyiségének megállapítására vagy hasonló témában íródott dokumentumok összehasonlítására is.



Miután kiválasztottuk a megfelelő keresést, például a magyar Wikipédiát, **elindíthatjuk a plágiumkeresést**. Erről a rendszer egy kis üzenetben tájékoztat minket.



A kereséseket a kereső beérkezési sorrendben dolgozza fel. A rendszer leterheltségétől függően az eredmény pár perc vagy pár óra múlva jelenik meg az üzenetek között, és ha a felhasználó nem tiltotta le, akkor az **eredményről egy email üzenetet** is **kap**.



A Wikipédiával történő összehasonlításkor az üzenet tartalmazza a Wikipédia szócikk nevét, a szócikkben talált mondatokat, valamint azokat a mondatokat, amelyekhez a dokumentumon belül hasonlított. Ez történhet egy nyelven is, mint a fenti példában, de lehet a cikk magyar nyelvű és a dokumentum angol nyelvű, vagy fordítva.


1. Pete Seeger (7)

Seeger was born in French Hospital, Midtown Manhattan, the youngest of three sons.

- Pete Seeger Manhattan közepén, a Midtown-nak is hívott városrész francia kórházában született.

His father, Charles Louis Seeger Jr. was a prominent musicologist, composer, and music professor.

- Apja, ifj. Charles Louis Seeger, zeneszerző és zenetudós volt, aki az elsők között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekből fakadó zenét.

His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the twentieth century.

- Nevelőanyja, Ruth Crawford Seeger egyike volt a huszadik század legkiemelkedőbb női zeneszerzőinek.

Fontos kiemelni, hogy a KOPI-t plágiumkeresőnek hívjuk, de tulajdonképpen **hasonlóságot keres**, azaz **nem különbözteti meg az idézetet a plágiumtól**, ennek eldöntését mindig a felhasználóra bízza a rendszer.

3. A többnyelvű kereső működése

A KOPI új, fordítási plágiumokat kereső algoritmusa négy lépcsős: első lépésben feldarabolja a dokumentumot mondatokra, majd azokat nyelvi feldolgozásnak veti alá, és minden mondatból készít egy összetett lekérdezést, melynek segítségével a hatalmas adatbázisból ki tudja keresni azokat a másik nyelven íródott mondatokat, amelyek lehetséges fordításai a keresettnek. Az utolsó előtti lépésben egy új hasonlósági metrika segítségével részletesen összehasonlítja az eredeti mondatot és az adatbázisban lévő mondatot, hogy azok mekkora valószínűséggel lehetnek egymásnak a fordításai. A hasonlósági metrika (Sim(x,y)) az alábbiak szerint kerül kiszámításra

$$Sim(x,y) = \min (\alpha \cdot | S_x \cap S_y | - \beta \cdot | S_x \setminus S_y | , \alpha \cdot | S_y \cap S_x | - \beta \cdot | S_y \setminus S_x |)$$

ahol S_x az egyik mondat S_y a másik mondat. A ∩ az egymásnak megfeleltethető szavaik száma, a \ az egyik mondat másikból hiányzó szavainak a száma és az α és β a jó fordítások és a hiányzó fordítások súlya, α tipikusan nagyobb mint β.

Utolsó lépésben a rendszer összegyűjti az összes találatot és kijelzi a felhasználónak azokat a szócikkeket, amelyekben megfelelő súlyú találat van. Ez lehet egy rövidebb, igen hasonló, szó szerinti fordítás, vagy egy hosszabb szakasz is, amelyben több kisebb egyezést talált a rendszer.

4. Találati arány

Mint minden keresőnél, a KOPI esetében is az a leglényegesebb kérdés, hogy mit képes megtalálni és mit nem. Az algoritmus teljesítményének számszerűsítéséhez, teszteléséhez a teljes feldolgozott angol Wikipédiát feltöltöttük egy adatbázisba, és ebben kerestünk, mind kézzel magyarra fordított, mind géppel fordított Wikipédia cikkeket. A két keresés között statisztikai különbséget nem találtunk, így a sokkal nagyobb mennyiségű, géppel fordított korpuszon elért eredményeket ismertetjük. A magyar mondatokra keresve 0,67 recall értéket kaptunk, azaz 67% annak az esélye, hogy a teljes Wikipédiából sikerült kiválasztanunk azt a mondatot, amelyiknek a megadott magyar mondat a fordítása. Ez annyit jelent, hogy egyenletes valószínűséget feltételezve a mondatoknál, annak az esélye, hogy egy-egy oldalas fordítást megtalálunk több mint 99.9%.

5. Konklúzió

Anélkül, hogy pontos útmutatást adnánk a rendszer kijátszására, elmondhatjuk, hogy a KOPI Plágiumkereső igen stabilan megtalálja a fordításokat, és nagyobb szövegrészek olyan átírása vagy fordítása, hogy az új algoritmus ne találja meg, olyan nagy munkabefektetést igényelnek az esetleg plagizálni vágyó hallgatótól, hogy ugyanannyi energiával már akár meg is írhatná maga is a dolgozatát. Ezzel a KOPI reményeink szerint el is érte a célját a plágiumok visszaszorítását a hazai felsőoktatásban.

Az új szolgáltatás már most igen népszerű felhasználóink körében, legtöbben a magyar dokumentumokat hasonlítják össze az angol vagy a magyar Wikipédiával. Ezért döntöttünk úgy, hogy a következő évben egy webes keresőt is hozzákapcsolunk a rendszerhez, így lehetővé téve, hogy magyar vagy angol internetes oldalakról átvett részeket is képes legyen megtalálni a rendszer.