A plágiumkereső szoftverek kiskapui

© Kiss András Károly
Pécsi Tudományegyetem, Természettudományi Kar, Pécs
andreasz@gamma.ttk.pte.hu

Plágiumról beszélünk, minden olyan esetben, amikor a más szerzőtől származó tartalmat használunk fel munkánkhoz anélkül, hogy annak igazi szerzőjét megjelölnénk. Az idegen szerző munkáját minden esetben jelölni kell, a tartalmat megfelelő hivatkozással kell ellátni. Alapvetően két esetet különítünk el. Szó szerinti idézés esetén a szerző megjelölésének elmulasztása, illetve az idézőjelek hiánya már plágiumnak minősül. Parafrazálás – más munkájának tartalmilag hű átírása – esetén pedig kötelesek vagyunk megjelölni az átfogalmazott szöveg forrását, ennek hiánya szintén plágium.

Célul tűztük ki, a jelenleg használatban levő plágiumkereső szoftverek hiányosságainak feltérképezését. Eddigi munkánk során tesztelt programok számos hibát vétettek és engedtek át plágiummal megtöltött dokumentumokat. Mint ahogy azt már említettük a tévedés lehetőségét csak növeli, ha az eredeti dokumentum nyelve különbözik a plágiummal élő hallgató nyelvétől, plágiummentesnek tüntetve fel akár teljesen értelmetlen, idegen nyelvről fordító segítségével alkotott szövegeket is. Persze ez alap hibának minősül, de további problémák is felmerültek. Az általunk vizsgált szoftverek közül kettő meglehetősen könnyedén félrevezethető a szavak sorrendjének megváltoztatásával, míg egy harmadikat a szöveg hosszával sikerült megtéveszteni.

Rendkívül fontos természetesen a bírák és a felsőoktatási intézmények egyhangú támogatása a plágium elleni küzdelemben, hogy csak azok munkái kerüljenek elfogadásra, akiknek leadott dolgozata valóban saját kutatásra épül, továbbá önálló gondolatokat rejt. Joggal feltételezzük, hogy a vizsgálat elmaradása esetén a hallgatók egy része élni fog ezzel a lehetőséggel. A vizsgálat elmulasztása, egy későbbi lebukás az egyetem hírnevére is rossz fényt vethet. Az ellenőrzésnek meg kell történnie a szakdolgozat készítését megelőző évek során leadott dolgozatok estében is, sőt a középiskolás házi dolgozatok esetében is javasolt. Mindez sok munkát és plusz energiát igényel, de ezen erőfeszítések megtétele biztosíthatja egy plágiummentes oktatás létrejöttét, megszüntetve ezzel napjaink intézményeinek egyik fő problémáját.

A plágiumszűrésnek meg kell történnie mindkét fél részéről. Az oktatóknak és a hallgatóknak egyaránt meg kell érteniük ennek fontosságát. Azért választottuk ezt a témát, mert célunk egy a hallgatók és az oktatók számára egyaránt hozzáférhető szoftver kifejlesztése.

Anyag és módszer

Kutatásunk kezdeti szakaszán 3 ingyenesen elérhető szoftvert teszteltünk, hogy mennyire megbízhatóan és hitelesen dolgoznak. A következő programokat vizsgáltuk:

- KOPI Plágiumkereső (1. ábra)
- Plagiarism Detector (3. ábra)
- Plagiarism Checker (4. ábra)

A kutatásunk alapját három dokumentum képezte, melyeket mindhárom programmal megvizsgáltunk. Az első dokumentum egy Wikipédiáról kimásolt szöveget tartalmazott, a második egy nagy terjedelmű, 38 oldalas Internetről származó cikk, míg a harmadik egy fordítószoftver által készített IEEE által indexelt cikk, ami a fordítás után értelmetlen szöveggé alakult. Ahol szükséges volt, ott más dokumentumokat is felhasználtam a teszteléshez, kettő teljesen megegyező, egy részben egyező és egy az előbbiektől teljesen különböző dokumentumot a saját művek közti egyezőség ellenőrzésének teszteléséhez. A dokumentumokban közös, hogy nem tartalmaz saját munkát, tartalmuk az internetre mások által feltöltött anyagból származik. Ezeket a dokumentumokat azért készítettem, hogy alátámasszam segítségükkel azon állításomat, hogy a plágiumkereső szoftverek működése nem elégíti ki a mai kor igényeit.

KOPI Plágiumkereső

Az általunk vizsgált első szoftver az MTA SZTAKI Elosztott Rendszerek Osztálya által fejlesztett és üzemeltetett KOPI online Plágiumkereső. A kutatások 2001-ben kezdődtek el, majd 3 évnyi munka után, 2004-ben elindult a szolgáltatás magyar és angol nyelven, 2011-ben pedig fordítási plágiumok felderítésére képes algoritmussal bővült. A program az egyetlen magyar fejlesztésű plágiumkereső szoftver.

Használatához regisztráció szükséges. A bejelentkezés után a vizsgálandó dokumentumokat fel kell tölteni, megadni a címét és szerzőjét. Pár perces konvertálást követően, a számunkra megfelelő keresés kiválasztása után kezdhetjük is az ellenőrzést. Itt több lehetőség közül választhatunk. Egynyelvű keresés esetén az általunk feltöltött dokumentumok között kereshetünk átfedést, továbbá összehasonlíthatjuk a mi dokumentumunkat a mások által feltöltöttekkel. Többnyelvű keresés esetén a dokumentumunkat a magyar, illetve az angol Wikipédia szövegével vethetjük össze. Ez csak egy szűk halmaza az Interneten fellelhető dokumentumoknak, másolni pedig – mint azt bizonyítjuk is- sok más helyről is lehet.

A KOPI Plágiumkereső működése négy lépésből áll, legyen az egy saját dokumentum, vagy a Wikipédia valamely cikke. A forrás szöveget a program feldarabolja 40-60 karakternyi töredékekre. A töredékek tárolása ebben a formában nagyméretű adatbázist eredményezne, ezért a töredékekből úgynevezett ujjlenyomatokat hoz létre egy tömörítési eljárással. Ennek két előnye is van, egyrészt ebben a formában jóval kisebb helyet foglal, másrészt ezekből az ujjlenyomatokból nem állítható elő az eredeti dokumentum. A program az adatbázisba felkerülő ujjlenyomatok között keres egyezést egy lekérdezés futtatásakor.

A korábban említett dokumentumok ellenőrzése meglehetősen érdekes eredményeket adott. Elsőként a Wikipédiáról kimásolt szöveget ellenőriztettük. A felhasználók dokumentumaival való összevetés még kielégítő eredményt adott. Valaki már előttünk töltött fel az miénkkel egyező anyagot, ezt a hasonlóságot könnyedén kimutatta. Ez után újra kellett futtatni a programot, hogy a magyar Wikipédia szövegével is összevethessük. A plágiumkereső 26 mondatnyi egyezést talált csak, ami a szöveg 68%. Ez elenyésző, ahhoz képest, hogy a dokumentum teljes tartalma a Wikipédiáról származott. Az angol Wikipédiával való összevetés értelemszerűen nem hozott eredményt.

A második dokumentum az internetről másolt hosszú szöveg volt. A feltöltött munkákkal való összevetés itt is jól működött, 100%-os egyezést mutatva, egy másik

általunk feltöltött dokumentummal. A másik két ellenőrzés viszont nagyon lassan futott le, a futás majd egy órát vett igénybe, az eredmény pedig az, amire előzetesen számítani lehetett. A program nem talált egyezést, mivel a szöveg nem a Wikipédiárol származott.

Harmadikként az IEEE által indexelt cikken végeztük el az ellenőrzést. Ez egy értelmetlen szöveggé vált a fordítást követően, így még csak hasonlót sem talált az adatbázisban, plágiummentesnek nyilvánítva a szöveget. Ezzel megegyező eredményt adott a kereső a magyar és az angol Wikipédiával való összehasonlítás eredményeként, ami nem meglepő, hiszen a forrásszöveg nem ezekről az oldalakról származott.

Itt szükségesnek éreztük tesztelni a saját dokumentumok összehasonlítása funkciót. Azért gondoltuk így, mert ez a funkció csak ennél a programnál jelenik meg, ráadásul rendkívül hasznosnak találtuk, mivel gyakori, hogy a hallgatók egymás írását másolják le. Ehhez két teljesen egyező, egy részben egyező és egy teljesen különböző dokumentumot töltöttünk fel. A dokumentumok összehasonlításának eredménye hűen tükrözte a valóságot, ez a funkció helyesen működött.

A program részletes vizsgálata során igen hamar falakba ütköztünk. A program 10 ellenőrzés lefuttatása után nem engedett több keresést lefuttatni, azzal az indokkal, hogy a napi keret lejárt. Folytatva a munkát a program már a havi keret elhasználását írta ki. A napi korlát megegyezik a havi korláttal, míg éves szinten 120 lekérdezést engedélyez. Ez nagyon kevés, hiszen egy dolgozat összevetését a többi dolgozattal és a Wikipédiával nem lehet egyidejűleg elvégezni. Egy évfolyam dolgozatainak kijavítása, ellenőrzése akár több évet is igénybe vehetne. Kétségeink vannak afelől, hogy bármely egyetem valóban használná a programot.

Plagiarism Detector

A következő vizsgált szoftver a Skyline, Inc. által fejlesztett Plagiarism Detector volt, melyet a készítők számos egyetemmel együttműködve hoztak létre. A program fejlesztése 2005-ben kezdődött el, majd több díjat is nyert 2008 és 2010 között. Számos nyelvet támogat, köztük több világnyelvet, de sajnos a magyart nem. Számos felhasználóval rendelkezik a világ minden pontjáról, főleg az Európai Unió és az Amerika Egyesült Államok területén terjedt el.

Használat előtt a programot le kell tölteni, majd telepíteni, ezután kezdhetjük el használni. Regisztrációt nem igényel. Az ellenőrzés kivitelezését egy varázsló segíti. Először meg kell adnunk a vizsgálni kívánt dokumentum elérési helyét, majd el kell döntenünk, hogy az adatbázisban, az interneten, esetleg mindkettőben egyszerre szeretnénk keresni. A keresés lefutása után a program egy másik ablakban mutatja meg az eredményt. Itt a listázás feltételeit változtathatjuk, annak függvényében, hogy mire vagyunk kíváncsiak. Megadhatjuk például, hogy milyen típusú kereséseket mutasson, a keresések dátumát, miszerint legyenek rendezve, hány százalék felett mutassa, illetve megváltoztathatjuk a nézetet is. A program ábrákkal szemlélteti, hogy a szöveg hány százalékát találta eredetinek illetve plágiumnak, továbbá megadja a linkelt és az idézett sorok százalékos arányát is. A Plagiarism Detector működése:

1. A dokumentum szövegének feldarabolása
2. A darabok elküldése egy keresőmotornak
3. A keresés találatainak letöltése
4. A lehetséges források összevetése a szöveggel

5. Az eredetiségvizsgálat eredményének elkészítése

A program működése 5 lépésből áll. A megadott dokumentumot a KOPI Plágiumkeresőhöz hasonlóan apró 40-60 karakteres darabokra vágja, ezeket a darabokat pedig elküldi egy keresőmotornak. A program a Google-t, a Bing-et és az AltaVista-t használja. Sosem tárolja a keresett dokumentumot és nem is publikálja azokat. A keresés eredményeit letölti a gépre, ezek a lehetséges források. Ezeket a program összehasonlítja az általunk megadott fájllal, majd ezekből elkészíti az eredetiség vizsgálat eredményét. A böngészővel megnyitva még részletesebb adatokat kaphatunk a keresés részleteiről. A program minden esetben a saját erőforrásainkat használja, gyorsabbá téve ezzel a keresést, hiszen nem kell osztoznunk ezeken más felhasználókkal.

Elsőként itt is a Wikipédiáról másolt szövegen futtattuk le a keresést. A program valóban – a gyártók oldalán említetteknek megfelelően – egy-két percen belül lefutott. Ez jóval gyorsabb, mint a korábban vizsgált KOPI Plágiumkereső futási ideje. A Plagiarism Detector az eredményben is felülmúlta vetélytársát, noha az eredmény továbbra sem tökéletes. A program 81%-ban határozta meg a másolt szöveg mennyiségét, ami nagyságrendekkel jobb a KOPI 68%-os eredményéhez képest, de még mindig rossz, ahhoz viszonyítva, hogy a dokumentum 100%-a plágium, illetve, hogy a készítők a honlapon 99%-os eredményt ígérnek. Az eredmény már csak azért is érdekes, mivel a forrásszöveget megtalálta az Interneten, az összehasonlítás során mégsem találta teljesen egyezőnek őket.

Folytatva a tesztelést, elvégeztük a plágiumkeresést a 38 oldalas dokumentumon is. Elsősorban arra voltunk kíváncsiak, hogy mennyivel hosszabbodik a keresés időtartama, továbbá, hogy mennyiben befolyásolja az eredményességet a szöveg hossza. A futamidő várakozásainknak megfelelően hosszabbodott, de a növekedés nem volt drasztikus. A program 3 perc alatt lefutott. Visszaemlékezve a KOPI Plágiumkereső által produkált, megközelítőleg egy órás ellenőrzésre megállapíthatjuk, hogy a Plagiarism Detector nagy terjedelmű szöveg esetén is gyorsan lefut. A keresőmotoroknak köszönhetően a program talált egyezést, de közel sem mindet. A szöveg 48%-át eredetinek találta, és csak 52%-nál fedezett fel plágiumgyanút.

Utolsóként az IEEE által indexelt idegen nyelvről fordított cikket vetettük ellenőrzés alá. A futtatás után a program által készített jelentésben szereplő eredmény kiábrándító. A kereső átengedett a szűrésen egy teljesen értelmetlen, fordító segítségével kreált szöveget, 100%-ban eredetinek titulálva azt, nem találva egy mondatnyi plágiumot sem.

A Plagiarism Detector a plágiumkeresésen felül lehetőséget ad az idézések ellenőrzésére is. Ezt az általunk készített dokumentumok minimális átalakítása után végeztük el. A változtatás lényege abban rejlett, hogy a szöveget lehivatkoztuk. Az idézések vizsgálata lényegében mennyiségi ellenőrzés, melynek eredménye megmutatja, hogy a tartalom mekkora hányada van helyesen lehivatkozva. Ez a funkció nem működik jól. Helyes és egyforma hivatkozással ellátott dokumentumok esetében, ez egyiknél felismeri azt, míg a másiknál nem. Ez súlyos hiba, hiszen közel sem mindegy, hogy szövegünk plágium vagy idézet.

Vizsgálatainkat a program demó verzióján végeztük el. Munkánkat ez két ponton korlátozta. Egyrészt a demo program tíz futtatást engedélyez, de ez alatt korlátlan mennyiségű ellenőrzést végezhetünk, így csak arra kellett figyelnünk, hogy ne zárjuk be az alkalmazást idő előtt. Másrészt ebben a verzióban le van tiltva az adatbázisban

való, illetve a kombinált keresés. Kutatásunk során ezekre nem volt szükségünk, hiszen az általunk használt dokumentumok az Interneten is fellelhetőek voltak.

Plagiarism Checker

A harmadik vizsgált plágiumkereső szoftver a Plagiarisma.Net által készített Plagiarism Checker volt. A program vásárlói neves, főleg amerikai középiskolák és egyetemek, mint a Benjamin Franklin High School, The University of Michigan, Harvard University stb. A programhoz van letölthető alkalmazás, de elérhető online is. Több tucat nyelvet támogat, de magyarul sajnos nem elérhető.

Ha a programot online szeretnénk használni, akkor a http://plagiarisma.net/ megnyitása után már kezdhetjük is a plágiumkeresést. Az ellenőrizendő szöveget be kell másolni a megfelelő ablakba, majd a számunkra tetszőleges keresőmotor – Google, Babylon vagy Yahoo – kiválasztása után indíthatjuk is a keresést. A kereső egyesével vizsgálja meg a mondatokat, az eredmény az eredeti mondatok és az összes mondat aránya százalékban kifejezve. Az egyes szövegrészek eredményeit külön megjeleníti. Az alkalmazás működése hasonló. Itt meg is nyithatjuk a szöveget, nem feltétlen kell nekünk bemásolni. Az ellenőrzést lefuttatva az online kereséshez hasonló módon írja ki az eredményt. A keresés folyamata mindenben megegyezik az előbb vizsgált Plagiarism Detector esetében bemutatottakkal.

Az ellenőrzést – mint mindenhol máshol – itt is a Wikipédiáról másolt szöveggel kezdtük. A futási ideje megegyezett az korábban vizsgált Plagiarism Detector futási idejével. A keresést a Yahoo keresőmotor kiválasztásával lefuttatva egy eredeti mondatot talált, a szöveg 91%-át pedig plágiumnak minősítette. Ugyanezt a szöveget ellenőrizve a Google is eredetinek minősítette az előbb említett mondat túlnyomó részét, de a mondat utolsó két szavát már másoltnak vélte, 92%-ban plágiumnak minősítve a szöveget. Végül lefuttattuk a Babylon-nal is a keresést. Ez már a teljes szöveget plágiumnak vélte, tökéletes eredményt adva.

Ezután következett a 38 oldalas dokumentum. A szöveg importálása kisebb nehézségeket okozott, igen lassan történt meg, de ez indokolható a szöveg hosszával. A futási idő mindhárom keresőmotor esetében hasonló volt, illetve megközelítőleg ugyanolyan hosszú volt, mint a Plagiarism Detector esetében. Érdekes, hogy a különböző keresőmotorokkal elvégzett keresések eredménye itt is eltérést mutat. A Google segítségével végzett ellenőrzés 94%-ban találta plágiumgyanúsnak a szöveget. Ezután elvégeztük az ellenőrzést a Yahoo használatával is, melynek eredménye, hogy a teljes szöveg plágium. Ugyan ezt az eredményt kaptuk a Babylon keresőmotorral végzett ellenőrzés eredményeként is.

Utolsóként az IEEE által indexelt angol nyelvről fordított szövegen végeztük el a plágiumkeresést, mindhárom keresőmotor segítségével. Kijelenthetjük, hogy a Plagiarism Checker nincs felkészítve fordítási plágiumok felderítésére. A Google kiválasztása után elvégzett keresés nem találta meg a szöveg angol megfelelőjét, így a szöveget teljes egészében eredetinek találta. Ugyanezen feladatnál a Yahoo segítségével elvégzett keresés sem fedezett fel plágiumot, majd a Babylon is plágiummentesnek minősítette a szöveget.

Összegzés

Ezek után már tisztán láthatjuk, hogy ezek a szoftverek tökéletesen kielégítik azok vágyait, akik még mindig más munkáiból szeretnének szakdolgozatot írni. A programok használatát nem ajánlom azoknak, akik érdemben kívánnak tenni a plágium ellen, mivel a vizsgált programok meglehetősen könnyen kijátszhatóak. Az elvégzett tesztek eredményei alapján már tudjuk, hogy milyen kiskapukat kell bezárnunk. Az ellenőrzések során megbízhatóan és helyesen működő funkciók esetében programunk célja a vizsgált programoknál tapasztalt eredményesség megtartása. A fontos fejlesztések a keresések futtatása során gyengén teljesítő funkciók terén lesznek. Programunk fő célja a tapasztalt hibák kijavítása, a hiányosságok pótlása, illetve egy könnyen kezelhető, mindenki számára átlátható grafikus kezelői felület készítése.

A KOPI Plágiumkeresőnél vizsgált, adatbázissal támogatott megoldást hasznosnak gondoljuk. A fejlesztett program is rendelkezni fog egy adatbázissal, ami a korábban vizsgált dokumentumokat fogja tárolni. Úgy gondoljuk, a korábbi években beadott munkákat sem szabad hagyni feledésbe merülni, hiszen a diákok előszeretettel vesznek alapul olyan szakdolgozatokat, amelyek korábban már megfeleltek a követelményeknek.

A KOPI által kínált, feltöltött dokumentumok összehasonlítása egymással funkciót mindenképpen szeretnénk megtartani. A tesztek során rendkívül megbízhatóan teljesített. Fontosságát annak köszönheti, hogy úgy gondoljuk meglehetősen magas az olyan esetek száma, amikor a hallgatók egymásról másolnak. Természetesen ez a funkció nem is annyira a szakdolgozatok, mint inkább a beadandó feladatok esetében lehet hasznos, ahol többen kapják ugyanazt a témát.

Programunkba szeretnénk beilleszteni, egy az idézés mennyiségét mérő algoritmust. Praktikus mivolta miatt készítjük el programunkhoz ezt a funkciót. A szakdolgozatokhoz szükséges igénybe vennünk mások munkáját, de nem mindegy, hogy ezt miként és milyen mértékben tesszük. Egyrészt a forrás megjelölésének elmulasztása plágium, másrészt a tanulmányi és vizsgaszabályzat minden esetben meghatározza, hogy a szakdolgozat hány százaléka lehet más munkásságának az idézése. A helyesen lehivatkozott forrásból származó szöveg mennyiségét fogja mérni algoritmusunk, megkönnyítve ezzel az oktató munkáját.

A potenciális források keresését meglehetősen szűk halmazon végzi el a KOPI Plágiumkereső, a magyar és angol Wikipédián kívül számos helyről származhat a forrásszöveg. A másik két szoftver már a teljes webet vizsgálja, úgy gondoljuk ez a követendő megoldás a források keresésére. Az általunk kapott eredmények gyengék, ezért ezen a téren sok fejleszteni való van. A források keresésénél a programnak nem csak szó szerinti másolás, hanem parafrazálás esetén is működni kell, csak ebben az esetben lehetünk biztosak abban, hogy a szakdolgozat valóban önálló munka eredménye.

A legelhanyagoltabb területnek a fordítási plágiumok vizsgálata mutatkozott. A Plagiarism Detector és a Plagiarism Checker teljesen inkompetens idegen nyelvű forrás esetén, míg a KOPI Plágiumkereső csak az angol Wikipédia szövegéről fordított szövegeket képes megtalálni. Világunkban, ahol a nyelvtudást megkövetelik és elvárják, a fordítási plágiumok száma növekvő tendenciát mutat. Úgy gondoljuk, hogy ez a fordítók fejlődésével tovább fog emelkedni. Éppen ezért, a fejlesztésben kulcsfontosságú és kiemelt szerepet fog kapni a plágium e formájának kiszűrésére alkalmas algoritmus elkészítése.