MTA SZTAKI DSD

Department
of Distributed
Systems

KOPI
A fordítási plágiumok keresője

Pataki Máté
Kovács László

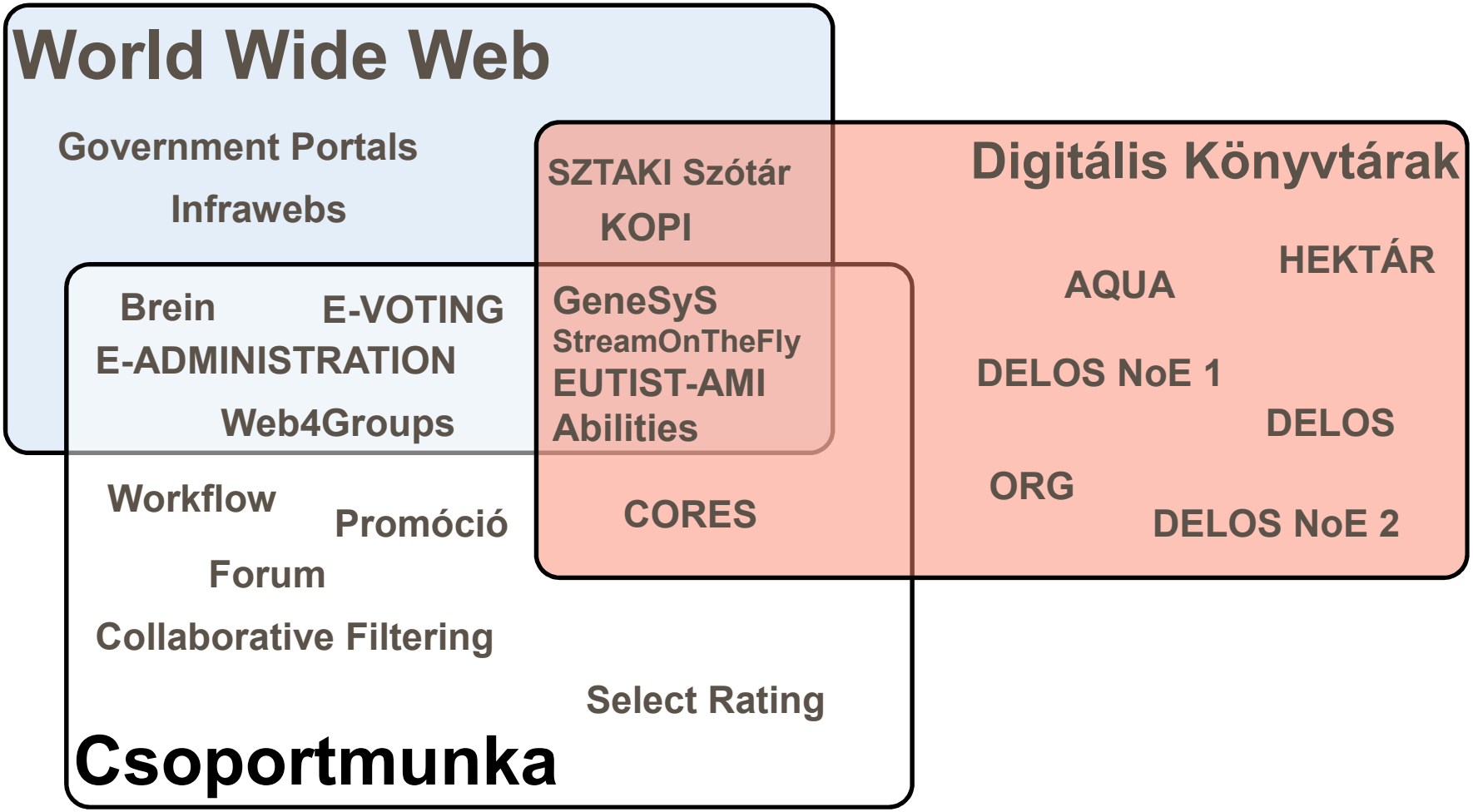
Department of Distributed Systems

DSD

MTA SZTAKI DSD

- MTA SZTAKI Elosztott Rendszerek Osztály
- 1995. óta létezik
- 12 teljes állású munkatárs, és diákok
- Kutatás, fejlesztés, (online) szolgáltatások
- Munkák eloszlása:
  - 80% EU-s k+f pályázatok
  - 15% Hazai pályázatok
  - 5% Belső projektek és szolgáltatások
- Három fő terület:
  - Digitális könyvtárak és archívumok
  - Csoportmunkát támogató technológiák
  - Webes rendszerek

K+f projektek



További információk

http://dsd.sztaki.hu

Probléma

- A plágium probléma a
  - Felsőoktatás területén
  - Középiskolában is egyre inkább
  - Tudományos életben
  - Digitális könyvtárak számára
  - Könyvkiadóknak
  - Cégek esetében is (pl. honlapok tartalma)
  - Wikipedia

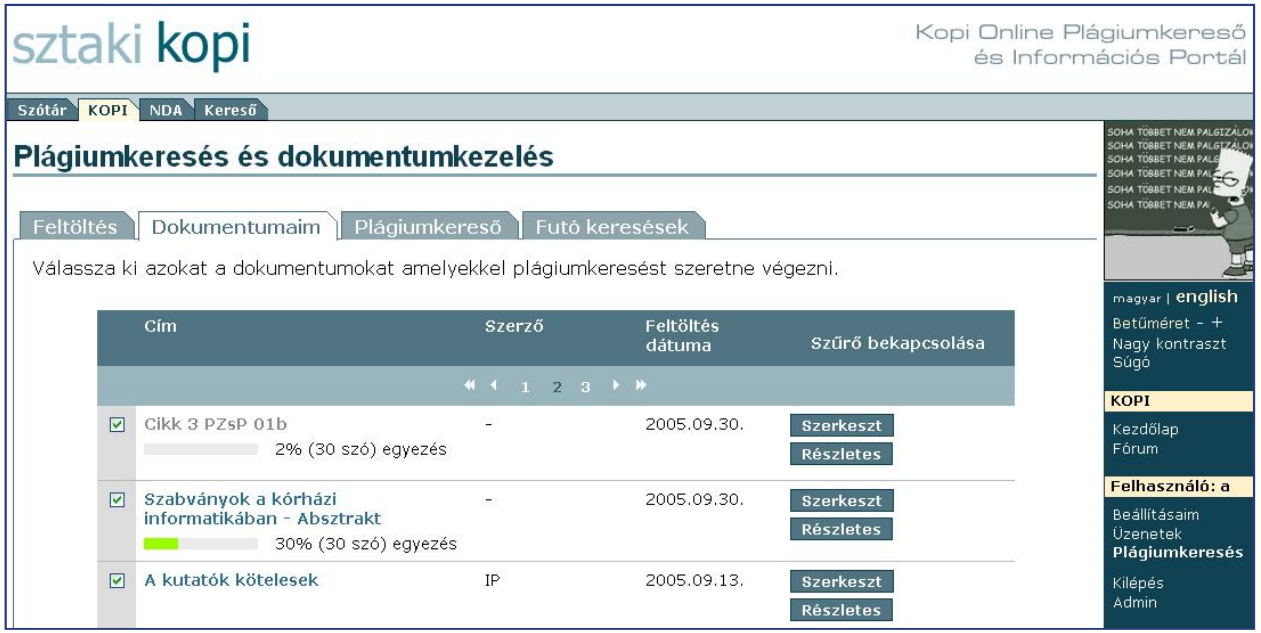
Probléma

1. Sok a diák
2. Hasznos anyagok az interneten
3. Digitális szakdolgozatok
4. Jó nyelvtudás

- 1-3 → könnyű plagizálás
  - Plágiumkeresők
  - KOPI
- +4 → fordítási plágiumok
  - ???

A KOPI Plágiumkereső ismertetése

- KOPI Online Plágiumkereső és Információs Portál - internetes **hasonlóság** és **plágium**kereső szolgáltatás
- MTA SZTAKI Elosztott Rendszerek Osztály
- http://kopi.sztaki.hu/



A KOPI Plágiumkereső története

- **2001**: elkezdődtek az **alapkutatások**, hogy miként lehetne egy nyelvfüggetlen, magyar nyelven is jól működő plágiumkeresőt elkészíteni
- **2003**: állami támogatással elkezdődött a KOPI Portál **fejlesztése** (ITEM pályázat, IHM-OM)
- **2004**: elindult a **publikus plágiumkereső szolgáltatás** magyar és angol nyelven
- **2006**: számos tanár elkezdi használni a szolgáltatást, **több mint 1000 felhasználó**

A KOPI Plágiumkereső története

- **2007**: **fejlesztés**ek az első három év tapasztalatai alapján
- **2009**: felhasználóink száma már több mint **10 000**, egyes **egyetemi karok** használják már a KOPI Plágiumkeresőt rendszeresen
- **2010**: **új kutatás**ba kezdtünk, hogy miként lehetne fordítási plágiumokat felismerni és megtalálni
- **2011**: a **világon elsőként** beépítettük a KOPI Plágiumkeresőbe **fordítási plágiumok** megtalálására képes algoritmust, amely a teljes angol **Wikipédiában** keres

Tapasztalat

- Sok esetben nem szándékos a plagizálás
- Nem oktatják az egyetemeken a helyes idézés módját
- Mekkora hasonlóságot várunk el
  - 0% - nincs irodalomkutatás
  - 10%
  - 50%
  - 100% - egyértelműen plágium
- Diákok és tanárok egyaránt használják a KOPI Plágiumkeresőt

Tapasztalat

- BME, 400-500 diák, 5 feladat, 6 év
  - 2007: 9 pár, 2009: 4 pár, 2010: 2 pár
- Statisztika
  - Közel 20 000 felhasználó
  - 30 000 dokumentum
  - Körülbelül 25 000 000 dokumentumrészlet
    - Ebből 20 000 000 magyar

Plágiumkeresés nyelvek között

- Feladat
  - Működő szolgáltatás magyaroknak
  - Az angol eredeti szöveg megtalálása a magyar fordítás ismeretében
- Egyéb felhasználási területek
  - Párhuzamos korpusz építése
  - Létező fordítások keresése
  - Hírek, cikkek, anyagok terjedésének a vizsgálata
  - Idézetkereső

Irodalom – fordítási plágiumok

- Test cases for plagiarism detection software, Debora Weber-Wulff, HTW Berlin, 2010
  - 48 különböző plágiumkereső, 42 teszt
  - *The biggest gap in all the plagiarism checkers was the **inability to locate translated plagiarism**. While this is widely expected as the technology to make such detections simply is not there.*

Irodalom – fordítási plágiumok

- CLEF 2010
- Potthast: Overview of the 2nd International Competition on Plagiarism Detection
  - *After analyzing all 17 reports, certain algorithmic patterns became apparent to which many participants followed independently. ... In order to simplify the detection of cross-language plagiarism, non-English documents in D are* ***translated to English using machine translation*** *(services).*

Irodalom – fordítási plágiumok

- Európában fontos téma
- Az algoritmusok nyelvpár-függők
- Magyar nyelvben három fő akadály
  - nem kötött szórend
  - ragozás
  - jelentős nyelvtani különbözőség az angol nyelvtől
  - rosszak az automatikus fordítók (erre)

Az új algoritmus

- Mondatalapú
  - szó, n-szó, tagmondat, bekezdés, dokumentum

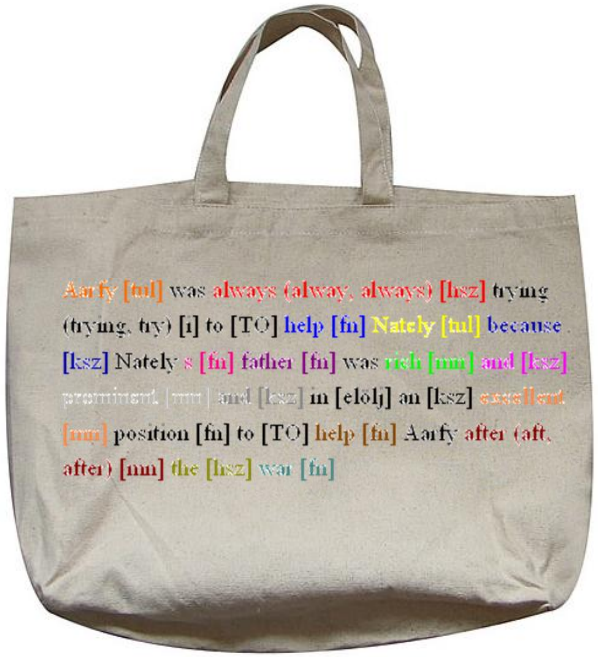
- Hasonlósági metrika

$$\mathrm{Sim}(x,y) = \min\left(\alpha \cdot |\, S_x \cap S_y \,| - \beta \cdot |\, S_x \setminus S_y \,|,\ \alpha \cdot |\, S_y \cap S_x \,| - \beta \cdot |\, S_y \setminus S_x \,| \right)$$

  - Lapos szótár, szószedet

Az új algoritmus

- Bag of words jellegű algoritmus
  - előnyök
    - nem kell szóegyértelműsítést alkalmazni
    - nem kell szinonimaegyértelműsítést / -szűrést alkalmazni
    - nem érzékeny a szavak sorrendjére
  - hátrányok
    - keresési tér nagy
    - lineáris keresési idő



Tesztkörnyezet

- Angol Wikipedia
  - 31GB XML
  - 3 800 000 szócikk
  - SZTAKI Desktop GRID
  - Letölthető szöveges változat több nyelven: **http://kopiwiki.dsd.sztaki.hu/**

- Google Translate
  - Csak teszteléshez
  - Találati arány egyezik a kézi fordításéval

Demó

- http://www.wikipedia.org
- http://translate.google.com
- http://kopi.sztaki.hu

Statisztikák

		találatok száma →				
		1	**2**	**3**	**4**	**5**
mondatok száma →	**1**	0,555709				
	2	0,802606	0,308813			
	3	0,9123	0,583218	0,17161		
	4	0,961035	0,766092	0,400344	0,095365	
	5	0,982688	0,874424	0,603594	0,264845	0,052995
	6	0,992309	0,934587	0,754096	0,453091	0,170722
	7	0,996583	0,966664	0,854397	0,620362	0,327637
	8	0,998482	0,98329	0,916785	0,750418	0,490307
	9	0,999325	0,991732	0,953742	0,842869	0,634853
	10	0,9997	0,995952	0,974854	0,904482	0,750449

Találati arány

Demó

sztaki kopi

Szótár | KOPI | NDA | Kereső

Teszt

A plágiumkeresőt úgy tesztelheti, hogy Petőfi Sándor verseiből szúr be egy-két versszakot ide. A rendszer már párszor tíz szavas egyezést is képes kijelezni. Petőfi verseket az alábbi oldalon találhat: http://mek.oszk.hu

magyar | english

Betűméret - +

Nagy kontraszt

Súgó

KOPI

Kezdőlap

Plágiumkeresés

Feltöltés

Dokumentumaim

Plágiumkereső

Futó keresések

Üzenetek

Fórum

Felhasználó:

Beállításaim

Kilépés

Dokumentumok

Jogszabályok

Kezdőlap

Tartalom: **Mit tud**, **Kinek szánjuk**, **Hol keres**, **A KOPI használatáról röviden**, **Történet**, **Kapcsolat**

Üdvözöljük a KOPI plágiumkereső portálon!

KOPI - A fordítási plágiumok keresője

"plágium: szellemi tolvajlás, más művének közlése saját név alatt, a mű alapgondolatának vagy részleteinek felhasználása a szerzőre való hivatkozás nélkül" (Magyar Értelmező Szótár)

Napjainkban egyre gyakrabban találkozunk szó szerint lemásolt, plagizált tartalommal. Ennek felkutatására számos megoldás született már, ezek közül magyar nyelven a SZTAKI Elosztott Rendszerek Osztálya által üzemeltetett **KOPI Plágiumkereső a legismertebb**.

Demó

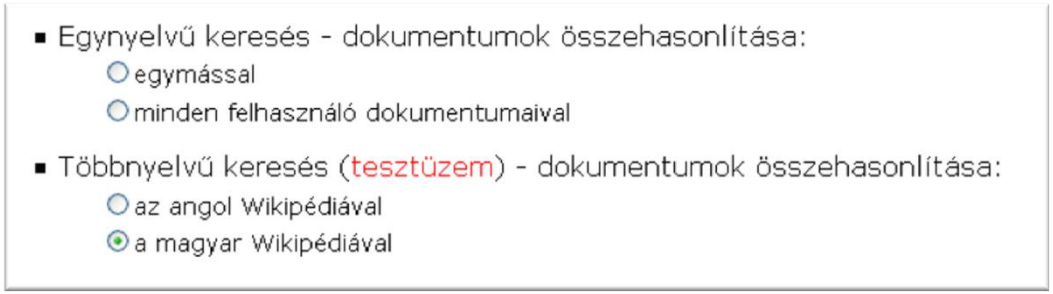
1.


Cím:
Szerző:
Dokumentum:
Browse...
Feltöltés

2.


Miért kell akadálymentesíteni?
Pataki Máté
2010.11.14.
Szerkeszt
Részletes

3.


Egynyelvű keresés - dokumentumok összehasonlítása:
egymással
minden felhasználó dokumentumaival
Többnyelvű keresés (tesztüzem) - dokumentumok összehasonlítása:
az angol Wikipédiával
a magyar Wikipédiával

Demó

From: KOPI
Date: 2012.01.24.
Subject: 1 dokumentum összehasonlítása az angol Wikipédiával.

[Üzenet törlése]

2 hasonló mondatot talált a rendszer 3 Wikipédia cikkben:

1. Rövidítés (3)
   Rövidítésnek (latinul abbreviatura) nevezzük közszavak és tulajdonnevek rövidített formáit, melyek szinte kizárólag írott formában élnek, azaz amelyeket kiejtve teljes alakjukban használunk.

   - rövidítés és mozaikszó egy szó, kifejezés vagy név rövidített formája
     megjegyzés 1: rövidítésnek n
     formáit, melyek szinte kizáról
     teljes alakjukban használunk.

   (utca), km (kilométer), É (észak

   - (utca), km (kilométer), É (ész

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.

Demó

12 cikk magyar	kopiwiki	2011.11.17.	Szerkeszt	
8% (8 mondat) egyezés			Részletes	
Prince Sisouk GER	Wiki	2011.11.30.	Szerkeszt	Tömör
28% (2 mondat) egyezés				

Kiadó:	-
dokumentum:	Eredeti: **PrinceSisoukGer.docx** Szöveges: **PrinceSisoukGer.docx.txt** Dokumentum hossza: 80 szó Csak én láthatom
Nyelv:	Német
Megjegyzés:	
Kulcsszavak:	
Kivonat:	
Hasonló dokumentumok:	**2 mondat** egyezés az angol Wikipédiával: **Sisouk na Champassak**

Prince Sisouk HUN	Wikipedia	2011.11.30.	Szerkeszt	
80% (4 mondat) egyezés			Részletes	

Demó – Nem talált mondatok

HUN: A Carnatic ez? - robbant ki.
ENG: Am I on the Carnatic?" -1

HUN: A detektívnek minden oka megvolt, hogy így okoskodjon.
ENG: The detective was not far wrong in making this conjecture. -2

HUN: A detektív is hasztalanul fáradozott, hogy ő legyen a nézeteltérésben a főszereplő.
ENG: As vainly did the detective endeavor to make the quarrel his. -3

HUN: - A hídon.
ENG: "On the bridge." 2

HUN: - Addig óvadék ellenében szabadlábra helyezem mindkettőjüket.
ENG: "Meanwhile, you are liberated on bail." -3

HUN: A gentlemannek különben is kész volt a terve a továbbiakra.
ENG: Mr. Fogg's course, however, was fully decided upon. -6

HUN: A gépész azonban történetesen épp e napon felment a fedélzetre, megkereste Mr. Foggot, és meglehetősen élénk vitát folytatott vele.
ENG: On this day the engineer came on deck, went up to Mr. Fogg, and began to speak earnestly with him. -4

KOPI Portal

http://kopi.sztaki.hu

Köszönöm a figyelmet!

Web: *http://dsd.sztaki.hu*

Email: *Mate.Pataki@sztaki.hu*