MTA SZTAKI DSD

Department
of Distributed
Systems

Fordítási plágiumok keresése

MSZNY 2011

Pataki Máté

Probléma

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

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

Feladat

- 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

- 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)

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**.*

Megoldás

- Lehetséges megoldások
  - Jelenlegi egynyelvű algoritmus továbbfejlesztése
    - n-gram
    - 3-6 szó: 2-4 nagyságrend
    - + szórend
  - Teljesen új algoritmus
    - ???

Az új algoritmus

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

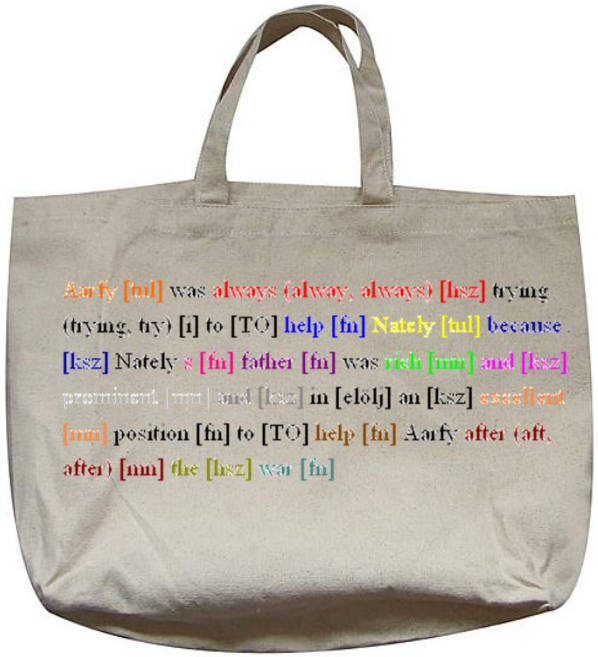
- Hasonlósági metrika

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

- 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



Az új algoritmus

- Fontosabb változók
  - Stopszavak
  - Tulajdonnevek / ismeretlen szavak
  - Hasonlóság mértéke / hol vágunk (f+ / f-)
  - Szótár mérete
  - Talált/nem talált szavak értéke (szófajonként)

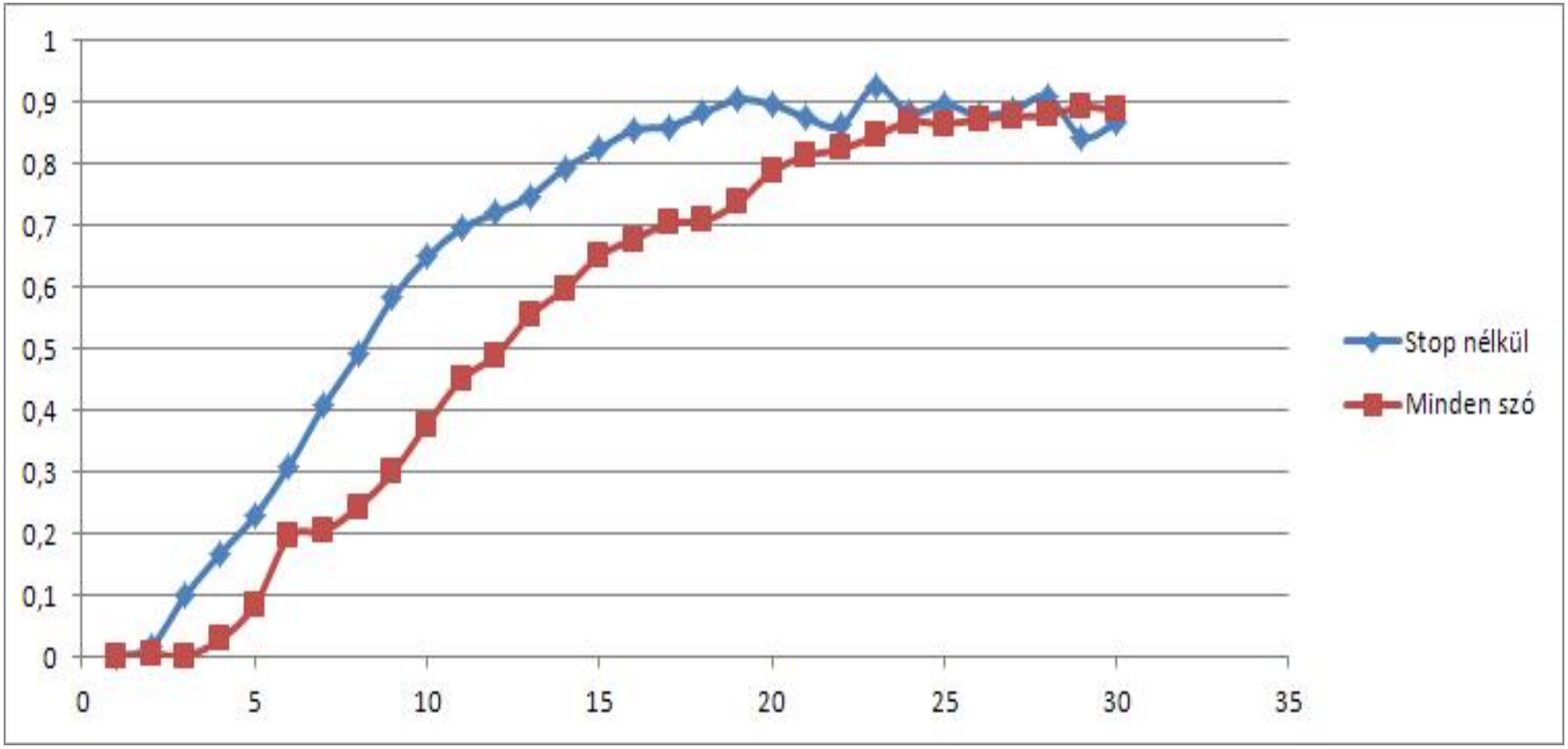
Aarfy [tul] was always (alway, always) [hsz] trying (trying, try) [i] to [TO] help [fn] Nately [tul] because [ksz] Nately s [fn] father [fn] was rich [mn] and [ksz] prominent [mn] and [ksz] in [elölj] an [ksz] excellent [mn] position [fn] to [TO] help [fn] Aarfy after (aft, after) [mn] the [hsz] war [fn]

Aarfy mindig (mindig, mind) igyekezett (igyekezik, igyekezett) Natelyn segíteni (segít) mert (mert, mer) Nately apja (apa) gazdag és befolyásos ember volt (volt, van) aki kitűnő állása (állás) révén (révén, rév) segíthetett (segít) volna (van) Aarfyn a háború után

Tesztkörnyezet

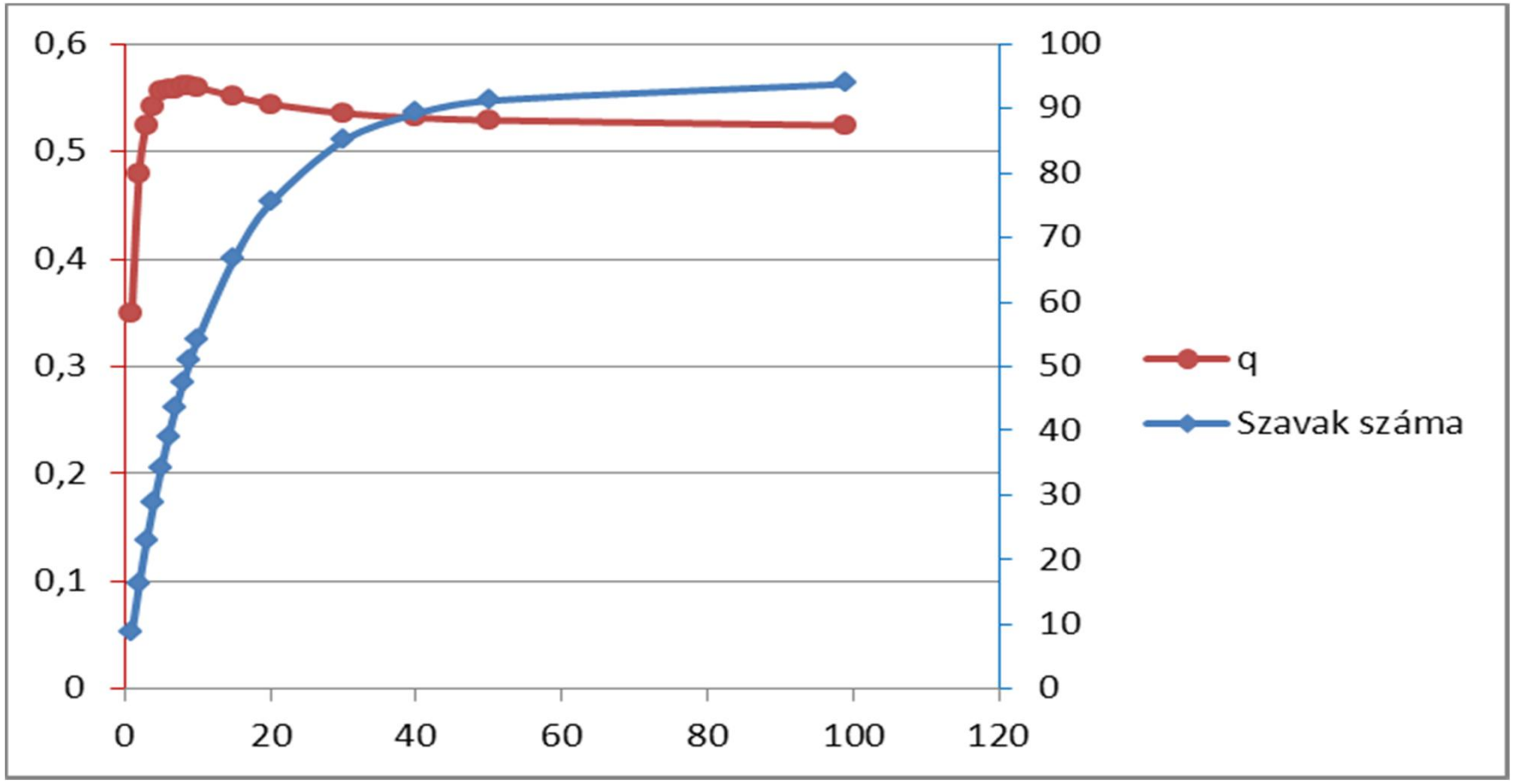
- Angol Wikipedia
  - 31GB XML
  - 3 800 000 szócikk
  - SZTAKI Desktop GRID
  - Lesz letölthető szöveges változat több nyelven
- Hunspell
- Google Translate
  - Csak teszteléshez
  - Találati arány egyezik a kézi fordításéval

Statisztikák



Mondat hossza és a megtalált fordítások aránya

Statisztikák



Szózsák mérete és találati arány viszonya

Demó – Wikipedia random article

Minitel

From Wikipedia, the free encyclopedia

The **Minitel** is a Videotex online service accessible through the telephone lines, and is considered one of the world's most successful pre-World Wide Web online services. It was launched in France in 1982 by the PTT (Poste, Téléphone et Télécommunications; divided since 1991 between France Télécom and La Poste). From its early days, users could make online purchases, make train reservations, check stock prices, search the telephone directory, have a mail box, and chat in a similar way to that now made possible by the Internet.

In February 2009, France Telecom indicates the Minitel network still has 10 million monthly connections, among which 1 million on the 3611 (directory). France Telecom is planning to retire the service on the 30th of June 2012.[1]

Minitel 1. Built 1982

Contents [hide]

Business model [edit]

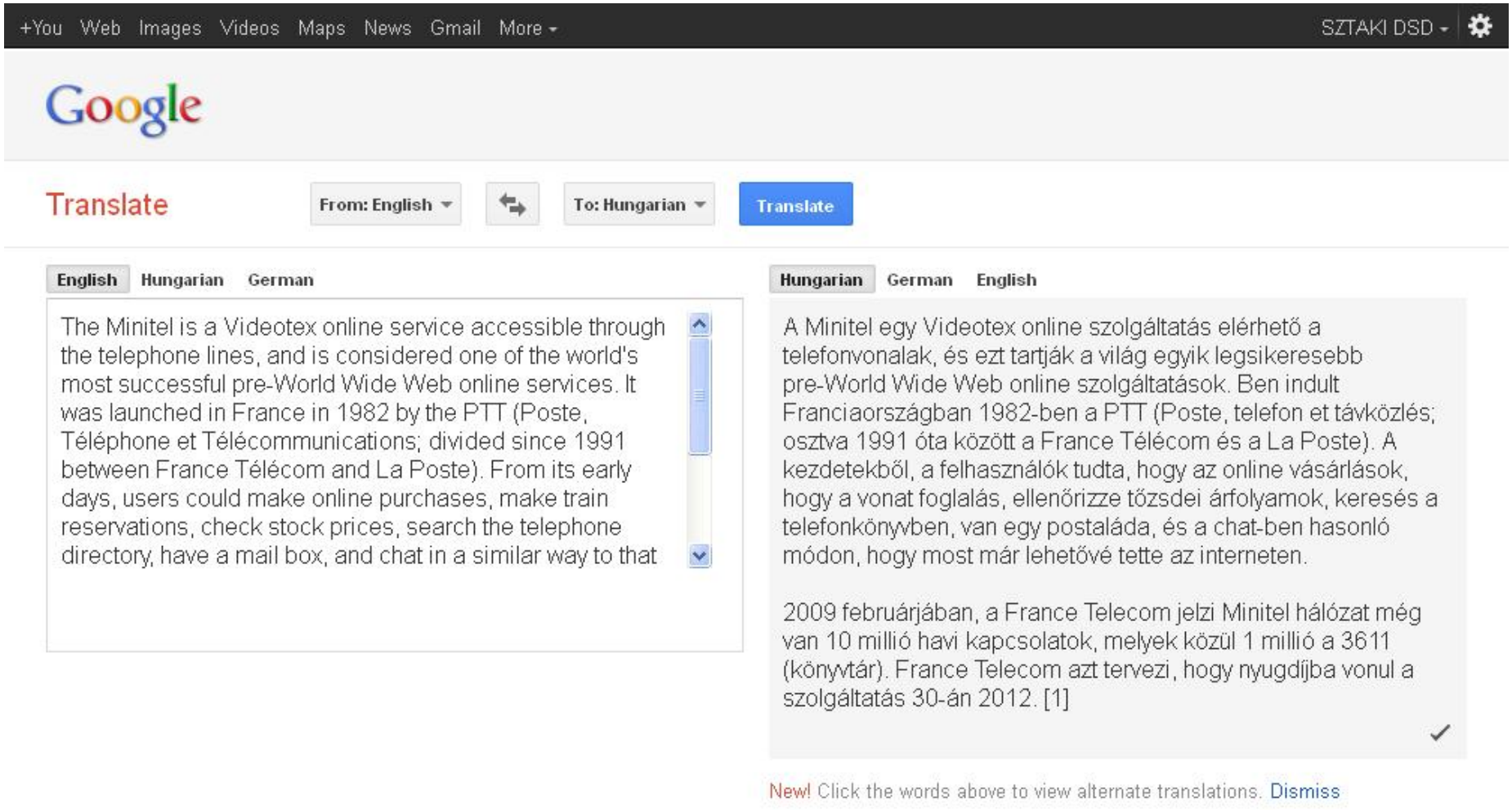
Millions of terminals were handed out free to telephone subscribers, resulting in a high penetration rate among businesses and the public. In exchange for the terminal, the possessors of Minitel would not be given free "white page" printed directories (alphabetical list of residents and firms), but only the yellow pages (classified commercial listings, with advertisements); the white pages were accessible for free on Minitel, and they could be searched by a reasonably intelligent search engine; much faster than flipping through a paper directory.

France Télécom estimates that almost 9 million terminals — including web-enabled personal computers (Windows, Mac OS, and Linux) — had access to the network at the end of 1999, and that it was used by 25 million people (of a total population of 60 million).

The Minitel allowed access to various categories of services:

- phone directory (free)
- mail-order retail companies
- airline or train ticket purchases
- information services
- databases
- message boards

Demó – Google translate



Demó – Keresés

Keresés az angol Wikipédiában

Magyar/angol/német szöveg:

A Minitel egy Videotex online szolgáltatás elérhető a telefonvonalak, és ezt tartják a világ egyik legsikeresebb pre-World Wide Web online szolgáltatások. Ben indult Franciaországban 1982-ben a PTT (Poste, telefon et távközlés; osztva 1991 óta között a France Télécom és a La Poste). A kezdetekből, a felhasználók tudta, hogy az online vásárlások, hogy a vonat foglalás, ellenőrizze tőzsdei árfolyamok, keresés a telefonkönyvben, van egy postaláda, és a chat-ben hasonló

Keres

A bemenet nyelve **magyar**.
6 mondatot találtam.

0. A Minitel egy Videotex online szolgáltatás elérhető a telefonvonalak, és ezt tartják a világ egyik legsikeresebb pre-World Wide Web online szolgáltatások..

The Minitel is a Videotex online service accessible through the telephone lines, and is considered one of the world's most successful pre-World Wide Web online services. (**15**, Minitel). .

Pre-World Wide Web online services are online service providers that predate the creation of the world wide web in the early 1990s and did not (initially) use TCP/IP. (**0**, Category:Pre-World Wide Web online services).

*Where to find Web Services on the Web: Investigating Web Services on the World Wide Web (2008) (**0**, Web service).

* Where to find Web Services on the Web: Investigating Web Services on the World Wide Web. (**0**, WS-Discovery). .

Online Matrimony Services in India This World Wide Web based service has millions of users Datta, Damayanti. (**0**, Matrimonial websites). . .

A list of the holdings is available on the World Wide Web. (**0**, Wells Coates)

Demó – Keresés

Missing translations: France, online, Poste, Telecom, Minitel, ben, chat, pre, án, Télécom, World, web, PTT, Wide, Ben

Találatok:

1. Minitel (59)
   The Minitel is a Videotex online service accessible through the telephone lines, and is considered one of the world's most successful pre-World Wide Web online services.
   - A Minitel egy Videotex online szolgáltatás elérhető a telefonvonalak, és ezt tartják a világ egyik legsikeresebb pre-World Wide Web online szolgáltatások. (15)

   It was launched in France in 1982 by the PTT (Poste, Téléphone et Télécommunications; divided since 1991 between France Télécom and La Poste).
   - Ben indult Franciaországban 1982-ben a PTT (Poste, telefon et távközlés; osztva 1991 óta között a France Télécom és a La Poste). (13)

   From its early days, users could make online purchases, make train reservations, check stock prices, search the telephone directory, and chat in a similar way to that now made possible by the Internet.
   - A kezdetekből, a felhasználók tudta, hogy az online vásárlások, hogy a vonat foglalás, ellenőrizze tőzsdei árfolyamok, keresés a telefonkönyvben, van egy postaláda, és a chat-ben hasonló módon, hogy most már lehetővé tette az interneten. (28)

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

Plágiumkeresés és dokumentumkezelés

Feltöltés | Dokumentumaim | Plágiumkereső | Futó keresések | Üzenetek

Kérem válasszon plágiumkeresési formát:

- Egynyelvű keresés - dokumentumok összehasonlítása:
  - minden felhasználó dokumentumaival
- Többnyelvű keresés (tesztüzem) - dokumentumok összehasonlítása:
  A dokumentum túl hosszú (2980 szó), a tesztüzemben maximum 2500 szavas dokumentumokkal lehet többnyelvű keresést indítani.

Plágiumkeresés indítása >>

Cím	Szerző	Feltöltés dátuma	
12 cikk	kopiwiki	2011.11.17.	Törlés

Amennyiben valamelyik dokumentumot mégse szeretné, hogy résztvegyen a keresésben, eltávilíthatja a "törlés" gombbal a listából.

Demó

From: KOPI
Date: October 7, 2011
Subject: 1 dokumentum összehasonlítása az angol Wikipédiával.

[Delete message]

6 hasonló mondatot talált a rendszer 2 Wikipédia cikkben:

1. Pete Seeger (6)
   Seeger was born in French Hospital, Midtown Manhattan.
   - Pete Seeger Manhattan közepén, a Midtown-nak is hívott városrész francia kórházában született. (4)

   His parents were living with his grandparents in Patterson, New York, from 1918 to 1920.
   - Szülei 1918 és 1920 között a nagyszül?kkel együtt a New York állambeli Pattersonban éltek. (11)

   His father, Charles Louis Seeger Jr., was a composer and pioneering ethnomusicologist investigating both American folk and non-Western music.
   - Apja, ifj. Charles Louis Seeger, zeneszerz? és zenetudós volt, aki az els?k között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekb?l fakadó zenét. (3)

   His mother, Constance de Clyver Edson, was a classical violinist and teacher.
   - Édesanyja, Constance de Clyver Edson klasszikus heged?m?vész és tanár volt. (9)

   His stepmother, Ruth Crawford Seeger, was one of the most significant female composers of the twentieth century.
   - Nevel?anyja, Ruth Crawford Seeger egyike volt a huszadik század legkiemelked?bb n?i zeneszerz?inek. (1)

   His half-sister, Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.
   - Féltestvére, Peggy Seeger, aki szintén ismert népzenei el?adó volt, hosszú évekig Ewan MacColl brit folkénekessel élt házasságban. (7)

2. Mike Seeger (2)
   His father, Charles Louis Seeger Jr., was a composer and pioneering ethnomusicologist, investigating both American folk and non-Western music.
   - Apja, ifj. Charles Louis Seeger, zeneszerz? és zenetudós volt, aki az els?k között vizsgálta mind az amerikai népzenét, mind a nem-európai gyökerekb?l fakadó zenét. (3)

   His sister Peggy Seeger, also a well-known folk performer, was married for many years to British folk singer Ewan MacColl.
   - Féltestvére, Peggy Seeger, aki szintén ismert népzenei el?adó volt, hosszú évekig Ewan MacColl brit folkénekessel élt házasságban. (9)

Demó

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

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 80% (4 mondat) egyezés	Wikipedia	2011.11.30.	Szerkeszt 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*