

2 CD**MOZI-JÁTÉK: X-MEN 2 PREMIER ELŐTT!**

1190 Ft • Előfizetéssel 833 Ft

MAGYAR

PC
MAGAZIN**5**II. évfolyam, 5. szám • 2003. május
<http://www.pcmagazin.hu>**Ingyenes
VLC**
DVD-lejátszó

Programok PDA-kra

Dev-C++ beta 5 -
C++ fejlesztőeszköz
Windowsra

Xfree86 4.3.0

Kaspersky Anti-Virus
Lite PC Magazin
Edition**Mozielőzetes:** Matrix
Reloaded, X-MEN 2**Fénypontjaink****A hónap nethelyei:**
www.monotone.hu •
www.geocaching.hu**DVD-je:**
xXx**játékai:**
Enclave • Praetorians**CD-ROM-ja:**
Tótfalusi István:
Operamesék**CeBIT 2003:** a mobilitás bővületében
Új csaták 3D-ben: **S3, Ati, Nvidia, SiS**
Stabil alapok: **Intel Granite Bay**
SMS-jövő: drágító Unicode?
Athlon XP **3000+** és Pentium 4 HT **3,06 GHz****Fotók
mega
pixelek****TESZTEK**

- 12** digitális fényképezőgép
- 12** 2.1-es hangrendszer
- 12** fekete-fehér SOHO lézernyomtató
- 7** nForce2 chipkészletes alaplap
- 12** belső modem



A PC Magazinban ebben a számban kettő CD található. Ha hiányoznának belőle, kérje az eladótól!

Elgépeled? Se baj!

Az elektronikus szótáraknak sok előnyük van papíros nagy testvéreikhez képest. Legfontosabb talán, hogy könnyen és gyorsan kereshetünk bennük – de csak akkor, ha a szót helyesen írjuk be a keresőbe. Ugyanakkor joggal várnánk el okosnak kikiáltott számítógépünktől, hogy akkor is megtalálja a keresett szót, ha elvettünk egy betűt.

Pataki Máté (Pataki.Mate@szta.hu)

Emmiképpen sem képtelen elvárás. Az MTA SZTAKI Elosztott Rendszerek Osztálya által hét éve üzemeltetett hálózati szótárba (<http://szotar.sztaki.hu>) bekerült egy ilyen szolgáltatás, amelynek vázlatos működési elvét szeretném bemutatni az érdeklődő olvasónak.

Az eljárás lényege: hasonlóságokat keresünk a szavak között. Amikor elkezdjük fejleszteni a hasonlóságkeresőt, elemeztük azokat a szavakat, amelyeket a szótárhasználók kerestek, de nem találtak meg a szótárban. Néha mulatságos dolgokat is olvashattunk, ebből egy kis gyűjtemény található keretes mellékletünkben. Most vegyük sorra a leggyakoribb gépelési hibákat!

- felcserélünk két betűt (ablak → albak);
- nem megfelelő betűt kettőzünk (kettő → keető);
- kihagyunk egy betűt (kocsi → koci);
- kétszer írunk be egy betűt (magazin → magaaizin);
- rossz billentyűt nyomunk (alma → alna);
- két billentyűt nyomunk le egyszerre (ajtó → asjto);

A gép összehasonlíthatná a beírt szót az adatbázisban lévő összes szóval – ez egyszerű megoldás volna, de lassú.

ABCDEF GHI JKLMNOPQRSTU VWXYZ
 NAGYJABOL → 11000010010101100000000010
 NAGGYABOL → 11000010000101100000000010
 0000000001000000000000000000

Hol a hiba?

Az angol–magyar szótárban például csaknem kétszázezer szópár van. Ha ehhez hozzávesszük, hogy napközben szinte másodpercenként jön olyan eredménytelen lekérdezés, amely után a hasonlóságkeresőt kell futtatni, láthatjuk: más utat kell keresnünk. Le kell szűkítenünk a felmerülő szavak listáját. A teljes folyamat a következő:

1. a gép eltávolítja az ékezeteket a szóból;
2. alkalmazza a kétlépcsős szűrőt (rögvést szólunk róla);
3. szavanként elvégzi az összehasonlítást (Oliver's method).

Szűrők

Szavak szűrésekor az első szempont a szó hossza lehet. A bemutatott gépelési hibák közül a betű kihagyása egy betűvel csökkenti, a betűkettőzés és két billentyű egyidejű lenyomása egy betűvel növeli a szó hosszúságát. (A többi esetben a

szavak hossza nem változik.) A két vagy annál több betűvel rövidebb, illetve hosszabb szavakat tehát nem is kell ellenőrizni.

A módszer főként hosszú, illetve nagyon rövid szavak esetén válik be. De például az angol szavak jelentős része négy, öt, hat betűből áll – ilyen szavak esetén még mindig sok szó akad fenn a rostán.

A szűrés második szempontja a betűk előfordulása. Ez sem változik jelentősen elgépeléskor. Képzeljük el az angol ábécét egy bináris számként!

A szóban előforduló betűk helyére 1-et írunk, az ábécé többi betűjét – amelyek nincsenek a szóban – 0-val jelöljük. Így minden szóhoz hozzárendelhetünk egy számot. Előtte a magyar nyelv ékezetes betűit a nekik megfelelő angol betűvé alakítjuk át.

Persze, ha két szóban azonos betűk szerepelnek, még nem jelenti azt, hogy

Gyakori hibák okozta eltérések

Hiba	Szóhossz eltérése	Eltérés a felhasznált betűkben
betűcsere	ugyanannyi betű más sorrendben (0)	ugyanazon betűk más sorrendben (0)
rossz betű kettőzése	az egyik betű eggyel többször, a másik eggyel kevesebbszer szerepel (0)	felhasznált betűknél nem számít a kettőzés (csak azt jelöljük, hogy van-e benne olyan betű) (0)
betű kihagyása	eggyel rövidebb lesz (1)	ha olyan betű, ami szerepel máshol a szóban, nem látunk különbséget; különben eggyel eltér (0-1)
betűkettőzés	eggyel hosszabb lesz (1)	betűkettőzésre nem érzékeny (0)
rossz billentyű lenyomása	ha egy betű helyett másik szerepel, az nincs hatással a szó hosszára (0)	ha a rossz és az oda illő betű is szerepel másutt a szóban, nem tapasztalunk különbséget; ha egyik sincs benne, akkor két helyen is eltér a helyes szótól (0-2)
két billentyű lenyomása	eggyel nő a szóhossz (1)	ha a véletlenül lenyomott billentyű betűje szerepel már a szóban, nincs változás; egyébként eggyel eltér (0-1)

Nem talált szavak, kifejezések

Nem talált szavak:

bocsájtja
czukraszda
dohny
elejtem
elelnjegyz
feldolgox
felszálltom
gyoggytorna
hatátalanít
himusz
hivalkods
igénybevestz
információhöz
kapcsolsi rajz
katólikus
kivégy

kötelezttség
marihuna
megfolt
megölele
mellyek
millárd
msogató
msoogat
nagyából
nyífe
ötvenkilenc
papagaly
papírsbeknedo
partfis
pirszing
portörölgetni
pszichologia

sanjos
szembesz
veyetés
welkome

Nem talált kifejezések:

augusztus végére eso
Budapestnek rengeteg
lakoso van
és mi van akkor
hatszög alakú kúp
Hová mentél
tegnap?
Kérem fájadjon bejeb
szabadon választás
teljesség iglyne nélkül

hasonlítanak egymásra. Fordítva viszont már igaz az állítás: két szó nem hasonlíthat egymásra, ha az őket helyettesítő számok több mint két helyértéken eltérnek. Rossz billentyű lenyomása esetén is maximum két betűben különbözik a két szó, például az „alna” szóban van n, míg az „alma” szóban nincs, helyette m olvasható.

Azt, hogy a fentebb említett hibák milyen eltéréseket okoznak a szavak hosszában, illetve az előforduló betűk körében, táblázatban foglaltuk össze. Két körben szűkítettük a szavak halmazát – így már kezelhető mennyiségű szót kapunk. Kezdődhet a szóról szóra történő hasonlóságvizsgálat. Ez már figyelembe veszi a betűk helyét. Például a budapesti és a

tepsidbe szavak hossza csak eggyel tér el, a bennük használt betűk is kettő híján azonosak – tehát átjutottak az első két szűrőn, pedig nem valószínű, hogy felcseréljük őket beírásakor. A részletes hasonlóságvizsgálat tehát elkerülhetetlen

Halmazott hibák

Felmerül a kérdés: mi történik, ha két hiba is van a beírt szóban. Ilyenkor előfordulhat, hogy az algoritmus nem találja a megfelelő szót, de szerencsére nagyon ritkák a kettős hibák. Az elírt szavakat böngészve kevés ilyen találtunk. Ennek az lehet az oka, hogy már gépeléskor feltűnő az ekkora eltérés.

Tapasztalat

Igen jó eredményeket sikerült elérni a fentebb vázolt algoritmussal, átlagban a szavak 80%-ához talál hasonlókat, ami nagyban növeli a szótár használhatóságát. Az utóbbi hónapokban szerencsére jelentősen nőtt a szotar.sztaki.hu forgalma, ezért elkerülhetetlenné vált az átköltöztetése egy új szerverre. Ez március közepén meg is történt, így a szótár újult erővel várja a szavakat kereső látogatókat.

H

a jót könnyű megszokni

professzionális adathordozók

Memorex

No.1
BRAND IN DIGITAL MEDIA WORLDWIDE

www.memorexlive.com

CD-R/RW
700MB
52X

CD-R
700MB
52X

8 cm-es CD-R
210MB

PÖCKET CD-R

DVD+R/RW
4.7GB
120MIN
16X

DVD+R
4.7GB
120MIN
16X

Memória kártyák
64MB
128MB
256MB

DVD-R/RW
4.7GB
120MIN
16X

DVD-R
4.7GB
120MIN
16X