Domáce zadanie na prvé zápočtové cvičenie

April 10, 2012

1 Zadanie

Napíšte bashovský skript `search.sh`, ktorý pošle svoje parametre vyhľadávaciemu servru `http://duckduckgo.com` a to, čo mu vráti prefiltruje a vypíše na svoj štandartný výstup tak, aby vo výstupe boli iba URL relevantných stránok a nič iné. Poradie vypísaných URL je irelevantné, ale nesmú sa opakovať.

2 Príklady vstupu a výstupu

```
$ ./search.sh slovak republic
http://data.worldbank.org/country/slovak-republic
http://fulbright.state.gov/participating-countries/europe-and-eurasia/slovak-republic.html
http://slovakrepublic.com/
http://slovensko.com/
http://topics.bloomberg.com/slovak-republic/
http://travel.state.gov/travel/cis_pa_tw/cis/cis_1019.html
http://www.amcham.sk/home
http://www.ebrd.com/pages/country/slovakrepublic.shtml
http://www.embassy.org/embassies/sk.html
http://www.indprop.gov.sk/?introduction
http://www.manufuture.org/manufacturing/?p=412
http://www.myspace.com/slovakrepublic#!
http://www.nationsonline.org/oneworld/slovakia.htm
http://www.oecd.org/country/0,3377,en_33873108_33873781_1_1_1_1_1,00.html
http://www.raileurope.com/europe-travel-guide/slovak-republic/index.html
http://www.slovakia.org/
http://www.slovak-republic.org/
http://www.state.gov/j/drl/rls/hrrpt/2001/eur/8338.htm
http://www.thefreedictionary.com/Slovak+Republic
http://www.un.org/esa/agenda21/natlinfo/countr/slovakia/natur.htm
```

```
$ ./search.sh slovak university technology
http://academic.research.microsoft.com/Organization/15762/slovak-university-of-technology
http://bratislavahotels.travelslovakia.sk/sites/slovak-university-of-technology-bratislava
http://stuba.academia.edu/
http://www.centrope-tt.info/slovak-university-of-technology-en
http://www.dievrinsis.gr/el/technical-study/98-slovakia-bratislava-slovak-university-of-technology-faculty-of-civil-engineering-civil-engineering
http://www.evri.com/organization/slovak-university-of-technology-0xbd552
http://www.gradschools.com/program-details/slovak-university-of-technology/graduate-course-in-information-technology-225972_1
http://www.innovationeu.org/news/innovation-eu-vol1-1/0062-slovak-university-of-technology.html
http://www.internationalgraduate.net/adverts/slovak-university-of-technology-in-bratislava.htm
http://www.ist-world.org/OrgUnitDetails.aspx?OrgUnitId=193c00e787264a15a3412467b4fd2e90
http://www.linkedin.com/company/slovak-university-of-technology
http://www.mastersportal.eu/students/browse/university/1361/slovak-university-of-technology-in-bratislava.html
http://www.minaretltd.com/slovak_university_of_technology.php
http://www.mladiinfo.com/2010/02/09/slovak-university-of-technology-bratislava/
http://www.mtf.stuba.sk/docs/doc/public_relations/MTF_publikacia_web.pdf
http://www.myuniversity-project.eu/index.php?option=com_content&amp;view=article&amp;id=37&amp;Itemid=71&amp;lang=en
http://www.quora.com/Slovak-University-of-Technology
http://www.slovakia.culturalprofiles.net/?id=4105
http://www.stuba.sk/new/generate_page.php?page_id=132
http://www.4icu.org/reviews/4205.html
```

3 Pomôcky, poznámky atď

- Nemám nič proti Google, ale jeho výstup je dnes nepoužiteľný na podobné účely; priveľa JavaScriptu.
- Pre sťahovanie vhodné použiť program `curl`. Ak ho nemáte nainštalovaný, nainštalujte si ho.
- Je treba sa tváriť ako prehliadač `lynx`. Povedzme v prvom príklade takto:

```
curl -A Lynx http://duckduckgo.com/?q=slovak+republic
```

- Je dobré predspracovať html tak, aby na každom riadku bol práve jeden html element. Ja som to spravil tak, že som v rúrovej sekvencii najprv zmazal všetky newline

```
tr -d '\n'
```

a potom všetky < nahradil *newline*:

```
tr '<' '\n'
```

To umožní pohodlne použiť `grep` `sed` atď.

- `curl` vypisuje kdečo zbytočné na stderr, vo finálnej verzii ho umlčte.