Domáce zadanie na prvé zápočtové cvičenie

22. októbra 2024

1 Zadanie

Napíšte bashovský skript `search.sh`, ktorý pošle svoje parametre vyhľadávaciemu servru `http://www.bing.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 môže byť hocijaké, ale nesmú sa opakovať. Na výstupe by nemali byť (spätné) odkazy na stránky Microsoftu, Bing, Creative Commons – tie tam sú vždy, a teda sú irelevantné.

2 Príklady

Príkaz

```
./search.sh slovak republic
```

Výstup

```
https://en.wikipedia.org/wiki/Slovakia
https://en.wikipedia.org/wiki/Slovakia#Culture
https://en.wikipedia.org/wiki/Slovakia#Demographics
https://en.wikipedia.org/wiki/Slovakia#Economy
https://en.wikipedia.org/wiki/Slovakia#Etymology
https://en.wikipedia.org/wiki/Slovakia#Geography
https://en.wikipedia.org/wiki/Slovakia#Government_and_politics
https://en.wikipedia.org/wiki/Slovakia#History
https://en.wikipedia.org/wiki/Slovak_Republic_(1939%E2%80%931945)
https://european-union.europa.eu/principles-countries-history/country-profiles/slovakia_en
https://linguaforum.sk/slovakia-alebo-the-slovak-republic/
http://slovakia.travel/en/25-years-of-the-slovak-republic
https://sk.wikipedia.org
https://sk.wikipedia.org/wiki/Slovensko
https://twitter.com/Rep_Slovakia
https://www.britannica.com/place/Slovakia
https://www.gettyimages.com/
https://www.nsud.sk/the-supreme-court-of-the-slovak-republic/
https://www.oecd.org/slovakia/
https://www.officeholidays.com/holidays/slovakia/day-of-the-establishment-of-the-slovak-republic
https://www.opengovpartnership.org/members/slovak-republic/
https://www.slovakia.com/
https://www.teachaway.com/teach-in-slovak-republic
https://www.vlada.gov.sk/government-of-the-slovak-republic/
```

Príkaz

```
./search.sh slovak university technology
```

Výstup

```
https://en.wikipedia.org/wiki/Slovak_University_of_Technology_in_Bratislava
https://interavers.com/en/slovak-university-of-technology-in-bratislava/
https://msmstudy.sk/en/universities/stu-bratislava/
https://sk.wikipedia.org
https://sk.wikipedia.org/wiki/Slovenská_technická_univerzita_v_Bratislave
https://spectator.sme.sk/c/20044411/stu-recognised-in-global-university-ranking.html
https://spectator.sme.sk/c/20844493/three-slovak-universities-ranked-among-1000-best.html
https://spectator.sme.sk/c/22828897/slovak-university-of-technology-in-bratislava.html
https://www.dreamstime.com/slovak-university-technology-bratislava-cars-parked-front-slovak-university-technology-population-image107919357
https://www.facebook.com/univerzita/
https://www.fchpt.stuba.sk/english.html?page_id=782
https://www.gotouniversity.com/university/slovak-university-of-technology-bratislava/tuition-fee
https://www.infoma.sk/en/company-adress-activity.php?firma=73722
https://www.instagram.com/stubratislava/
https://www.stuba.sk/
https://www.stuba.sk/english/degree-students/legislation.html?page_id=2021
https://www.stuba.sk/english/ects/ects-information-package/information-on-degree-programmes/all-programmes.html?page_id=5552
https://www.stuba.sk/english.html?page_id=132
https://www.stuba.sk/english/university/faculties.html?page_id=2989
https://www.topuniversities.com/universities/slovak-university-technology-bratislava
https://www.univerna.com/program-detail/slovak-university-of-technology/chemical-engineering/351
https://www.youtube.com/user/STUvBratislave
https://www.youtube.com/watch?v=pDJsUjoMebQ
https://www.youtube.com/watch?v=-sFIDfCDk9M
https://www.youtube.com/watch?v=Y5OZT5YFENA
https://www.youtube.com/watch?v=5Z-UyH8tKSw
```

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 -s -L -A Lynx https://www.bing.com/search?q=slovak+republic
```

  viď `man curl` pre význam prepínačov.

- 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 pre vyfiltrovanie linkov.

- To, že sa URL nesmie opakovať zabezpečíte pomocou idiómu

```
... | sort | uniq
```

- Irelevantné odkazy zmažete pomocou `grep -v`.