Webscraping

Než začneme webscraping …

- Potřebujeme znát
- Strukturu url linků
- Strukturu webových stránek

URL

- Označení místa v internetu, kde se stránka nachází
- Is.muni.cz
- Is – server
- Muni – doména
- Cz doména vyššího řádu

- Pro scraping nutný kompletní link
- https://is.muni.cz/...

- Na začátku HTTP nebo HTTPS

Linky

- Pro webscraping potřebujeme získat linky
  - Stáhnutí z nějaké stránky
  - Generovaní v R

Nezbytné základy

- Webové stránky jsou psány v html
  - Hypertext markup language
- Software „vidí“ stránku jinak než člověk
- Obsah je formátován pomocí značek (tagů)
- Každá tag má danou funkci
- Tagy vytvářejí hierarchii
- Tagy mohou být doplněny o atributy
- Stránky mohou „naplňeny“ jinak pragramovaným obsahem (java, xml, …)

Html

- Struktura tagu
- Začátek obsah konec
- <nejakytag> nejaky obsah</nejakytag>
- Existuje několik tagů, mezi které není vložen obsah </br>

Struktura stránek

- Celá stránka je vložena mezi <html> a </ html>
- Na úvod stránky je obvykle umístěna hlavička <head> a </ head >
  - Nezobrazuje se návštěvníkovi stránky, ale obsahuje důležité informace o stránce
  - Obvykle obsahuje odkaz na formátování (kaskádový styl)
- Samotný obsah stránky vnořen do <body> </body>
- Horní část obvykle v <header></header>
- Nadpisy určeny pomocí <h1></h1>
  - Číslo určuje úroveň
- Odstavce

Odkazy na další stránky

- text, na který se klikne
- Odkazy mimo stránku bývají vždycky kompletní
- Odkazy v rámci stránky jsou občas kratší a neobsahují kořenovou adresu

tabulka

- Tabulka uvozena <table></table>
- Řádek začíná <tr> a končí </tr>
- buňka začíná <td> a končí </td>

- <table>
- <tr> <td> a </td> <td> b </td> </tr>
- <tr> <td> 1 </td> <td> 2 </td> </tr>
- </table>

a	b
1	2

seznam

- Začíná <ul>, končí
- Každá položka
- Kombinacemi lze vytvářet různé úrovně

„bezvýznamné“ tagy

- Div
- Span

- Důležité pro formátování a orientaci ve stránce
- Div formátuje blok (odstavec), span nějakou část v rámci bloku

Problémy

- Data mohou být „zabalena“ v nějaké funkci
- Stránka jen zprostředkuje údaje z databáze, ale k databazi se nelze dostat
- Rozbalovací menu
- Definování požadavků ve formuláři

- V těchto případech obtížné stahování