IB111 – cv. 12
Práce s textem, daty

Miroslav Kadlec

Obsah

- Regulární výrazy
  - Syntaxe
  - Příklady
- Práce se soubory
  - Funkce open()
  - Objekt file
- Komplexnější příklad

Regulární výrazy

- Způsob, jak popsat množinu řetězců, jejichž některé části mají nějaké vlastnosti (výskyt dvojtečky, číslice, ...) a jiné jsou naopak libovolné.
  - Definujeme "pattern" (regulární výraz) popisující části řetězce s pomocí zástupných znaků
    - Některé znaky jsou tedy brány jako řídicí (tečka, závorka) a pokud je chceme použít v jejich normálním významu, vložíme před ni "\" (to platí i pro "\").
    - "\" v kombinaci s nějakým písmenem definuje třídu znaků (\d značí libovolnou číslici)
  - Regulární výraz pak můžeme použít k
    - Vyhledání/nahrazení všech výskytů nějakého podřetězce
    - Ověření, že daný řetězec vyhovuje danému schématu
      - URL, IP adresa, ...

Regulární výrazy

- Základní konstrukce
  - `.` libovolný znak
  - `*` libovolný počet opakování předchozího znaku
    - `+` libovolný nenulový počet opakování předchozího znaku
  - `[abc]` jeden ze znaků a, b, c
    - `[^abc]` cokoliv mimo znaky v závorce
    - `[a-h]` znaky v daném rozpětí
  - `(abc)` řetězec beroucí se jako celek (pro opakování atd.)
- Třídy znaků
  - **\d** (číslice), **\D** (ne-číslice) **\w** (alfanumerický znak), **\W** (nealfanumerický znak), **\s** (netisknutelný znak) **\S** (...)
- K vyzkoušení: pythex.org

Práce se soubory

- Vestavěnáfunkce open()
  - Slouží k otevření souboru
    - 1. parametr: jméno souboru
    - 2. parametr: režim otevření (zadán řetězcem)
      - "w" zápis do souboru (smazání obsahu, pokud existuje)
      - "a" zápis do souboru (ponechání obsahu, pokud existuje)
      - "r" čtení souboru (nic se nemaže)
      - "b" binární soubor
      - "w+" čtení a zápis (soubor je vytvořen pokud neexistuje)
      - "r+" čtení a zápis
    - 3. (volitelný) parametr: bufferování (default nám bude stačit)
  - Funce open() vrací objekt file, který slouží k přísupu k obsahu souboru.

Práce se soubory

- Objekt file má různé metody
  - .close() tím zavřeme soubor, až ho nebudeme potřebovat
  - .flush() vynutí uložení změn do souboru
  - .read(pocetBytu)
  - .readline(maxPocetBytu)
  - **.readlines(maxPocetBytu)**
  - .write("string"), .writelines(["line1", "line2"])
  - .tell() zjistí aktuální pozici v souboru
  - .seek(pozice, odkud) nastaví pozici v souboru (od zacatku, konce, aktualni pozice)

Komplexnější příklad

- Statistiky textu
  - Nejčastější slova
  - Délky vět
    - Průměrná délka věty
  - Korelace písmen
    - Počty případů, kdy za písmenem X následuje písmeno Y
    - Udržovat strukturu
  - Naivní imitace textu
    - Generovat text o podobné korelaci písmen
    - Začlenit i délku vět