Cvičení 12, skupiny IB113/06,07

Slovníky, zpracování textu ze souboru, regulární výrazy

Domácí úloha č. 6 (poslední :-))

- na 40 bodů, má 3 části (práce s obrázky bude náplní posledního cvičení)
- zadání v ISu: https://is.muni.cz/auth/el/fi/podzim2023/IB113/um/skupiny_06_07/cviceni_12/DomaciUkol6.pdf
- **odevzdání do 20.12.2023 23:55 (středeční půlnoc)**

Slovníky + zpracování textových souborů

Slovníky, řazení, filtrování – užitečné funkce – viz. cvičení před vnitrosemestrálkou: https://is.muni.cz/auth/el/fi/podzim2023/IB113/um/skupiny_06_07/cviceni_10/IB113-cvika10.pdf

Čtení ze souboru, zpracování textového souboru:

```
f = open(filename, "r")
for line in f:
    print(line)
f.close()
```

Módy otevření souboru: `"r"` čtení, `"w"` zápis (přepsat), `"a"` zápis (na konec) – zápis pomocí `f.write(text)`, po práci se souborem je potřeba jej zase zavřít `f.close()`

Textové soubory – stáhněte si do adresáře, kde máte své zdrojáky v Pythonu

- z minula: https://www.umimeinformatiku.cz/files/**alice.txt**, https://www.fi.muni.cz/~xpelanek/IB113/sbirka/_downloads/1dfa0d128643dc218db043d0c142cf42/**sherlock-holmes.txt**
- https://www.umimeinformatiku.cz/files/**tabulky-jmena.csv** – textový soubor ve formátu .csv – řádky souboru reprezentují řádky tabulky, hodnoty na řádcích jsou odděleny čárkou ,
- https://is.muni.cz/auth/el/fi/podzim2023/IB113/um/skupiny_06_07/cviceni_12/**slovnik.txt** textový soubor, řádek=slovo, budeme používat na úkoly s regulárními výrazy

Úkoly na slovníky + zpracování textových souborů:

1. Soubor *alice.txt* (případně *sherlock.txt*) - naprogramujte:
   - výpis 10 nejfrekventovanějších dlouhých slov ze souboru (délka slova větší než 4 písmena)
   - nalezení nejčastějších bigramů (dvojic písmen po sobě), výpis frekvence *k* nejčastějších
2. Soubor *tabulky-jmena.csv*:
   - načtěte si data do vhodné datové struktury a zjistěte:
     - jaké bylo nejčastější jméno dávané potomkům narozeným v roce 1960?
     - ve kterém roce dosáhlo své maximální popularity jméno Marek?
     - jaké je celkově nejčastější jméno za celou dobu zaznamenanou v souboru?

Regulární výrazy

Slidy z přednášky: http://www.fi.muni.cz/~xpelanek/IB113/slidy/regexp-texty.pdf
Cheatsheet: https://www.activestate.com/resources/datasheets/python-regex-cheatsheet/

`import re` – modul pro práci s regexpy

`re.match(regexp, my_string)` – odpovídá (začátek řetězce) `my_string` regexpu `regexp`?
`re.search(regexp, my_string)` – obsahuje řetězec někde kousek odpovídající regexpu?
`re.match, re.search` vrací None (žádná shoda) nebo Match objekt, ze kterého můžeme zjistit:
`.span()` – vrátí počáteční a koncovou pozici nalezené shody
`.string` – vrátí původní řetězec `my_string`
`.group()` – vrátí podřetězec, kde nastala shoda

`re.sub(regexp, replacement, my_string)` – nahraď v řetězci všechny shody *regexp*u řetězcem *replacement* (což může být i regexp používající kousky nalezené shody)
`re.split(regexp, my_string)` – rozdělí řetězec do seznamu slov podle shod s *regexp*em
"raw string" (nedochází k interpretaci speciálních znaků) – `r'výraz'`

Úkoly na regulární výrazy:

1. Vyzkoušejte si, co dělají regulární výrazy:
   ```
   [a-z]+@[a-z]+\.cz
   kocka|pes
   ^[Pp]rase$
   \d[A-Z]\d \d\d\d\d
   ^\s*Nadpis
   ^a.+a$
   \d{3}\s?\d{3}\s?\d{3}
   [a-z]+@[a-z]+\.cz
   ^To:\s*(fi|kit)(-int)?@fi\.muni\.cz
   ```
   Pomocné vyhodnocovátko: https://pythex.org/

2. Se souborem *slovnik.txt* a zdrojákem *fileio.py* – v dlouhé poznámce na konci zdrojáku jsou různé vlastnosti slov ("obsahuje oo", "začíná a končí na a" atd.), které zkuste zapsat regexpem a modifikujte zdroják (tj. zavolejte funkci `search_in_file` s tím správným parametrem *pattern*) tak, aby na výstup byla vypsána všechna slova ze souboru *slovnik.txt*, která splňují danou vlastnost.

Další úkoly

- Cokoliv, co ještě nemáte ze sbírky: https://www.fi.muni.cz/~xpelanek/IB113/sbirka/11-text.html
- Pro rychlé – zkuste si zpracovat textový zdroják stránky z wikipedie a (heuristicky pomocí regulárních výrazů) najít jména co nejvíce panovníků:

https://cs.wikipedia.org/wiki/Seznam_p%C5%99edstavitel%C5%AF_%C4%8Desk%C3%A9ho_st%C3%A1tu