České vysoké učení technické v Praze FIT

Programování v Pythonu

Jiří Znamenáček

Příprava studijního programu Informatika je podporována projektem financovaným z Evropského sociálního fondu a rozpočtu hlavního města Prahy.

Praha & EU: Investujeme do vaší budoucnosti

Množiny

1. Zjistěte, jaké různé *znaky* se vyskytují v kratším textu. Přitom postupně:

- uvažujte rozdíly mezi malými a velkými písmeny
- **ne**uvažujte rozdíly mezi malými a velkými písmeny

[-] nápověda

```
set( list() )
```

[-] řešení (*typy/mnoziny/01.py*)

```
text = ''
with open('example.2.txt', mode='r', encoding='utf-8') as f:
    text = f.read()

print( set(text) )
print( set(text.lower()) )
```

2. Zopakujte si předchozí analýzu na větším textu, ale tentokrát hledejte různá *slova*. Přitom postupně:

- uvažujte rozdíly mezi malými a velkými písmeny
- **ne**uvažujte rozdíly mezi malými a velkými písmeny (tj. *Auto* = *auto* apod.)
- odstraňte interpunkci (tj. *auto,* = *auto* apod.)

[-] nápověda

```
set( list() )
import string; string.punctuation
```

[-] řešení (*typy/mnoziny/02.py*)

```
text = ''
with open('example.2.txt', mode='r', encoding='utf-8') as f:
    text = f.read()

# a) A != a
text_po_slovech = text.split()
print( set(text_po_slovech), len( text_po_slovech ) )

# b) A == a
text_po_slovech_v_malych_pismenech = text.lower().split()
print( set( text_po_slovech_v_malych_pismenech ), len(
text_po_slovech_v_malych_pismenech ) )

# c) pryč s interpunkcí
import string
text2 = [ slovo.strip(string.punctuation) for slovo in
text_po_slovech_v_malych_pismenech ]
print( set(text2), len(text2) )
```

3. Proveďte předchozí analýzu pro dva různé texty a porovnejte výsledky následujícím způsobem:

- zjistěte, která slova se vyskytují v obou textech
- vytvořte množinu všech slov, která je možno v uvedených textech najít
- najděte slova, která obsahuje pouze první text a která obsahuje pouze druhý text
- vytvořte množinu slov, která oba texty nemají společná

[-] nápověda

průnik, sjednocení, rozdíl a rozdíl, symetrický rozdíl

[-] řešení (*typy/mnoziny/03.py*)

```
import string

text1, text2 = '', ''
with open('example.1.txt', mode='r', encoding='utf-8') as f:
    text1 = f.read()
with open('leacock-abc.txt', mode='r', encoding='utf-8') as
f:
    text2 = f.read()

# množiny slov bez interpunkce
text1 = { slovo.strip(string.punctuation) for slovo in
text1.lower().split() }
text2 = { slovo.strip(string.punctuation) for slovo in
text2.lower().split() }

# ANALÝZA
# a) slova v obou textech – průnik
print('text1 & text2 = ', text1 & text2)
# b) všechna možná slova v obou textech – sjednocení
print('text1 | text2 = ', text1 | text2)
# c) slova pouze v prvním textu a slova pouze v druhém textu
– rozdíl
print('text1 - text2 = ', text1 - text2)
print('text2 - text1 = ', text2 - text1)
# d) slova buď v jednom nebo pouze v druhém textu –
symetrický rozdíl
print('text1 ^ text2 = ', text1 ^ text2)
```

4. Ještě jednou zanalyzujte text na výskyt slov, tentokrát se ale ptejte po tom, jak různě dlouhá slova se v textu vyskytují.

[-] řešení (*typy/mnoziny/04.py*)

```
import string

text = ''
with open('example.1.txt', mode='r', encoding='utf-8') as f:
    text = f.read()

# množina slov bez interpunkce
text = { slovo.strip(string.punctuation) for slovo in
text.lower().split() }

# délky slov
delky = { len(slovo) for slovo in text }
print(delky)
```