Exercise Sheet 10 Solutions – Regular Expressions

Sebastian Höffner Aline Vilks

Deadline: Mon, 12 June 2017 08:00 +0200

Exercise 1: Requests

File: `bookreview.py`

```
import re
import os
import sys
import time

import parse
import requests


def write_file(content, filename):
    # This assures that the 'books' directory exists, not part of the sheet.
    try:
        os.makedirs('books', 0o755)
    except OSError:
        pass

    with open(os.path.join('books', filename), 'w') as file_handle:
        file_handle.write(content)


def download(id):
    # Bonus task: measure time. Start time.
    download_time = time.time()

    res = requests.get('https://www.gutenberg.org/cache/epub/{id}/pg{id}.txt'
                       .format(id=id))

    # Bonus task: measure time. End time.
    download_time = time.time() - download_time
    return res.text, download_time
```

```
def save_book(text, author, title):
    write_file(text, '{}-{}.txt'.format(author, title))


def save_counts(words, counts, author, title):
    text = '\n'.join('{}, {}'.format(w, c) for w, c in zip(words, counts))
    write_file(text, '{}-{}-words.csv'.format(author, title))


def save_sentences(sentences, author, title):
    text = '\n\n'.join(sentences)
    write_file(text, '{}-{}-sentences.txt'.format(author, title))


def strip_book_meta(text, title):
    tpl = '*** {} OF THIS PROJECT GUTENBERG EBOOK {} ***'

    start = tpl.format('START', title.upper())
    end = tpl.format('END', title.upper())

    idx_start = text.index(start) + len(start)
    idx_end = text.index(end)

    book_content = text[idx_start:idx_end]

    return book_content


def get_words(filename='wordlist.txt'):
    with open(filename) as file_handle:
        return file_handle.read().splitlines()


def find_sentences(content):
    words = get_words()
    counts = []
    sentences = []

    for word in words:
        results = re.findall(r'[^.!?]*\b{}\b[^.!?]*[.!?]'.format(word),
                             content, re.DOTALL | re.IGNORECASE)
        if results:
            counts.append(len(results))
            sentences.append(results[0])
        else:
            counts.append(0)
            sentences.append('')

    return words, counts, sentences


def get_meta(text):
    pgline = text.splitlines()[0]
    result = parse.parse(
        '\ufeffThe Project Gutenberg EBook of {title}, by {author}',
        pgline)
    return result['author'], result['title']


def print_info(id, author, title, download_time, words, counts, sentences):
    print('Downloaded {}, {}: {}, in {:.3f} s.'.format(id, author, title,
                                                       download_time))
    print('\nThe word counts are:')
    for word, count in zip(words, counts):
        print('  {: <8}{: >4}'.format(word, count))

    print('\nSome example sentences:\n')
    count = 0
    for sentence in sentences:
        if sentence:
            count += 1
            print(sentence + '\n')
        if count >= 2:
            break


def main():
    # Bonus task: Read ID from command line arguments.
    pgid = sys.argv[1]
    full_text, download_time = download(pgid)

    author, title = get_meta(full_text)

    book_content = strip_book_meta(full_text, title)

    words, counts, sentences = find_sentences(book_content)

    save_book(full_text, author, title)
    save_counts(words, counts, author, title)
    save_sentences(sentences, author, title)

    print_info(pgid, author, title, download_time, words, counts, sentences)


if __name__ == '__main__':
    main()
```

Output:

```
Downloaded 1228, Charles Darwin: On the Origin of Species, in 2.009 s.

The word counts are:
  he       120
  she       18
  love       3
  live      29
  hate       0
  food      48
  body      34
  wise       0
  plant     59
  rich       5
  legend     0

Some example sentences:

 Last year he sent to
me a memoir on this subject, with a request that I would forward it
to Sir Charles Lyell, who sent it to the Linnean Society, and it is
published in the third volume of the Journal of that Society.

 She can act on every internal organ, on every shade of
constitutional difference, on the whole machinery of life.
```

File: `Charles Darwin-On the Origin of Species-words.csv`

```
he, 120
she, 18
love, 3
live, 29
hate, 0
food, 48
body, 34
wise, 0
plant, 59
rich, 5
```

legend, 0

File: Charles Darwin-On the Origin of Species-sentences.txt

Last year he sent to
me a memoir on this subject, with a request that I would forward it
to Sir Charles Lyell, who sent it to the Linnean Society, and it is
published in the third volume of the Journal of that Society.

She can act on every internal organ, on every shade of
constitutional difference, on the whole machinery of life.

It may be difficult, but we ought
to admire the savage instinctive hatred of the queen-bee, which urges
her instantly to destroy the young queens her daughters as soon as born,
or to perish herself in the combat; for undoubtedly this is for the
good of the community; and maternal love or maternal hatred, though
the latter fortunately is most rare, is all the same to the inexorable
principle of natural selection.

No one ought to feel surprise at much remaining as yet unexplained in
regard to the origin of species and varieties, if he makes due allowance
for our profound ignorance in regard to the mutual relations of all
the beings which live around us.

Naturalists
continually refer to external conditions, such as climate, food, etc.

If it could be shown that our
domestic varieties manifested a strong tendency to reversion,--that
is, to lose their acquired characters, whilst kept under unchanged
conditions, and whilst kept in a considerable body, so that free
intercrossing might check, by blending together, any slight deviations
of structure, in such case, I grant that we could deduce nothing from
domestic varieties in regard to species.

In the case of the
misseltoe, which draws its nourishment from certain trees, which has
seeds that must be transported by certain birds, and which has flowers
with separate sexes absolutely requiring the agency of certain insects
to bring pollen from one flower to the other, it is equally preposterous

to account for the structure of this parasite, with its relations to several distinct organic beings, by the effects of external conditions, or of habit, or of the volition of the plant itself.

It is not that these countries, so rich in species, do not by a strange chance possess the aboriginal stocks of any useful plants, but that the native plants have not been improved by continued selection up to a standard of perfection comparable with that given to the plants in countries anciently civilised.