Exercise Sheet 10 – Regular Expressions

Sebastian Höffner Aline Vilks

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

Submission

By the end of this sheet you will have a number of different files to submit. In Stud.IP you will have a directory for your own group, please upload them there. It is easier for you if you just archive (preferably zip) all files and upload your archive, but it is okay if you upload them one by one.

Packages

Attention

For this exercise sheet you will need two additional Python packages. Install them using:

```
pip install requests parse
```

If you have any troubles with installing them, let us know as soon as possible so we can resolve issues fast.

You can find the documentations here:

- requests: http://docs.python-requests.org
- parse: https://github.com/r1chardj0n3s/parse/blob/master/README.rst

Exercise 1: Requests

Go to Project Gutenberg[1] and search for some books. Pick a book! We picked Darwin (1859), which has ID 1228[2].

Write a script `bookreview.py` which, given an ID, performs the following tasks, which are explained in more detail below:

[1] https://www.gutenberg.org/
[2] https://www.gutenberg.org/ebooks/1228

1. Download the book and save it as `{author}-{title}.txt`. You can extract the required information from the first line of the downloaded file.
2. Preprocess the book by removing the preamble and license. Don't remove this information from the saved version.
3. Search for all sentences containing the words in the `wordlist.txt`. Store each first sentence in a file `{author}-{title}-sentences.txt`. Store the counts in a file called `{author}-{title}-words.csv`.

Running the program

To analyze Darwin (1859), we need to provide the ID 1228. You may pick another book if you like. Note that you may hard code the ID, but we recommend to use a variable – especially if you attempt the bonus tasks (see below).

Downloading the book

Use the `requests` library to download the book from Project Gutenberg. Find a way to create proper URLs to the text-only version of a Project Gutenberg book. When downloading the data, measure the time and present the user with some output about it.

You can extract the author and title information from the first line, in our book it is:

```
\ufeffThe Project Gutenberg EBook of On the Origin of Species, by
Charles Darwin
```

To extract the information, you can adapt this snippet:

```
import parse
pgline = ('\ufeffThe Project Gutenberg EBook of On the ' +
          'Origin of Species, by Charles Darwin')
result = parse.parse('\ufeffThe Project Gutenberg EBook of ' +
                     '{title}, by {author}', pgline)
title = result['title']
author = result['author']
print('{}: {}'.format(author, title))
```

Output:

```
Charles Darwin: On the Origin of Species
```

The weird `\ufeff` is the byte order mark, it tells programs how to read the data inside the file. It is counted as a single character, so you can either skip it (`pgline[1:]`) or parse around it like we do in the example. You can use `repr(...)` to make it visible. If you have troubles, try to leave it out, skip it, parse parts, ... You will be able to get this done.

Use the author and title information to write the contents you downloaded to a file `{author}-{title}.txt`.

Preprocessing the data

A Project Gutenberg file contains some data before and after the actual book contents. These contain meta data and license statements. "Clean" the data by removing everything before and including `*** START OF THIS PROJECT GUTENBERG EBOOK {} ***` and everything from `*** END OF THIS PROJECT GUTENBERG EBOOK {} ***`. Note that the placeholder is the title in uppercase letters, i.e. `title.upper()`. You can just use `.index` on a string to find a substrings occurrence. Use it to slice the string: `bookcontent[start:end]`.

Crawling the data

In the file `wordlist.txt` are some words. Load the list and search all sentences which contain the words, respectively.

Write one example sentence per word into the file `{author}-{title}-sentences.txt`. Write a csv file `{author}-{title}-words.csv` which contains the words and their respective counts.

Bonus tasks

1. Use `sys.argv` to read a custom book ID when running the program, e.g. `python bookreview.py 1228` would use the ID `1228` which can be found in the list `sys.argv` at position `1`.
2. Measure the download duration using `time`. You can measure time by taking difference between the start and end time of an action. `time.time()` gives you the current time in seconds since January 1st, 1970, 00:00:00 UTC.

Example output

Here is an example output for program call `python bookreview.py` (or, if you finished the bonus tasks, `python bookreview.py 1228`). Your output can differ.

Output:

```
Downloaded 1228, Charles Darwin: On the Origin of Species, in 7.118 s.

The word counts are:
  he        81
```

```
she       14
love       3
live      21
hate       0
food      21
body      16
wise       0
plant     33
rich       4
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.

 Though nature grants vast periods of time
for the work of natural selection, she does not grant an indefinite
period; for as all organic beings are striving, it may be said, to seize
on each place in the economy of nature, if any one species does
not become modified and improved in a corresponding degree with its
competitors, it will soon be exterminated.
```

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

```
he, 81
she, 14
love, 3
live, 21
hate, 0
food, 21
body, 16
wise, 0
plant, 33
rich, 4
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.

 Though nature grants vast periods of time
for the work of natural selection, she does not grant an indefinite
```

period; for as all organic beings are striving, it may be said, to seize on each place in the economy of nature, if any one species does not become modified and improved in a corresponding degree with its competitors, it will soon be exterminated.

If we admire the truly wonderful power of scent by which the males of many insects find their females, can we admire the production for this single purpose of thousands of drones, which are utterly useless to the community for any other end, and which are ultimately slaughtered by their industrious and sterile sisters? 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.

Nevertheless some slight amount of change may, I think, be attributed to the direct action of the conditions of life--as, in some cases, increased size from amount of food, colour from particular kinds of food and from light, and perhaps the thickness of fur from climate.

In the first place, I have collected so large a body of facts, showing, in accordance with the almost universal belief of breeders, that with animals and plants a cross between different varieties, or between individuals of the same variety but of another strain, gives vigour and fertility to the offspring; and on the other hand, that CLOSE interbreeding diminishes vigour and fertility; that these facts alone incline me to believe that it is a general law of nature (utterly ignorant though we be of the meaning of the law) that no organic being self-fertilises itself for an eternity of generations; but that a cross with another individual is occasionally--perhaps at very long intervals--indispensable.

```
 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.
```

References

Darwin, Charles. 1859. *On the Origin of Species*.