Counting words in Moby Dick

An exercise using multi-staged learning aids

Instruction for trainers

Summary

This exercise exemplifies how learning aid cards can be structured in such a way that they do not provide too many answers for lazy students. The exercise itself has been performed in practice without learning aid cards only – and no further material.

Learning Goals

- Use a dictionary data type for counting things.
- Use list data type to sort information.

Task

Write a program that counts how often each word occurs in the book "Moby Dick" by Herman Melville, and output the words, starting with the most frequent ones, and how often they occur.

Requirements

- basic knowledge of Python programming (www.python.org)
- basic operations on strings, dictionaries, and lists
- for loops
- reading text files

Material

- Herman Melville's "Moby Dick", available as a text file from http://www.gutenberg.org
- Python programming environment.

1) What’s the output like?

How should the output of the program look like? Write down a few sample lines of output .

2) Find a program structure

Which steps should the program execute, and in which order? Draw a small flowchart.

3) Finding the right data type

Which data type in Python is suited well to count things? Which operations on this data type will be necessary to

1) initialize the data type?

2) count a word?

4) Processing text data

Which functions can be used to

1) Read a text file?

2) Separate a string into words?

5) Sorting

Which data type in Python can be used to sort things?

How would you want to represent words and counts in this data structure?

6) Sorting by word counts, not words

How does Python sort integers, strings, tuples, and other lists?

7) Did it work?

Where would you expect words like ‘is’, ‘the’, ‘sea’, and ‘cerebellum’ to occur. Check whether the output of the program corresponds to your expectations.

Does ‘captain’ or ‘whale’ occur more often in the text?

8) Caveat

Special and uppercase characters may be a problem when separating words. Remove all special characters before starting counting.

How can this be done?

Program structure

- Read the file.
- Split it into words.
- Count each word.
- Sort the words by counts.
- Output the words and counts

Output example

```
2307   is
 228   through
   5   tobacco
```

Processing text data (reminder)

Reading a text file:

```
text = open(filename).read()
```

chopping up a string:

```
list = string.split()
```

Finding the right data type

Dictionaries can be used to count things.

```
counter = {}
counter.setdefault(‘fish’, 0)
counter[‘fish’] += 1
```

Sorting by word count, not words:

Try to sort on the command line these lists:

```
[ ( “aaa”, 100), ( “bbb”, 20) ]
```

and

```
[ ( 100, “aaa”), ( 20, “bbb”) ]
```

Sorting

In Python, lists can be sorted.

Lists can contain tuples, e.g.

```
my_list = [ (12, 34), (56, 78) ]
my_list.sort()
```

Caveat

Special characters can be removed by the str.replace() function – or more comfortably using the re module.

Did it work?

The first five places should be taken by of (6614), and (6433), a (4726), to (4625), and in (4173).

You have to check yourself whether ‘whale’ or ‘captain’ is first.