EPAM Python Software Engineer Training

Lesson 2: Python collections

Contents

1 Course

1.1 Collection types

In python there are four built-in collection types: *tuple*, *list*, *dict*, and *set* - mentioned in order of memory used (ascending). An ideology is that a *tuple* is an immutable array of data with an extremely fast access by index; a *list* is an ordered mutable set of data with equally fast index access to all elements; a *dict* is a hash-table dedicated to an extremely fast access by a hash value; and a *set* is an unordered set of data with no duplicates (a *set* uses either a *list* or a *dict* internally depending on its size).

Tuples are so fast because they are typically kept not in heap but in a stack, i.e. their elements behave just like local variables and in python it is normal to convert an array of local variables into a tuple and vise versa; in contrast, it makes no sense to have a tuple of 1000 elements as in this case all tuple benefits are lost.

In a *collections* module there are other useful primitive collection types, most important of which is a *deque* - a thread-safe ordered mutable collection with extremely fast addition/deletion operation at both ends. It is used in many builtin libraries demanding a queue-like functionality.

2 Tasks

2.1 Packing and unpacking

Write a function which returns a tuple of first 10 and last 10 characters of an input string. Use an output of that function to print the first 10, the last 10 characters of some string and their concatenation.

2.2 Performance

Using a *timeit* module compare a speed of `1 000 000` operations of creating a *tuple*, a *list*, and a *set* of all integer digits (numbers `0-9`).

2.3 Alphabet

Print a number of times that each English character is used in `alice.txt`.

> ***Hint***
>
> For this task and below a character case shall be ignored.

2.4 Words Count

Print a statistics about `alice.txt` like a *wc* Unix command (number of lines, words, and characters).

> ***Hint***
>
> For this task and below assume that an English word cannot be split over multiple lines using a hyphen; and that a hyphen is a normal separator between words (e.g. a phrase `in-line` consists on two words).

2.5 Unique Words

Print a number of unique English words in `alice.txt`.

2.6 Words Usage

Print the first `N` most used English words in `alice.txt`; `N` shall be a positive number taken from a command line.