Advanced Python: Homework set 6

2023/2024

You must complete one of these assignments for class. Each task is worth 5 points.

Problem 1

Program a function `crawl(start_page, distance, action)` that reads the page at `start_page`, calls an `action` function whose argument is the content of the page, and then executes this function for other pages linked to that page.

Since there may be many links, the search depth is limited by the `distance` parameter. Also, make sure not to process the same page twice.

Implement the solution in the form of an iterator that returns tuples of the form `(url, action_function_result)`

```
for url, wynik in crawl("http://www.ii.uni.wroc.pl", 2,
                        lambda tekst : 'Python' in tekst):
    print(f"{url}: {wynik}")
```

Demonstrate the use of the function to search for sentences containing the word Python on websites.

Problem 2

Program a website monitoring program that checks whether any website has changed its content. We assume that our program can monitor more than one website; we also assume that checking takes place from time to time (e.g. every 1 minute).

In this task, we assume that the page layout rarely changes, only individual elements of the layout change, so if the program detects a change, it should only return what has changed.

Problem 3

Write your own web page indexing system that creates an index of words appearing on pages. The list of pages for which we create an index is given as an argument to the call.

Also program a function that will return, for a given index and word, on which website a given word is the most popular.