Assignment 1 - Information Retrieval

Luiz Philippe Pereira Amaral

[1]Departamento de Ciência da Computação - Universidade Federal de Minas Gerais (UFMG)

Abstract. *This project is a sample of a simple web crawler which takes two arguments: a starting URL and a number of links to visit from the starting page. At the end of the execution, the program outputs a summary of the time spent (on average per page as well as total time).*
Keywords: *C++, Chilkat, web crawler*

1. Implementation

The code is written in C++ 17 and utilizes the Chilkat library for C++. Instructions for installing Chilkat can be found at `https://www.chilkatsoft.com/downloads_CPP.asp`.

1.1. Compiling and running

You can use make to build the project by simply typing ”make”on your terminal. The output is an executable file inside the build directory. You will need, however, to install the Chilkat library on your environment.

2. Results

The crawler was tested in two different scenarios, starting at a institutional website for a university which has its servers at the same city were the test was made, and starting at a web page with servers located in another continent.

In the first case, the starting web address was https://ufmg.br crawling through 15 pages, which produced the following output:

```
Duration: 0.874 seconds
Average time spent per page: 0.054625 seconds
```

On the second example, the starting page was https://kurzgesagt.org/crawling through 10 pages. This address belongs to a page hosted in Germany (the tests were made in Brazil), is a server-side rendered page filled with images and animations, so it is expected to have a bigger latency to the requests. This was the output:

```
Duration: 13.644 seconds
Average time spent per page: 1.24036 seconds
```

Referências

1. https://www.chilkatsoft.com/