RUG_Information_Retrieval

Our application runs on **Python**.

Assignment 5

Tne web crawler is an application that automatically surfs the web. In order to develop such an application we have used several additional libraries.

URLLIB which helps to operate with the urls, allowing to divide the url into parts such as scheme, netloc, to join the urls and etc.

BS4 BeautifulSoup allows us to be able to get tags from the HTML pages.

REQUESTS is a library that simply operates with HTTP requests.

NETWORKX library used to draw the graph.

The workflow of our web crawler can be described as following : First of all we make a GET request to the first input URL, after which we extracted the URL's HTML content. As the next step, we convert this HTML content in a BeautifulSoup object, so that later on we can extract HTML tags and their contents. We are interested in anchor tag, and especially in "href" parameter, since we are looking to retrieve all links present on the page. Getting the links, we extract their domain, and look if it is the same as the initial, to know if our crawler is still on the same domain or it went further. The idea to separate the different domains links would be good when displaying data on the graph, so the graph would't be overcrowded. Since URL is formed as a tree structure, on the first level we have the starting URL, while on the next level we have all the URLs that can be accesed from inside of the input URL and so on, we have used Breadth First Search (BFS) traversal. Our web crawler is retrieving data following depth parameter, in our condition we must crawlto the depth=2, meaning that the application must crawl all urls in the given web page and in turn it must crawl all the urls within them.

Application Usage:

Our program is very easy to use, user must make sure he/she has all the libraries installed if not this can be done by the command :

```
$ pip install <library_name>
```

Also user must change the value of starting_url variable to his desired URL to start with and of course run the program.