Universitat Pompeu Fabra
Barcelona

Web Intelligence 2019-2020

Lab: Web Crawler

Goal

The goal of this lab is to develop a simple web crawler that collects web pages by following links between those web pages. This lab is part of the search engine lab that will be delivered on March 13th.

Procedure

1. Pick a URL
2. Parse the HTML to extract links to other URLs (jsoup or Beautiful Soup)
3. Add the links to a repository following *breadth-first* approach.
4. For each link in the repository:
    a. Check if it has not been already crawled
        i. Confirm that it agrees to be checked (robots.txt, crawling frequency)
        ii. Store the page in a local folder
        iii. Go to step 2

You should put a limit in the number of crawled web pages in order to stop the crawler.

Tools

The crawler can be developed in Java (jsoup library) or Python (beautifulSoup library).

Submission

The crawler will be run during the labs defense: March 13th.