Introduction to Web Science

Tutorial (Assignment 5)

Olga Zagovora

WeST

People and Knowledge Networks

Who am I

2006-10: B.Sc. CS
2010-11: M.Sc. IT of Design
2010-13: Software dev. (in CAD domain)

2014-16: M.Sc. Web Science
2015-16: Research Assistant at GESIS

My scientific interests:
Computational Social Science

- **gender bias,**
- **altmetrics**

Olga Zagovora [**zagovora@uni-koblenz.de**](mailto:zagovora@uni-koblenz.de)
Office hours: by request (?B122)

Exercise 1

What does client do when it wants talk to server?
it sends a http request

Can server start a talk with client?
no

How server sends us messages?
it makes a response to our request

How can server response any time?
always leave one pending request from client

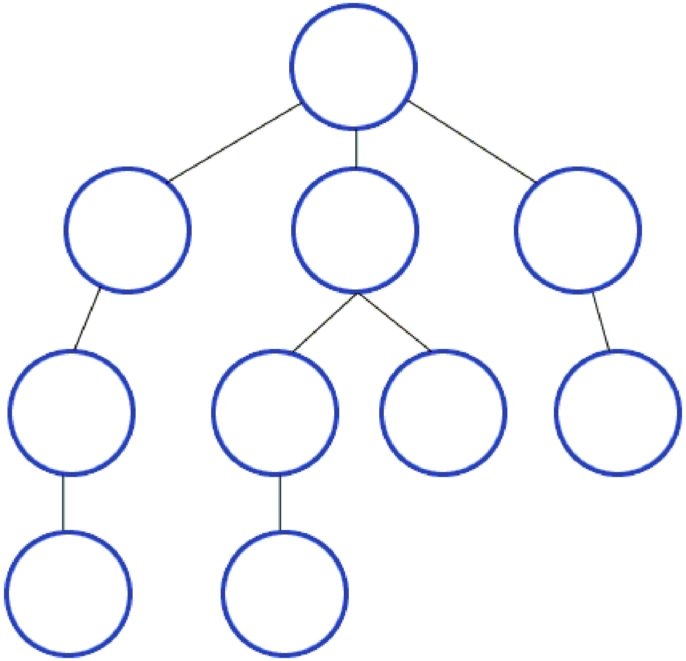
Exercise 1

Live Demo!

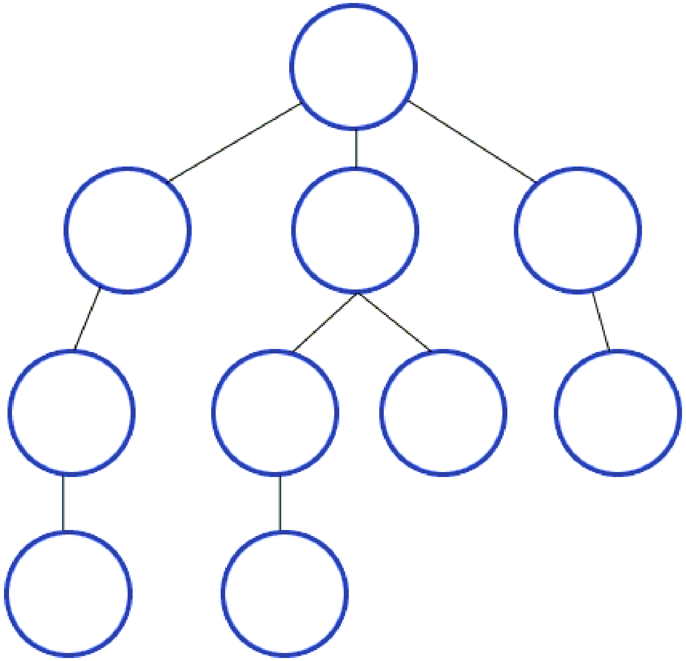
Exercise 2

Algorithms for traversing tree and graph data structures

Breadth-first search



Depth-first search



Exercise 2

Demo -> ipython notebook

Exercise 2

How to scale and make you crawler efficient?

1. Do not write your data every time. Dump your data after storing around 5000/10000 urls
2. Do not use different configurations (simple-> faster). Use log files for resume your crawler
3. Distributed tasks:
   1. Downloader
   2. Link extractor
   3. Visited urls
4. Use efficient data structures (e.g., pandas DataFrame)
5. %timeit

Exercise 3

Demo -> ipython notebook

Questions?

✉ zagovora@uni-koblenz.de

Images

1. Slide 5: By Mre (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons
2. Slide 5: By Mre (Own work) [GFDL (http://www.gnu.org/copyleft/fdl.html) or CC BY-SA 3.0 (http://creativecommons.org/licenses/by-sa/3.0)], via Wikimedia Commons