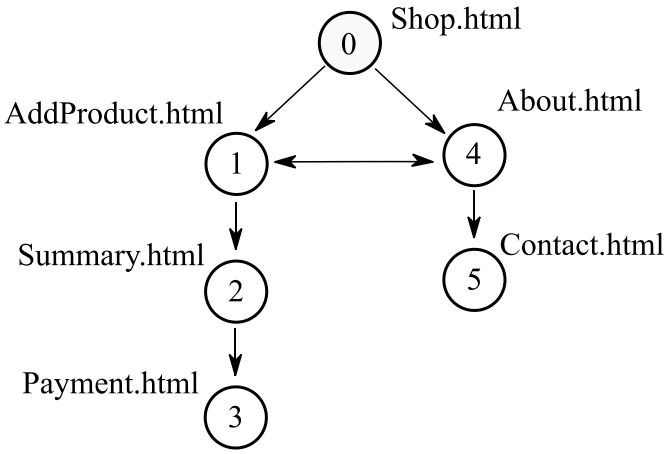
LAB 11: Log Files

Programming assignment (Python notebook, deadline +1 week)

Given is the following web structure:



Download

- notebook that contains exercises http://www.cs.put.poznan.pl/mtomczyk/ir/lab11 /Lab11_Logs.ipynb.
- Server log file: http://www.cs.put.poznan.pl/mtomczyk/ir/lab11 /log.txt.

The log file contains requests in the following form:

141.243.1.172 [01/Jun/2018:03:09:21 -0600] "GET /Shop.html HTTP/1.0" 200 1497

Your task is to identify users and sessions. Then, you have to analyze the collected data (e.g., identify the most common entry pages; see the notebook file). It is assumed that one user uses only one computer (IP). In order to identify the sessions, use a combination of some of the following heuristics:

1. Total session duration may not exceed a threshold θ. Given t_0, the timestamp for the first request in a constructed session S, the request with timestamp t is assigned to S, iff $t - t_0 \leq \theta$,
2. Total time spent on a page may not exceed a threshold δ. Given t_1, the timestamp for request assigned to constructed session S, the next request with timestamp t_2 is assigned to S, iff $t_2 - t_1 \leq \delta$,
3. Given two consecutive requests p and q, q is assigned to S, if the referrer for q was previously invoked in S (in case of a conflict, assign q to the last open session).