Lab 9 – Log Files

Deadline: +1 week

1   Exercise

Given is the following web structure:
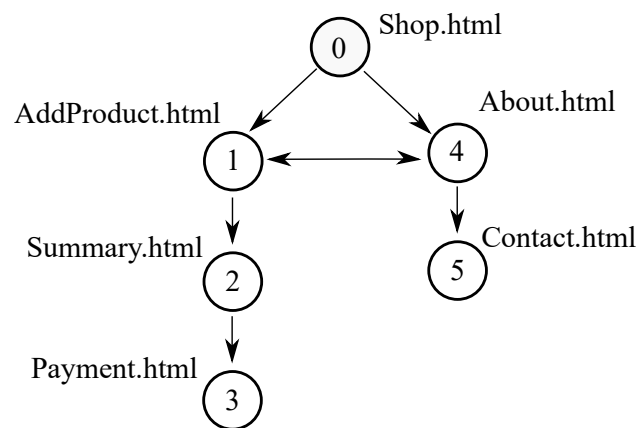

Figure 1: Web structure

Download:

- Notebook with exercises [HTML version]

- Server log file

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. Let:

- $S = [r_1, r_2, ..., r_n]$ be a session being currently constructed,

- r_i be the ith request in S ($i = 1 \rightarrow$ the oldest request; $i = n \rightarrow$ the newest request),

- r_{new} be a request being currently processed,

- $t(r)$ be a timestamp for a request r

Consider the following heuristics:

1. A total duration of S should not exceed a threshold θ, i.e., r_{new} is added to S if $t(r_{new}) - t(r_1) \leq \theta$.

2. Total time spent on a single page (in S) should not exceed a threshold δ, i.e., r_{new} is added to S if $t(r_{new}) - t(r_n) \leq \delta$.

3. r_{new} is added to S if $\exists_{(r \in S)} r \xrightarrow{\text{refers to}} r_{new}$

4. r_{new} is added to S if $\exists_{(r \in S)} r = r_{new}$